REVSTAT – Statistical Journal
Volume 4, Number 1, March 2006, 67–80

STATISTICAL EVALUATION OF METHODS FOR THE ANALYSIS OF DYNAMIC PROTEIN EXPRESSION DATA FROM A TUMOR STUDY

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

Ognjan Apostolov
– Medical Proteom-Center, Ruhr-University Bochum, Germany

Alexander Schramm
– University Children's Hospital of Essen,
Division of Hematology, Oncology and Endocrinology, 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

Abstract:

- In this article, we analyze time dependent protein expression data obtained from a proteome study of a neuroblastoma cell line. Neuroblastoma are common solid tumors which occur in early childhood. The expression data was obtained by difference gel electrophoresis (DIGE). It is known that the clinical outcome of neuroblastoma depends on the activation of different neurotrophin receptors by their ligands. Here, we are looking for proteome changes resulting from the activation of Tyrosine Kinase (TrkA) receptors by their ligand NGF (nerve growth factor). Before analyzing the data by longitudinal data analysis we do data preprocessing and apply a method for the imputation of missing values.

Key-Words:

- *protein expression data; proteome; data preprocessing; missing values imputation; longitudinal data analysis; neuroblastoma.*

1. INTRODUCTION

The term ‘proteome’ stands for all proteins, which are coded by a genome at specific time points and under certain conditions. It is known that in addition to the analysis of the genome investigation of the highly complex and dynamic proteome will provide a far more detailed description of biological processes. To this end a number of proteomic techniques have been developed which allow the analysis of complex protein mixtures. Currently the two-dimensional electrophoresis (2-DE) is the separation method with highest resolution power for protein samples. Up to 10,000 proteins can be separated in one gel and therefore are accessible for quantitative analysis (cf. Klose and Kobalz ([11])). Statistical methods for the analysis of protein expression data from 2-DE comprise data preprocessing, multiple hypothesis testing and nearly the whole spectrum of multivariate techniques. In Jung et al. ([8]) we reviewed and presented some methods for data preprocessing, missing values imputation and longitudinal data analysis. Here, we apply and evaluate these techniques by analyzing protein expression data from a proteome study of the neuroblastoma cell line SY5Y. Neuroblastoma are common solid tumors which occur in early childhood. The proteome of neuroblastoma depends on the activation of different neurotrophin receptors (TrkA and TrkB) by their ligands (cf. Nakagaware et al. ([13])). In this article, we compare proteome samples of the SY5Y cell line when the TrkA receptors are activated by their ligand NGF (nerve growth factor) and when they are not activated. Hence, we have a treatment and a control group. The experiment is detailed in Sitek et al. ([16]). The protein expression in the two groups was measured at 5 time points (0, 0.5, 1, 6 and 24h) with 4 biological replicates at time 0 and 5 biological replicates at each of the other time points. The data was obtained using the latest improvement of 2-DE, the so called Difference Gel Electrophoresis (DIGE). This technique allows one to put up to three different samples on the same gel. These samples (usually treatment, control and an internal standard) are tagged by different fluorophores (Cy2, Cy3 and Cy5). The internal standard is used to standardize all gels to the same level. 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 where fluorescence intensity of a spot can be regarded as a measure of expression for its respective protein. For quantitative proteome analysis image analysis software (DeCyder V5.0, Amersham Biosciences ([3])) automatically determines the boundaries and sizes of the spots.

Our article is organized as follows. In section 2 we analyze the performance of the preprocessing methods like calibration, normalization and standardization. In section 3 we evaluate the k nearest neighbour method for the estimation of missing values with respect to an estimation error. Furthermore, we apply an analysis of variance model for longitudinal data to the neuroblastoma data in section 4 and discuss the biological implications. Finally, we will mention future challenges in statistical proteomics.

2. DATA PREPROCESSING

Before starting the actual statistical analysis of expression values from 2-D fluorescence difference gel electrophoresis (DIGE) several preprocessing steps are required. In this chapter we examine procedures for calibration, normalization and standardization of such expression values. In particular, we evaluate the performance of the preprocessing methods that were proposed by Karp et al. ([10]). The figures in this chapter are based on the measurements taken from the 'master gel' of the TrkA experiment, i.e. the gel with the greatest number of detected spots (3562, here). Nevertheless, we obtained the same results from all other gels of the experiment.

