Methods for Identifying Candidate Genes for Cardiovascular Diseases by using Microarrays

Adášková, Jana
2008

Methods for Identifying Candidate Genes for Cardiovascular Diseases by Using Microarrays

Post-Graduate Student:
MGR. JANA ADÁŠKOVÁ

Department of Medical Informatics
Instutite of Computer Science of the ASCR, v. v. i.
Pod Vodárenskou věží 2

182 07 Prague, Czech Republic

adaskova@euromise.cz

Supervisor:
PROF. RNDR. JANA ZVÁROVÁ, DRSC.

Department of Medical Informatics
Instutite of Computer Science of the ASCR, v. v. i.
Pod Vodárenskou věží 2

182 07 Prague, Czech Republic

zvarova@euromise.cz

Field of Study:
Biomedical Informatics

The work was supported by the grant 1M06014 of the Ministry of Education, Youth and Sport of the Czech Republic.

Abstract

Microarrays present new powerful technique for high-throughput, global transcriptomic profiling of gene expression. It permits to investigate the expression levels of thousands of genes simultaneously. The global snapshots of gene expression, both among different cell types and among different states of a particular cell type can help in identifying candidate genes that may be involved in a variety of normal or disease processes. This promises to provide insight into the pathophysiology of human syndromes such as cardiovascular diseases, whose etiologies are due to multiple genetic factors and their interaction with the environment.

Microarrays also present new statistical and bioinformatical problems because the data are very high dimensional with very little replication. Almost all research employing microarray expression analysis depends heavily on statistical analysis to extract the most useful information from the huge number of data points generated.

The aim of this paper is to present possibilities of use of microarrays for identifying candidate genes for cardiovascular diseases and specially attention is devoted to statistical methods for identifying differentially expressed genes from microarray data.

Keywords: microarray, gene expression, cardiovascular diseases, microarray data, SAM, Bayes T-test, samroc, Zhao-Pan method.

1. Introduction

Identification of genetic determinants that predispose to common diseases such as cardiovascular diseases is a major challenge for current biomedical research. Despite recent advances in molecular and statistical genetics and the availability of complete genome sequences of humans and animal models, however, the underlying molecular pathogenic mechanisms for these disorders are still largely unknown. Nowadays a valuable tool for increasing our understanding of the regulatory and functional complexity of the molecular basis of multifactorially determined diseases is expression profiling.

Gene expression profiling is a logical next step after sequencing a genome: the sequence tells us, what the cell could possibly do, while the expression profile tells us, what it is actually doing now. Genes contain the instructions for making messenger RNA (mRNA), but at any moment each cell makes mRNA from only a fraction of the genes it carries. If a gene is used to produce mRNA, it is considered ”on”, otherwise ”off”. Expression profiling experiments involve measuring the relative amount of mRNA expressed in two or more experimental conditions. This is because altered levels of a specific sequence of mRNA suggest a changed need for the protein coded for by the mRNA, perhaps indicating a homeostatic response or a pathological condition. Therefore gene expression profiling can help in identifying candidate genes that may be involved in a variety of normal or disease processes. Additionally, characterization of genes abnormally expressed in diseased tissues may lead to the discovery of genes that can serve as diagnostic markers, prognostic indicators or targets for therapeutic intervention.

The development of several gene expression profiling methods, such as comparative genomic hybridization (CGH), differential display, serial analysis of gene expression (SAGE) and gene microarray, together with the sequencing of the human genome, has provided an opportunity to monitor and investigate the complex cascade of molecular events leading to cardiovascular diseases [2]. High-throughput technologies can be used to follow changing patterns of gene expression over time. Among them, gene microarray has become prominent because it is easier to use, does not require large-scale DNA sequencing, and allows for the parallel quantification of thousands of genes from multiple samples. Nowadays gene microarray technology is rapidly spreading worldwide and has the potential to drastically change the therapeutic approach to patients affected with cardiovascular or others complex diseases [3]. Therefore, it is important to know the principles underlying the analysis of the huge amount of data generated with microarray technology.

