Send Orders for Reprints to reprints@benthamscience.net

Data Visualization and Feature Selection Methods in Gel-based Proteomics

Tomé S. Silva[1,2,3,*], Nadège Richard[1], Jorge P. Dias[3] and Pedro M. Rodrigues[1]

[1]CIMAR/CCMAR, Centre of Marine Sciences of Algarve, University of Algarve, Campus de Gambelas, 8005-139, Faro, Portugal; [2]DTU Food, National Food Institute, Technical University of Denmark, Søltofts Plads, Building 221, 2800 Kgs. Lyngby, Denmark; [3]SPAROS Lda., CRIA, University of Algarve, Campus de Gambelas, 8005-139, Faro, Portugal

Abstract: Despite the increasing popularity of gel-free proteomic strategies, two-dimensional gel electrophoresis (2DE) is still the most widely used approach in top-down proteomic studies, for all sorts of biological models. In order to achieve meaningful biological insight using 2DE approaches, importance must be given not only to ensure proper experimental design, experimental practice and 2DE technical performance, but also a valid approach for data acquisition, processing and analysis. This paper reviews and illustrates several different aspects of data analysis within the context of gel-based proteomics, summarizing the current state of research within this field. Particular focus is given on discussing the usefulness of available multivariate analysis tools both for data visualization and feature selection purposes. Visual examples are given using a real gel-based proteomic dataset as basis.

Keywords: Independent component analysis, multidimensional scaling, partial least squares regression, principal component analysis, self-organized maps, two-dimensional gel electrophoresis.

1. INTRODUCTION

Proteomics is increasingly seen as an essential set of approaches in the systematic analysis of biological systems, providing an assessment of the changes in protein abundance that occur in living organisms. In this field, two-dimensional gel electrophoresis (2DE) is still one of the most important techniques, mostly due to its high performance regarding the separation of complex mixtures of full-length proteins.

A typical 2DE-based workflow is composed of several steps, with the final purpose (usually) being the identification of proteins that display abundance variations in response to some experimental factor. In order for biologically meaningful results to be obtained, importance must be given to a correct undertaking of all stages of the process (Fig. **1**).

This paper focuses on some of these issues, particularly at the level of experimental design and data processing/analysis, extensively discussing some of the statistical and machine learning tools available for data visualization and feature selection purposes.

2. EXPERIMENTAL DESIGN

Regardless of the context, the type and quality of the statistical analysis are constrained by the nature of the data, and therefore by the experiment itself. Because of that, understanding the issues regarding the proper design of experiments is highly relevant. Several publications within the context of proteomic data analysis thoroughly discuss these matters [1-6].

*Address correspondence to this author at the CIMAR/CCMAR, Centre of Marine Sciences of Algarve, University of Algarve, Campus de Gambelas, 8005-139, Faro, Portugal; Tel: +351 289800051; Fax: +351 289800051; E-mail: tome@tomesilva.com

Fig. (**2**) illustrates a typical and simple situation, in which we want to compare two groups (group 1 *vs.* group 2) through a number of samples which should be as representative of their respective groups as possible. The first issue concerns the number and type of replicates to use. While the standard seems to be a minimum of 4-6 replicates per experimental group, the optimal number of replicates can only be determined through a power analysis (*i.e. a posteriori*). This is important in terms of resource management, particularly when a high number of replicates is feasible, since it might be better in those cases to perform two experiments with less replicates rather than a single large experiment (or, even, to consider introducing more factors into the design). Regarding the type of replication employed, a distinction should be made between "biological replication" and "technical replication". While it is important to understand the magnitude and structure of the noise introduced throughout the technical steps, the general consensus is that biological replication is more important than technical replication, since technical variability has no biological meaning (unlike biological variability). This implies that, in general, technical replication should never be performed at the expense of biological replication. Another consensus is that one should also try to avoid, as much as logistically possible, pooling biological samples, to prevent losing relevant information on biological variability. On the other hand, there are situations in which sample pooling can be justified: an example would be when the number of biological units per treatment/group are either below or above the total number of replicates one intends to run per treatment/group. In these cases, it might make sense to run one (or more) pooled samples as they do provide better estimates of the multivariate centroid of each treatment/group (*i.e.* sample distribution location in feature space), even though they provide scarce to no

1875-5550/14 $58.00+.00

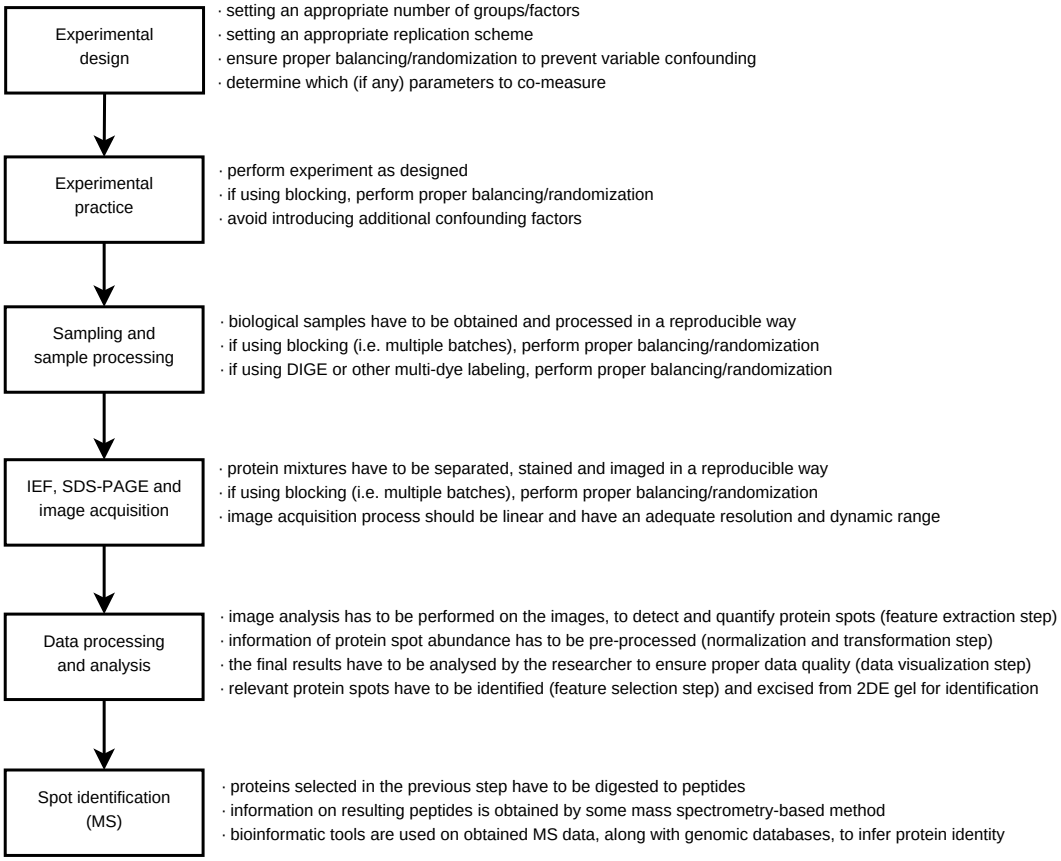


Fig. (1). Diagram depicting a high-level overview of a typical gel-based proteomic workflow. Some of the required sub-steps and important concerns that the experimenter should have in mind throughout the different stages are also described.