2.1. Calibration

An impression of the necessity of calibration can be obtained from figure 1 were the raw background subtracted spot volumes (that have been obtained from the DeCyder software) of the Cy2, Cy3 and Cy5 labelled samples are plotted against each other. The plots show linear dependencies between the different labelled samples. However, the point clouds appear not on the line of gradient unity, so it can be assumed that the scatter is not only due to biological variation but also to some dye effect. To remove this technical variation given by these dye effects Karp et al. ([10]) and Kreil et al. ([12]) proposed to use the calibration model

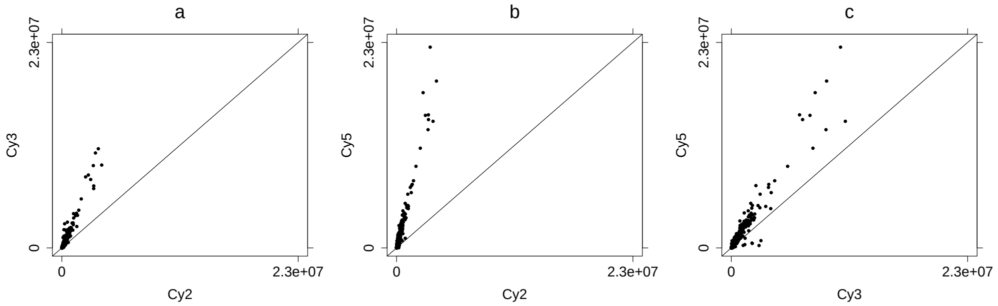

Figure 1: Raw background subtracted spot volumes spot volumes of the Cy2, Cy3 and Cy5 labelled samples plotted against each other.

$$y_{ij} = a_j + b_j \tilde{y}_{ij} \ , \tag{2.1}$$

separately for each gel, with $i=1,...,n$ and $j=1,2,3$, where $\tilde{y}_{ij}$ is the measured background subtracted spot volume of the ith spot from the sample that has been labelled with the jth dye. The calibrated value of this spot volume is y_{ij}. The dye effects are adjusted by the scaling factors b_j and the additive offsets a_j compensate for any constant additive bias present after background subtraction. This calibration model was developed by Huber et al. ([7]) for the calibration of DNA microarrays. A corresponding software package, called 'vsn', for the open source statistic software R (available at `http://cran.r-project.org`) uses a robust version of maximum likelihood estimation for the estimation of the model parameters. We will call this preprocessing method the 'vsn-method', here. After calibration the spot volumes scatter around the bisecting line (figure 2) and the scatter should now represent only the biological variation. This calibration method raises the question whether the dye effects were the same for all gels, so we compared the estimated parameters when calibrating each gel of the TrkA experiment. Table 1 shows the mean and its percentage deviation of the calibration factors and offsets for all gels of the experiment. As we can see the percentage deviations from the means are higher than 100%, so there are obviously different dye effects from gel to gel. Hence, the calibration has to be done separately for each gel.

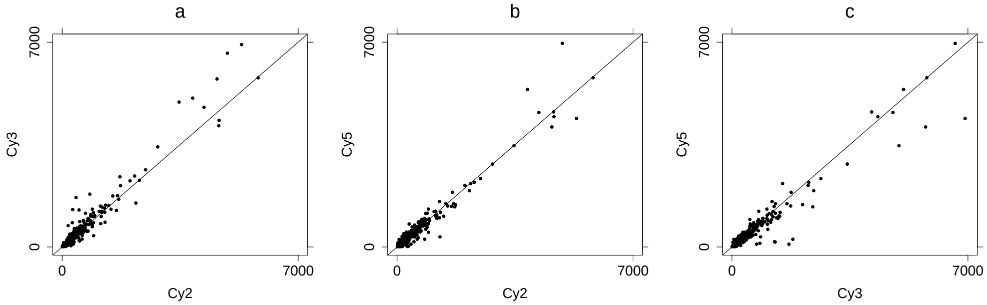

Figure 2: Spot volumes, calibrated by the vsn-method, of the Cy2, Cy3 and Cy5 labelled samples plotted against each other.

