Bioinformatics Advance Access published February 24, 2014

BIOINFORMATICS ORIGINAL PAPER

2014, pages 1–7
doi:10.1093/bioinformatics/btu062

Gene expression

Advance Access publication January 30, 2014

Adaption of the global test idea to proteomics data with missing values

Klaus Jung[1,*], Hassan Dihazi[2], Asima Bibi[2], Gry H. Dihazi[2] and Tim Beißbarth[1]

[1]Department of Medical Statistics, and [2]Department of Nephrology and Rheumatology, University Medical Center Göttingen, Göttingen 37099, Germany

Associate Editor: Ziv Bar-Joseph

ABSTRACT

Motivation: Global test procedures are frequently used in gene expression analysis to study the relationship between a functional subset of RNA transcripts and an experimental group factor. However, these procedures have been rarely used for the analysis of high-throughput data from other sources, such as proteome expression data. The main difficulties in transferring global test procedures from genomics to proteomics data are the more complicated way of obtaining functional annotations and the handling of missing values in some types of proteomics data.

Results: We propose a simple mixed linear model in combination with a permutation procedure and missing values imputation to conduct global tests in proteomics experiments. This new approach is motivated by protein expression data obtained by means of 2-D gel electrophoresis within a mouse experiment of our current research. A simulation study yielded that power and testing level of the mixed model alone can be affected by missing values in the dataset. Imputation of missing values was able to correct for a bias in some simulation settings. Our new approach provides the possibility to rank Gene Ontology (GO) terms associated with protein sets. It is also helpful in the case in which a specific protein is represented by multiple spots on a 2-D gel by considering these spots also as a protein set. Analysis of our data points at correlations between the deficiency of the protein 'calreticulin' and protein sets related to biological processes in the heart muscle.

Availability and implementation: Our proposed approach is included in the R-package 'RepeatedHighDim', which already contains a global test procedure for gene expression data. The package can be retrieved from http://cran.r-project.org/.

Contact: klaus.jung@ams.med.uni-goettingen.de

Received on October 4, 2013; revised on December 20, 2013; accepted on January 26, 2014

1 INTRODUCTION

Can certain molecular functions, cellular components or biological processes be related to the levels of the grouping factor of a study or an experiment? This question is frequently studied in the analysis of expression data from high-throughput experiments (e.g. with DNA microarrays). To answer this question, expression levels of those features that are connected to a particular molecular function can be compared between the different study groups. With this approach, Grond-Ginsbach *et al.* (2008), for example, detected a relation between a set of genes involved in the molecular mechanisms of inflammation and acute ischemic stroke. In another study, Groene *et al.* (2006) detected a relation between genes involved in the p53 pathway and the Union for International Cancer Control (UICC) stages of colorectal cancer. Statistically, the previously mentioned questions can be analyzed by means of global test procedures. In contrast to methods that analyze the expression levels of all features from a high-throughput experiment individually (Smyth, 2004), a global test focuses on sets of features that are all involved in the same biological function or cellular pathway. Although a number of global test approaches for the analysis of functional gene sets have been published (Goeman *et al.*, 2004; Jung *et al.*, 2011; Mansmann and Meister, 2005), the idea was rarely considered in protein expression data from proteomics experiments. In such experiments, protein expression is usually measured in a high-throughput manner by either mass spectrometry (MS) (Aebersold and Mann, 2003) or 2-D gel electrophoresis (2-DE) (Klose and Kobalz, 1995). For protein expression data measured by MS, a global test approach based on Hotelling's T^2-statistic to relate specific phenotypes to functionally related sets of proteins was proposed by Chen *et al.* (2011). Besides, an approach for testing protein set enrichment in MS experiments was presented (Louie *et al.*, 2010).

In this article, we present a new approach for global testing of functional protein sets in the case where expression levels are measured by 2-DE. Although the MS-based protein expression data referred to by Chen *et al.* (2011) present as a complete $(d \times n)$-matrix, i.e. without missing values, up to 30% of entries may be empty in the matrix of 2-DE-based expression data (Jung *et al.*, 2006). However, missing values are also not atypical in MS-based data because of missing peaks in the mass spectra (Käll and Vitek, 2011; Karpievitch *et al.*, 2012; Smith *et al.*, 2006). Here, d denotes the number of proteins and n the sample size. Missing values in protein expression data from 2-DE are because features on a 2-D gel are not placed on an ordered grid like the probe sets on a DNA microarray where the location of each particular feature is known. In contrast, spots of labeled proteins appear at more or less different locations on 2-D gels, and spot matching algorithms must be used to bring the information of the experimental replications in line (Xin and Zhu, 2009). For a number of protein spots the matching fails, and missing values 'occur'.

