Data analysis methods for detection of differential protein expression in two-dimensional gel electrophoresis

Bruno Meunier, Julien Bouley, Isabelle Piec, Carine Bernard, Brigitte Picard, Jean-François Hocquette *

INRA, Clermont-Ferrand Research Center, Herbicore Research Unit, Muscle Growth and Metabolism Group, 63122 St-Genès-Champanelle, France

Received 7 October 2004
Available online 14 March 2005

Abstract

The recent development of microarray technology has led statisticians and bioinformaticians to develop new statistical methodologies for comparing different biological samples. The objective is to identify a small number of differentially expressed genes from among thousands. In quantitative proteomics, analysis of protein expression using two-dimensional gel electrophoresis shows some similarities with transcriptomic studies. Thus, the goal of this study was to evaluate different data analysis methodologies widely used in array analysis using different proteomic data sets of hundreds of proteins. Even with few replications, the significance analysis of microarrays method appeared to be more powerful than the Student’s t test in truly declaring differentially expressed proteins. This procedure will avoid wasting time due to false positives and losing information with false negatives.

© 2005 Elsevier Inc. All rights reserved.

Keywords: Data analysis; Differential expression; Significance analysis of microarrays; Two-dimensional gel electrophoresis

In quantitative proteomics, a classical approach is to detect protein spots on two-dimensional polyacrylamide gel electrophoresis (2DPAGE) and to compare spot volumes between different biological samples, for example, condition A versus condition B [1–3]. Commonly, more than 500 protein spots can be visualized on 2DPAGE and then quantified by image analysis in an accurate manner [4]. However, because 2DPAGE analysis is a time-consuming method, only a few replications are generally made in these experiments (between three and seven in 90% of published proteomic studies). For a reliable differential expression analysis, a statistical analysis is then carried out on the whole data set to filter out interesting protein spots whose volume difference (A – B) is significant at a predefined level $\alpha = 5\%$ (the type I error cutoff currently used in biology). To this end, the Student’s t test has been employed in more than 60% of the recent literature, but it is often misused (violated assumptions). This is a reason why processing of proteomic data is a subject of active research [5,6]. Moreover, among those proteins declared to be “statistically” significant, biochemists generally retain those that are “biologically” significant, that is, those presenting a volume ratio (A/B) greater than a predefined fold change (FC) level [3]. In this respect, the problem of proteomic data analysis appears to be very similar to that encountered in transcriptomic approaches (several thousands of genes studied simultaneously, fewer than 10 replications). Here, microarray data analysis methods have been generally compared and improved by new ap-
proaches [7,8] to minimize false positives (FP) and to avoid losing information with false negatives (FN).

Thus, the focus of this article is on determining whether one of the analysis methods developed for array technologies can be used to improve differential quantitative proteomic analyses, especially when few replications are available. We decided to use simple approaches that need neither sophisticated computer environments nor particular statistical packages so that our methods could be easily used by biochemists. To this end, the widely used significance analysis of microarrays (SAM) method [9] is compared with the usual Student’s t test using two different proteomic data sets. The SAM method has been identified as potentially interesting [5] and as being used in proteomics [10], but it has not yet been evaluated in any real differential proteomic study.

Materials and methods

Data set description

The first data set (DS1) results from a differential proteomic analysis that was performed on bovine skeletal muscle. Triplicate 2DPAGE were carried out on each muscle sample representing conditions A and B (replications per condition \( n = 3 \)). Colloidal blue-stained gels and image analysis were performed as described previously [10]. The second data set (DS2) concerns a differential proteomic analysis that investigated rat skeletal muscle atrophy [11]. 2DPAGE was performed on five protein extracts for each condition A and B (replications per condition \( n = 5 \)). Colloidal blue-stained gel images were acquired using a GS-800 imaging densitometer, and image analysis was performed using PDQuest 2D analysis software (Bio-Rad Laboratories). Image analysis allowed protein spot detection (segmentation) and then quantification of the spots’ abundance in calculating an integrated optical density (IOD) after local background subtraction. IOD corresponds to a volume \((V = \text{area} \times \text{mean optical density})\). After analysis of all individual images, protein spots were matched on a chosen reference gel. To take experimental variations into account, each raw IOD value corresponding to one protein spot on a gel was divided by the sum of all matched protein spot IODs on that gel. Thus, for each \( i \) protein spot among a total of \( P_{\text{max}} \) detected, and for each \( j \) gel among the \( n \) replications, normalized volume values \( V_{Aij} \) and \( V_{Bij} \) are available, respectively, for conditions A and B and for each data set. To obtain a complete data matrix, a protein spot is kept for further analysis and declared reliable if at least \( n - 1 \) volume values are available in each condition. Then a missing value in a condition is replaced by the mean of the existing values for that protein spot. If less than \( n - 1 \) volume values are available in each condition, we consider that the amount of this protein cannot be determined accurately and must be permanently excluded. Finally, unreliable protein spots in one condition \((< n - 1 \text{ values})\) but reliable in the other condition are analyzed separately by visual inspection. Then, after this filtering step, each \( i \) protein among a total of \( P \) filtered can be characterized by its mean volume in each condition A and B, \((V_{Ai} \text{ and } V_{Bi}, \text{ respectively})\).