Table 1: The mean and its percentage deviation of the calibration factors and offsets, respectively, when using calibration model (2.1) for each gel of the TrkA experiment.

j	$\mu_1 = \text{mean}(a_j)$	deviation(μ_1)	$\mu_2 = \text{mean}(b_j)$	deviation(μ_2)
1	0.0006	128.0%	4.45	166.7%
2	0.0003	134.3%	6.67	155.1%
3	0.0001	125.8%	7.34	154.9%

2.2. Variance stabilization and normalization

In figure 2 it can be seen that the deviations of the spot volumes from the different labelled samples calibrated by the two methods is bigger for big values than the deviation for small values. For all five gels that have been prepared with the samples taken at time five (24h) we calibrated the expression values by the above method. From these values we calculated the mean and the variance of each spot. We analyzed only those spots which have been detected on at least three gels of time five, i.e. 1910 spots. The ranks of the means are plotted against the variances in figure 3a. Here, it can also be seen that the variance for big values is larger than the variance for small values. For this reason, in the standardisation process (cf. next section) where the internal standard is subtracted from the treatment and from the control, respectively, we also apply a transformation to stabilise the variance. One can either apply the logarithm or the asinh on the calibrated values to get a uniformly distributed variance. Figure 4 shows the calibrated spot volumes with the logarithm applied on them. However,

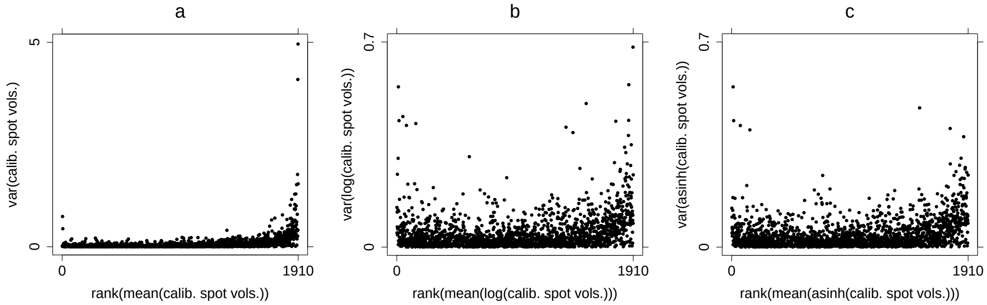

Figure 3: Rank of the mean versus the variance of a) the calibrated spot volumes, b) the calibrated and log-transformed spot volumes and c) the calibrated and asinh-transformed spot volumes.

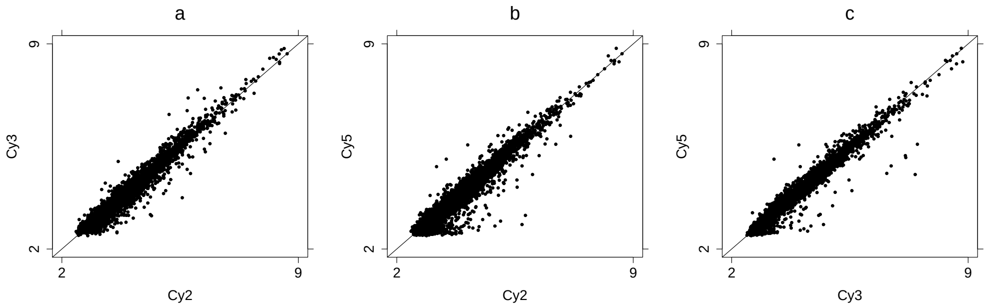

Figure 4: Calibrated and log-transformed spot volumes.

the logarithm goes very fast to $-\infty$ for small values and can thus causes a bias for small values. Instead of the logarithm one can also use the asinh. This is a function that is similar to the logarithm but smoother for small values. The calibrated and asinh-transformed values are plotted in figure 5. The effect of these transformations on the variance-mean-dependencies can be seen in figure 3. Fig. 3b and c show that after applying the logarithm or the asinh transformation to the calibrated values the variance is stabilised with respect to the mean.