Therefore, a global test procedure for proteomics data from 2-DE has to deal with two main difficulties. On the one hand, there is an issue of missing values, and on the other hand, it is more complicated to get functional annotations for the protein

*To whom correspondence should be addressed.

spots, as these spots must first be removed from the gel and the underlying protein must be identified by MS (Koenig *et al.*, 2008)—unlike DNA microarray data, where the identity of a probe set is given directly. Not specific for 2-DE data, the problem for all high-throughput data is that the number d of features in a functional set can be much larger than that of the sample size n. The new approach we present here is based on a simple mixed linear model, which is able to handle missing values, in combination with a permutation algorithm to account for the high dimensionality. We evaluate this approach when being applied on the incomplete data on the one hand and on data filled up by a missing value imputation procedure proposed by Troyanskaya *et al.* (2001) on the other hand.

Our article is structured as follows. We first detail the mixed linear model, the permutation procedure and the missing value imputation algorithm, followed by a description of our proteomics experiment. Next, we evaluate the methods in a simulation study and apply them to the data example. Finally, we provide a discussion of the results and give some conclusions.

2 METHODS AND EXAMPLE DATA

2.1 Simple mixed linear model

Our approach is based on a simple mixed linear model of the following form:

$$Y_{ijk} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} + U_i + \varepsilon_{ijk}, \quad (1)$$

where Y_{ijk} represents the expression level of protein k in group j and individual i after normalization and variance stabilization. The model is composed of an overall mean μ, the effect α_j of group j ($j = 1, 2$), the effect β_k of protein k ($k = 1, \ldots, d$) and a (group × protein)-interaction $(\alpha\beta)_{jk}$. In addition, the model contains a random effect $U_i \sim N(0, \sigma_j)$ for the ith individual ($i = 1, \ldots, n$) and an overall random error $\varepsilon_{ijk} \sim N_d(0, Z_j)$. Classically, the aforementioned model would assume that there are less proteins than individuals, i.e. $d < n$.

To test the hypothesis that the mean expression profile of the d proteins is the same in both groups, one would test the null hypothesis of no interaction effect, i.e. $H_0 : (\alpha\beta)_{jk} = 0$ for all j, k. For example, this hypothesis can be tested by means of likelihood ratio tests (Faraway, 2006).

To ensure that the data meet the linearity assumption of the model raw 2-DE data should be pre-processed by normalization and transformed by some variance stabilizing function such as the logarithm or the arsinh (Huber *et al.*, 2002; Kreil *et al.*, 2006).

2.2 Permutation procedure

Under unequal covariance matrices for the random errors, i.e. if $Z_1 \neq Z_2$, or if sample sizes are unbalanced, the application of model (1) can fail to maintain the pre-specified level of significance, particularly if the number d of proteins is much larger than that of the samples size n, as typical for high-throughput data. As a consequence, the true testing level would often become too much liberal yielding to many false-positive test results. Therefore, we embed this model into a permutation procedure to correct for this bias. The working principle of the permutation procedure is as follows. Assume, the test of H_0 on the unpermuted data yields the P-value p_{data}, whereas the tests on B permuted datasets yield the P-values p_b ($b = 1, \ldots, B$). The permutation P-value is then given by $p_{perm} = \#\{p_b < p_{data}\}/B$ (Efron and Tibshirani, 1993).

2.3 Missing values imputation

As incomplete datasets are typical in gel-based proteomics experiments, a loss of power is to be expected for the aforementioned testing procedures. Therefore, we propose an approach of combining the mixed model not only with a permutation procedure but also with a missing values imputation algorithm. For that purpose we use an existing algorithm based on the idea of nearest neighbors. This algorithm determines neighbors of the protein with the missing value for subject i in the sense of correlated expression levels among all other individuals. The missing value is then determined, for example, by the mean expression levels in subject i from the neighboring proteins. This approach has originally been proposed for missing values imputation of DNA microarray data (Troyanskaya *et al.*, 2001) and has already been applied to gel-based proteomics data (Jung *et al.*, 2006).

