On the statistical analysis of the GS-NS0 cell proteome: 
Imputation, clustering and variability testing

Norhaiza Ahmad a,1, Jian Zhang a,1, Phillip J. Brown a, David C. James b, John R. Birch c, Andrew J. Racher c, C. Mark Smales d,*

a Institute of Mathematics and Statistics, University of Kent, Canterbury, Kent, UK
b School of Engineering, University of Queensland, QLD 4072, Australia
c Lonza Biologics plc, 228 Bath Road, Slough, SL1 4DY, UK
d Research School of Biosciences, University of Kent, Canterbury, Kent CT2 7NJ, UK

Received 18 November 2005; received in revised form 2 May 2006; accepted 3 May 2006
Available online 19 May 2006

Abstract

We have undertaken two-dimensional gel electrophoresis proteomic profiling on a series of cell lines with different recombinant antibody production rates. Due to the nature of gel-based experiments not all protein spots are detected across all samples in an experiment, and hence datasets are invariably incomplete. New approaches are therefore required for the analysis of such graduated datasets. We approached this problem in two ways. Firstly, we applied a missing value imputation technique to calculate missing data points. Secondly, we combined a singular value decomposition based hierarchical clustering with the expression variability test to identify protein spots whose expression correlates with increased antibody production. The results have shown that while imputation of missing data was a useful method to improve the statistical analysis of such data sets, this was of limited use in differentiating between the samples investigated, and highlighted a small number of candidate proteins for further investigation.

© 2006 Elsevier B.V. All rights reserved.

Keywords: 2D-PAGE; proteomic profiling; NS0 cells; imputed values; hierarchical clustering; rank correlation

1. Introduction

The proteome is a dynamic entity defined as the ‘protein complement of a genome’ [1]. Analysis of the proteome allows the simultaneous monitoring and quantitative measurement of thousands of cellular proteins; and despite the emergence of new methodologies, analysis by two-dimensional gel electrophoresis (2-DE) remains one of the methods of choice for such investigations. For expression analysis studies, whereby the differences in protein abundance are measured, it is essential that we can compare protein levels accurately and with confidence. Furthermore, this problem is exacerbated by the fact that there is a vast range of protein abundances in any given 2-DE gel, with previous reports estimating that there is at least six orders of magnitude difference between the least and most abundant proteins [2]. Despite the advances in 2-D technology in the last decade, obtaining consistent, reproducible 2-D gels and data still remains a technically demanding exercise [3–5]. As a result gel-to-gel variation does occur, and this must be considered in any study scrutinizing digital images of 2-D gels for statistically significant changes in protein expression [6].

Here we have investigated the variability in the proteome of murine myeloma NS0 cells stably producing a recombinant chimeric monoclonal antibody at different specific production rates. Specifically, the cell lines under investigation (transfectant blank, 4O, 4R, 2X, 2P, 2N2) produce antibody in increasing amounts with the control (blank) producing no antibody and 2N2 producing the most. In an attempt to understand why some cell lines secrete more recombinant antibody than others we have analyzed the proteome of the six cell lines above at mid-exponential phase of culture. We have previously shown that the difference in the observed productivities is not related to the relative abundance of recombinant antibody mRNA and the limiting factor must therefore be downstream of this (i.e., post-
transcriptional) [7]. This justifies the proteomic approach used in this study.

The major limitation in our proteomic approach lies in assessing whether the relevant data (concerning the dynamic levels of the proteins between cell lines) are statistically relevant. Further, almost all proteomic analyses undertaken to date compare two sample groups (e.g., diseased and healthy samples) or are concerned with detecting a specific stage of a cell in time-series experiments [8], whereas this study requires comparison of a series of samples on a graduated scale (individual protein expression for each cell line against increasing productivity). We wanted to investigate whether we could discriminate between these cell lines by statistical analysis (i.e., does the global expression pattern of the proteome correlate with changes in antibody production)? Here we report the results and validity of our statistical approach and the implications for future gel-based proteomic investigations.

2. Materials and methods

2.1. Cell lines and two-dimensional electrophoresis

All reagents were of analytical reagent grade or better. NS0 cells engineered to secrete a recombinant, humanized IgG4 under glutamine synthetase (GS) selection/amplication were a generous gift from Lonza Biologies plc. NS0 cells were grown and sampled for proteome analysis as previously described [7]. Two-dimensional electrophoresis (2-DE) separation of NS0 cell protein extracts was performed using immobilized pH gradient strips for isoelectric focusing and a vertical system for the SDS-PAGE separation as previously described [6]. Following the completion of SDS-PAGE, gels were fixed in a solution of 40% ethanol, 10% acetic acid overnight. Protein spots were then detected using the silver stain plusone™ protocol (GE Healthcare) apart from the omission of glutaraldehyde in the sensitizing step. The resulting gel images were captured at 200 dpi using a calibrated Powerlook III prepress color scanner (GE Healthcare).

2.2. Image analysis

Gel images were analyzed using ImageMaster™ 2-D Elite software (version 4.01; GE Healthcare). Spot detection was undertaken as previously described [6]. Following spot detection, background subtraction was automatically achieved using the mode of non-spot option and a margin of 45. Each sample was run in triplicate and then an ‘average gel’ created from the combination of the triplicate runs. The spot volume of the average gel was the mean value of the spots (pixel intensity) in the gels used to construct it. In order for a given spot to be included in the average gel it was required to be present in at least 2 of the 3 gels. These average gels were then compared to each other. Matching (or gel comparison) was undertaken after the placement of user seeds across the gel. All matches were manually checked. The software automatically calculates and reports variation or error for each spot and columns to cell lines. For simplicity we refer to those spots that were observed in all cell lines as complete spots and the remainder as incomplete.