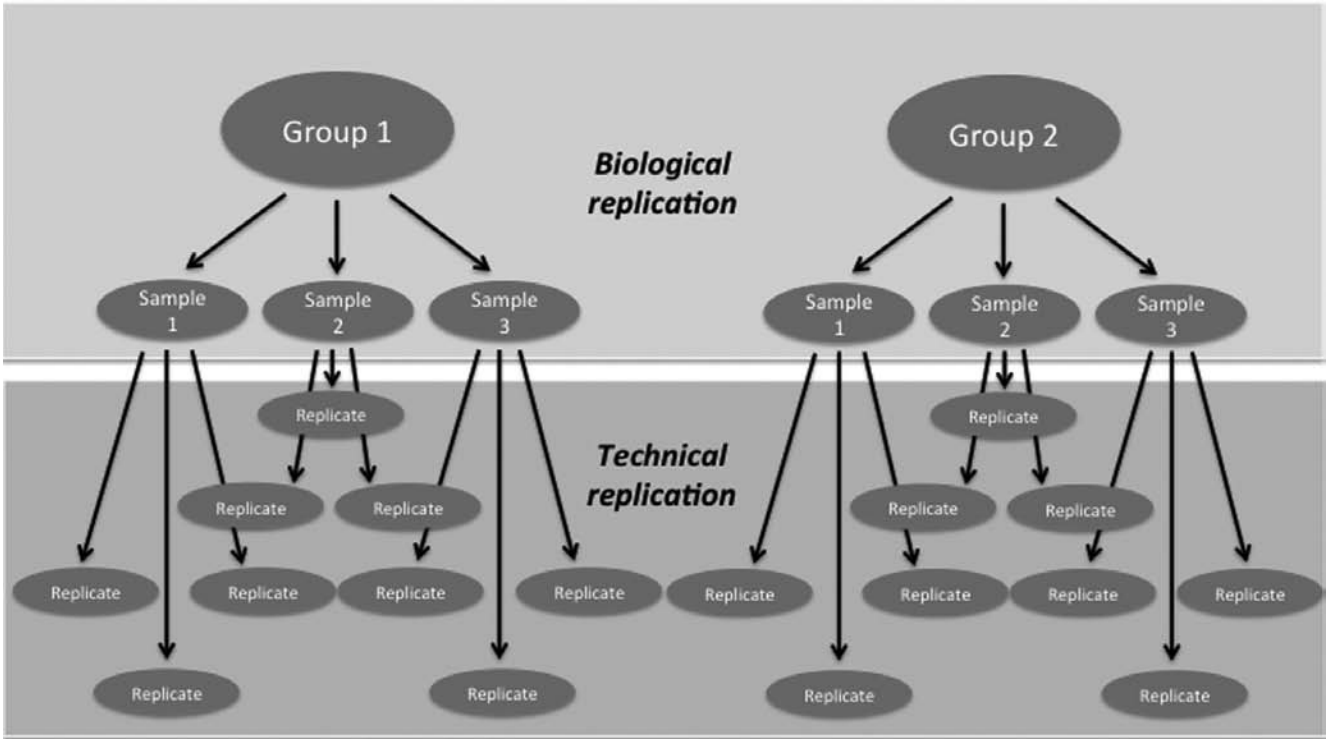


Fig. (2). Diagram showing a simple experiment in which two groups are being compared, illustrating the distinction between "biological (or experimental) replication" and "technical replication".

information about biological variability (*i.e.* sample distribution spread in feature space). This information can assist in the identification of possible biological/technical outliers, since a more accurate estimate of the central location of the distributions of each treatment/group can be calculated when pooled samples are included. There might also be logistical reasons that justify sample pooling, such as in the case in which single organisms do not provide enough protein material to run a single gel. In fact, it is important to remember that, in any case, whether bacteria or eukaryotes are used as model organism, a single protein extract will usually come from a (possibly heterogeneous) pool of cells, which implies that, at some level and to some degree, pooling is almost always a necessity.

Another consideration is how to deal with known and unknown sources of variation (both at the level of experimental practice, as in downstream steps) that can negatively impact the ability to observe a biological "signal". An important technique, called "blocking", consists in distributing experimental units between similar groups, by applying proper balancing of known sources of variation. A generic example would be (assuming one wants to test a *treatment vs. control* situation), to consider "sex" as a blocking factor and distribute half of the males to the treatment group and half to the control group (doing the same for females). In the context of 2DE, blocking should also be considered in cases such as the use of multi-dye labeling and/or when performing several extraction, IEF and/or SDS-PAGE batches. After accounting for blocking, sample distribution is often assigned using randomization to ensure that the confounding effect of even unknown sources of variation on the experimental factors is minimized.

Finally, it is relevant to stress the importance of experimental context for a proper and thorough analysis of 2DE data. Not only is it important to know and keep track of the factors (*i.e.* independent variables) associated with each gel/sample, but also other possible co-measured parameters (*i.e.* dependent variables other than 2DE data) that can, for example, provide a better insight into any eventual unexpected sample heterogeneity encountered later on.

3. IMAGE ANALYSIS/FEATURE EXTRACTION

After running a 2DE-based experiment, gels are digitalized using an appropriate scanner (depending on the type of staining used) and some form of image analysis is performed, in order to obtain information on how the different gel features vary between gels. While there is some research on 2DE gel analysis methods that work directly at the pixel level (*e.g.* [7-9]), practically all known 2DE gel analysis software consider "spots" as the basic feature and attempt to summarize all the image information by providing measures of "spot abundance" for each spot, across all gels. This is usually done by matching each spot in each gel to the equivalent spot in all other gels and then integrating the volume (area × intensity) of each spot either explicitly or by modeling each spot as a smooth curve (*e.g.* a bivariate Gaussian function) and integrating the fitted curve instead. Although this may seem trivial, in practice, 2DE gel scans often display characteristics (*e.g.* uneven background, fused or smeared spots, artifacts due to the presence of dust or bubbles) that present challenges to spot detection and image segmentation (*i.e.* defining the boundaries of each spot), spot matching and spot quantification algorithms. Some of these issues are also common with other separation techniques (*e.g.* analysis of LC-MS or MudPIT runs)

Regarding the specific software used, the overwhelming majority of gel-based proteomic studies use some form of proprietary software, which are usually classified either as taking a "classical 2DE analysis approach" (*e.g.* PDQuest™, Phoretix™, DeCyder™) or a "4th generation 2DE analysis approach" (*e.g.* SameSpots™, Delta2D™), with the difference being that the "classical" 2DE software packages usually perform spot detection before matching, while newer software packages usually perform image matching/warping before spot detection. There is still a third possible approach, which avoids working with "spots" as the basic feature, using instead the lowest-level possible features (*i.e.* pixels) as basis. Despite the fact that this paper focuses on the "spot-based" approaches, most of the described data visualization and feature selection methods also apply to "pixel-based" approaches. Fig. (**3**) shows an overview of these different used approaches for image analysis.

While a deeper view on the issues of image analysis is outside the scope of this paper (see [8, 10-15] for reviews on the subject), it is important to note that some variability in the results (*i.e.* introduction of random errors) can be attributed to this step. It is expected that improved, increasingly automated, image analysis pipelines will help mitigate these problems, in the near future.

4. DATA PRE-PROCESSING

Regardless of the software used, the end result of 2DE image analysis is usually given as a table or a matrix, where each row represents a sample/gel, while each column represents a variable/spot (illustrated in Fig. **4**). Geometrically, the matrix of samples is often transposed, so that each sample represents a (column) vector in a p-dimensional ("feature") vector space that has as many dimensions as the number of spots.

Before statistical analysis can be performed, several types of pre-processing algorithms are applied to correct for some of the technical issues. The first point concerns the possible existence of missing values. Particularly in the case of software packages that use a "classical 2DE analysis approach" (*i.e.* where spot detection is performed before matching), it is very common to have wide variations in the number of spots detected per gel (often more than 10%, mostly due to gel-gel technical variability), which implies the existence of "empty cells" (or "missing values"). This problem can be solved in several ways [1, 5, 16-19]. First, it is important to take into account that these missing values can sometimes simply be a symptom of a reduction in protein abundance below the threshold of detection (either of the staining or of the peak detection algorithm). In this case, it would make sense to replace the missing value either by a zero or by some estimate of the local background intensity. On the other hand, more often than not, these missing values actually result from inconsistent spot detection results and/or incorrect spot matching (as explained above, due to technical variability). These situations can be sometimes (but not always)

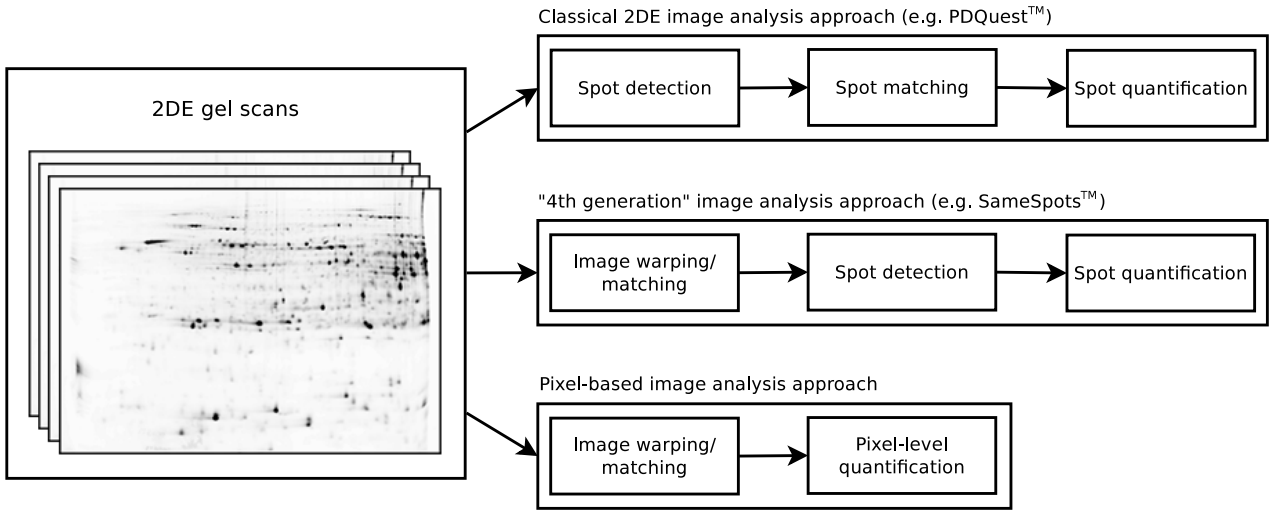
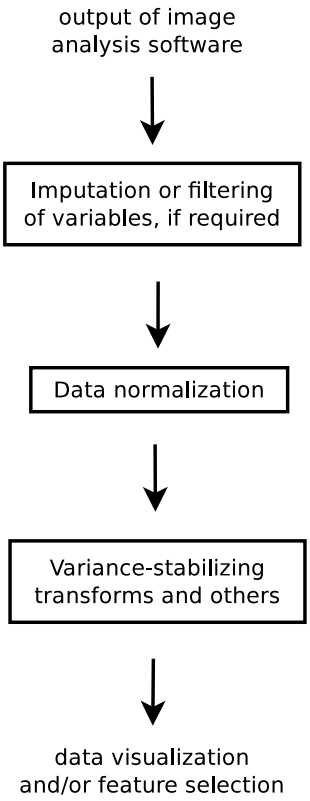


Fig. (3). Diagram showing the different type of approaches used for analysis of 2DE gel images. Some common steps (like "background subtraction") have been omitted for clarity.