2.4 Case study: proteomics data from 2-DE

As a case example we studied protein expression data from 2-DE, generated as part of our current proteomics research. Specifically, we focus on the role of the protein 'calreticulin' in the heart muscle. To this end, we compared three calreticulin heterozygote (Calr+/−) and three wild-type (WT) littermate mice in identical C57BL/6J genetic background. Animals were obtained from Prof. Marek Michalak, University of Alberta, Edmonton, Alberta, Canada. Mice were bred under specific pathogen-free housing conditions and genotyped as previously described in Michalak *et al.* (1999). All experimental procedures were performed according to the German animal care and ethics legislation and were approved by the local government authorities at the University Medical Centre Götttingen.

Immediately after cervical dislocation, the freshly excised hearts from adult mice (WT, Calr+/−) were quickly removed, cleaned, washed in sterile saline solution and weighed. Mice hearts were homogenized in buffer containing 50-mmol/l Tris-HCl (pH 7.4), 1% Triton X-100, 100-mmol/L NaCl and protease inhibitors. After incubation for 30 min at 4 °C, heart tissue homogenates were centrifuged two times at 14 000 rpm for 30 min, and the supernatant was collected. To reduce the salt contamination and to enrich the proteins, protein precipitation was performed. Whole tissue homogenate was precipitated using methanol-chloroform.

Next, processed samples were analyzed by means of 2-DE. In total, six gels were run where three of them were prepared with WT samples and the other three with Calr+/− samples. After scanning and image analysis (using the software Delta2D, version 4, Decodon), 103 matched gel spots were identified by MS and assigned a UniProt accession number (UniProt Consortium, 2007). Probably because of alternative splicing and post-translational modifications, some of the spots carried the same protein so that 63 proteins could be related to the 103 spots. In detail, 37 proteins were represented by only 1 spot, 16 by 2 spots, 8 by 3 spots, 1 ('ATP synthase subunit beta') by 4 spots and 1 ('electron transfer flavoprotein subunit alpha') by 6 spots.

Table 1. Frequency of functional protein sets sizes in the proteomics case example

Ontology domain	Set size d (absolute frequency)
Biological process	1 (136), 2 (73), 3 (26), 4 (6), 5 (6), 6 (5), 7 (2), 8 (1), 9 (4), 10 (2), 11 (1), 12 (1), 30 (1)
Cellular component	1 (29), 2 (31), 3 (9), 4 (1), 5 (1), 6 (2), 7 (2), 8 (1), 9 (4), 11 (1), 12 (1), 13 (1), 17 (1), 18 (1), 33 (1), 38 (1)
Molecular function	1 (47), 2 (16), 3 (16), 4 (4), 5 (2), 6 (6), 7 (3), 8 (1), 9 (2), 11 (1), 23 (1)

Note: Most of the functional terms were related to only small number of gel spots. For example, Each 136 biological processes were related to only one gel spot, while the largest set was given by a cellular component that was related to 38 gel spots.

Via the accession number, the identified spots could further be linked to 264 different biological processes, 87 different cellular components and 99 different molecular functions. The most frequent biological process 'transport' was related to 30 gel spots; the second frequent one 'glycosis' was related to 12 gel spots. Concerning cellular components, the most frequent component 'mitochondrion' was related to 38 gel spots. The most frequent molecular function 'ATP binding' was related to 23 gel spots. The complete distribution of functional set sizes is given in Table 1.

Approximately 3% of values were missing in the (103×3)-data matrix of the untreated samples, and ~15% were missing in the matrix of the treated samples.

3 RESULTS

3.1 Simulation studies

To study the effect of missing values on the testing level and the power of the proposed global test approach, we performed two different series of simulation studies. In the first one, we simulated expression data with certain types of arranged covariance matrix (autoregressive or unstructured scenarios A1–A4), and we used a shrinkage covariance estimate (Schäfer and Strimmer, 2005) from the example data in the second one (scenarios B1–B6). The shrinkage method allows the estimation of a covariance matrix in the case of high-dimensional data. Besides, we varied, in both simulation studies, the proportion of missing values, the size d of the protein sets and the sample sizes per group. Expression levels were drawn from the multivariate normal distribution.

The off-diagonal entries of the autoregressive covariance matrices were calculated by $\sigma_{jj'} = \rho^{\lfloor i-i' \rfloor}$ $(i, i' = 1, \ldots, d)$ with $\rho = 0.5$ and those of the unstructured covariance matrices were drawn from the standard normal distribution. In both cases, the diagonal elements were increased evenly from 1 to 2 to have unequal variances for the proteins.