In the dataset 839 spots (30% of the total number observed) were complete. Further, the missing values in the whole data matrix constitute approximately 37% of the entries. If we only considered the information from the 839 complete data spots, information from the remaining 1978 spots would be lost. We therefore imputed missing data values to account for the absent data [9]. However, imputing a data set with a missing rate as high as 37% is likely to be statistically unreliable. In microarray data it is generally advised to impute a data set only when the missing rate is less than 15%. We therefore considered only a subset of the incomplete spots, imputing or calculating data for data point entries whereby a data point was observed in at least four of the 6-cell lines. This approach results in a 1577 by 6 data matrix, X consisting of 1577 spots, amongst which 836 spots were initially complete and with a missing rate of ~11%.

We used the information from the complete spots on relations between the 6 cell lines, which we assumed to be shared by the incomplete spots, to infer the missing information. For example, to impute the missing values in column 4O in spot 5631 in Table 1, we built a regression model from the training dataset with column 4O as the response and the other cell lines as predictors. This model was then used to predict the missing values in column 4O and impute them. When values were missing in 2 cell lines, we imputed the missing values of each cell line in turn. For instance, to impute the missing values of 2 cell lines (4r and 2N2) in spot 5631, we applied a two-step imputation whereby we first built a regression model by taking 4r as the response and every other column except column 4O as the predictors. This model was then used to impute the missing values in column 4O and impute them. When values were missing in 2 cell lines, we imputed the missing values of each cell line in turn. For instance, to impute the missing values of 2 cell lines (4r and 2N2) in spot 5631, we applied a two-step imputation whereby we first built a regression model by taking 4r as the response and every other column except column 2N2 as the predictors. This model was then used to impute the missing values in column 4r.

To assess the reliability of the imputation methodology, in particular the effect on the clustering of cell lines, we introduce a ‘bootstrap’ procedure. In this

![Fig. 1. Q–Q plot of log-transformed normalized expression intensities for 839 complete spots. Normally distributed log-transformed normalized expression intensities are expected to plot near a curve close to the straight line. The figure demonstrates that apart from very few outlying spots, a majority of the log-transformed normalized expression intensities are normally distributed.](image-url)
procedure we first generated $K$ bootstrap data matrices, which have the same structure as the original data matrix. Each of these matrices was produced by randomly deleting the entries in the above imputed original data matrix, one at a time, until the missing rate reached $\sim 11\%$ subject to the following criteria; (i) no more than two of the entries in each row are allowed to be deleted, and (ii) no entries in the $839$ complete spots from the original data were deleted. We then imputed the missing entries in these ‘manufactured’ data matrices using the same predictive regression method described above.

Using the singular value decomposition (SVD) based distance calculated as described below, we calculated the similarity matrix $S^{(k)}$ of the $6$-cell lines for the $k$th imputed bootstrap data matrix and $S$ for the imputed original data matrix. The difference between the similarity matrices $S^{(k)}$ and $S$ is measured by the average Canberra-distance

$$d_k = \frac{1}{15} \sum_{i=1}^{15} \sum_{j=1}^{6} \frac{S_{ij} - S^{(k)}_{ij}}{S_{ij} + S^{(k)}_{ij}}, \quad k=1, \ldots, K,$$

where $S_{ij}$ and $S^{(k)}_{ij}$ are the $(i, j)$ elements in $S^{(k)}$ and $S$, respectively. As a result, we end up with $d_1, \ldots, d_K$ distances. Averaging these distances yields the following bootstrap estimate of the mean error rate of our imputation procedure, that is;

$$d = \frac{1}{K} \sum_{k=1}^{K} d_k,$$

which quantitatively demonstrates the effects of our imputation procedure on any similarity matrix based clustering methods.

2.4. Global statistical analysis

We first investigated whether we could discriminate between the $6$-cell lines from their complete 2D-PAGE expression patterns. For this purpose we measured the similarity between two cell lines by the Euclidean distance $d$ between their corresponding vectors of the normalized log-expression values in the $1577$ spots. We then applied average linkage hierarchical clustering to merge the true configuration, and (ii) there should exist an ‘elbow’ between the significant and the less significant eigenvalues in the scree plot. The SVD based distance between two cell lines is then defined as the Euclidean distance between the corresponding columns in $X$. Finally, based on this new distance, we applied the average linkage clustering method to group the cell lines.

2.5. Local statistical analysis

2.5.1. Expression variability test

We applied local statistical analysis to define a list of protein spots from the $580$ cell lines that potentially correlate with the antibody production rates of each cell line. Here we describe a novel method termed ‘an expression variability test’ for the rapid screening of the dataset for such protein spots. This method is based on the assumption that a protein spot associated with antibody production should exhibit significant expression variability across the cell lines investigated.

Although some of the systematic variation in the proteomic data may have been removed by way of normalization with the software as described above, there still exists background noise. As a result we assumed that each log-expression level consists of the underlying expression and the background noise,

$$\tilde{X} = U' \tilde{X}_c,$$

where $\tilde{U}$ is made up by the first $m$ columns of $U$ [11]. Here the $m$ principal components are selected according to the following rules of thumb [12]; (i) the components explain at least $75$ to $80\%$ of the available variation before the subspace representation by these components was deemed to be an adequate approximation to the true configuration, and (ii) there should exist an ‘elbow’ between the significant and the less significant eigenvalues in the scree plot. The SVD based distance between two cell lines is then defined as the Euclidean distance between the corresponding columns in $X$. Finally, based on this new distance, we applied the average linkage clustering method to group the cell lines.

