REVSTAT – Statistical Journal
Volume 3, Number 2, November 2005, 99–111

ANALYSIS OF DYNAMIC PROTEIN EXPRESSION DATA

Authors: Klaus Jung
– Department of Statistics, University of Dortmund, Germany
(klaus.jung@uni-dortmund.de)

Ali Gannoun
– Equipe de Probabilités et Statistique, Université Montpellier, France

Barbara Sitek
– Medical Proteom-Center, Ruhr-University Bochum, Germany

Helmut E. Meyer
– Medical Proteom-Center, Ruhr-University Bochum, Germany

Kai Stühler
– Medical Proteom-Center, Ruhr-University Bochum, Germany

Wolfgang Urfer
– Department of Statistics, University of Dortmund, Germany
(urfer@statistik.uni-dortmund.de)

Received: May 2005 Accepted: June 2005

Abstract:

- Difference gel electrophoresis (DIGE) is the new gold standard analysing complex protein mixtures in proteomics. It is used for measuring the expression levels of proteins in different mixtures on the same two-dimensional electrophoresis (2-DE) gel. In this paper we review a method for the calibration and normalization of those protein expression measurements. Further we show how to find treatment effects and time-treatment-interactions in longitudinal data obtained from DIGE experiments. A problem in those data sets is the existence of a lot of missing values. Therefore, we propose a method for the estimation of missing data points.

Key-Words:

- *difference gel electrophoresis; data calibration; mixed linear model for longitudinal data; missing values; proteomics.*

1. INTRODUCTION

While the focus of biochemical research was addressed on the genome in the last decade the view is now turned onto the proteome. Big data sets of gene expression obtained from DNA-microarrays made the development of statistical methods necessary to make correct inferences from these measurements. For quantitative protein expression analysis either mass spectrometry (cf. Aebersold and Goodlett ([1]) and Gygi et al. ([7])) or two-dimensional gel electrophoresis (2-DE) (cf. Westermeier et al. ([14])) is applied. In this paper we focus on the analysis of protein expression data obtained from a new detection method (Difference Gel Electrophoresis, DIGE) based on fluorescence labelling before 2-DE. 2-DE separates the proteins of a mixture by their isoelectric point (pI) and molecular size to distinct spots. After separation the proteins are detected using a confocal fluorescence scanner whereas fluorescence intensity of a spot can be regarded as a measure of expression for its respective protein. DIGE enables the user to put up to three different mixtures of proteins on the same gel. The different mixtures are labelled by different fluorescence dyes (Cy2, Cy3 and Cy5). For quantitative proteome analysis image analysis software automatically determines the boundaries and sizes of the spots. Usually, a DIGE experiment is designed such that m independent replications of treatment and control mixtures are put on the same m gels. The internal standard, a mixture of same amounts of all m treatment and m control probes, is also put on each gel. This internal standard allows high accuracy calibration of the expression values. Calibration and normalization of protein expression data is reviewed in section 2. In order to obtain information about interactions of treatment and control with the time, DIGE experiments often include measurements over several time points. Known statistical methods for the analysis of longitudinal data can be used to analyze those experiments. One possible method for such an analysis is detailed in section 3. Often, 2-DE data contains up to 50% of missing values. The missing values occur because not each protein is visible on each gel when replicating probes on several gels. For example, on gel number one there are 1732 protein spots and 1967 spots are on gel number two, but only 1447 of these spots belong to proteins commonly represented on both gels. Some statistical methods, however, need complete data sets, for example, some methods for the detection of differentially expressed genes (cf. Gannoun et al. ([6])) or the correspondence analysis for microarray data (cf. Fellenberg et al. ([5])). These methods could also be applied to protein expression data if the data sets were complete. One possible method to overcome this problem is to estimate the missing values by using the available measurements from other proteins. In section 4, we investigate how the k nearest neighbor method behaves when being applied to DIGE data. This method was also applied for the estimation of missing values in gene expression data by Troyanskaya et al. ([13]). The idea of this method is that there are groups of proteins with similar expression profiles. A missing value of a protein can then be estimated by available values from the proteins of the same group.