In the case of the arranged covariance matrices, log-fold changes were drawn from normal distributions $N(0, \delta)$, where δ started with 0 and was increased until a power of 1 was reached in each simulation scenario. In the simulations with shrinkage covariance matrices, the log-fold changes were all zero when simulating the null hypothesis and were drawn from the fold change distribution of the real data example when simulating global effects. For smaller group effects, log fold changes were simulated as a fraction of the distribution of the real data, and for larger effects, log fold changes were a multiple of this distribution.

In each simulation setting, the 'optimal' number of neighbors was assessed individually for each proportion of missing values within a further simulation loop, where the imputed matrix was compared with the original complete matrix using the normalized root mean squared error as proposed by Troyanskaya *et al.* (2001). In general, the number of neighbors increased with the proportion of missing values. For 10% of missing values, the mean number of neighbors over all settings was 24 (34 and 41 in the cases of 20% and 30% of missing values, respectively).

3.1.1 Arranged covariance structures Four different scenarios with arranged covariance structures were simulated. Under scenarios A1 and A2, autocorrelated covariance structures were simulated, where group sizes were equal in scenario A1 (i.e., $n_1 = n_2 = 10$, $n_1 + n_2 = n$) and unequal in scenario A2 ($n_1 = 5$ and $n_2 = 15$). Expression data for scenarios A3 and A4 were simulated with unstructured covariance matrices with sample sizes $n_1 = 5$ and $n_2 = 15$ in scenario A3 and to study especially the case of small sample sizes $n_1 = n_2 = 4$ in scenario A4. The set size $d = 100$ was fixed in all scenarios A1–A4.

In these four scenarios, the simple model without permutation procedures yielded completely unacceptable testing levels ranging from 32% to 39% false-positive rejections of H_0 in scenarios A3 and A4 and ~70% in scenarios A1 and A2. Applying instead the mixed model in combination with the permutation procedure, the pre-specified testing level of 5% was sufficiently maintained in all scenarios, in the case of no missing values (Table 2). The introduction of missing values did not seriously affect the testing levels. Imputation of missing values by the k-nearest neighbor method, however, led to a decrease of testing levels in scenarios A2 and A3, down to 2.7% in A2 and down to 3.6% in A3.

To study the influence of missing values onto the power, we selected that alternative δ for the log-fold changes where each scenario reached 80% power in the case when there were no missing values in the data (Figs. 1 and 2). Under scenario A1, there was only a small loss of power because of the missing values, and imputation did not correct for this loss. In scenarios A2 and A3 (unequal sample sizes), the loss of power through the 'existence' of missing values was much stronger, but the imputation procedure could correct for this loss. A strong power decrease was also observed in scenario A4 (small group sizes); however, the imputation approach could not help to compensate for this loss.

3.1.2 Shrinkage covariance estimate An overview of samples sizes for simulation scenarios B1–B6 is provided in Table 3. In these scenarios, the set size d was 10, 50 or 100. When covariance matrices were estimated by the shrinkage approach from the real data and no permutation procedure was applied to linear model analysis, simulated testing levels were either too small or too large in most settings (Table 3). Only in the cases with set sizes of $d = 50$, simulated testing levels were near the required 5%.

Table 2. Simulated testing levels for simulation scenarios with arranged correlation structures

Scenario	Imputation	Covariance structure	Sample sizes	Missing values (%)			
				0	10	20	30
A1	False	Autoregressive	$n_1 = n_2 = 10$	0.046	0.050	0.053	0.049
	True				0.049	0.053	0.053
A2	False	Autoregressive	$n_1 = 5, n_2 = 10$	0.049	0.051	0.046	0.040
	True				0.045	0.038	0.027
A3	False	Unstructured	$n_1 = 5, n_2 = 10$	0.053	0.052	0.051	0.048
	True				0.044	0.042	0.036
A4	False	Unstructured	$n_1 = n_2 = 4$	0.054	0.058	0.054	0.054
	True				0.047	0.045	0.043

Note: Each scenario was either simulated with or without missing values imputation. Only results for the combined approach (linear model plus permutation procedure) are shown.