Data set visualization and correction

A common representation of the total information is a scatterplot of ratio value, \( R_i = \log_{10}(V_{Ai}/V_{Bi}) \) [12], as a function of the global volume intensity, \( I_i = \log_{10}(V_{Ai} \times V_{Bi}) \). This representation, called the \( R-I \) plot [13] for ratio versus intensity (Fig. 1), is quite

![Fig. 1. Representative R-I plot before (A) and after (B) intensity-dependent normalization.](image.png)
similar to $M \cdot A$ plot ($M = R = \log_2(V_A / V_B)$ and $A = \log_2(V_A \ast V_B)^{1/2}$) described by Dudoit et al. [14] to improve microarray data visualization and data correction. As is often observed with microarrays, artifacts visible on $R-I$ plots are observed; the cloud of points may be off-center locally relative to the $x$ axis. This phenomenon has also been recently identified and described in 2D difference gel electrophoresis (DIGE) [6,15], SYPRO Ruby [5], and silver-stained 2DPAGE [16,17]. Chang et al. [5] recently proposed a quantile normalization for their SYPRO Ruby-stained gel data. Image Master 2D Platinum 5.0 software (Amersham Biosciences) proposes an offset/scale normalization [6,18], but this method should be used with care [17,18] essentially because it is based on a nonlogged intensity scatterplot. In the data sets presented here, low-volume proteins deviate from zero, and this can be revealed by the best-fitting curve on an $R-I$ plot (logarithmic regression $c = f(I)$, Fig. 1A). Then a simple intensity-dependent normalization consists of subtracting the $c$ value (calculated from the $I$ value from the $c(I)$ equation) from the logged ratio values $R_i$ of each $i$ protein spot at volume intensity $I_i$ [13]. Correction of raw data volumes gives $V_{AIj} \rightarrow V_{AIj} \ast \sqrt{2}^{f(I)}$, $V_{BIj} \rightarrow V_{BIj}/\sqrt{2}^{f(I)}$. The result of this correction is represented on the corrected $R-I$ plot (Fig. 1B). This normalization successfully eliminates the small between-condition bias observed in the two data sets. Moreover, it was appropriate because a small percentage of differentially expressed protein spots was expected ($\sim 10\%$). In the presence of a higher intensity-dependent bias with numerous outliers, it could be useful to use robust regressions that have been proposed for microarrays [13,19]. Locally weighted linear regression (lowess) is widely used but needs specific statistical tools (MathSoft S+, R package).

Statistical methods

Once data are correctly filtered and normalized, a statistical test is possible for declaring a protein spot significantly different between the two conditions. In proteomics, the Student’s $t$ test is widely used but is questionable because strong parametric assumptions are often violated [18]. The nonparametric Mann–Whitney $U$ test is sometimes preferred, but only with large numbers of replications ($n > 7$) [18]. Because a small sample size is used ($n = 3$ or 5), a two-sample $t$ test (unpaired) adapted to unequal variances is preferred here to any nonparametric test, as advised in transcriptomics [7,8] and widely used in proteomics [2,3]. In a multiple-test problem (one test per protein spot detected), a Bonferroni correction normally is necessary, whereby the confidence level ($\alpha = 5\%$) is adjusted with the number of tests ($\alpha' = \alpha/P$). Because this correction has the reputation of being too conservative [5,20], it is not used, as can be observed in the majority of published proteomics studies.