2.5.2. Expression variability test

We applied local statistical analysis to define a list of protein spots from the $580$ cell lines that potentially correlate with the antibody production rates of each cell line. Here we describe a novel method termed ‘an expression variability test’ for the rapid screening of the dataset for such protein spots. This method is based on the assumption that a protein spot associated with antibody production should exhibit significant expression variability across the cell lines investigated.

Although some of the systematic variation in the proteomic data may have been removed by way of normalization with the software as described above, there still exists background noise. As a result we assumed that each log-expression level consists of the underlying expression and the background noise, $\tilde{X} = U' \tilde{X}_c$, where $U$ is made up by the first $m$ columns of $U$ [11]. Here the $m$ principal components are selected according to the following rules of thumb [12]; (i) the components explain at least $75$ to $80\%$ of the available variation before the subspace representation by these components was deemed to be an adequate approximation to the true configuration, and (ii) there should exist an ‘elbow’ between the significant and the less significant eigenvalues in the scree plot. The SVD based distance between two cell lines is then defined as the Euclidean distance between the corresponding columns in $X$. Finally, based on this new distance, we applied the average linkage clustering method to group the cell lines.

2.5. Local statistical analysis

2.5.1. Expression variability test

We applied local statistical analysis to define a list of protein spots from the $580$ cell lines that potentially correlate with the antibody production rates of each cell line. Here we describe a novel method termed ‘an expression variability test’ for the rapid screening of the dataset for such protein spots. This method is based on the assumption that a protein spot associated with antibody production should exhibit significant expression variability across the cell lines investigated.

Although some of the systematic variation in the proteomic data may have been removed by way of normalization with the software as described above, there still exists background noise. As a result we assumed that each log-expression level consists of the underlying expression and the background noise,
that is, the observed log-expression level of spot \( i \) at cell line \( j \), \( x_{ij} \) follows the model

\[
x_{ij} = f(i, j) + \epsilon_{ij}
\]

where \( f(i, j) \) is the underlying log-expression level and \( \epsilon_{ij} \) is the background noise. For spot \( i \) we let \( \bar{x}_i \) and \( f_i \) denote the averages of observed log-expression levels and underlying log-expression levels over the 6-cell lines, respectively. Then the expectation of the sample variance of log-expressions, \( S_i^2 = \frac{1}{6} \sum_{j=1}^{6} (x_{ij} - \bar{x}_i)^2 \) for spot \( i \) can be decomposed into two parts, namely

\[
E[S_i^2] = \frac{1}{6-1} \sum_{j=1}^{6} (f(i, j) - \bar{f})^2 + E\left(\frac{1}{6-1} \sum_{j=1}^{6} (\epsilon_{ij} - \bar{\epsilon}_i)^2\right),
\]

where the first expression corresponds to the variability of the underlying log-expression levels across the cell lines while the second expression refers to the background noise levels across the cell lines. We first considered the case whereby the underlying log-expression levels of spot \( i \) do not change across the 6-cell lines, in other words where there is no expression variability across the 6-cell lines. In this case we have \( f(1, 1) = f(1, 2) = \ldots = f(1, 6) \). Assuming that all spots have the same background noise level \( \sigma^2 \), we can reduce the equation in (b) into

\[
E[S_i^2] = \sigma^2,
\]

since the first expression in (b) is essentially zero. In contrast, for the case whereby the underlying log-expression level of spot \( i \) does change across the cell lines, that is, there exists \( j, f(i, j) \neq f(i, j + 1) \), we have a positive value of the expression variability in the equation (b), thus having a larger expectation of the sample variance than in the previous case, i.e.,

\[
E[S_i^2] > \sigma^2.
\]

This gives the following expression variability test in the form of

\[
S_i^2 > c,
\]

where \( c \) is the critical value. Note that under the null hypothesis, \( (6-1)S_i^2/\sigma^2 \) follows \( \chi^2_{(6-1)} \), a chi-square distribution with 5 degrees of freedom. Accordingly, \( c = \chi^2(5)(\sigma^2)/(6-1) \) depends on the background noise level \( \sigma^2 \) as well as the significance level \( u \). The \( P \)-value of \( S_i^2 \) can be calculated by \( P(\chi^2_{(6-1)}>S_i^2/\sigma^2) \). The inequality (d) is equivalent to

\[
P(\chi^2_{(6-1)}>S_i^2/\sigma^2) < p_i,
\]

where \( p_i \) is a \( P \)-value threshold depending on \( u \).

To estimate \( \sigma^2 \), we considered spots whose underlying log-expression levels did not change significantly across the 6-cell lines. These spots are classified for convenience as ‘housekeeping spots’. The sample variances observed from the identified housekeeping spots was considered to arise from background noise. Note that in practice the background noise levels may be slightly varying with the log-expression levels of the spots. To contend this problem, using a worst scenario argument, we estimated \( \sigma^2 \) by the upper bound of the confidence interval for \( \sigma^2 \), i.e., by the \( v \)-quartile of the sample variances of these housekeeping spots, where \( v \geq 0.50 \) will be determined later.