2. Microarray technology

Microarray technology takes advantage of hybridization properties of nucleic acid (DNA or RNA) and uses complementary molecules attached to a solid surface, referred to as probes, to measure the quantity of specific nucleic acid transcripts (mRNA) of interest that are present in a sample, referred to as the target. The molecules in the target are labelled, and specialized scanner is used to measure the amount of hybridized target at each probe, which is reported as an intensity. The raw or probe-level data are the intensities of each spot on the hybridization array, from which the initial concentrations of the corresponding transcripts are inferred.

Various manufacturers provide a large assortment of different platforms. The different platforms can be divided into two main classes that are differentiated by the data they produce. The high-density oligonucleotide array platforms produce one set of probe-level data per microarray with some probes designed to measure specific binding and others to measure non-specific binding. The two-color spotted platforms produce two sets of probe-level data per microarray (the red and green channels), and local background noise levels are measured from areas in the glass slide not containing probes [4]. Despite the differences among the different platforms, the steps of microarray data analysis are similarly to all microarray technology.

3. Microarray data analysis

Microarray experiments produce a huge amount of data. A single microarray run can produce between 100,000 and a million data points, and a typical experiment may require tens or hundreds of runs [5]. Microarray data analysis consist of three parts: (i) data preparation, in which data are adjusted for the downstream algorithms; (ii) algorithm selection for data analysis; and (iii) interpretation, in which the results from the algorithms are explained in a biological context. In Fig. 1 are shown the major phases of microarray data analysis (colored icons) and their connectivity (arrows) in the microarray workflow process.

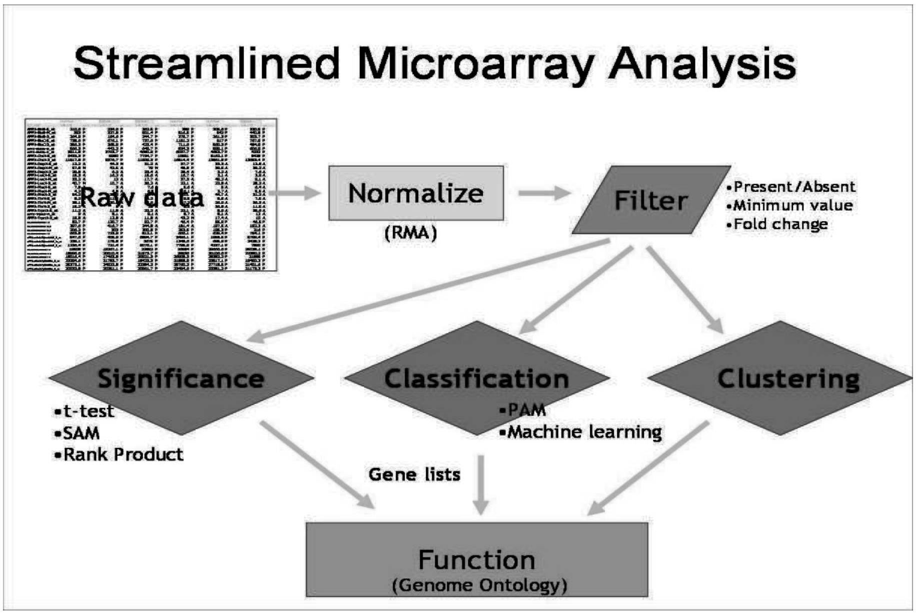

Figure 1: Microarray data analysis.

3.1. Low-Level analysis

Primary image data having been collected from a microarray experiment. The aims of the first level of analysis, so-called low-level analysis or data preprocessing, are image analysis, background elimination, filtration, normalization and data transformation, all of which should contribute to the removal of systematic variation between chips, enabling group comparisons.

Image analysis permits us to convert pixel intensities in the scanned images into probe-level data. Many image-processing approaches have been developed, among which the main differences relate to how spot segmentation, distinguishing foreground from background intensities, is carried out [4]. Another important preprocessing step is normalization. Normalization involves comparing different microarrays relative to some standard intensity value. This could be the overall intensity of the microarray, the overall intensity of all of the genes on the microarray, the intensity of so-called housekeeping genes (the expression of which are supposedly constant), or spiked targets, containing a known and constant amount of a labelled control. Negative normalization controls might be represented by target sequences from a different organism. Several normalization approaches have been introduced, and are discussed elsewhere [4]. Data are often then subjected to log transformation to improve the characteristics of the distribution of the expression values.