	X76n	X109n	X106n	X103n	X100n	X1 ...
Bt1-1	17.14493	19.21325	18.33822	16.63203	16.3086	16
Bt2-1	16.64908	18.83419	18.07880	16.40755	16.04035	15
Bt3-1	17.58192	18.77949	17.84658	16.01431	15.94653	14
Bt1-2	17.93956	19.37274	18.46946	17.19217	16.50839	15
Bt2-2	16.92856	19.34234	18.40109	17.00316	16.70143	16
Bt3-2	17.14513	19.25894	18.28177	16.92022	16.27011	15
Bt1-3	17.52932	17.63848	18.15603	16.97973	17.01202	15
Bt2-3	16.75193	18.98537	18.15102	16.60695	15.96461	15
Bt3-3	14.89414	19.24115	18.16069	15.19019	15.09952	15
Bt1-4	17.19295	18.74706	18.07524	16.59247	16.33450	16
Bt2-4	17.50601	19.29449	18.27639	16.47854	16.46002	15
Bt3-4	16.79147	19.28416	18.20944	16.80281	16.06193	15
Bt1-5	16.42859	19.12906	18.08399	16.54993	15.80660	16
Bt2-5	17.77268	19.05522	18.05079	16.04618	16.02540	16
Bt3-5	17.80109	19.34112	18.49924	16.79071	16.72975	15
At1-1	16.05098	18.51913	17.51376	16.56271	15.79523	16
At2-1	15.71247	18.29564	17.45454	15.79927	15.94605	15
At3-1	17.69884	18.70175	17.90718	15.99844	15.96705	15

Fig. (4). On the left, overview of a typical data pre-processing stage in a gel-based proteomic analysis pipeline. It is relevant to note that some image analysis software packages already internally perform these steps as part of their workflow. On the right, an illustration showing how spot data is usually represented: as a matrix in which every row represents a sample/gel, while every column represents a variable/spot across all gels.

mitigated by time-consuming manual editing of each spot. A common strategy here involves removal of columns (*i.e.* spots) that contain missing values. Another solution, that should be applied after removal of spots with a high proportion of missing values, involves the use of imputation methods to estimate the missing values, which seems to have a good performance in the cases where missing values are few (always less than 25%, but preferably less than 5% of the total number of cells) and randomly distributed. Nonetheless, care should be taken if using these imputation methods for anything other than exploratory data visualization. Examples of possible imputation procedures include approaches based

on k-nearest neighbors, least squares or partial least squares methods [18].

Another issue relates to the need for "data normalization" before analysis, to correct for such factors as variations in protein load per gel, staining variations (both in terms of spot intensity as in terms of background intensity), scanning variations and/or distinct fluorophore efficiencies (in the case of DIGE approaches [20, 21]). There are many possibilities here and different 2DE analysis software packages often apply distinct normalization approaches. The simplest solution, for classical 2DE datasets, involves expressing each spot abundance value as a ratio between that value and the sum of the abundances of the spots in the same gel (*i.e.* the total spot volume for each gel). Though this solution should mostly solve biases related to variations in protein load, there are often nonlinearities (*e.g.* high-intensity, possibly saturated spots) that may compromise the validity of "total spot volume"-based normalization methods. A possible way of addressing this point would be to remove the highest-intensity spots from the calculation of total spot volume or simply use a smaller set of mostly-constant protein spots (which can be chosen through a preliminary analysis) as reference. Other, less often used, normalization methods include LOESS regression-based [22] and quantile-based normalization [23-25]. More recently, the use of normalization methods inspired by the ones used in transcriptomics (*e.g.* SameSpots' normalization method and the VSN method [26, 27]) have become increasingly prevalent. This is mostly due to the fact that these methods truncate differentially-expressed spots from the regressions, implicitly restricting normalization to be performed based on mostly-constant protein spots, while explicitly modeling both multiplicative (*e.g.* differences in protein load/spot staining) and additive biases (*e.g.* background staining). In the case of DIGE approaches, normalization is performed based on the presence of a common internal standard (usually labeled with Cy2) in all gels, so protein quantity in the other channels (Cy3 and Cy5) are commonly given as ratios in relation to the internal standard (Cy3/Cy2 and Cy5/Cy2). Nevertheless, corrections are required to account for multiplicative biases, not only due to differences in protein load, but also due to differences in fluorescence efficiency between dyes, which implies that the use of methods such as VSN is also relevant in this context.

Finally, there's the issue of data transformation, which attempts to address the positive correlation that exists between a spot's abundance and its variance. This can be attributed to the fact that high-abundance spots tend to have larger areas and, therefore, volume calculations for high-abundance spots are affected by a larger amount of noise, since the errors of single pixels add up. This is a problem for *e.g.* ANOVA-based analyses, which usually assume not only that spot abundances are normally distributed and independent, but also that variances are homogeneous (*i.e.* homoscedasticity). To assess if such a transformation is required, it is common to plot each spot's standard deviation as a function of its mean abundance, in order to visually assess whether there is any specific trend. The classical way of correcting these trends is through a log-transformation, which usually displays a strong variance stabilization effect on highly-abundant spots. On the other hand, log-transformation cannot handle negative values and is unstable for values near zero, which is probably why other transformations (*e.g.* arsinh and cubic root) seem to provide better results [5]. There are also suggestions that more generalized transformations should be used (for example, the Box-Cox transform or other types of power transforms), with the specific transform function being chosen through (*e.g.* maximum-likelihood) fitting [28]. Many 2DE analysis software packages, particularly in the case of DIGE, already return log-transformed data (and often perform normalization in log-space), which is sensible given that (relative) abundance results often constitute ratios.

5. DATA VISUALIZATION/EXPLORATORY DATA ANALYSIS

The first step after applying the appropriate normalization and transformation methods should be to look at the data itself, as it constitutes an important diagnostic step in the assessment of the quality of data (and of the underlying experiment). This helps to avoid erroneously taking statistical artifacts as meaningful biological signal. Fig. (**5**) illustrates this by showing a situation where a confounding variable (in this case, sex) induces what's known as the Yule-Simpson paradox: the answer to the question "is this protein's expression positively or negatively correlated with body weight?" depends on whether the population is treated as one or as two sub-populations. It is particularly due to these types of situations that an assessment of sample heterogeneity is important before any statistical tests are applied.

The problem with proteomic datasets (and other –omics datasets) is that the common tools for exploratory data analysis (*e.g.* boxplots, scatter plots) are inadequate for visualization of points/vectors in a p-dimensional space (where p is often on the order of hundreds or thousands of variables/spots). On the other hand, the intrinsic dimensionality of the subspace implied by the samples (minus noise) is usually quite low, particularly with simple experimental setups (*e.g.* one homogeneous group is compared against another homogeneous group), which justifies the use of multivariate dimensionality reduction approaches that attempt to preserve most of the high-dimensional information through either projections (*e.g.* principal component analysis, PCA) or embeddings (*e.g.* multidimensional scaling, MDS) in low-dimension spaces (usually 2 or 3). Due to the low number of samples and high number of features in typical 2DE experiments, these datasets display specific properties (*e.g.* multicollinearity and sparse sampling of the feature space) which can be problematic for some multivariate approaches [29, 30]. Several papers within the field of proteomic data analysis explore, discuss and review some of these methods [1, 5, 7, 11, 14, 17, 31-36].

5.1. Linear Projection Methods

Linear projection methods approach the issue using linear algebra, often formulating the problem under the form of a singular value decomposition problem, for which there are several known reasonably efficient (*i.e.* polynomial complexity) algorithms. Principal Component Analysis (PCA) is the most well-known example of such methods, but they also include variants for categorical data, such as Correspondence Analysis, or supervised variants, such as Partial Least Squares Regression.

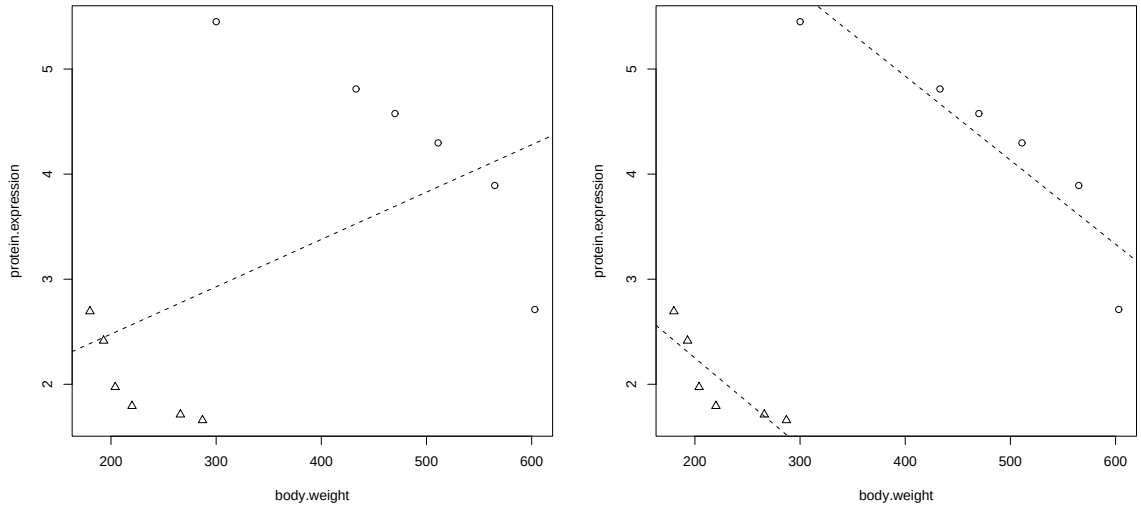


Fig. (5). Hypothetical example showing the expression of a protein in function of an individual's body weight. Samples from females are represented with a triangle and samples from males are represented with a circle. On the left, linear regression fit using all points (positive slope); on the right, linear regression fit performed separately for the two sub-populations (negative slopes).