2. CALIBRATION, NORMALIZATION AND STANDARDIZATION OF DIGE DATA

A usual DIGE experiment results in three values for each spot on a gel, i.e. treated, untreated and internal standard. From the DeCyder^{TM} software one can obtain the background subtracted spot volumes (cf. Amersham Biosciences ([2])). In this software, a borderline for each spot is automatically detected and the sum of the pixel intensities within the spot boundary is the spot volume. The background is subtracted by excluding the lowest 10th percentile pixel values on the spot boundary. As we will see in this section the statistical analysis cannot be done with this raw data material. Data obtained from analytical instruments are always affected by technical and biological variation. To make correct inferences on the biological variation preprocessing of data is necessary. In this section we discuss the features of the background subtracted spot volumes and describe how to calibrate and transform the values for further actual analysis. One source of technical variation comes from the different dyes. In figure 1 the Cy5 and Cy3 spot volumes of a DIGE gel are plotted against each other.

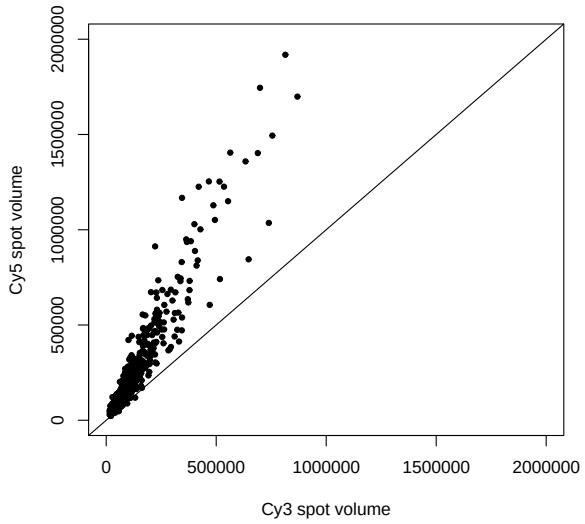

Figure 1: Scatterplot of the Cy5 versus the Cy3 spot volumes of a DIGE gel.

It can be seen that the Cy5 dye causes higher volume values than the Cy3 dye. To calibrate the spot volumes Karp et al. ([9]) proposed to use a scaling factor which adjusts for the dye-specific gain, and to use an additive offset which compensates for any constant additive bias present after background subtraction.

The additive offset is used because the different dyes result in different background fluorescence. This calibration method was originally introduced by Huber et al. ([8]) for the preprocessing of DNA-microarray data. Having n spots on a gel with three different mixtures (internal standard, treated, untreated) this calibration can be modelled by

$$\tilde{y}_{ih} = a_h + b_h y_{ih} \tag{2.1}$$

with $i = 1, ..., n$ and $h = 1, 2, 3$. For $h = 1$ we have the value for the treated probe, $h = 2$ for the untreated probe and $h = 3$ for the internal standard. In this model $\tilde{y}_{ih}$ are the measured background subtracted spot volumes, a_h are the additive offsets and b_h are the scaling factors. Hence, $2 * 3$ parameters have to be estimated. How to do this will be explained below. Some more features of the raw data require a second transformation. The scanning of the fluorescent gels results in lognormal distributions of the spot volumes. However, a normal distribution would be more appropriate for most statistical applications so the data has to be normalized. Furthermore, the variance of the spot volumes is dependent on the mean of the spot volumes. This is illustrated in figure 2.