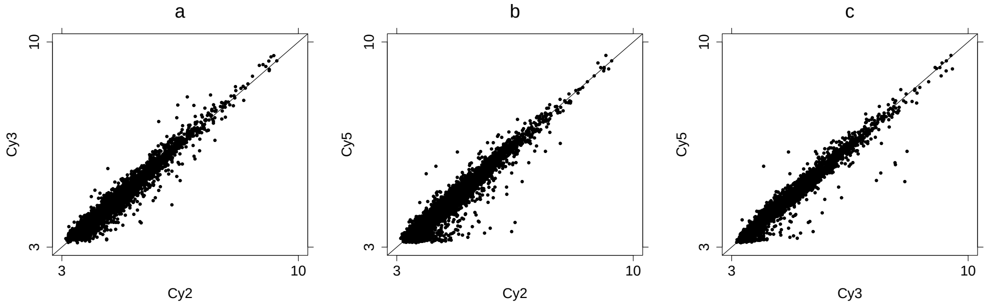

Figure 5: Calibrated and asinh-transformed spot volumes.

2.3. Standardization

The benefit of the DIGE method is to have an internal standard on each gel. The internal standard is a sample consisting of aliquots from all other samples of the experiment. Subtracting the values of the internal standard from values from the treated and untreated samples brings all gels on the same level and thus reduces the gel-to-gel variance. The complete preprocessing for the treatment values is thus given by either

$$\log(a_2 + b_2\,\tilde{y}_{i2}) - \log(a_1 + b_1\,\tilde{y}_{i1})\ , \tag{2.2}$$

or by

$$\operatorname{asinh}(a_2 + b_2\,\tilde{y}_{i2}) - \operatorname{asinh}(a_1 + b_1\,\tilde{y}_{i1})\ , \tag{2.3}$$

and similarly for the control values. In figure 6 the density histogram of the vsn-processed and standardized values for the treatment values is given. This distribution is symmetric and nearly normally distributed as can also be seen in the QQ-plot.

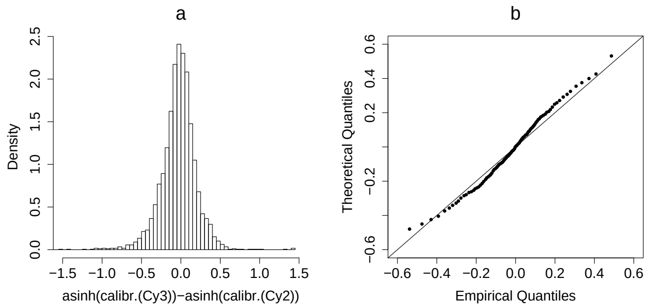

Figure 6: a) Density histogram of the preprocessed spot volumes from the treatment sample. b) QQ-plot from these values.

3. ESTIMATION OF MISSING VALUES

Many statistical methods, especially those for multivariate data, are based on the assumption that the data set to be analyzed is complete. However, in 2-D DIGE studies with gel replicates between 10 to 30 % of the values are missing. This is due to the fact that not all spots are detected or matched on each gel. In this section we compare two methods for the estimation of missing values, the 'row-mean method' and the 'k-nearest neighbour ($k$nn) method'. The latter one has already been successfully applied to microarray data (cf. Troyanskaya et al. ([17])). To illustrate these two methods we consider a simple example with artificial data of six spots on four gels (cf. table 2). In this example spot 2 on gel 3 has not been detected, so the value is missing. The row mean method simply uses the average of all available measurements in the row where the value is missing as estimator for this missing value. For the example in table 2 the estimated value for spot 2 on gel 3 is then $(26.43 + 18.07 + 23.84)\,/\,3 = 22.78$. The underlying idea of the $k$nn method is that there is a relationship between the expression profiles of some proteins. So, if a value for spot x is missing, the

Table 2: Artificial gel data with a missing value for spot 2 on gel 3.