Here we present a rank invariant algorithm for identifying housekeeping spots as an extension of the work in [13]. The algorithm is based on the assumption that housekeeping spots should have most highly but not differentially express across the cell lines. In other words, housekeeping spots should have relatively high expressions as well as smaller expression variability compared to non-housekeeping spots. More specifically, the algorithm depends on the following rank invariant score for the \( i \)-th spot:

\[
R_i = \frac{\sum_{j=1}^{6} | r_{ij} - \bar{r}_i |}{r_{ii} + \bar{r}_i},
\]

where \( r_{ij} \) and \( \bar{r}_i \) represent the ranks of spots in the control cell line and the \( i \)-th cell line, respectively. In this score, we first ranked spots in each cell line and then for each spot we calculated its average rank-difference between the expressions in the control and non-control cell lines, weighted by the ranks of the expressions in these cell lines. As a result the higher the rank of a spot in the cell line, the smaller weight it will receive. The lower the rank invariant score for a spot \( i \), the more likely that spot is a member of the housekeeping spots. We thus decided whether the \( i \)-th spot is housekeeping according to whether \( R_i \leq r \), where \( 0 \leq r \leq 1 \) is a pre-specified threshold.

To assess the performances of the above algorithm and the variability test under the different \((\nu, \alpha)\)-settings, we first generated 20 data matrices, each with 839 spots and 6 cell lines. These datasets mimicked certain features of our real 2D gel data: (1) the numbers of housekeeping spots and differentially expressed spots were much smaller than the number of noise spots; (2) the spots were highly correlated. In each of these simulated dataset, the first 12 spots, drawn from a two-dimensional normal, are pair-wisely correlated. These spots showed increasing differential expressions as antibody production rate was increasing across the cell lines. The next 16 pairs of correlated spots, drawn from a two-dimensional normal, had constant expression across the cell lines. These spots stand for the housekeeping spots. The remaining 795 spots were merely noise, consisting of 397 pairs of correlated spots drawn from a two-dimensional normal distribution. Specifically, if letting \( X = (X_{ij}) \) denote such a kind of data matrix, for \( j = 1, \ldots, 6 \), we have

\[
\begin{pmatrix}
X_{2a-1}/X_{2k}/
X_{2b}/X_{2k}/
\end{pmatrix} = \begin{pmatrix} 1 & 0.5 \\ -1 & 0.5 \\ \end{pmatrix} + N\left(0, \begin{pmatrix} 1/2 \\ 0 \end{pmatrix}\right),
\]

\[
\begin{pmatrix}
X_{2a}/X_{2k}/
X_{2b}/X_{2k}/
\end{pmatrix} = \begin{pmatrix} 10 & 0.5 \\ -10 & 0.5 \\ \end{pmatrix} + N\left(0, \begin{pmatrix} 1/2 \\ 0 \end{pmatrix}\right),
\]

\[
\begin{pmatrix}
X_{2a-1}/X_{2k}/
X_{2b}/X_{2k}/
\end{pmatrix} = N\left(0, \begin{pmatrix} 1/2 \\ 0 \end{pmatrix}\right),
\]

and \( X_{10} \sim N(0,1) \). Here \( N(0,0) \) denotes a two-dimensional normal with mean 0 and covariance matrix \( A \) and \( N(0,1) \) is a standard normal. We then applied our rank invariant algorithm and variability test to these 20 datasets and gauged their average performances under different settings of \((\nu, \alpha)\) in terms of false discovery rate (FDR),
sensitivity and specificity defined in the next section. This will give an optimal \((r,v)\)-setting for the variability test.

2.5.2. False discovery rate

Note that both the selection of housekeeping spots and the testing of expression variability above involve several hundreds of the tests. It is important to take into account the potential effects of multiple testing on the power of these procedures. Note that protein spots on 2D gels are rarely independent measurements, but rather are correlated in amount with several others [14]. This is true for our 2D gel dataset, where the number of mutually uncorrelated spots could be less than 38 if taking 0.8 as the threshold for the correlation coefficient of two spots being significant. Thus, many of these spot related tests are dependent. Due to this dependence, a simple Bonferroni correction would be too conservative. We therefore opted to use the false discovery rate procedure in [15] to address this issue. In this procedure a statistical significance measure termed the \(q\)-value was defined to show the minimum false discovery rate incurred in a multiple testing. The false discovery rate (FDR) is the expected proportion of false positives among the tests found to be significant. We can also evaluate the performance by use of the so-called sensitivity and specificity of the tests. Sensitivity measures the proportion of true positives over the total null features, whereas specificity measures the proportion of true negatives over the total alternative features. In particular, let \(p\) be the total number of tests, \(FP\) the number of false positives, \(TP\) the number of true positives, \(TN\) the number of true negatives and \(FN\) the number of false negatives. Then \(FP+TP\) is the total number of features called significant, \(FP+TN\) is the total number of null features, \(TP+FN\) is the number of truly alternative features, and

\[
\text{FDR} = \frac{FP}{FP+TP} ; \quad \text{Sensitivity} = \frac{TP}{TP+FN} ; \quad \text{Specificity} = \frac{TN}{TN+FP}
\]

There exists a trade off between sensitivity and specificity. However, due to the nature of proteomic experiments, we expect only a handful of significant spots