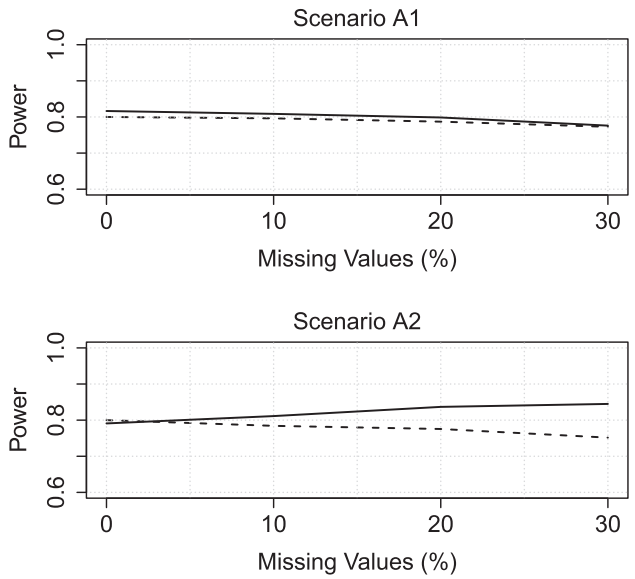

Fig. 1. Power versus proportion of missing values in simulation scenarios A1 and A2 with arranged covariance matrices. Power was studied under that alternative where it reached the 80% level in the case of no missing values. Dashed line: without missing values imputation; solid line: after missing values imputation

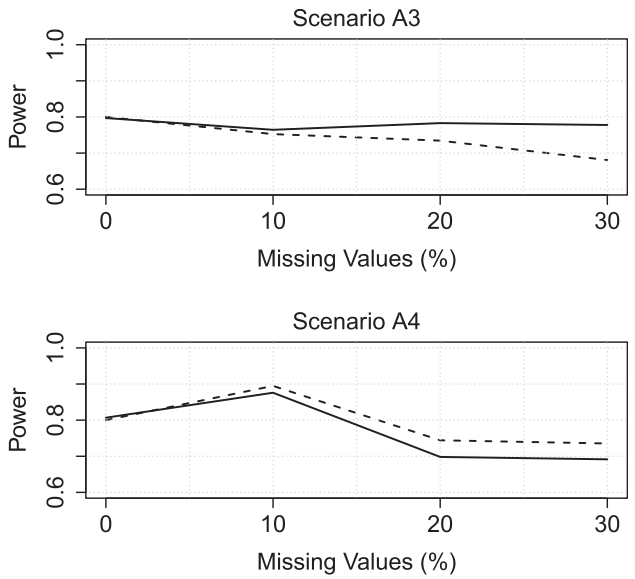

Fig. 2. Power versus proportion of missing values in simulation scenarios A3 and A4 with arranged covariance matrices. Power was studied under that alternative where it reached the 80% level in the case of no missing values. Dashed line: without missing values imputation; solid line: after missing values imputation

In most cases without permutation procedure, simulated testing levels increased with an increasing proportion of missing values. Imputation of these missing values could not correct for this bias.

When the linear model analysis was combined with the permutation procedure, the pre-specified testing level was maintained in most cases (Table 3). There were only some smaller deviations, yielding testing levels of ~4% (B6) or 5.5% (B2). The introduction of missing values led either to small decreases (B1, B2, B4) or to small increases (B3, B5, B6) of the testing levels. Imputation of these missing values led to a clear improvement in scenarios B2 and B6. However, for scenarios B4 and B5 (i.e. with unequal sample sizes and small or moderate set sizes d), the imputation of missing values was counterproductive.

In most of the scenarios with shrinkage covariance estimates, power was not seriously affected by the introduction of missing values. A strong decrease in power could, however, be observed in scenario B5, i.e. with a set size of $d = 50$ and unequal samples sizes. Missing value imputation could not adjust for this bias (Fig. 3).

3.2 Analysis of case study

To detect differences between the Calr+/− and the WT mice, we first compared each protein spot individually between the two experimental groups by means of Welch's t-test, yielding no significant result.