	gel 1	gel 2	gel 3	gel 4
spot 1	24.21	28.87	21.59	22.79
spot 2	26.43	18.07		23.84
spot 3	238.42	270.97	258.74	233.63
spot 4	27.53	30.05	25.35	28.50
spot 5	132.58	152.61	144.09	148.82
spot 6	250.41	277.93	273.65	264.53

method uses the values from those spots which are strongly related to this spot x. To determine the relationships of spot x to all other spots of the data set one can use a distance measure like the Euclidean, the Mahalanobis or the Chebyshev distance (cf. Jung et al. ([8])). If the value for spot x is missing on gel y, these distances are calculated by using only the values from the gels other than gel y. In the example, we use the values from gel 1, 2 and 4 to calculate the distances of spot 2 to all other spots. Spots 1 and 4 have a very short distance to spot 2, here, so we take the values from spot 1 and 4 on gel 3 to estimate the missing value, for example by taking the average of these values: $(21.59 + 25.35) / 2 = 23.47$. Other possible estimators are the median or some weighted mean. An important question that appears when using the $k$nn method is, how many neighbours should be used for the estimation. To determine the estimation error we used the five gels from the fifth time point in the TrkA experiment, removed all rows with missing values, so that a complete data set A with 526 rows and 5 columns remained. From this data set we generated 4 incomplete data sets B1 to B4 with 5, 10, 20 and 30 % of randomly chosen missing values, respectively. Then we applied the $k$nn method using different numbers k of neighbours. The resulting filled up data sets C1 to C4 were then compared to the original complete data set A by calculating the normalized root mean square (RMS) error:

$$\text{(3.1)} \qquad \textit{normalized RMS error} = \frac{\sqrt{\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{m}(A_{ij} - C_{ij})^2/(n * m)}}{\sum\limits_{i=1}^{m}\sum\limits_{j=1}^{m} A_{ij}/(n * m)} ,$$

where n is the number of spots and m is the number of replicates. A plot of this error is given in figure 7. Plot 7a shows the error when using the $k$nn method

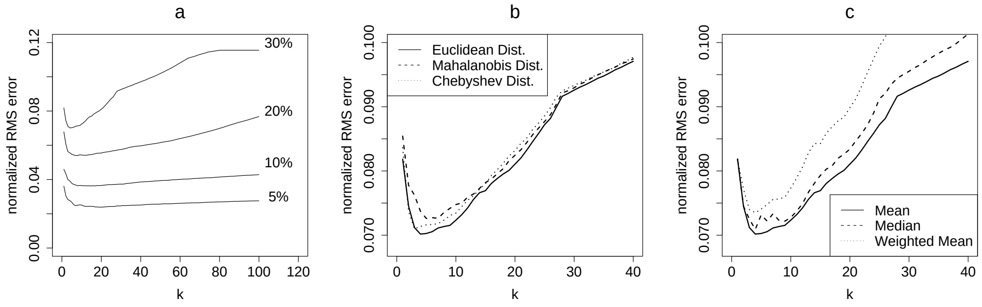

Figure 7: a) Normalized RMS error in dependence of k. The $k$nn method applied a) to data with different proportions of missing values, b) with different distance measures and c) with different missing values estimators.

with the Euclidean distance and the mean applied to data sets with different proportions of missing values. The error increases with increasing proportion of missing values and the minimum of the curves is between 5 and 20 neighbours. We compared also the performance of the difference measures (figure 7b) and of the estimators (figure 7c). For both plots a data set with 30% of missing values was analyzed. Furthermore, for figure 7b the mean was used as missing values estimator and for figure 7c the Euclidean distance was used as distance measure. These plots show that using the Euclidean distance is slightly better than using the Mahalanobis or the Chebyshev distance and that the mean is a better estimator than the median or a weighted mean. We obtained the same results from plots with other combinations of difference measures and estimators. Compared to the row-mean method the $k$nn method results in smaller errors. The minima of the error-curves of figure 7a and the errors of the row-mean method are given in table 3. As further research activity it would also be of interest to compare the $k$nn method to other imputation methods given in Nguyen et al. ([14]). They use for example Partial Least Square (PLS) regression to impute missing values.