Table 2

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Estimated FDR (r=0.01)</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Estimated FDR (r=0.05)</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.3%</td>
<td>90.6%</td>
<td>99.9%</td>
<td>1.0%</td>
<td>100.0%</td>
<td>99.6%</td>
</tr>
<tr>
<td>2</td>
<td>0.0%</td>
<td>87.5%</td>
<td>99.9%</td>
<td>5.9%</td>
<td>100.0%</td>
<td>99.8%</td>
</tr>
<tr>
<td>3</td>
<td>0.0%</td>
<td>100.0%</td>
<td>100.0%</td>
<td>8.6%</td>
<td>100.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td>4</td>
<td>6.1%</td>
<td>96.9%</td>
<td>99.8%</td>
<td>5.9%</td>
<td>100.0%</td>
<td>99.8%</td>
</tr>
<tr>
<td>5</td>
<td>0.0%</td>
<td>84.4%</td>
<td>100.0%</td>
<td>8.6%</td>
<td>100.0%</td>
<td>99.9%</td>
</tr>
<tr>
<td>6</td>
<td>3.3%</td>
<td>90.6%</td>
<td>99.9%</td>
<td>11.1%</td>
<td>100.0%</td>
<td>99.9%</td>
</tr>
<tr>
<td>7</td>
<td>5.9%</td>
<td>100.0%</td>
<td>99.9%</td>
<td>8.6%</td>
<td>100.0%</td>
<td>99.9%</td>
</tr>
<tr>
<td>8</td>
<td>0.0%</td>
<td>87.5%</td>
<td>99.9%</td>
<td>11.1%</td>
<td>100.0%</td>
<td>99.9%</td>
</tr>
<tr>
<td>9</td>
<td>3.3%</td>
<td>90.6%</td>
<td>99.9%</td>
<td>11.1%</td>
<td>100.0%</td>
<td>99.9%</td>
</tr>
<tr>
<td>10</td>
<td>0.0%</td>
<td>90.6%</td>
<td>99.9%</td>
<td>11.1%</td>
<td>100.0%</td>
<td>99.9%</td>
</tr>
<tr>
<td>11</td>
<td>3.4%</td>
<td>87.5%</td>
<td>99.9%</td>
<td>13.5%</td>
<td>100.0%</td>
<td>99.4%</td>
</tr>
<tr>
<td>12</td>
<td>9.1%</td>
<td>93.8%</td>
<td>99.9%</td>
<td>13.5%</td>
<td>100.0%</td>
<td>99.4%</td>
</tr>
<tr>
<td>13</td>
<td>0.0%</td>
<td>84.4%</td>
<td>100.0%</td>
<td>5.9%</td>
<td>100.0%</td>
<td>99.9%</td>
</tr>
<tr>
<td>14</td>
<td>3.4%</td>
<td>87.5%</td>
<td>99.9%</td>
<td>5.9%</td>
<td>100.0%</td>
<td>99.8%</td>
</tr>
<tr>
<td>15</td>
<td>0.0%</td>
<td>87.5%</td>
<td>99.9%</td>
<td>11.1%</td>
<td>100.0%</td>
<td>99.9%</td>
</tr>
<tr>
<td>16</td>
<td>0.0%</td>
<td>90.6%</td>
<td>99.9%</td>
<td>11.1%</td>
<td>100.0%</td>
<td>99.9%</td>
</tr>
<tr>
<td>17</td>
<td>0.0%</td>
<td>93.8%</td>
<td>99.9%</td>
<td>5.9%</td>
<td>100.0%</td>
<td>99.9%</td>
</tr>
<tr>
<td>18</td>
<td>5.9%</td>
<td>100.0%</td>
<td>99.9%</td>
<td>11.1%</td>
<td>100.0%</td>
<td>99.9%</td>
</tr>
<tr>
<td>19</td>
<td>3.4%</td>
<td>87.5%</td>
<td>99.9%</td>
<td>15.8%</td>
<td>100.0%</td>
<td>99.9%</td>
</tr>
<tr>
<td>20</td>
<td>3.3%</td>
<td>90.6%</td>
<td>99.9%</td>
<td>5.9%</td>
<td>100.0%</td>
<td>99.9%</td>
</tr>
<tr>
<td>Average</td>
<td>2.5%</td>
<td>91.1%</td>
<td>99.9%</td>
<td>10.0%</td>
<td>100.0%</td>
<td>99.6%</td>
</tr>
</tbody>
</table>
Table 3
The average FDR, sensitivity and specificity of the expression variability test, with 12 different \((r,v,p)\)-settings, over 20 simulated datasets