3.2. Statistical analysis

Microarrays present new statistical problems because the data are very high dimensional with a very small number of replications. A common task in analyzing microarray data is to determine which genes are differentially expressed across two tissue samples or samples obtained under two experimental conditions.

In early days, the simple method of fold changes was used. Simple and intuitive, this method, involves the calculation of a ratio relating the expression level of a gene under control and experimental conditions. An arbitrary ratio (usually 2-fold) is then selected as being ”significant.” Because this ratio has no biological merit, this approach amounts to nothing more than a blind guess. The selection of an arbitrary threshold results in both low specificity (false positives, particularly with low-abundance transcripts or when a data set is derived from a divergent comparison) and low sensitivity (false negatives, particularly with high-abundance transcripts or when a data set is derived from a closely linked comparison) [6]. It is now accepted that the use of the fold change method should be discontinued. Since then, many more sophisticated methods have been proposed (e.g. Chen et al 1997, Efron et al 2000, Ideker et al 2000, Newton et al 2001, Tusher et al 2001, Lin et al 2001, Pan et al. 2001) [3]. It has been also noticed that data based on a single array may not reliable and may contain high noises. As the technology advances, microarray experiments are becoming less expensive, which make the use of multiple arrays feasible. Most, if not all, statistical tests can be modified accordingly for a multiple comparison adjustment.

In this section I would like to review more in detail two types of parametric methods (such as T-test and Bayes T-test) and three types of non-parametric methods (such as samroc, SAM, and a modified mixture model proposed by Zhao and Pan) recently used for identifying differentially expressed genes in microarray data. Suppose that the experimental data consist of measurements y_{gi} under two conditions, where i ($i = 1, 2, ..., k$) denotes the i-th array, g ($g = 1, 2, ..., G$) denotes the g-th gene, and k_1 and k_2 are the number of arrays for each condition, that is, $k = k_1 + k_2$. Let the sample means and the sample variances of y_{gi}'s for gene g under two conditions be denoted as $\overline{y}_{g1}$, s^2_{g1} and, $\overline{y}_{g2}$, s^2_{g2} respectively. Here, diff is the difference between $\overline{y}_{g1}$ and $\overline{y}_{g2}$, and s_g and Se_g represent the pooled standard deviation and the standard error of the diff across the replicates for the gene, respectively.

3.2.1 T-statistics: The two sample T-statistics with two independent normal samples without assuming the equal variances between two samples could be written as follows:

$$t_g = \frac{diff}{Se_g}, Se_g = \sqrt{\frac{s^2_{g1}}{k_1} + \frac{s^2_{g2}}{k_2}}$$

A gene with very small variance due to its low expression level contributes to have large absolute t-value regardless of the mean difference under two conditions, and thus this gene can be selected as the differentially expressed gene although it is not truly differentially expressed. To overcome this problem of the traditional T-test, various methods have been proposed. Among these methods, there are SAM and samroc (see below).

3.2.2 Bayes T-test: Baldi and Long [7] developed a Bayesian probabilistic framework for microarray data analysis. Their statistics is used to solve small variance problems in low expression level and uses the parametric Bayesian method to have the parameters (mean, standard deviation and so on.)

for T-statistics. This statistics is well known for its effectiveness in analyzing the samples having small size, but it still heavily depends on the parametric assumption. Bayes T-test uses the estimate of parameters such as population mean (μ) and variance (σ^2) by Bayesian method instead of sample mean and sample variance of the traditional T-statistics. The mean of posterior estimate in each group is given as

$$\mu_j = \mu_{nj}, \sigma_j^2 = \frac{\nu_j \sigma_{nj}^2}{\nu_j - 2},$$

where the mean of the posterior estimate (μ_{nj}) is a convex weighted average of the prior mean (μ_{0j}) and the sample mean $\overline{y}_j$ for group j, $j = 1, 2$, that is,

$$\mu_{nj} = \frac{\lambda_{0j}}{\lambda_{0j} + k_j}\mu_{0j} + \frac{k_j}{\lambda_{0j} + k_j}\overline{y}_j$$