Table 3: Comparison of the normalized RMS error when using the row mean method and the $k$nn method, respectively.

proportion of missing values	5%	10%	20%	30%
row mean error	0.13	0.19	0.26	0.32
min ($k$nn error)	0.02	0.04	0.05	0.07

4. LONGITUDINAL DATA ANALYSIS

4.1. Analysis of variance

Before doing the statistical analysis we preprocessed the data by the vsn-method described in chapter 2. We also filled up the missing values by the $k$nn method described in chapter 3. The interest of the statistical analysis was to find those proteins for which the expression profiles over the time were different in the treated and untreated sample, respectively. In order to find differences in the temporal course of the treated and untreated samples we used an analysis of variance model for longitudinal data (cf. Jung et al. ([8]) and Diggle et al. ([5])). Such a model should take the time dependence of the measurements into account. Using F-tests one can detect time/treatment-interactions of spots. For our analysis we used only those spots for which at least three values were available at each time point. The detected significant spots are presented in table 4. The p-values in this table are not corrected for multiple testing (cf. Dudoit et al. ([6])), because the number n of spots is not clearly fixed in a 2-DE experiment. Biochemist often decide to exclude a great number of spots still after the statistical analysis.

Table 4: Spots with a time/treatment interaction.

rank	spot-no.	p-value
1	1136	0.0023
2	910	0.0055
3	988	0.0075
4	1669	0.0255
5	1054	0.0428

The expression profiles of the two most significant spots are plotted in figure 8. Both spots have a similar expression at the beginning of the experiment and the profiles drift at the end.

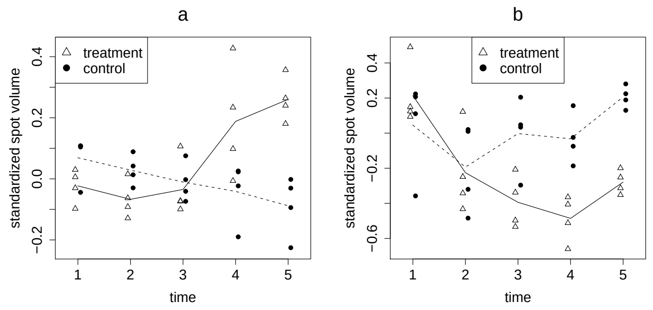

Figure 8: a) Temporal courses of the expression values of the spot numbers 910 (a) and 1136 (b) in the treated (solid line) and untreated (dashed line) samples.

4.2. Biological implications

This biological experiment was performed to identify candidate proteins contributing to neuroblastoma clinical outcome. We identified 5 proteins with a time/treatment-interaction upon addition of neurotrophin in SY5Y-TrkA cells (table 4). These proteins were identified using MALDI (matrix assisted laser desorption and ionization) mass spectrometry. For instance protein 910 with a changed temporal course after neurotrophin receptor activation consists to a family of heat shock proteins known to be involved in a number of cellular processes. Regarding cancer research the increased expression of heat shock protein 70 (Hsp 70) has been reported in a variety of tumor tissues. Hsp 70 has also been detected in plasma and therefore could potentially be used as a biomarker for diagnosis. It has been demonstrated, that patients suffering from prostate cancer have an increased level of Hsp 70 in the blood plasma (cf. Abe et al. ([1])). Based on this knowledge Hsp 70 could be a candidate tumor marker for neuroblastoma. To test this hypothesis further experiments have to be performed.

5. FUTURE CHALLENGES IN STATISTICAL PROTEOMICS

The statistical analysis of protein expression data is similar to the analysis of gene expression data from DNA microarrays. A future challenge for statisticians is the adaption of the methods for the analysis of gene expression data to be applicable to protein expression data. An important question of genomics was to find genes with differential expression in samples from different tissue types (cf. Jung et al. ([9])). Statistical tools for this purpose can also be applied to protein expression data from 2-DE when having estimated the missing values before. Of interest are also protein expression data from mass spectrometry (cf. Aebersold and Goodlett ([2]) and Pusch et al. ([15])). Statistical applications for those data span the whole range of multivariate methods like classification problems or multivariate outlier detection.

Furthermore, interactions between biomolecules are important in many important processes, such as cell proliferation and cell signalling. When pathogens (e.g. bacteria) attack our body, it responds by producing many antibodies. They bind to a part of pathogen, called antigen. Biochemists have studied how and where a given antibody binds to an antigen by investigation of a single point mutant of the antibody. Andersson ([4]) describes a different strategy for such mutation experiments. Instead of mutating each antibody at one position only, several modifications are made in the same antibody. Using statistical tools like Partial Least Squares regression he could find out which modification was relevant for establishing the binding. Also the investigation of the impact of environmental changes on the binding strength of an antibody-antigen interaction is important for antibodies used in diagnostic tests for cancer.