A data transformation is nevertheless essential to compensate for dependence between variance and mean [16] as well as the possibility that data are not normally distributed. Because logarithmic transformation has positive effects [5,12,15], we transform raw data as follows: $V_{AIj} \rightarrow \log_2(V_{AIj})$, $V_{BIj} \rightarrow \log_2(V_{BIj})$ before application of the $t$ test. Unfortunately, this approach takes into account individual variability that is poorly estimated with so few replications. Moreover, assuming $5\%$ error with 500 proteins could lead to 25 (0.05 × 500) FP, which could be unacceptable [20]. A recent development of microarray data analysis methods aims to control the false discovery rate (FDR), which is the proportion of falsely declared significant (DS) proteins among the declared differential ones. For example, if 50 protein spots are expected to be differentially expressed (10% of 500 protein spots), an FDR cutoff of 5% will lead to fewer than 3 (0.05 × 50) FP, which would then be acceptable.

SAM [9] was designed for this purpose. SAM (version 1.21) is available as an add-in for Microsoft Excel (www-stat.stanford.edu/~tibs/sam/index.html). SAM is a modified $t$ test using a data permutation technique that enables FDR to be estimated. Because it does not use strong parametric assumptions [7], a two-class unpaired SAM is carried out on each nonlogged data set with the following parameters: FDR cutoff $= 5\%$ and number of permutations $= 1000$.

Finally, we associate each statistical test ($t$ test and SAM) with the basic method, which consists of applying a fixed FC threshold to ratio values $R_i$. Then for each $i$ protein spot

if $R_i > \log_2(FC)$ and the test is positive
   → the protein is declared up-expressed;

if $R_i < -\log_2(FC)$ and the test is positive
   → the protein is declared down-expressed;

otherwise → the protein expression is declared unchanged.

We deliberately chose the current level, $FC = 2.0$ (or $\log_2(FC) = 1$), even if an arbitrary and fixed threshold is statistically criticized [8] but nevertheless is considered biologically important [3].

Comparison of data analysis methods

To compare the two methods studied ($t$ test & FC and SAM & FC), a visual basic application under Excel has been developed for automatic data computation. Then the number of up- or down-expressed protein spots DS with each method is automatically available for each data set. In addition, the number of FP (FP $\iff$ type I error – controlled by the test) and FN (FN $\iff$ type II error) may be evaluated manually after identification
by different operators based on a visual inspection of the raw volume distribution for each protein. As a consequence, it is possible to calculate an empirical false negative rate (FNR) and an empirical FDR as defined in multiple-hypothesis testing approaches [20,21]. These error types and error rates are explained in Table 1. FNR and FDR are preferred here to the false nondiscovery rate [21] and false positive rate to give the same importance to FN and FP errors, and so these two empirical rates appear as indicators of the global efficiency of each statistical method in terms of reliability (few FP or low FDR) and sensitivity (few FN or low FNR).

### Results and discussion

For DS1, among a total of \( P_{\text{max}} = 510 \) protein spots detected, only \( P = 409 \) remained after the data filtering and data correction steps. Table 2A summarizes all of the results obtained with the \( t \) test & FC and SAM & FC methods on these 409 reliable protein spots. In addition, the results of the simplistic FC method are reported. Hence, the \( t \) test & FC method does not seem to be sensitive enough (too many FN) and is characterized by an FNR that is too high (22.4%). The \( t \) test & FC method appears to be too sensitive to high variability in condition A even if the ratio is high and if all of the volume values for condition A are higher than those for condition B. FNR for the FC method is null, but this method generates too many FP, leading to high FDR (14.0%). Finally, the SAM & FC method is the only one that presents a reasonable balance between FP (FDR = 8.0%) and FN (FNR = 6.1%), and so it appears to be more efficient. However, perfect results cannot be expected with a low number of replications, as is the case in this experimental design (\( n = 3 \)).

The same data analysis was carried out on DS2. A total of 617 protein spots were detected, and \( P = 571 \) remained after the filtering step. Table 2B summarizes the overall results obtained with the different methods after the appropriate data correction presented previously. Here the FC method is clearly unsatisfactory because it produces too many FP. With an FNR of 5.8% and a null FDR, the \( t \) test & FC method is powerful here. This is certainly due to the fact that there are a sufficient number of replicates (\( n = 5 \)), and this generally improves the power of statistical tests. Nevertheless, with an FNR of 1.9% and a null FDR, the SAM & FC method again seems to be the best solution. It should be noted that these low empirical rates (lower than cutoffs) might incite a lowering of the FC level. In fact, an FC level of two is generally recommended when only three replicates are available [22].

Besides SAM, other powerful microarray tools would be useful in proteomics. Thus, the generalized likelihood ratio test [8] also appears to be more powerful than the \( t \) test and seems to give good results on microarray data with only two replications.