The hyperparameters μ_{0j} and σ_j^2/λ_{0j} can be interpreted as the location and the scale of μ_j, respectively, and k_j is the sample size for each group. σ_{nj}^2 is posterior variance component and posterior sum of squares is

$$\nu_j \sigma_{nj}^2 = \nu_{0j}\sigma_{0j}^2 + (k_j - 1)s_j^2 + \lambda_{0j}k_j/(\lambda_{0j} + k_j)(\overline{y}_j - \mu_{0j})^2,$$

and the posterior degree of freedom is $v_j = v_{0j} + k_j$. In Bayes T-test, the hyperparameters for the prior v_{0j} and σ_{0j}^2 can be interpreted as the degree of freedom and scale of σ_j^2, respectively [7]. Owing to the complicated theoretical background, I will not discuss it here in more detail. This statistics is currently implemented in the Limma software package [8] as part of project Bioconductor accessible at www.bioconductor.org .

3.2.3 Significant analysis of microarrays (SAM): To avoid the small variance problem of T-test, SAM uses a statistics similar to T-statistics and the permutation of repeated measurements to estimate the false discovery rate [9]. At low expression levels, the absolute value of t_{sam} can be high because of small values in Se_g. The shortcoming of the traditional T-test is that genes with small sample variances due to the low expression levels have high chance of being declared as the differentially expressed genes. Thus SAM added a small positive constant a to alleviate this problem. The SAM statistics is

$$t_{sam} = \frac{diff}{Se_g + a}, Se_g = s_g\sqrt{\frac{1}{k_1} + \frac{1}{k_2}},$$

where the value for a is chosen to minimize the coefficient of variation. SAM is similar to the method by Efron et al. [10], which use a to be equal to the 90th percentile of the standard errors of all the genes. SAM assigns a score based on changes that is related to the standard deviation of repeated measurements for that gene. Genes with scores greater than a cutoff value are determined to be significant.

3.2.4 Samroc: Broberg [11] proposed a method for ranking genes in the order of likelihood of being differentially expressed, which is often called as samroc. The main purpose of this method is to estimate the false negative (FN) and false positive (FP) rates. The procedure sets out to minimize these errors. The samroc method is similar to SAM, although an added constant in the denominator of the statistics is different. The proposed statistics is

$$t_{sam} = \frac{diff}{Se_g + b}.$$

Main interest is to find the optimal constant b for given significance level of α. This procedure proposed a criterion, which is the distance of points on the curve to the origin, for choosing a good receiver operating characteristic (ROC) curve. ROC curve allows users to compare the FP error rate and FN error rate of various test statistics without involving P-values. This minimizes the number of genes that are falsely declared positive and falsely declared negative for a given significance level of α and a value b [11].

3.2.5 Zhao-Pan method: Zhao and Pan [12] adopted a modified non-parametric approach to detect the differentially expressed genes in replicated microarray experiments. The basic idea of this non-parametric method lies in estimating the null distribution of test statistics, say Z_g, by directly constructing a null statistics, say z_g, such that the distribution of z_g is the same as the distribution of Z_g under the null hypothesis. This avoids the strong assumptions about the null distribution of the parametric methods. A common problem with these methods is that the numerator and the denominator of z_g and Z_g are assumed to be independent of each other. In practice, this independency is violated by z_g, and z_g and Z_g are used to overcome this problem. For more details refer to the Zhao and Pan [12].

Method	Sample	Distributional	Equal variance assumption between groups
T-statistics	Large	Strong	Unequal
B-statistics	Small	Strong	Unequal
SAM	Small	None	Equal
samroc	Small	None	Equal
Zhao-Pan	Large	Weak	Equal

Table 1: The main features of the statistical methods .

Table 1 summarizes main features of the previous described methods in the context of sample size, distributional assumption, and variance condition between two groups. In general, SAM, samroc and Bayes T-test are known to work well with the small sample size, and T-statistics and Zhao-Pan method are known to perform well with large sample size. This difference may be related to the fact that SAM and samroc do not need any distributional assumption, whereas the others need distributional assumptions for the analysis. Of these five methods, SAM, samroc and Zhao-Pan method require the equal variance assumption between two groups.