In both situations the binding properties of the interaction of biomolecules can be characterized by association and dissociation rates. These parameters can be measured by surface plasmon resonance detectors. New multivariate methods should be developed to analyze the relationships between these kinetic parameters and all the factors that influence these measures and to predict the kinetics of biomolecular interactions for new combinations of explanatory variables. There is also need for new statistical tools which allow the inclusion of structural and sequence information from nuclear magnetic resonance spectra and Fourier transform ion cyclotron resonance mass spectra for generating new biological and clinical knowledge. Such extensions of available methods could be of considerable importance in drug development for improving the binding of a drug to the desired target and for decreasing unwanted side reactions.

ACKNOWLEDGMENTS

Klaus Jung is supported by a predoctoral fellowship of the German Research Foundation (DFG). This work was also sponsored by the 'Optimization Fund' within the German National Genome Research Network (NGFN) and the Ministry of Innovation, Science and Research and Technology in North Rhine-Westfalia.

REFERENCES

[1] Abe, M.; Manola, J.B.; Oh, W.K.; Parslow, D.L.; George, D.J.; Austin, C.L. and Kantoff, P.W. (2004). Plasma levels of heat shock protein 70 in patients with prostate cancer: a potential biomarker for prostate cancer, *Clinical Prostate Cancer*, **3**, 49–53.

[2] Aebersold, R. and Goodlett, D.R. (2001). Mass Spectrometry in Proteomics, *Chemical Reviews*, **1**, 269–295.

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

[4] Andersson, K. (2004). Characterization of Biomolecular Interactions Using a Multivariate Approach, *Acta Universitatis Upsaliensis, Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine 1363*, Uppsala.

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

[6] Dudoit, S.; Shaffer, J.P. and Boldrick, J.C. (2003). Multiple hypothesis testing in microarray experiments, *Statistical Science*, **18**, 71–103.

[7] Huber, W.; Heydebreck, A. von; Sueltmann, H.; Poustka, A. and Vingron, M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression, *Bioinformatics*, **18**, S96–S104.

[8] Jung, K.; Gannoun, A.; Stühler, K.; Sitek, B.; Meyer, H.E. and Urfer, W. (2005). Analysis of dynamic protein expression data, *RevStat-Statistical Journal*, **3**, 99–111.

[9] Jung, K.; Quast, K.; Gannoun, A. and Urfer, W. (2006). A renewed approach to the nonparametric analysis of replicated microarray experiments, *Biometrical Journal*, to appear.

[10] 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.

[11] Klose, J. and Kobalz, U. (1995). Two-dimensional electrophoresis of proteins: an updated protocol and implications for a functional analysis of the genome, *Electrophoresis*, **16**, 1034–1059.

[12] Kreil, D.P.; Karp, A.N. and Lilley, K.S. (2004). DNA microarray normalization methods can remove bias from differential protein expression analysis of 2D difference gel electrophoresis results, *Bioinformatics*, **20**, 2026–2034.

[13] Nakagawara, A.; Azar, C.G.; Scavarda, N.J. and Brodeur, G.M. (1994). Expression and function of Trk-B and BDNF in human neuroblastomas, *Molecular and Cellular Biology*, **14**, 759–767.

[14] Nguyen, D.V.; Wang, N. and Carrol, R.J. (2004). Evaluation of missing value estimation for microarray data, *Journal of Data Science*, **2**, 347–370.

[15] Pusch, W.; Flocco, M.T.; Leung, S.-M.; Thiele, H. and Kostrzewa, M. (2003). Mass spectrometry-based clinical proteomics, *Pharmacogenomics*, **4**, 463–476.

[16] Sitek, B.; Apostolov, O.; Stühler, K.; Pfeiffer, K.; Meyer, H.E.; Eggert, A. and Schramm, A. (2005). Identification of dynamic proteome changes upon ligand-activation of Trk-receptors using two-dimensional fluorescence difference gel electrophoresis and mass spectrometry, *Molecular and Cellular Proteomics*, **4**, 291–299.

[17] 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.