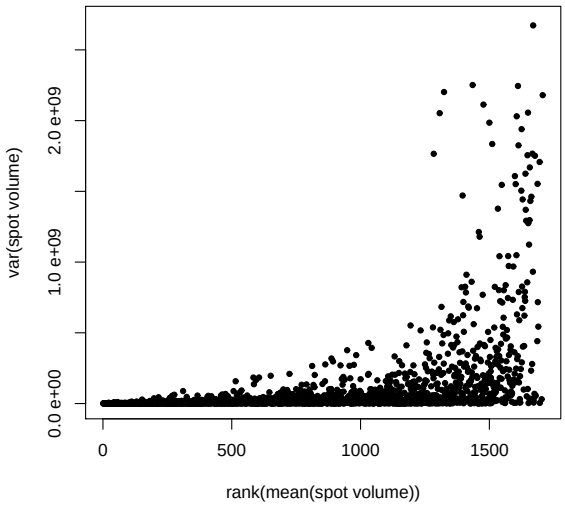

Figure 2: Scatterplot of the variance of the Cy3 and Cy5 spot volumes versus the rank their mean.

The variance of the spot volumes increases when the mean also increases. One possibility to normalize the data and to stabilize the variance would be to apply the logarithm on the data. But the logarithm results in a bias for low spot volumes as can be seen in in figure 3 where the Cy3 and Cy5 spot volumes with the logarithm applied on them are plotted against each other. Instead of using the logarithm we will use the arsinh for normalization and variance stabilization.

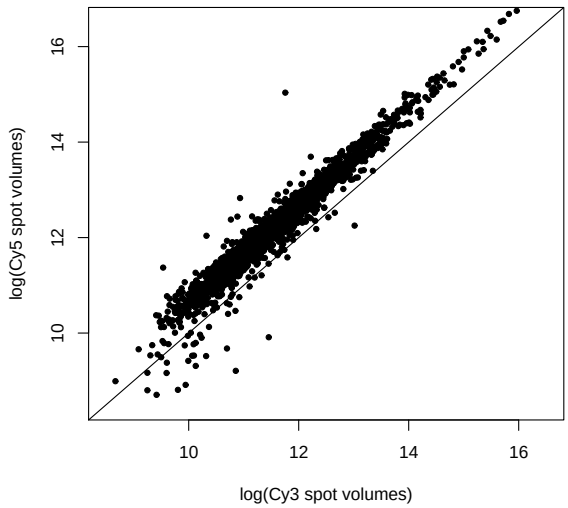

Figure 3: Scatterplot of the log-transformed Cy5 spot volumes versus the log-transformed Cy3 spot volumes.

The graphs of the logarithm and the arsinh are plotted in figure 4.

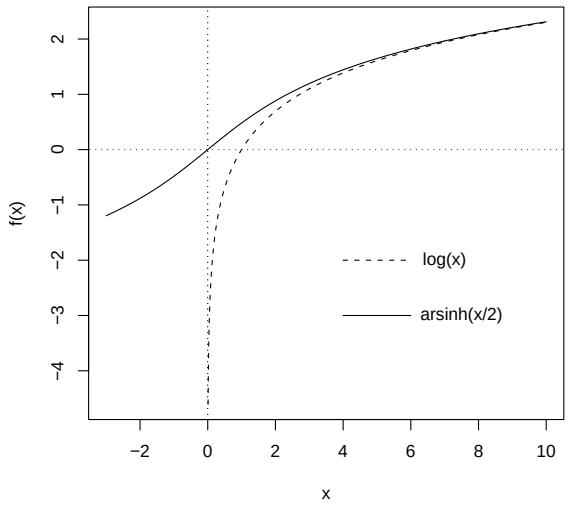

Figure 4: Graphs of the arsinh and the logarithm.

The relationship between the two functions can be expressed by

$$\lim_{\xi\to\infty}(\operatorname{arsinh}\xi - \log\xi - \log 2) = 0\ .$$

Hence, for big values the arsinh is equivalent to the logarithm, but it has not a singularity at zero and it is smooth for small values. Now, using the calibration

transformation and the arsinh, the true protein abundance x_{ih} can be modelled by

$$\text{arsinh}\,\frac{\tilde{y}_{ih} - a_h}{b_h} = x_{ih} + \varepsilon_{ih} \tag{2.2}$$

where $\varepsilon_{ih} \sim N(0, \sigma_\varepsilon)$. To estimate $(a_1, a_2, a_3, b_1, b_2, b_3)$ Huber et al. ([8]) proposed a robust version of maximum likelihood estimation. The robust version is necessary because maximum likelihood estimation itself is very sensitive to deviations from the normal distribution and to the presence of differentially expressed proteins. The above estimation algorithm is implemented in the vsn-package for the software R (both free available at http://cran.r-project.org). The resulting benefits of calibration and normalization can be seen in the figures 5 and 6.