3.3. High-Level analysis

High-level microarray analysis is required to identify groups of genes that are similarly regulated across the biological samples under study. A variety of mathematical procedures have been developed that partition genes or samples into groups, or clusters, with maximum similarity, thus enabling the identification of gene signatures or informative gene subsets. Methods for classification are either unsupervised or supervised. Supervised methods use existing biological information about specific genes that are functionally related to ”guide” or ”test” the cluster algorithm. With unsupervised methods, no prior test set is required. The most commonly employed unsupervised classification methods are the clustering techniques [13]. However discussion of these techniques more in detail is beyond the scope of this paper.

Conclusion

Nowadays comprehensive gene expression approaches like microarrays have fundamental role in providing basic information integral to biological and clinical investigation of complex diseases such as cardiovascular diseases. The statistical analysis of microarray data is probably the most difficult problem associated with the use of these technique. We can see, that the selection of the significant genes heavily depends on the choice of the testing methods. We can also see that the performance of the testing methods is affected by sample size, distributional assumption, the variance structure and so on (see Table 1). Therefore, to obtain the reliable testing results for detecting significant genes in microarray data analysis, we first need to explore the characteristic of the data and then apply the most appropriate testing method under the given situation. It is also important to choose the measure of differential expression based on the biological system of interest and particular problem specification. In a situation where the most reliable list of genes is desirable, the best approach may be to examine the intersection of genes identified by more methods.

In our future work we would like to apply the statistical methods described in this paper to the real microarray dataset from project of Centre of Biomedical Informatics *(The goal of this experiment is to identify genes that are differentially expressed in acute myocardial infarction patients and cerebrovascular accident patients)* and compare selected top significant genes by each of testing methods and also compare it with reference selected candidate genes (from well-curated publicly available databases), which are believed to be truly differentially expressed.

References

[1] S. Archacki, Q. K. Wang, “Expression profiling of cardiovascular disease”, *Human Genomics*, vol. 1, pp. 355–370, 2004.

[2] Q.K. Wang, S. Archacki, “Cardiovascular diseases”, *Humana Press*, vol. 129, pp. 1–13, 2007.

[3] J. L. Haines, M. Pericak-Vance, “Genetic analysis of complex diseases”, *John Wiley and Sons Publisher*, 2006.

[4] R. Gentleman, V. J. Carey, W. Huber, R. Irizarry, S. Dudoit, “Bioinformatics and Computational Biology Solutions Using R and Bioconductor”, *Springer Publisher*, 2005.

[5] D. B. Allison, X. Cui, G. P. Page, M. Sabripour, “Microarray data analysis: from disarray to consolidation and consensus”, *Nature Reviews*, vol. 7, pp. 55–65, 2006.

[6] D. Murphy , “Gene expression studies using microarrays: Principles, problems and prospects”, *Advan. Physiol. Edu.*, vol. 26, pp. 256–270, 2002.

[7] P. Baldi, A. D. Long, “A Bayesian framework for the analysis of microarry expression data: regularized t-test and statistical inferences of gene changes”, *Bioinformatics*, vol. 17, pp. 509–19, 2001.

[8] G. K. Smyth, “Linear models and empirical bayes methods for assessing differential expression in microarray experiments”, *Statistical Applications in Genetics and Molecular Biology 3*, vol. 1, Article 3, Epub. 2004.

[9] V. Tusher, R. Tibshirani, G. Chu, “Significance analysis of microarrays applied to the ionizing radiation response”, *Proceedings of the National Academy of Sciences*, vol. 98, pp. 5116–21, 2001.

[10] B. Efron, R. Tibshirani, J. D. Strey, V. Tusher, “Empirical Bayes analysis of a microarray experiment”, *Journal of the American Statistical Association*, vol. 96, pp.: 1151–60, 2001.

[11] P. Broberg, “Ranking genes with respect to differential expression”, *Genome Biology*, vol. 3: preprint0007.1-0007.23, from http://genomebiology.com/2002/3/9/preprint/0007 , 2002.

[12] Y. Zhao, W. Pan, “Modified nonparametric approaches to detecting differentially expressed genes in replicated microarray experiments”, *Bioinformatics*, vol. 19, pp. 1046–54, 2003.

[13] R. B. Altman, “Whole-genome expression analysis: challenges beyond clustering”, *Curr Opp Structural Biol.*, vol. 11, pp. 340-347, 2001.