<table>
<thead>
<tr>
<th>(r)</th>
<th>(v)</th>
<th>(p)</th>
<th>FDR</th>
<th>Specificity</th>
<th>Sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1%</td>
<td>85%</td>
<td>0.1%</td>
<td>0.00</td>
<td>0.256</td>
<td>1.00</td>
</tr>
<tr>
<td>75%</td>
<td>85%</td>
<td>1%</td>
<td>0.034</td>
<td>0.375</td>
<td>1.00</td>
</tr>
<tr>
<td>50%</td>
<td>85%</td>
<td>5%</td>
<td>0.346</td>
<td>0.692</td>
<td>0.992</td>
</tr>
<tr>
<td>75%</td>
<td>1%</td>
<td></td>
<td>0.060</td>
<td>0.999</td>
<td>0.404</td>
</tr>
<tr>
<td>50%</td>
<td>75%</td>
<td>5%</td>
<td>0.246</td>
<td>0.995</td>
<td>0.629</td>
</tr>
<tr>
<td>75%</td>
<td>5%</td>
<td></td>
<td>0.702</td>
<td>0.960</td>
<td>0.842</td>
</tr>
<tr>
<td>85%</td>
<td>0.1%</td>
<td></td>
<td>0.335</td>
<td>0.991</td>
<td>0.708</td>
</tr>
<tr>
<td>50%</td>
<td>0.1%</td>
<td></td>
<td>0.571</td>
<td>0.975</td>
<td>0.808</td>
</tr>
<tr>
<td>50%</td>
<td>85%</td>
<td>5%</td>
<td>0.885</td>
<td>0.877</td>
<td>0.921</td>
</tr>
<tr>
<td>5%</td>
<td>85%</td>
<td>0.1%</td>
<td>0.00</td>
<td>0.158</td>
<td>1.00</td>
</tr>
<tr>
<td>75%</td>
<td>85%</td>
<td>1%</td>
<td>0.037</td>
<td>0.276</td>
<td>1.00</td>
</tr>
<tr>
<td>50%</td>
<td>85%</td>
<td>5%</td>
<td>0.23</td>
<td>0.625</td>
<td>1.00</td>
</tr>
<tr>
<td>75%</td>
<td>1%</td>
<td></td>
<td>0.016</td>
<td>1.00</td>
<td>0.254</td>
</tr>
<tr>
<td>50%</td>
<td>75%</td>
<td>5%</td>
<td>0.128</td>
<td>0.998</td>
<td>0.450</td>
</tr>
<tr>
<td>75%</td>
<td>5%</td>
<td></td>
<td>0.617</td>
<td>0.974</td>
<td>0.808</td>
</tr>
<tr>
<td>85%</td>
<td>0.1%</td>
<td></td>
<td>0.163</td>
<td>0.998</td>
<td>0.558</td>
</tr>
<tr>
<td>50%</td>
<td>0.1%</td>
<td></td>
<td>0.394</td>
<td>0.988</td>
<td>0.742</td>
</tr>
<tr>
<td>50%</td>
<td>85%</td>
<td>5%</td>
<td>0.851</td>
<td>0.908</td>
<td>0.904</td>
</tr>
</tbody>
</table>

Here \(r\) is the threshold for the rank invariant statistics in selecting housekeeping spots, the \(v\)-quartile of the sample variances of the selected housekeeping spots is taken as the estimate of the background noise level \(\sigma^2\) in the variability tests, and \(p\) is the threshold for the \(P\)-values of the variability tests.

### 3. Results

#### 3.1. Clustering of cell lines

We have employed the methods described above to initially impute the data matrix of 1577 spots selected from the average gels (as depicted in Fig. 2), using the predictive regression method. We choose to impute the data after spot analysis and the generation of average gels. We could have chosen to impute data at the individual gel level and then created average gels for each cell line. We could have chosen to impute data at the average gels. We could have chosen to impute data after spot analysis and the generation of the predictive regression method. We choose to impute the data after spot analysis and the generation of average gels. We could have chosen to impute data at the individual gel level and then created average gels for each cell line. We choose to impute the data after spot analysis and the generation of average gels. We could have chosen to impute data after spot analysis and the generation of average gels. We could have chosen to impute data after spot analysis and the generation of average gels.

For the 20 datasets simulated in Expression variability test, the rank invariant algorithm appeared to correctly identify the majority of the housekeeping spots planted in the data. In particular, if taking \(r=0.01\) as the threshold for the \(R_i\)-values, the algorithm, over the 20 datasets, gave an average FDR of ~2.5% with high sensitivity and specificity of 91.1% and 99.9%, respectively. If taking \(r=0.05\) as the threshold for the \(R_i\)-value, the algorithm attained an average FDR of ~10% with high specificity at 99.6% and perfect sensitivity at 100%. These results imply that the rank invariant algorithm performed reasonably well even when the spots were over-dominated by noisy spots. See Table 2.

#### 3.2. Testing for expression variability

##### 3.2.1. Performance of the rank invariant algorithm

For the 20 datasets simulated in Expression variability test, the rank invariant algorithm appeared to correctly identify the majority of the housekeeping spots planted in the data. In particular, if taking \(r=0.01\) as the threshold for the \(R_i\)-values, the algorithm, over the 20 datasets, gave an average FDR of ~2.5% with high sensitivity and specificity of 91.1% and 99.9%, respectively. If taking \(r=0.05\) as the threshold for the \(R_i\)-value, the algorithm attained an average FDR of ~10% with high specificity at 99.6% and perfect sensitivity at 100%. These results imply that the rank invariant algorithm performed reasonably well even when the spots were over-dominated by noisy spots. See Table 2.

##### 3.2.2. Performance of expression variability test

In order to select an appropriate \((r,v)\), we compared the performance of the variability test under different \((r,v)\)-settings. Table 3 shows the FDR, sensitivity, and specificity of the expression variability test under 12 different settings of \((r,v,p)\): \((0.01,0.50,0.01), (0.01,0.50,0.05), (0.01,0.75,0.01), (0.01,0.75,0.05), (0.01,0.85,0.01), (0.01,0.85,0.05), (0.05,0.50,0.01), (0.05,0.50,0.05), (0.05,0.75,0.01), (0.05,0.75,0.05), (0.05,0.85,0.01), (0.05,0.85,0.05)).

From Table 3, we can observe that the lower the quartile used to estimate the background noise level, the higher the sensitivity relative to the insignificant spots. Therefore, ideally we would expect the test to achieve high sensitivity and greater specificity while maintaining a low level of FDR.

their SVD distances, we obtained the grouping of the 6-cell lines. The grouping presented in Fig. 4 demonstrates that the cell lines 4O and 4R are grouped together. This is followed by the control, 2P and 2N2. Cell line 2X appears to be different from the other cell lines, consistent with the observed biology and proteome that shows major differences relative to the others.