In the case of PCA, the procedure can be seen as an orthogonal linear transformation of the set of vectors that represent the samples, such that most of the observed variance aligns with the first few axes [37]. This can be efficiently achieved by finding the direction in the p-dimensional space along which most variation occurs (called Principal Component 1, or PC1), then looking for the direction orthogonal to PC1 along which most of the residual variation (*i.e.* after removal of variation along PC1) occurs (called PC2), then looking for the direction orthogonal to PC1 and PC2 along which most residual variation (*i.e.* after removal of variation along PC1 and PC2) occurs (called PC3), and so on, until one has a number of principal components equal to the original dimensionality of the dataset (or, alternatively, until the amount of residual variation is below a certain threshold). Then, the projection of every sample vector on these "principal axes" is calculated, effectively performing a basis change (from using the set of variables/features/spots as basis to using the set of principal components as basis) and, optionally, dimensionality reduction (if the vectors are only projected unto a subset of the calculated principal components). An assessment of the quality of the projection is usually performed by looking at the variance explained by each PC: a successful projection usually retains most of the total variance along the first 2 or 3 components, enabling accurate visualization of the samples using (at most) 3 scatter plots. Since, for simple gel-based experimental setups (*e.g.* one homogeneous group compared against another homogeneous group), the first two components often retain 50-90% of observed variation, one can safely assume that, in these cases, samples which are distant in a 2D PCA projection are also distant in the original p-dimensional "feature space" (*i.e.* in terms of 2DE gel features).

The output of the PCA procedure is usually given by a "loadings plot" and a "scores plot". The "loadings plot" shows the weight of each variable (*i.e.* spot) for the first few principal components (which are simply linear combinations of variables), enabling the visualization of which variables (*i.e.* spots) contribute the most to observed variance and how these are generally correlated between themselves. The "scores plot" displays the projected samples themselves (usually on the subspace defined by the first two principal components), allowing a visualization of the relations of similarity between samples within the subspace where most variation occurs. An option that is often employed is the use of "biplots", which show both the weights of variables for each PC, as well as the projection of each sample on each PC. These plots in which "loadings" and "scores" are overlaid can be quite useful, for example, to observe if most variation can be attributed to a large or a small set of spots, or which spots are responsible for a particular outlier, constituting a useful diagnostic tool for the assessment of sample distribution, regardless of other factors or class memberships. For this reason, and given its intuitive geometric interpretation, the use of PCA in gel-based proteomics seems quite established, with generally positive results.

Given that PCA aims to display the subspace along which most variation occurs, the relative scaling of variables (*i.e.* of the p dimensions of the feature space) is relevant and often highly determinant in terms of the obtained projection. As mentioned before, proteomic data from gel-based experiments often display a correlation between spot abundance and variance (*i.e.* larger, more intense spots tend to display larger variations), which implies that a PCA will naturally give more importance to higher abundance spots over lower abundance spots. Although variance-stabilizing transforms (like the log or arsinh transforms) usually mitigate this effect to some degree, it is common for "scaling methods" to be applied prior to PCA (and other projection-based methods). A typical choice, dubbed "autoscaling", is to simply divide each abundance value by the standard deviation of the spot's abundance after mean centering, effectively forcing variance homogeneity across all variables. Although this ensures that every variable is "treated equally" by the PCA, it also implies that spots displaying very low variances are greatly amplified, often making it difficult to distinguish noise from signal. Suggested alternatives include using a correction parameter (*i.e.* dividing by the standard deviation plus a constant, rather than just by the standard

deviation), to prevent the amplification of irrelevant spots with very small variations, or the use of "group scaling" (which can be seen as a form of supervised filtering of the variables) [32]. Finally, it is important to have into account that both atypical samples (*i.e.* outliers) and atypical variables can have a disproportionate weight in determining the obtained projection. Although appropriate scaling might mitigate this to some degree, removal of atypical samples (*i.e.* gels) or variables (*i.e.* spots) from the dataset prior to PCA may also improve the quality of the obtained projection.

An illustration of the effect of scaling on the obtained projections can be seen in (Fig. **6**). This example is based on an actual 2DE dataset (VSN-normalized) in which two treatments ("A" and "B") were being compared, with biological samples being taken at three time-points (0h, 6h and 48h). It is straightforward to see that simple autoscaling already provides a significant improvement over no scaling, clearly showing that samples can be roughly clustered in three groups: 48h samples from both treatments, 0h+6h samples from treatment A and 0h+6h samples from treatment B.

An interesting variation on PCA, dubbed Independent Component Analysis (ICA), also looks for orthogonal (*i.e.* independent or, at least, uncorrelated) directions in the original space, but instead of aligning them along the directions of maximum variance, it aligns them (usually) along the directions of maximum non-Gaussianity (which is assessed using measures such as kurtosis and negentropy) [38]. The underlying logic is that, when measuring the proteome across different homogeneous groups, one is often interested in the spots for which there are differences between the means of the groups (*i.e.* there is a statistically-significant effect). Assuming the distribution of each group is Gaussian-like, when there are no differences between means for a certain spot, the resulting overall distribution will also be Gaussian-like. On the other hand, when there are significant differences between the means of the groups, the sum of their distributions will result in a (probably less Gaussian-like) polymodal distribution. Fig. (**7a**) shows an example in which three groups/treatments are being compared. It can be seen that, for spots which display little to no differences between groups, the resulting overall distribution will be Gaussian-like (upper panel), while the same cannot be said when there is actually a significant difference between groups/treatments (lower panel). In Fig. (**7b**), ICA was applied to the above-mentioned example proteomic dataset for illustrative purposes. Looking at the obtained projection, we can see that it is actually quite close to the one obtained via PCA, with the three "main clusters" of samples shown reasonably well separated. In situations of equal (and relatively low) group variances, a spreading of the group/treatment distribution means also implies that the overall variance of the distribution should increase, which is why the directions of maximum non-Gaussianity often coincide with the directions of maximum variance. In this sense, it is not very surprising that PCA and ICA might provide consistent projections, specifically when most of the variation in the dataset is due to the treatment/group effects (rather than just technical noise or intra-group biological variability).

Another type of linear projection-based visualization method is Partial Least Squares regression (or Projection to Latent Structures), which can be seen as a supervised variant of PCA [39]. As such, it does not constitute an unbiased diagnostic tool in the assessment of data quality (as PCA does), but nonetheless provides insightful visual information on the structure of the dataset and its relation with the experimental factors, which is why it is mentioned here as a data visualization tool, rather than strictly as a feature selection tool. This method requires the preparation of two matrices: the X matrix (or matrix of predictors, which contains the spot data across all samples), which is the usual input to a PCA, and the Y matrix (or matrix of responses, which contains the factors/class-membership across all samples). Unlike PCA, which takes the X matrix and attempts to find the directions of greatest variation (regardless of any eventual sample heterogeneity or classification differences), PLS regression provides the directions in the X-matrix space that better explain the variation in the Y-matrix (or, alternatively, the X-matrix vectors are linearly projected onto the subspace where prediction of the factors/class-memberships is optimal). In this sense, PLS can be seen as a supervised variant of PCA where we want to focus on the variables (*i.e.* spots)

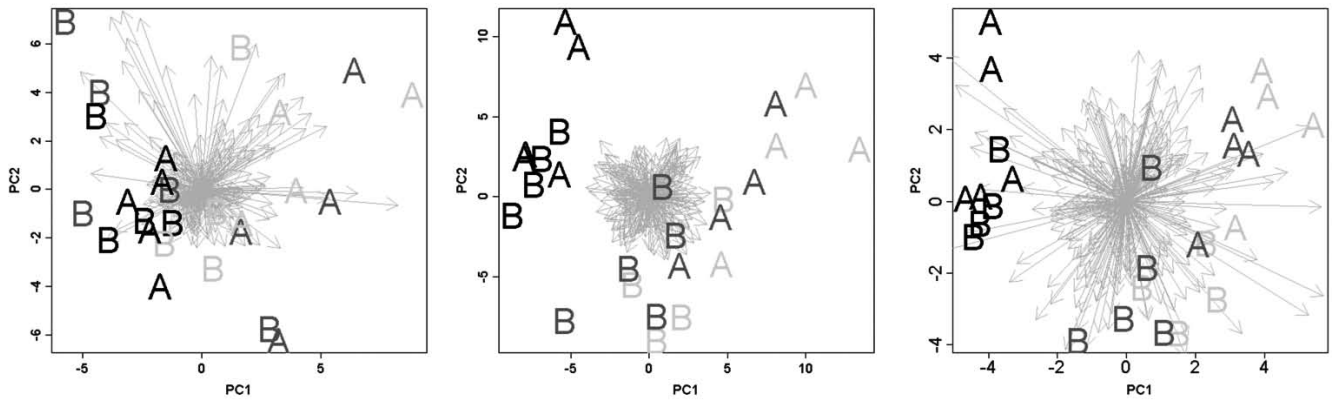


Fig. (6). Biplots showing the effect of applying different types of variable scaling prior to PCA. From left to right, plots represent either "no scaling", "autoscaling" or "group scaling". The projection of all samples on PC1 and PC2 is represented, with the symbol coding for treatment group (either "A" or "B") and color coding for sampling time (light grey, dark grey and black for 0h, 6h and 48h, respectively). Arrows represent variable loadings.