Table 3. Simulated testing levels for scenarios with shrinkage covariance estimated from the real data example

Scenario	Set size	Sample sizes	Missing values (%)	With permutation		Without permutation	
				With imputation	Without imputation	With imputation	Without imputation
B1	10	$n_1 = n_2 = 5$	0		0.050		0.083
			10	0.058	0.054	0.080	0.069
			20	0.056	0.047	0.098	0.108
B2	50		0		0.055		0.055
			10	0.052	0.052	0.046	0.041
			20	0.050	0.045	0.045	0.044
B3	100		0		0.054		0.007
			10	0.058	0.057	0.006	0.006
			20	0.064	0.062	0.015	0.015
B4	10	$n_1 = 5, n_2 = 10$	0		0.054		0.092
			10	0.063	0.043	0.109	0.089
			20	0.072	0.046	0.109	0.075
B5	50		0		0.047		0.053
			10	0.065	0.054	0.059	0.065
			20	0.084	0.066	0.078	0.079
B6	100		0		0.040		0.011
			10	0.052	0.045	0.013	0.013
			20	0.060	0.062	0.015	0.006

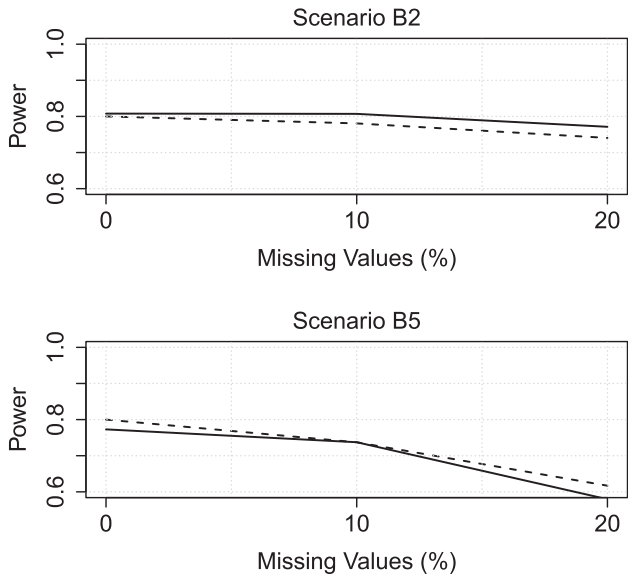

Fig. 3. Power versus proportion of missing values in simulations with shrinkage covariance matrices. Power was studied under that alternative where it reached the 80% level in the case of no missing values. Dashed line: without missing values imputation; solid line: after missing values imputation

As next step, each protein set, according to the functional terms of the different ontology domains, was analyzed individually. The top five terms in each ontology domain, i.e. with the smallest *P*-value, are listed in Table 4. The sizes of the listed sets were rather small, ranging from 2 to 11 protein spots. Missing values were only 'present' in the Calr+/− group with proportion from 11% to 50%. In the biological process domain, the top three terms were related to cardiac processes. Among the cellular components, only the term 'actomyosin actin part' yielded a $P < 5\%$ level. None of the molecular function terms reached a $P < 0.05$.

As several gel spots were assigned to the same protein, these spots were also analyzed as set (lower part of Table 4). Among these, 'Actin, alpha cardiac muscle 1' yielded a *P*-value of 0.02. Missing values were again only on the treatment group.

4 DISCUSSION

The 2-DE is widely used in proteomics research. To date, gel spots are typically compared between the experimental groups only individually; also, several spots may belong to the same protein or to the same functional term related to some biological processes, cellular component or molecular function. Like in DNA microarray analysis, global tests for comprehensive testing of functional subsets of genes are not available for gel-based proteomics data. Furthermore, most global tests used for the analysis of genomics data do not allow for missing values as typical in gel-based or MS-based protein expression data. In this regard, we studied the applicability of a simple global test procedure to such proteomics data and evaluated the effect of missing values onto the testing level and the power.

Our simulations with arranged and real-world covariance matrices have shown, that in combination with a permutation procedure, a pre-specified testing levels can be maintained in diverse scenarios. Permutation approaches were also successfully used in global test procedures in genomics to correct for biases in the testing level (Goeman *et al.*, 2004; Mansmann and Meister, 2005). With regard to our simulation results, we would generally