### 3.2. Testing for expression variability

#### 3.2.1. Performance of the rank invariant algorithm

For the 20 datasets simulated in Expression variability test, the rank invariant algorithm appeared to correctly identify the majority of the housekeeping spots planted in the data. In particular, if taking \(r=0.01\) as the threshold for the \(R_i\)-values, the algorithm, over the 20 datasets, gave an average FDR of ~2.5% with high sensitivity and specificity of 91.1% and 99.9%, respectively. If taking \(r=0.05\) as the threshold for the \(R_i\)-value, the algorithm attained an average FDR of ~10% with high specificity at 99.6% and perfect sensitivity at 100%. These results imply that the rank invariant algorithm performed reasonably well even when the spots were over-dominated by noisy spots. See Table 2.

#### 3.2.2. Performance of expression variability test

In order to select an appropriate \((r,v)\), we compared the performance of the variability test under different \((r,v)\)-settings. Table 3 shows the FDR, sensitivity, and specificity of the expression variability test under 12 different settings of \((r,v,p)\): \((0.01,0.50,0.01), (0.01,0.50,0.05), (0.01,0.75,0.01), (0.01,0.75,0.05), (0.01,0.85,0.01), (0.01,0.85,0.05), (0.05,0.50,0.01), (0.05,0.50,0.05), (0.05,0.75,0.01), (0.05,0.75,0.05), (0.05,0.85,0.01), (0.05,0.85,0.05)).

From Table 3, we can observe that the lower the quartile used to estimate the background noise level, the higher the sensitivity...
we will attain but at the risk of having a very high false discovery rate. Thus, in order to achieve a low FDR, we have to reduce the power of detecting true differentially expressed spots. In the following sections, we chose the two settings for reducing the power of detecting true differentially expressed spots including the imputed data (Fig. 4A and B). Then using the rank invariant algorithm, we obtained two subsets of housekeeping spots at \( r = 0.01 \) and \( r = 0.05 \), respectively. There are 38 and 193 housekeeping spots in these two sets, respectively.

The 85\% quartile of the sample variances in the first set of housekeeping spots is 0.092. According to the expression variability test defined in Local statistical analysis, the expression variability of the spot \( i \) would be significantly different from the background noise at the 0.01 significance level if \( s^2_i > c \), with the critical value

\[
c = \frac{0.092 \times 15.09}{6-1} = 0.278.
\]

This analysis highlighted 146 spots with significant expression variability (Fig. 5). According to Storey and Tibshirani [15], the estimated FDR of these 146 spots is 0.042. We further identified 21 spots with significant variability \( S^2_i > 0.06 \) at the significance level of \( 10^{-6} \) (Fig. 5), which are, respectively, spot reference numbers 5646, 5655, 5733, 5759, 5782, 5879, 5918, 6010, 6188, 6229, 6347, 6369, 6478, 6566, 6841, 6894, 7082, 7644, 8046, 8099, and 8137. The method of Storey and Tibshirani [15] can be applied here again to show that spots with \( P \)-values \( < 10^{-6} \) had a very negligible estimated false discovery rate of \( \sim 2.94144 \times 10^{-5} \). A similar analysis can be carried out using the 75\% quartile, 0.0588, of the sample variances in the second set of housekeeping spots. The details are omitted here.

4. Discussions and conclusions

We have investigated the feasibility of calculating and imputing missing data points from gel-based proteomic expression studies in order to generate and analyze a larger component of observable protein spots in multidimensional data sets. Specifically we used this approach to investigate if there were conserved changes in protein expression that correlated with increasing heterologous antibody production rates across six NS0 cell lines. To our knowledge this is one of the first proteomic studies to address the statistical significance of observed changes in protein expression whereby there is a graduated response in the sample population rather than an 'on-off' response (i.e., comparison of protein expression against increasing antibody productivity as opposed to, for example, a comparison of healthy vs. diseased state). Whilst we successfully imputed the data for missing spot values within the criteria of the spot being observed in at least 4 of the 6 samples (i.e., only one or two absences were allowed), imputation of these data points was of no benefit in differentiating between the data sets. This is the result of comparing a range of samples with increasing/decreasing productivity rather than on/off type samples and the method of calculating the imputed data. Thus, only the initial complete data points were useful in differentiating between the cell lines with respect to productivity. However, the following simulation study demonstrates that this is not always the case.

We simulated an original data matrix \( X \) with 1577 pair-correlated normal spots and 6 cell lines such that for \( j = 1, \ldots, 6 \),

\[
\begin{align*}
\frac{X_{2k-1} - \mu_{2k}}{X_{2k}} &= N \left( 0, \frac{1}{2} \right), \\
\frac{X_{2k-1} - \mu_{2k}}{X_{2k}} &= N \left( 0, \frac{(-1)^k}{2} \right), \\
\frac{X_{2k-1} - \mu_{2k}}{X_{2k}} &= N \left( 0, \frac{1}{2} \right), \\
\frac{X_{2k-1} - \mu_{2k}}{X_{2k}} &= N \left( 0, \frac{(-1)^k}{2} \right),
\end{align*}
\]

\( k = 1, \ldots, 6, \quad 7, \ldots, 22, \quad 23, \ldots, 788 \),

and

\( X_0 \sim N(0, 1), \quad i = 1577 \).