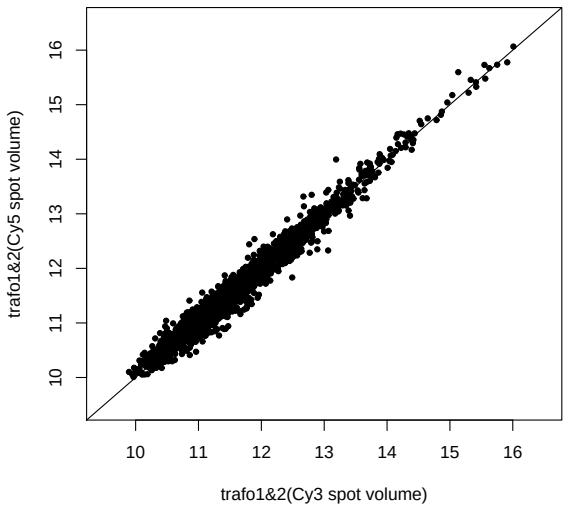

Figure 5: Calibrated and transformed Cy3 spot volumes versus calibrated and transformed Cy5 spot volumes.

In figure 5 it is shown that there is no more dye-specific gain for the calibrated and transformed spot volumes. Further, the bias for low spot volumes has disappeared. The variance of the calibrated and transformed volumes versus the rank of their mean is plotted in figure 6. It can be seen that there is no more dependence between variance and mean. Now, after calibration and normalization, we can use the benefit of the internal standard to reduce the gel-to-gel variation and bring all gels on the same level. This means we set the calibrated and arsinh-transformed treatment and control values in relation to the internal standard value. More precisely we have to subtract the internal standard from the treatment and control value, respectively, because ratios become differences when the logarithm or the arsinh is applied on them. Hence, the standardized treatment value is $x_{i1} - x_{i3}$ and the standardized control value is $x_{i2} - x_{i3}$.

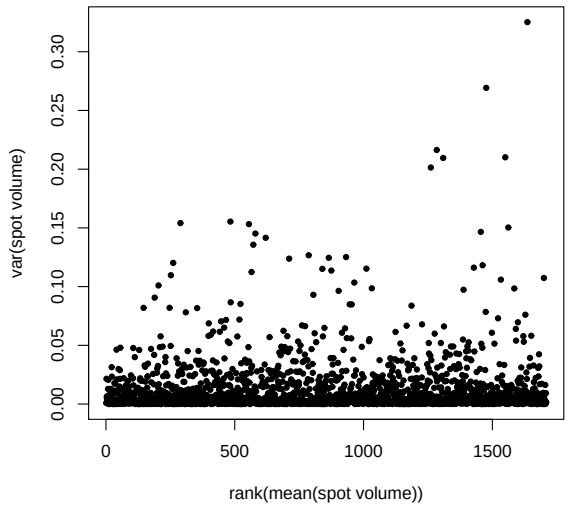

Figure 6: Variance of the calibrated and normalized spot volumes versus the rank of their mean.

3. ANALYSIS OF LONGITUDINAL DIGE DATA

A frequent subject of DIGE studies is the comparison of the temporal course of the protein expression in some treated probes to the temporal course of the protein expression in some untreated probes. Since there are only a few time points to be regarded such a study can be analyzed by using methods of longitudinal data analysis. Here, we adapt such a method, given in Diggle et al. ([4]), to the situation of a DIGE experiment. The design for a time dependent DIGE experiment is given in table 1. For each spot, which has been detected

Table 1: Design of a time dependent DIGE experiment.