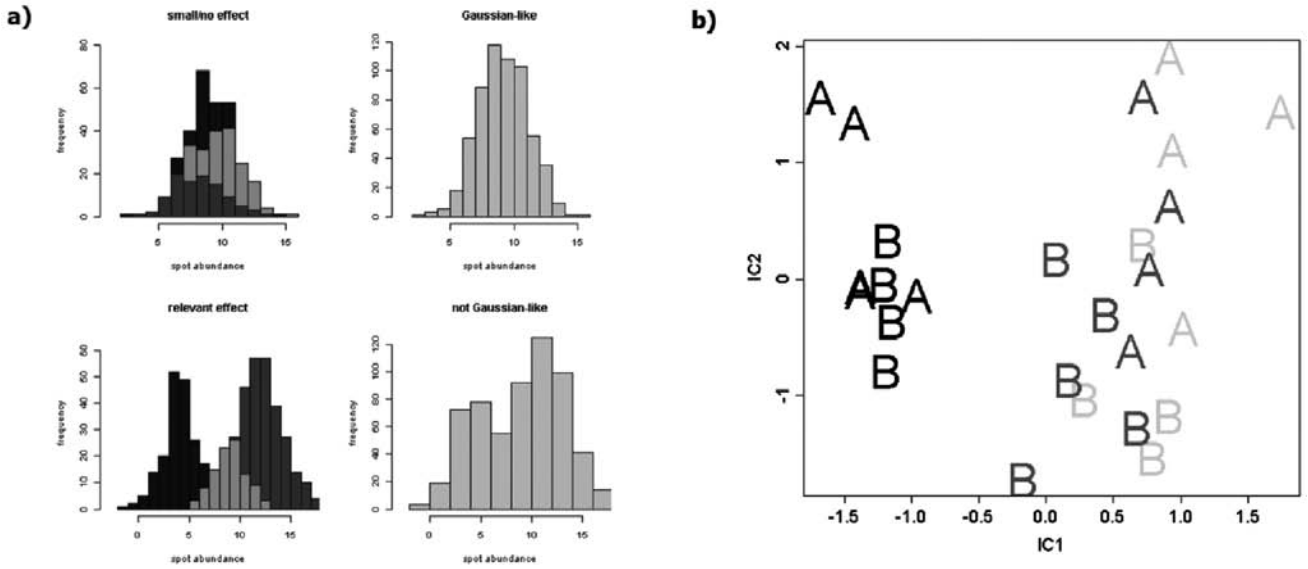


Fig. (7). a) Example histograms illustrating the underlying reason for the use of non-Gaussianity as a measure of how "interesting" a variable/spot is. **b**) Results of ICA on the example proteomic dataset, using the FastICA algorithm. Samples/gels are plotted according to their projections along IC1 and IC2, with the symbol coding for treatment group (either "A" or "B") and color coding for sampling time (light grey, dark grey and black for 0h, 6h and 48h, respectively). The dataset was pre-processed by mean-centering and autoscaling.

which correlate better with the other experimental variables (factors, groupings, co-measurements), regardless of whether those variables contribute most to the total variance among all spots. Therefore, the usefulness of PLS (and related methods) in data visualization extends beyond simple group-based experimental setups, enabling easy visualization of eventual correlations between proteomic data and other co-measured variables (*e.g.* transcriptome, metabolome).

Similarly to PCA, a PLS regression is usually presented as a set of plots: X-loadings, X-scores, Y-loadings and Y-scores. Since PLS is usually performed under the constraint that the covariance between X-scores and Y-scores is maximized, these are often represented as a single "scores plot". In addition, plots of the X-loadings and Y-loadings are often overlaid to better display correlations between predictor (*i.e.* proteomic) and response variables (within the projected subspace). Also, as for PCA, looking at the distribution of explained X-matrix variance by the latent variables (*i.e.* the PLS equivalent of the principal components) provides a good assessment of the proportion of variance that is being retained/represented in the projection. Unlike PCA, it is also important to assess the distribution of explained Y-matrix variance, to determine if the observed data (*i.e.* X-matrix) does indeed contain information that linearly relates to all the experimental factors and/or co-measurements (*i.e.* Y-matrix).

Furthermore, it is important to take into account the often disproportionately high impact of outliers in the obtained projections, as well as the possibility of overfitting, which can imply poor PLS model stability and low reliability of estimated regression coefficients. For this reason, it is common to apply cross-validation (*e.g.* k-fold cross-validation, jack-knifing or repeated random sub-sampling) to guide PLS model selection. Cross-validation is generically performed by separating samples into "calibration samples" (used for PLS model construction) and "validation samples" (used to assess the quality of the PLS models), usually repeating the procedure with different calibration/validation groupings. Such procedures provide parameters that quantify the predictability and reliability of PLS models, such as the prediction error sum of squares (PRESS) or the root-mean-square value of prediction error (RMSPE), which can be used *e.g.* to select the optimal number of latent variables. Outliers can often be detected by *e.g.* looking for samples with both high Studentized residual and high leverage (which implies that the sample is both atypical and has high impact in the final model), providing an alternative to PCA-based methods of outlier detection.

Finally, as with PCA, the relative scaling of variables in PLS (for both the X and Y matrices) is a very important issue, since it directly affects the obtained projection. In the case of the X-matrix (representing the proteomic data), the above-mentioned statements regarding PCA equally apply to PLS regression. In addition, it is also important to take into account the scaling of the factors (*i.e.* Y-matrix). Even if we assume mean-centering and autoscaling of the different factors, to ensure that they are given equal weight during the PLS model generation, there are many situations where a certain factor can be encoded in several non-equivalent ways, which are likely to result in distinct PLS models. An example would be the encoding of a "time" variable: if we have an experiment in which samples were taken at 4 different time-points (*e.g.* at 0 h, 1 h, 6 h and 24 h), the "time" variable can be encoded (at least) in 3 different ways: as 4 (binary) variables or as 1 variable (either encoded as {1, 2, 3, 4} or as {0, 1, 6, 24}, depending on whether we consider the sampling points to be "equally-spaced in feature space" or not).

The use of linear projection-based methods (like PCA and PLS) has obvious advantages, such as the fact that the dimensions of the target space are explicitly defined and

straightforward to interpret (the PCs/LVs are simply linear combinations of the variables/spots), which probably explains their high popularity and success in approaching multidimensional datasets within the context of gel-based proteomics. On the other hand, these methods implicitly assume linearity, that the p-dimensional feature space is metric and that the Euclidean distance is an appropriate measure of dissimilarity between points/vectors (which is the underlying reason for, among other issues, the sensitivity of these methods with regard to the relative scaling of the variables). In cases of very high dimensional feature spaces, it has been shown that the Euclidean distance can be a suboptimal measure of dissimilarity. This is due to the fact that, as the number of dimensions goes to infinity, the spread of the distribution of Euclidian distances between (random) points goes down, implying that, as the number of dimensions goes up, points/vectors become increasingly "equidistant", when using the Euclidian distance as dissimilarity measure. An explanation of this phenomenon can be seen in the context of k-means clustering [40]. This suggests the need to explore other (possibly nonlinear) methods of projection, or even distinct dissimilarity measures, particularly in the case of complex/highly structured experimental designs (*e.g.* involving several factors, using heterogeneous groups).

5.2. Distance/Embedding Methods

Embedding methods, of which Multidimensional Scaling [41, 42] is the best-known example, take a distinct, two-step approach: first, a dissimilarity matrix (which contains the distance between every two objects) is calculated from the high-dimensional data, using a chosen metric or non-metric measure of dissimilarity between objects (*i.e.* gels or spots); then, the objects (gels or spots) are embedded in a low-dimensional space in a way that some measure of reconstruction error is minimized. In the case of Classical/Metric Multidimensional Scaling (MDS), the configuration of points is chosen so that it minimizes a strain/stress function (usually, the sum of squared differences between the actual calculated distances and the distances achieved in the low-dimensional embedding). Alternatively, other type of reconstruction error criteria can be used, as in the case of Non-metric Multidimensional Scaling (NMDS), where the embedding is performed so that the ranks of inter-gel or inter-spot dissimilarities are (as far as possible) preserved in the low-dimensional embedding.