Table 4. Global test results in the real data example

Ontology domain	Description	Set size	P	Missing values (%)	
				WT	Calr+/−
Biological process	Cardiac muscle contraction	2	0.01	0	0
	Cardiac muscle tissue morphogenesis	2	0.01	0	0
	Cardiac myofibril assembly	2	0.03	0	0
	Skeletal muscle thin filament assembly	2	0.03	0	0
	Actin–myosin filament sliding	2	0.05	0	0
Cellular component	Actomyosin actin part	2	0.03	0	0
	I band	2	0.05	0	0
	Pyruvate dehydrogenase complex	3	0.09	0	11
	Mitochondrial matrix	11	0.10	0	18
	Nucleolus	2	0.24	0	50
Molecular function	Cysteine-type endopeptidase inhibitor activity involved in apoptotic process	3	0.10	0	22
	Creatine kinase activity	3	0.25	0	0
	Isocitrate hydro-lyase cis-aconitate-forming activity	3	0.31	0	0
	Citrate hydro-lyase cis-aconitate-forming activity	11	0.33	0	15
	Nucleolus	3	0.36	0	0
Individual protein	Actin, alpha cardiac muscle 1	2	0.02	0	0
	Alpha-enolase	3	0.11	0	11
	Creatine kinase M-type	3	0.18	0	0
	Stress-70 protein, mitochondrial	2	0.21	0	50
	Beta-enolase	3	0.30	0	11

Note: The upper part lists the top five terms in the different ontology domains. In the lower part, gel spots that were assigned to the same protein were tested as a set.

recommend the permutation approach for global tests on gel-based proteomics data with missing values. Without permutation procedures, the global test performed poor in each of our simulated scenarios.

Our simulations also yielded that scenarios with unequal sizes of the experimental groups (A2, A3, B4–B6) are critical with regard to testing level and power. However, in one of these cases (B6), a bias through missing values could be corrected by means of a missing values imputation approach. In more detail, an inappropriate decrease of simulated levels was observed in scenarios A2 and A3, whereas an inappropriate increase of simulated levels was observed in scenarios B4 and B5. From these results, we would recommend that users should use the mixed model without imputation procedure in the case of unequal sample sizes.

Interestingly, the missing values were more frequent in the Calr+/− than those in WT group in our example. From this, one could conclude that the missingness is group-specific here. However, it is known that in multiplex gel approaches, where a gel is prepared with more than one sample (e.g. one treatment and one control sample), missing values 'appear' parallel in both groups.

Although, the *P*-values in our analysis of the real data from our mouse experiment were not adjusted for multiple testing (Dudoit *et al.*, 2003), the generation of *P*-values by the global test procedure allowed at least for a ranking of functional terms. Therefore, we are cautious with the biological conclusions in this concrete example. Nevertheless, the analysis of our mouse data points at correlations between the deficiency of the protein 'calreticulin' and protein sets related to biological processes in the heart muscle. These correlations were also reported in the context of other experiments (Lee *et al.*, 2013; Li *et al.*, 2002). Moreover, our analysis detected a correlation between the protein 'Actin, alpha cardiac muscle 1' and the experimental group factor 'calreticulin'. Likewise, this correlation was reported earlier in another experiment (Papp *et al.*, 2010). Both reproductions of known results show, additional to the simulation results, that our methods produced reasonable results.

It should also be pointed out that the protein 'Actin, alpha cardiac muscle 1' was not detected in the analysis of the individual spots but detected only when analyzed as a set of multiple spots. This might be explained by an increased statistical power when testing a set of spots globally in comparison to individual testing.

An alternative to global tests can be enrichment analysis. In microarray analyses, for example, enrichment tests are used to see whether a certain functional term is overrepresented among the differentially expressed features compared to the non-significant features. In the context of gel-based expression data, such an enrichment approach appears to be less reasonable because not each gel spot is identified and related to a functional annotation. An enrichment procedure, proposed for mass spectrometric protein expression data by Louie *et al.* (2010), is, therefore, not adequate for gel-based data. Thus, our proposed global test procedure is a more useable approach for the biological interpretation of group comparison in gel-based proteomics.

Although we focused specifically on global tests for gel-based proteomics data, other proteomics data, e.g. from Liquid chromatography-MS/MS (LC-MS/MS) experiments, are also concerned with missing values. For these cases, further methods

for missing values imputation have been proposed (Karpievitch *et al*., 2012; Smith *et al*., 2006). Because the frequency of LC-MS/MS experiments has overtaken that of 2-DE experiments in proteomics, the adaption of the global test idea to these types of data presents an interesting challenge for subsequent research.

5 CONCLUSIONS

Our proposed global test procedure can detect differences between experimental groups that would be omitted by standard protein-wise testing in proteomics experiments. Like in other already published global test approaches, we used a permutation procedure to correct for a bias in the testing level. As missing values are frequent in protein expression data, our simulation results show that their imputation can adjust for a loss in power. In addition, the presented method allows for the ranking of GO terms related to certain protein sets and thus facilitates the biological interpretation of a proteomics experiment.