	replication 1	replication 2	...	replication m
time 1	gel_{11}	gel_{12}	...	gel_{1m}
time 2	gel_{21}	gel_{22}	...	gel_{2m}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
time p	gel_{p1}	gel_{p2}	...	gel_{pm}

on each of the pm gels, the analysis is done separately. Recall, that for each spot and each gel we get a standardized volume value for the treated probe and a

standardized value for the untreated probe. We denote y_{hiq} as the standardized volume value for the spot in question on the jth gel at the qth time point within the hth group (treated, untreated), where $j = 1, ..., m$, $q = 1, ..., p$ and $h = 1, 2$. Since we analyze the same protein over the time we need a model which heeds the time-dependence of the values. Therefore, we regard the mixed linear model

$$y_{hjq} = \beta_h + \gamma_{hq} + U_{hj} + Z_{hjq} \tag{3.1}$$

where β_h is the main effect of the hth group, γ_{hq} is the interaction between group and time, $U_{hj} \sim N(0, \nu^2)$ is the random effect of the jth replication and $Z_{hjq} \sim N(0, \sigma^2)$ are the random errors. With the given distribution assumptions for the random effects the vector $Y_{hj} = (Y_{hj1}, Y_{hj2}, ..., Y_{hjp})$ is normally distributed with covariance matrix $V = \sigma^2 I + \nu^2 J$. That means that the correlation between two time points is given by $\rho = \nu^2/(\nu^2 + \sigma^2)$. At first we want to test the null hypothesis that there is no treatment effect, i.e. testing $\beta_h = \beta$ for $h = 1, 2$, meaning that the temporal course for the protein in the treated and untreated probe are on the same level. The F-statistic for testing this hypothesis is given by $F_1 = \{BTSS_1/(2-1)\}/\{RSS_1/(2m-2)\} \sim F_{(2-1),(2m-2)}$. The sums of squares are given in the corrected ANOVA table 2 below. We are further interested in the question if there is a treatment-time interaction, i.e. the temporal courses are not parallel. This can be answered by testing the null hypothesis $\gamma_{hq} = \gamma_q$ for $h = 1, 2$ and for $q = 1, ..., p$. This null hypothesis means that the response profiles of the group means are parallel. The according test statistic is given by $F_2 = \{ISS_2/[(2-1)(p-1)]\}/\{RSS_2/[(2m-2)(p-1)]\} \sim F_{(2-1)(p-1),(2m-2)(p-1)}$.

Table 2: ANOVA table for the Analysis of longitudinal DIGE data.

source of variance	sums of squares	d.o.f.
between treatment	$BTSS_1 = p\sum_{h=1}^{2} m(y_{h\cdot\cdot} - y_{\cdot\cdot\cdot})^2$	$2-1$
whole plot residual	$RSS_1 = TSS_1 - BTSS_1$	$2m-2$
whole plot total	$TSS_1 = p\sum_{h=1}^{2}\sum_{j=1}^{m}(y_{hj\cdot} - y_{\cdot\cdot\cdot})^2$	
between time	$BTSS_2 = 2m\sum_{q=1}^{p}(y_{\cdot\cdot q} - y_{\cdot\cdot\cdot})^2$	$p-1$
treatment-time interaction	$ISS_2 = \sum_{q=1}^{p}\sum_{h=1}^{2} m(y_{h\cdot q} - y_{\cdot\cdot\cdot})^2 - BTSS_1 - BTSS_2$	$(2-1)\times (p-1)$
split plot residual	$RSS_2 = TSS_2 - ISS_2 - BTSS_2 - TSS_1$	$(2m-2)\times (p-1)$
split plot total	$TSS_2 = \sum_{h=1}^{2}\sum_{j=1}^{m}\sum_{q=1}^{p}(y_{hjq} - y_{\cdot\cdot\cdot})^2$	$2pm-1$