Besides the method of embedding, one has to be aware of the impact of the chosen dissimilarity measure on the final result. The most intuitive type of dissimilarity measure would be the Euclidean distance over the p-dimensional "feature space", which (as already mentioned) is the distance implicitly used when linear projection-based methods (*e.g.* PCA) are applied. In fact, when Classical MDS is performed using Euclidean distance as dissimilarity measure between samples, the results obtained are similar to the results obtained when performing a PCA on those samples. In this sense, Classical MDS can be seen as a generalization of PCA, which allows the use on non-Euclidean dissimilarity measures, while other types of MDS can be seen as generalizations of Classical MDS, allowing the use of different embedding criteria. The immediately obvious advantage of MDS-based methods over PCA-based methods for data visualization is the possibility of applying alternative dissimilarity measures that can address some of issues regarding the use of Euclidean distance (*e.g.* high dependence on the relative scaling of variables, sub-optimality for very high dimensional datasets).

The number of constraints on the applicable dissimilarity measures is small: the measure should be strictly non-negative, symmetric (*i.e.* the distance between A and B should be the same as between B and A) and zero if (and only if) the two objects/samples are the same (formally, it should constitute a semimetric). Specific examples include variations and generalizations of the Euclidean distance, such as the Harmonically Summed Euclidean, Mahalanobis or Minkowski distances, but also other types of dissimilarity measures, such as cosine distances or correlation-based measures. The latter are particularly interesting choices since, in practice, the correlation between spot intensity in two different gels and the correlation between two spots across all gels are often a good measure of how similar these are.

In Fig. (**8**), the effect of performing MDS on different dissimilarity matrices is displayed, using the same example as before. As expected, the configuration obtained when using the Euclidean distance as dissimilarity metric is consistent to the one obtained using PCA. On the other hand, other dissimilarity measures also seem to perform equally well, in terms of achieving good separation between known clusters (particularly the measures based on Kendall's correlation and Minkowski distances with $p < 2$). It is particularly interesting to note that no scaling is required before calculating correlation-based measures, which underlines their general scale-independence and robustness when dealing with high dimensional datasets. Also, the generally better performance for norms with a low power parameter compared to norms with a high power parameter suggests that, at least in this particular case, the number of affected spots is more informative than the magnitude of the individual effects of the spots, for the purpose of establishing sample similarity.

Another interesting generalization of PCA, allowing for nonlinear principal components, is known as Kernel PCA [43]. It bears resemblance to "distance-based methods", however it operates on the samples indirectly by representing them as a matrix of similarities (*i.e.* a kernel matrix, containing dot products between vectors), rather than dissimilarities (*i.e.* a distance matrix). Of course, as in the case of MDS, there are also different possible ways of representing similarities, which generate different (possibly nonlinear) projections. Kernel generalizations of PLS regression also exist [44].

Finally, it is worth to talk separately about a popular and useful subset of classification/embedding methods generally dubbed "clustering methods". These also involve the estimation of a dissimilarity matrix from a dataset, which is then used to separate either gels or spots according to how similar they are, either through agglomerative or partitioning methods. In the case of partitioning methods, like k-means clustering, segregation of samples is performed in a top-down fashion (*i.e.* samples are "separated into clusters"). On the other hand, the very popular (hierarchical) agglomerative clustering methods perform classification in a bottom-up

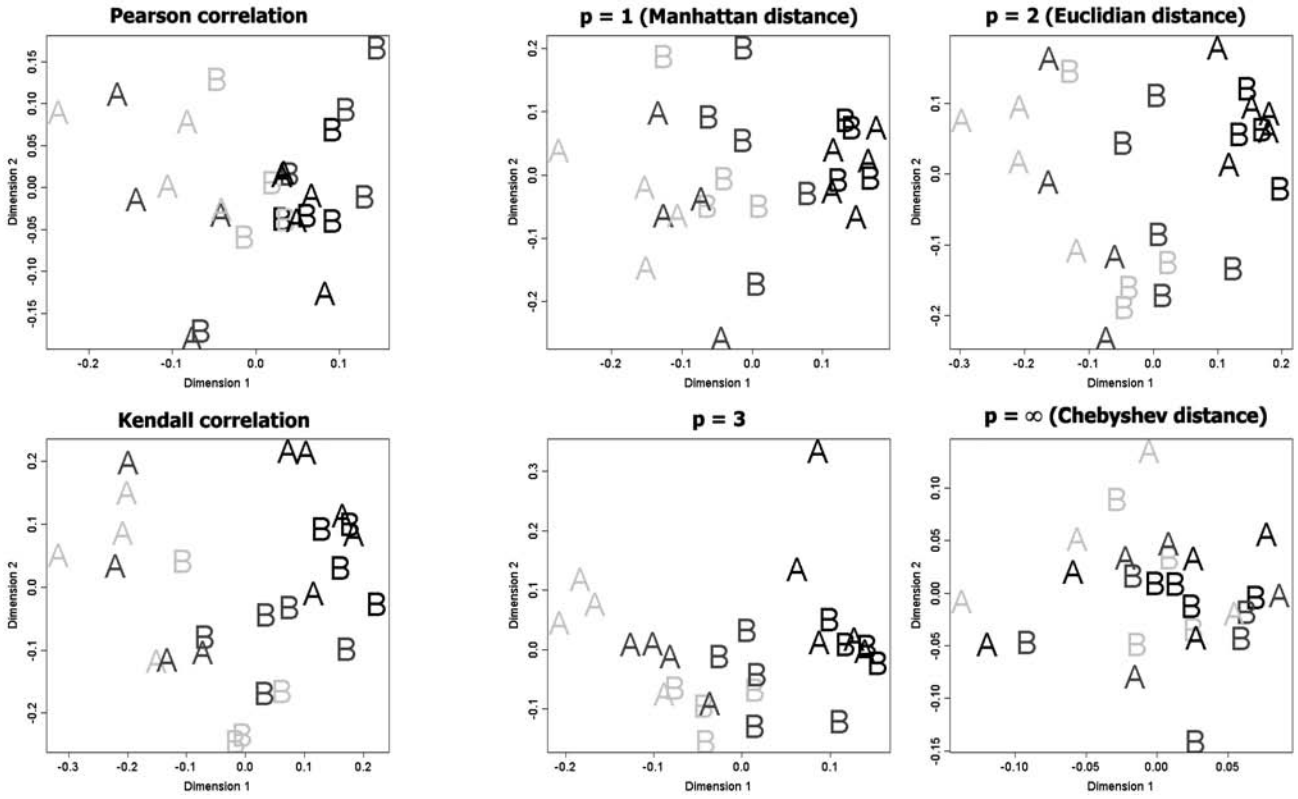


Fig. (8). Plots showing the impact of using different dissimilarity measures on the obtained MDS embeddings for the example dataset, with the symbol coding for treatment group (either "A" or "B") and color coding for sampling time (light grey, dark grey and black for 0h, 6h and 48h, respectively). On the left, correlation-based dissimilarity measures were used ($1-\rho^2$ above and $1-\tau^2$ below). On the right, MDS embeddings based on Minkowski distances (with different values for the *power* parameter), which is, in specific cases, equivalent to the use of distances such as Euclidean, Manhattan and Chebyshev. Mean-centering and autoscaling of the dataset was performed before calculating dissimilarities based on Minkowski distances.

fashion (*i.e.* samples are "aggregated into clusters"). The usual way of representing the results of clustering, in the case of hierarchical clustering, is through dendrograms, which provide a clear and easy way to interpret embedding of the gels/samples as a tree structure that attempts to reproduce the dissimilarity structure of the dataset. Also, for all these clustering methods, it is important to consider not only the type of dissimilarity measure used but also the underlying criterion for clustering, as all these are bound to affect the resulting embedding. Finally, it is worth mentioning that recent developments in the research field of clustering algorithms has provided many new approaches that could also be explored as being potentially useful in the context of gel-based proteomics (*e.g.* DBSCAN, SUBCLU, OPTICS, BIRCH, CURE and FLAME algorithms).

5.3. Other Methods

An often-used plot in transcriptomic experiments is the "heatmap" (Fig. **9**), which can be also successfully applied to proteomic data. This type of plot attempts to directly display the values of each feature (*i.e.* spot) across all samples (*i.e.* gel) using a color gradient to encode protein abundance. Some form of seriation (*i.e.* ordering of gels and ordering of spots) is usually applied so that particular clusters of gels/spots become visually apparent. This seriation is often achieved using hierarchical clustering to provide a "support tree". Overall, the interpretation of these types of plots is very intuitive, providing immediate information on global protein abundance trends across all spots and gels.

Another interesting group of visualization methods are Self-Organizing Maps (SOM), which provide a low-dimensional discrete representation of the original p-dimensional "feature space" (*i.e.* a map) obtained through unsupervised training, using a type of adaptive vector quantization that iteratively optimizes a topology-preserving mapping based on the distribution of samples in the original feature space [45]. Unlike *e.g.* PCA, self-organizing maps do not assume that the distribution of sample mostly lies along a plane (or some equally simple manifold) in the original feature space, as these are optimized to (as far as possible) represent the whole range of samples regardless of the shape of their distribution in the original feature space.