### Conclusion

An FC approach is an attractive method for quantitative determination of differential expression, but it must be associated with a suitable statistical test and it requires perfect data normalization. This may be achieved after a close examination of \( R-I \) plots (cloud of points spread and centered on abscissa). Following this, the classic Student’s \( t \) test is able to give valid results so long as volumes are logged (previous \( t \) test application on nonlogged data led to spurious results) and the number of replications is strictly more than three. This experimental result is consistent with current sample size recommendations where a minimum of three replications is considered acceptable and five is generally

<table>
<thead>
<tr>
<th>Data analysis method</th>
<th>DS</th>
<th>FP</th>
<th>FN</th>
<th>FDR (%)</th>
<th>FNR (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A) FC</td>
<td>57</td>
<td>8</td>
<td>0</td>
<td>14.0</td>
<td>0.0</td>
</tr>
<tr>
<td>( t ) test &amp; FC</td>
<td>38</td>
<td>0</td>
<td>11</td>
<td>0.0</td>
<td>22.4</td>
</tr>
<tr>
<td>SAM &amp; FC</td>
<td>50</td>
<td>4</td>
<td>3</td>
<td>8.0</td>
<td>6.1</td>
</tr>
<tr>
<td>(B) FC</td>
<td>78</td>
<td>26</td>
<td>0</td>
<td>33.3</td>
<td>0.0</td>
</tr>
<tr>
<td>( t ) test &amp; FC</td>
<td>49</td>
<td>0</td>
<td>3</td>
<td>0.0</td>
<td>5.8</td>
</tr>
<tr>
<td>SAM &amp; FC</td>
<td>51</td>
<td>0</td>
<td>1</td>
<td>0.0</td>
<td>1.9</td>
</tr>
</tbody>
</table>

Note. For abbreviations, see note under Table 1. (A) Of a total of \( P = 409 \) detected and filtered protein spots in DS1 (\( n = 3 \)), the number of DS with each method is reported. The numbers of FP and FN have been evaluated manually. Empirical FDR and FNR are inferred. (B) Results were obtained identically from a total of \( P = 571 \) detected and filtered protein spots in DS2 (\( n = 5 \)).

---

Table 1

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Test</th>
<th>Positive</th>
<th>Negative</th>
<th>( \Sigma )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( H_0 \ (A = B) )</td>
<td>FP</td>
<td>TN</td>
<td></td>
<td>( P_0 )</td>
</tr>
<tr>
<td>( H_1 \ (A &gt; B) )</td>
<td>TP</td>
<td>FN</td>
<td>( P_1 )</td>
<td></td>
</tr>
<tr>
<td>( \Sigma )</td>
<td>DS</td>
<td>( P - DS )</td>
<td></td>
<td>( \Sigma )</td>
</tr>
</tbody>
</table>

\[ \text{FDR} = \frac{FP}{DS \text{ if } DS > 0} \]
\[ \text{else FDR} = 0 \]

| Note. Of a total of \( P (P = P_0 + P_1) \) protein spots, the null hypothesis (\( H_0 \), where protein is similarly expressed between conditions A and B) is tested against the alternative hypothesis (\( H_1 \), where protein is differentially expressed between conditions A and B). \( P_0 \) are really true, and \( P_1 \) are really false. The statistical method (or test) reports the number of protein spots that are declared significant (DS, up- or down-expressed in one condition). Among those DS, some are false positives (FP, declared significant even if truly not) or true positives (TP = DS – FP). In a similar manner, among those declared unchanged (\( P - DS \)), some are false negatives (FN, falsely declared not significant) or true negatives (TN). The empirical false discovery rate (FDR) and false negative rate (FNR) are inferred. |
advised. For at least three replications, SAM clearly appears to be the best solution provided that it is used correctly. In fact, SAM is also very sensitive to small variations of data normalization methods. Finally, this method can be applied to data sets from different species (cattle and rats in this case) and different proteomic methodologies (colloidal blue staining technique in this case). Its generality relies on the fact that our approach is well known for other high-throughput approaches (arrays) and can be expanded to any other large-scale analysis. The major advantages of this method are avoidance of spurious results and detection of all the true positive proteins, minimizing FN ones. We have demonstrated here that the SAM method gave better results than the $t$ test method.

Acknowledgment

We acknowledge the AGENAE program for its financial support. AGENAE is a national French program related to structural and functional genomics applied to animal science.

References