4. MISSING VALUE ESTIMATION

As mentioned in the beginning missing values are a general problem in 2-DE data. In this section we present a method for the estimation of missing data, using the k nearest neighbor method. We begin with some notation. Let $E = (e_{ij})$ be the matrix of observations, where the rows are referred to protein spots and the columns are referred to replications (gels). Hence, e_{ij} is the expression value of protein i on gel j, with $i = 1, ..., n$ and $j = 1, ..., m$, as given below.

$$\begin{pmatrix} e_{11} & ... & e_{1m} \\ \vdots & & \vdots \\ e_{i1} & e_{ij} & e_{im} \\ \vdots & & \vdots \\ e_{n1} & ... & e_{nm} \end{pmatrix} \tag{4.1}$$

Now, we can define distances between each pair of rows of E $(E_i = (e_{i1}, ..., e_{im})'$, $E_{i'} = (e_{i'1}, ..., e_{i'm})')$. The Euclidean distance is given by

$$d_1(E_i, E_{i'}) = \sqrt{(e_{i1} - e_{i'1})^2 + (e_{i2} - e_{i'2})^2 + ...(e_{im} - e_{i'm})^2} \ , \tag{4.2}$$

the Tschebyscheff distance is given by

$$d_2(E_i, E_{i'}) = \sup |e_{ij} - e_{i'j}| \ , \tag{4.3}$$

$j = 1, ..., m,$ and the Mahalanobis distance is given by

$$d_3(E_i, E_{i'}) = \sqrt{(E_i - E_{i'})^T A^{-1} (E_i - E_{i'})} \ , \tag{4.4}$$

where A is the empirical covariance matrix of the m gels. The principle of the k nearest neighbor method is now the following. For the row E_i the k nearest neighbors are those rows of E with the k smallest distances to E_i. More details on the k nearest neighbor method can be found in Ripley ([11]). This method was used in nonparametric estimation of the density (see for example Rosenblatt ([12]) and regression (see for example Devroye ([3])) as well as in classification problems (see for example Ketskemety ([10])). With the above given notations missing protein measurements can be estimated as follows. Let E_i be the row where the value e_{ij} is missing. Let Q_i be the set of non missing values of E_i. We denote these values by e'_{ip}, $p = 1, ..., q$, and $E'_i = (e'_{i1}, ..., e'_{iq})^T$. Let E_s, $s \neq i$, be the row s of the Matrix E. We suppose that e_{sj} is available and at least q other e_{sp} are available, too, in the same columns as in E_i. Then we denote $E'_s = (e'_{s1}, ..., e'_{sq})^T$ and give the

Definition 4.1. E_i and E_s are neighbors if $d(E_i', E_s')$ is small.

and

Definition 4.2. The k rows E_s $(s \neq i)$ with the k smallest distances to E_i are the k nearest neighbors to E_i.

To estimate the missing value e_{ij} let $e_{s_1 j}, e_{s_2 j}, ..., e_{s_k j}$ be the e_{sj} such that E_s belongs to the k nearest neighbors of E_i. The missing value e_{ij} can now be estimated by

$$\hat{e}_{ij}^{\text{mean}} = \frac{1}{k} \sum_{l=1}^{k} e_{s_l j} \ , \tag{4.5}$$

$$\hat{e}_{ij}^{\text{wmean}} = \frac{1}{k} \sum_{l=1}^{k} w_{i s_l} e_{s_l j} \ , \tag{4.6}$$

with

$$w_{i s_l} = \frac{1}{d(E_i', E_{s_l}') \sum\limits_{t=1}^{k} \frac{1}{d(E_i', E_{s_t}')}} \ , \tag{4.7}$$

or by

$$\hat{e}_{ij}^{\text{median}} = \text{median}(e_{s_1 j}, e_{s_2 j}, ..., e_{s_k j}) \ . \tag{4.8}$$