Examples of such mappings can be seen in (Fig. **10**), in which SOM were trained with our example dataset. Interpretation of these mappings is generally intuitive and follows the same underlying logic of other projection/embedding methods: samples assigned to the same node/circle are very similar, samples assigned to directly adjacent nodes are less similar, but still somewhat similar, and so on (*i.e.* samples spatially close are similar, samples spatially far are dissimilar). On the other hand, unlike PCA/MDS, where the spatial

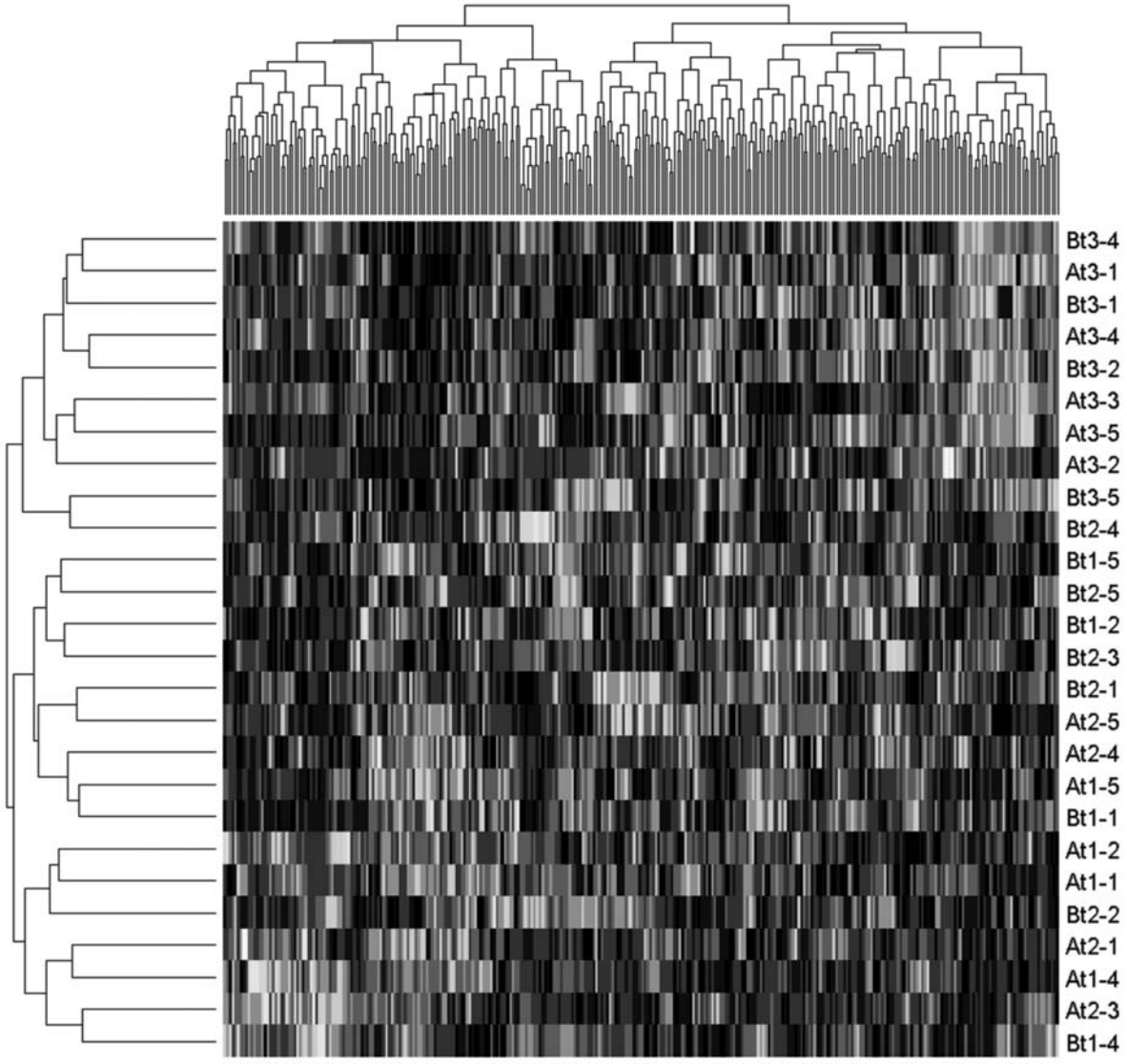


Fig. (9). Heatmap plot showing the abundance of each spot in the example dataset across all gels, encoded as color (brighter tones for "above average" abundance and darker tones for "below average" abundance). Each sample is identified according to the treatment and sampling time ("At1-X" is a sample taken at 0h from treatment A, for instance). Sample and variable ordering was performed using agglomerative hierarchical clustering as basis (with Euclidean distance as dissimilarity measure and following the "complete linkage" criterion), indicated by the presence of dendrograms for rows and columns. The dataset was pre-processed by mean-centering and autoscaling the variables.

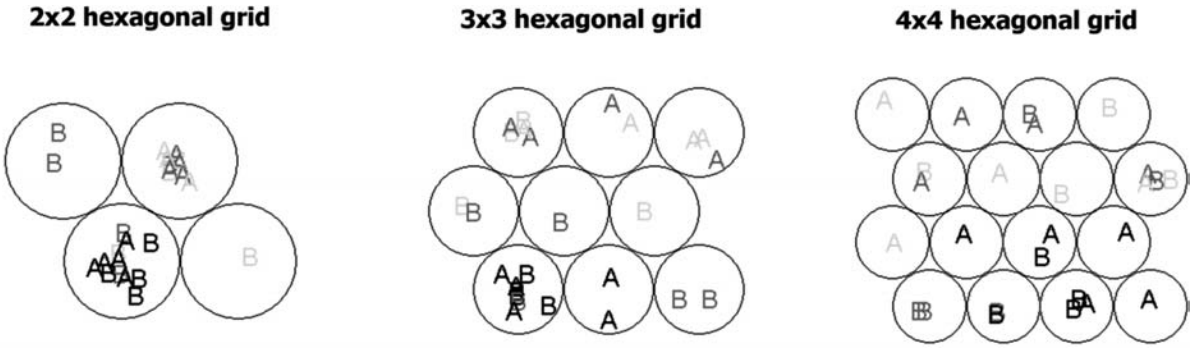


Fig. (10). Plots showing the use of unsupervised SOM to model the subspace spanned by the samples of the example proteomic dataset, using either a 2×2 (left), 3×3 (center) or 4×4 (right) hexagonal grid of nodes. Samples are directly mapped on the simplified 2D representations of the feature space, with the symbol coding for treatment group (either "A" or "B") and color coding for sampling time (light grey, dark grey and black for 0h, 6h and 48h, respectively).

distance between samples in the simplified space usually constitutes an estimate of the actual distance/dissimilarity in the original space, no such simple assumption can be made in the case of SOM. In a simplified way, SOM can be seen as a type of clustering (of the k-nearest neighbors type), with a pre-defined number of centroids, in a way that topological relations between clusters (*i.e.*, if they are "neighbors" or not) are retained. The use of rectangular or hexagonal grids therefore seems natural as topological constraints for the SOM, as these types of graphs can be readily plotted in 2D. Like other randomized/stochastic clustering methods, it is often necessary to repeat calculations a few times to ensure that stable and consistent configurations are obtained. Looking at the results of these three mappings, they seem to be generally consistent with the PCA and MDS results, showing the three main clusters of samples that are also apparent with other visualization methods ("early A samples", "early B samples" and "late A+B samples") as fairly separated.

There are also other possible (more advanced) methods that are sensitive to local structure/topology, such as the graph-based Isomap method [46] and Maximum Variance Unfolding [47], or very generic methods that work very well even for extremely complicated datasets (*e.g.* artificial neural network-based methods [23, 48]). The problem with some of these relates to the fact that they require dense sampling of the phase space (*i.e.* high number of replicates) for an accurate "manifold learning", which probably constrains their usefulness in the context of most gel-based proteomic experiments. Fig. (**11**) shows the low stability of an Isomap embedding with a low number of samples, in terms of configuration microstructure, given the observed differences as the *k* parameter (which defines each sample's "neighbourhood" as its k-nearest neighbors) is changed, even though the global trends (*e.g.* "late" samples are always separated from all the others) seem to be preserved.