Based on this original matrix, a matrix with some missing entries can be generated by randomly deleting the entries in the

Fig. 5. Histogram of the sample variances of the 839 spots. The dashed vertical lines show the critical values at \( S^2_i = 0.278 \) and \( S^2_i = 0.661 \) beyond which the expression variability is considered significantly different from the estimated background noise, \( \sigma^2 \) at the significance levels of 0.01 and 10^{-6}, respectively.
original matrix, one at a time, until the missing rate reached ~11% subject to the criteria (i) defined in Normalization and imputation. Repeating this procedure 20 times, we obtained 20 matrices, each with missing entries. There are two different strategies for tackling the missing entries in these 20 matrices: (1) drop the rows with missing entries, that is, ignore the incomplete spots, and (2) impute these missing entries using the predictive regression. The first strategy led to 20 sub-matrices, while the second strategy resulted in 20 imputed matrices. We then calculated the similarity matrix of the 6-cell lines for each of these data matrices, using SVD-based distance defined in Global statistical analysis. Finally, we computed the average Canberra-distances between the similarity matrices for the imputed and the original matrices and between the similarity matrices for the above 20 sub-matrices and the original matrix. Using these distances we demonstrated the average clustering errors of the 6-cell lines based on two different strategies of tackling missing entries. As we expected, the first strategy led an average clustering error of 10% much less than 18% derived from the second strategy. This implies that the imputation can be beneficial indeed in some circumstances.

In order to further our understanding of the mechanism(s) which limit recombinant protein productivity in NS0 cells using a 2D-PAGE approach, it is vital that protein(s) that exhibit statistically significant changes in abundance that mirror changes in recombinant protein production can be identified. Only then will we be able to develop novel cell engineering strategies based on the identification of those target proteins that appear to limit recombinant protein production. We are particularly interested in identifying conserved patterns of protein expression that are correlated with increasing or decreasing antibody production. In this study we tested the significance of these patterns at the 0.05 and 0.01 levels relying on the expression variability tests. The results indicate that the expression of the majority of proteins in the cells may not be correlated with antibody production, or influences/contributes towards the functional competence of these cells in terms of antibody production in a moderate manner. However, our analysis has highlighted a relatively small number of candidate proteins whose expression is significantly correlated with antibody production. We intend to further investigate these proteins to determine if they are potential targets for cell engineering strategies for improved recombinant protein production in mammalian cell systems.

The imputation, the SVD-based clustering, and variability test are novel developments for the specific type of experiment described in this paper. In literature, the principal component analysis (a SVD technique) and the Euclidean distance based hierarchical clustering have separately been applied to classify 2D gels and to detect mutations in case-control studies. In time-series experiments, the SVD and neural networks have been employed to detect a specific stage of growth. See [8,16,17]. In contrast, our hierarchical clustering, based on the SVD distance, allows one to combine the dimension-reduction technique with the classical cluster analysis. This could extract much more amount of information from the data while making the results still interpretable.

Our variability test is different from the traditional one-way and two-way analyses of variance (ANOVA) s used in [18]. In our procedure, we fitted a one-way random effect model for the $i$th spot—namely, $X_{ij} = \mu + e_{ij}, \mu \sim N(\theta, \sigma^2), e_{ij} \sim N(0, \sigma^2)$. For each $i$, we are interested in the size of the variation in the $\mu_i$, i.e., the expression variability of the $i$th spot. The novel features in our procedure are as follows: (1) we have applied multiple one-way random effect models to the data, allowing for the clustering structures in the protein spots without the assumption of any distribution structure for $\theta$; (2) we have considered the effect of multiple testing on the power of this procedure; (3) we have developed a housekeeping-spot based method to estimate the background noise level $\sigma^2$. Some simulation studies have been carried out to select the critical values in the test and to assess the performance of our procedure. These new developments made the inference more attractive.

Since the expression variability test identifies proteins with a considerable variability across the cell lines, it might pick up proteins with a generally high biological variability of expression. To tackle this issue, in this paper we ran the experiment in triplicate and applied the variability test to the ‘average expression levels’ created from the combination of the triplicate runs. This can reduce the effects of biological variability on the test and therefore give higher power to identify proteins correlated with antibody production.

The variability test above can be extended to the case where the spots have the different average background noise levels over the cell lines. For this purpose, assuming that the background noise $e_{ij}$ in the observed log-expression level of spot $i$ at cell line $j$ has mean zero and variance $\sigma^2$, we modified the equations (b) and (c) in Local statistical analysis, respectively, by

$$E[S^2]_i = \frac{1}{6-1} \sum_{j=1}^{6} (f(i,j) - f_i)^2 + \sigma^2, \quad E[S^2]_i = \sigma^2,$$

where $\sigma^2 = \frac{1}{6} \sum_{j=1}^{6} \sigma^2_j$ is the average background noise level of the $i$th spot. The $P$-value of $S^2_i$ could be calculated by $P \left( z^2 > \frac{S^2_i - \sigma^2_i}{\sigma^2} \right)$. Here we are unable to use the housekeeping spots to estimate these average background noise levels. However, as our 2DE experiment was run in triplicate, we had three replicates for estimating $\sigma^2$. Note that the quality of these estimates might be very poor because of the small number of the replicates. Applying the above extension to the 146 spots and the 21 spots identified in Results, we found that the maximum probabilities of type I errors of the above 146 spots and 21 spots are $\sim 0.045$ and $\sim 2.02 \times 10^{-7}$, respectively. Therefore, we concluded that these spots are significantly differentially expressed even taking into the account the potential effects of the varying background noise levels on the result.

Acknowledgements

We would like to thank the editor and referees for their very valuable comments that led to the improvements of our paper.
References