We applied the k nearest neighbor algorithm to protein expression data from a neuroblastoma DIGE study. To get an idea of how good the method works, we took a complete matrix A from which we generated an incomplete matrix B with 40% of randomly chosen missing values. The missing values where estimated with the k nearest neighbor method by using different combinations of distances (d_1, d_2, d_3) and estimators $(\hat{e}_{ij}^{\text{mean}}, \hat{e}_{ij}^{\text{wmean}}, \hat{e}_{ij}^{\text{median}})$ as well as different ks. For each estimated matrix B we calculated the normalized root mean square (RMS) error

$$\frac{\sqrt{\sum\limits_{j=1}^{m} \sum\limits_{i=1}^{n} (A_{ij} - B_{ij})^2 / (n * m)}}{\text{mean}(A)} \ , \tag{4.9}$$

to compare it to the complete matrix A. By comparing the errors for the different ways of estimation we came to the result that $\hat{e}_{ij}^{\text{mean}}$, $\hat{e}_{ij}^{\text{wmean}}$ and $\hat{e}_{ij}^{\text{median}}$ have a similar performance. Further, we found out that the error is nearly the same when the Euclidean or Mahalanobis distance is used, but it is higher when the *sup*-distance is used. For the appropriate number of neighbors, we saw that the error was smallest between 5 and 20 neighbors. We applied this missing value estimation to get a balanced data structure for the analysis of the longitudinal DIGE data using the mixed linear model described in section 3.

ACKNOWLEDGMENTS

Klaus Jung is supported by a predoctoral fellowship of the German Research Foundation.

REFERENCES

[1] Aebersold, R. and Goodlett, D.R. (2001). Mass spectrometry in proteomics, *Chemical Reviews*, **101**, 269–295.

[2] Amersham Biosciences (2003). *DeCyder Differential Analysis Software, Version 5.0, User Manual*, Amersham Biosciences, Sweden.

[3] Devroye, L.P. (1978). The uniform convergence of nearest neighbor regression function estimators and their application in optimization, *IEEE Trans. Inf. Theory*, **24**, 142–151.

[4] Diggle, P.J.; Liang, K.Y. and Zeger, S.L. (1994). *Analysis of Longitudinal Data*, Clarendon Press, Oxford.

[5] Fellenberg, K.; Hauser, N.C.; Brors, B.; Neutzner, A.; Hoheisel, J.D. and Vingron, M. (2001). Correspondence analysis applied to microarray data, *PNAS*, **98**, 10781–10786.

[6] Gannoun, A.; Saracco, J.; Urfer, W. and Bonney, G.E. (2004). Nonparametric analysis of replicated microarray experiments, *Statistical Modelling*, **4**, 195–209.

[7] Gygi, S.P.; Rist, B.; Gerber, S.A.; Turecek, F.; Gelb, M.H. and Aebersold, R. (1999). Quantitative analysis of complex protein mixtures using isotope-coded affinity tags, *Nature Biotechnology*, **17**, 994–999.

[8] Huber, W.; Heydebreck, A. von; Sueltmann, H.; Poustka, A. and Vingron, M. (2002), *Bioinformatics*, **18**, S96–S104.

[9] Karp, A.N.; Kreil, D.P. and Lilley, K.S. (2004). Determining a significant change in protein expression with DeCyder during a pairwise comparison and to the quantification of differential expression, *Proteomics*, **4**, 1421–1432.

[10] Ketskemety, L. (2004). Effectiveness of nearest neighbor classification with optimal scaling, *Alkalmazott Mat. Lapok.*, **21**, 81–97.

[11] Ripley, B.D. (1996). *Pattern Recognition and Neural Networks*, Cambridge University Press.

[12] Rosenblatt, M. (1979). Global measures of deviation for kernel and nearest neighbor density estimates. Smoothing techniques for curve estimation. *Proc. Workshop, Heidelberg, Lect. Notes Math.*, **757**, 181–190.

[13] Troyanskaya, O.; Cantor, M.; Sherlock, G.; Brown, P.; Hastie, T.; Tibshirani, R.; Botstein, D. and Altman, R.B. (2001). Missing value estimation methods for DNA microarrays, *Bioinformatics*, **17**:6, 520–525.

[14] Westermeier, R.; Gronau, S. and Beckett, P. (2004). *Electrophoresis in Practice*, Wiley-VCH, Weinheim.