Funding: This work was supported by the Deutsche Forschungsgemeinschaft (clinical research group KFO 179) and by the BMBF e:Bio project MetastaSys (0316173A).

Conflict of Interest: none declared.

REFERENCES

Aebersold,R. and Mann,M. (2003) Mass spectrometry-based proteomics. *Nature*, **422**, 198–207.

Chen,L.S. *et al.* (2011) A regularized Hotellings T 2 test for pathway analysis in proteomics studies. *J. Am. Stat. Assoc.*, **106**, 1345–1360.

Dudoit,S. *et al.* (2003) Multiple hypothesis testing in microarray experiments. *Bioinformatics*, **18**, 71–103.

Efron,B. and Tibshirani,R.J. (1993) *Permutation Tests*. An Introduction to the Bootstrap. Chapman and Hall, New York.

Faraway,J.J. (2006) *Generalized Linear Models*. Extending the Linear Model with R. Chapman and Hall, New York.

Goeman,J.J. *et al.* (2004) A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, **20**, 93–99.

Groene,J. *et al.* (2006) Transcriptional census of 36 microdissected colorectal cancers yields a gene signature to distinguishe UICC II and III. *Int. J. Cancer*, **119**, 1829–1836.

Grond-Ginsbach,C. *et al.* (2008) Gene expression in human peripheral blood mononuclear cells upon acute ischemic stroke. *J. Neurol.*, **255**, 723–731.

Huber,W. *et al.* (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, **18**, S96–S104.

Jung,K. *et al.* (2006) Statistical evaluation of methods for the analysis of dynamic protein expression data from a tumor study. *Rev. Stat. J.*, **4**, 67–80.

Jung,K. *et al.* (2011) Comparison of global tests for functional gene sets in two-group designs and selection of potentially effect-causing genes. *Bioinformatics*, **27**, 1377–1383.

Käll,L. and Vitek,O. (2011) Computational mass spectrometry-based proteomics. *PLoS Comput. Biol.*, **7**, e1002277.

Karpievitch,Y.V. *et al.* (2012) Normalization and missing value imputation for label-free LC-MS analysis. *BMC Bioinformatics*, **13** (**Suppl. 16**), S5.

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.

Koenig,T. *et al.* (2008) Robust prediction of the MASCOT score for an improved quality assessment in mass spectrometric proteomics. *J. Proteome Res.*, **7**, 3708–3717.

Kreil,D.P. *et al.* (2006) DNA microarray normalization methods can remove bias from differential protein expression analysis of 2D difference gel electrophoresis results. *Bioinformatics*, **20**, 2026–2034.

Lee,D. *et al.* (2013) Calreticulin induces dilated cardiomyopathy. *PLoS One*, **8**, e56387.

Li,J. *et al.* (2002) Calreticulin reveals a critical Ca2+ checkpoint in cardiac myofibrillogenesis. *J. Cell Biol.*, **158**, 103–113.

Louie,B. *et al.* (2010) The necessity of adjusting tests of protein category enrichment in discovery proteomics. *Bioinformatics*, **26**, 3007–3011.

Mansmann,U. and Meister,R. (2005) Testing differential gene expression in functional groups. Goeman's global test versus an ANCOVA approach. *Methods Inf. Med.*, **44**, 449–453.

Michalak,M. *et al.* (1999) Calreticulin: one protein, one gene, many functions. *Biochem. J.*, **344**, 281–292.

Papp,S. *et al.* (2010) Evidence for calreticulin attenuation of cardiac hypertrophy induced by pressure overload and soluble agonists. *Am. J. Pathol.*, **176**, 1113–1121.

Schäfer,J. and Strimmer,K. (2005) A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat. Appl. Gent. Mol. Biol.*, **4**, Article 32.

Smith,C.A. *et al.* (2006) XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem.*, **78**, 779–787.

Smyth,G.K. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, **3**, Article 3.

Troyanskaya,O. *et al.* (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17**, 520–525.

UniProt Consortium. (2007) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **35**, D193–D197.

Xin,H.M. and Zhu,Y. (2009) Multiple information-based spot matching method for 2-de images. *Electrophoresis*, **30**, 2477–2480.