For a final comparison of the different possible unsupervised data visualization approaches possible, we generated an extrapolated dataset based on our example dataset by modeling each of the 6 groups ("A-t0", "A-t6", "A-t48", "B-t0", "B-t6" and "B-t48") as multivariate Gaussian distributions and drawing random samples from them. This synthetic dataset was generated to simulate a (best-case) situation, in which 20 gels (n = 120 total gels) were ran for each group (and each of the groups is appropriately shaped, without any outliers), in order to assess the performances of the different visualization approaches. In Fig. (**12**), a comparison between using PCA, ICA, MDS with a Pearson correlation-based distance, MDS with a Kendell correlation-based measure, SOM and Isomap embedding on this hypothetical dataset is shown. It is interesting to note the particularly good performance (given that it consists of a purely unsupervised method) of Isomap embedding in this situation of dense sampling, with convex clusters, being able to perfectly separate the 6 groups. The rest of the methods provided generally

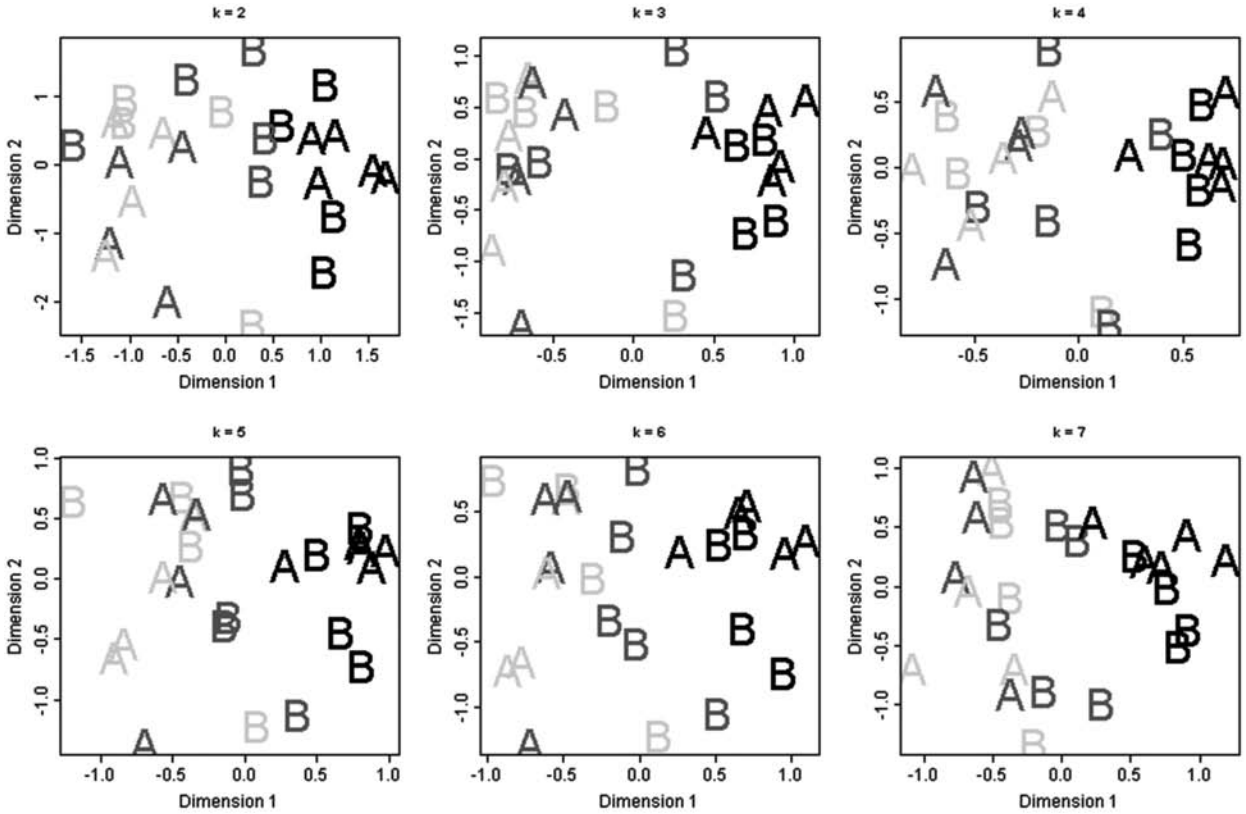


Fig. (11). Plots showing the use of Isomap embedding to display inter-sample similarity (Euclidean distance after autoscaling was used in this example), using different values for the *k* parameter (which defines the number of neighbors to consider). Samples are directly mapped on the simplified 2D representations of the feature space, with the symbol coding for treatment group (either "A" or "B") and color coding for sampling time (light grey, dark grey and black for 0h, 6h and 48h, respectively).

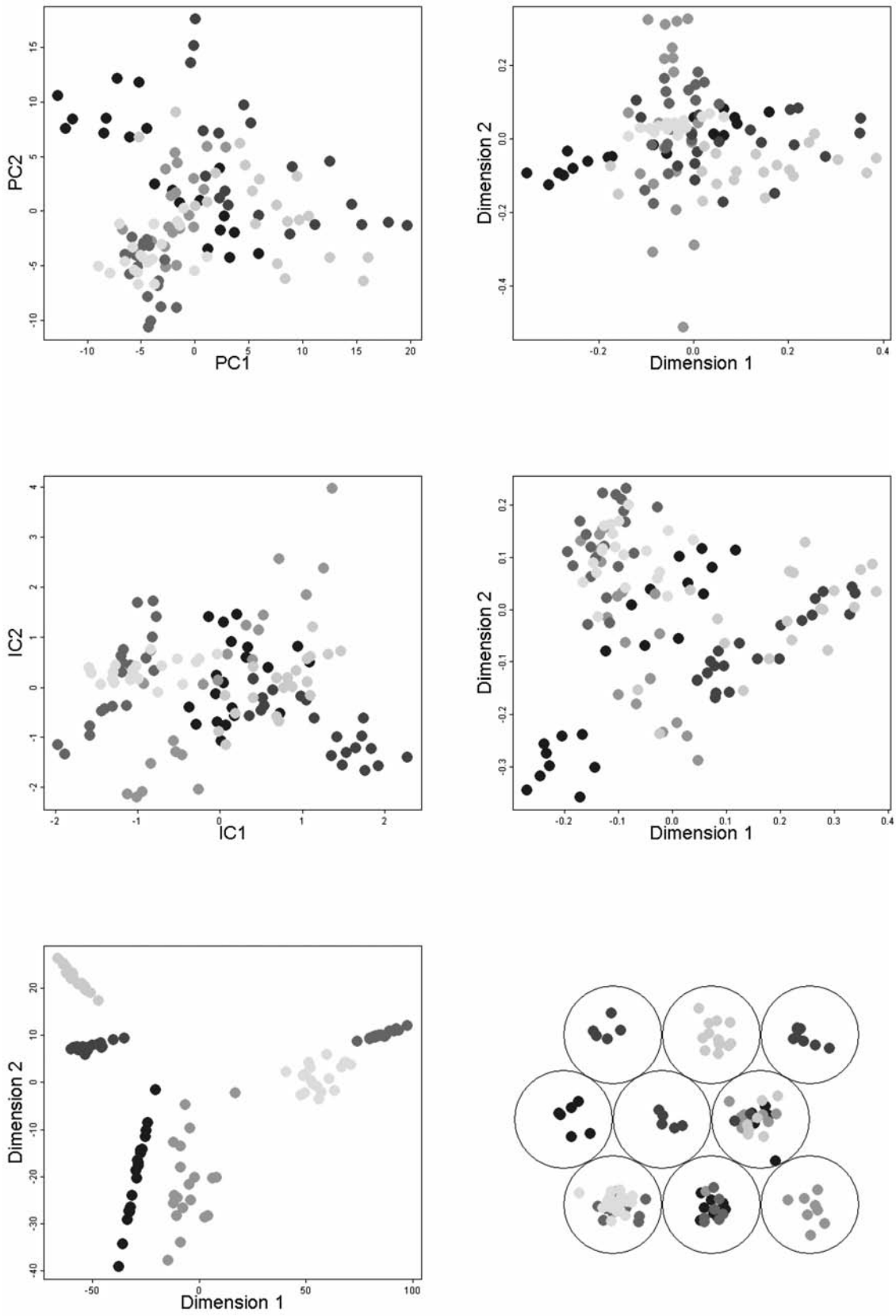


Fig. (12). Plots comparing the performance of some of the discussed visualization methods for a synthetic dataset based on our gel-based proteomic example (top row: PCA (with autoscaling) and MDS with Pearson correlation-based distance; middle row: ICA (with autoscaling) and MDS with Kendall correlation-based distance; bottom row: Isomap embedding (with autoscaled Euclidean distance and k=5) and 3×3 hexagonal grid SOM). Each group is identified with a different shade of grey.

acceptable projections/embeddings, apart perhaps from MDS with Pearson correlation-based distance, which fails to show any meaningful separation between groups.

6. FEATURE SELECTION

An essential step in the gel analysis workflow is "feature selection", which consists of separating relevant from irrelevant spots. In this specific context, the scientist is usually interested in selecting only the subset of spots that are "interesting" (*i.e.* display consistent differences in abundance between experimental treatments/groups or display a strong correlation with an experimental factor) for downstream MS-based identification. This step is therefore important to prevent waste of time and resources on the identification of features that are not affected by the experimental factors in any significant way. Choosing between the different types of possible approaches can depend not only on issues related to the purpose and design of the experiment, but also on the nature of the dataset, which is why the application of data visualization tools prior to feature selection is advisable.

6.1. Univariate Methods

The classical approach for feature selection in gel-based proteomics involves the application of univariate methods on each feature separately, usually to look for differences between experimental groups in terms of some measure of central tendency (such as the mean) through statistical hypothesis testing. This is done by formulating a null hypothesis (usually something like "there are no differences in means between groups") and performing a specific statistical test for which the distribution under the null hypothesis is known (or can be estimated/approximated). From this statistical test, a *p-value* is usually derived, which provides an estimate of the probability of obtaining a test statistic value equal or higher than the one obtained, under the assumption that the null hypothesis is true (*i.e.* an estimation of the likeliness of erroneously rejecting a true null hypothesis). Although it can be seen as somewhat arbitrary, it is common for researchers to set a threshold value (often $p = 0.05$ or $p = 0.01$) and only consider as "statistically significant" changes for which the estimated p-value is below this threshold.

Concerning types of tests used, these are generally separated into parametric and non-parametric tests. Parametric tests, like Student's t-test, Analysis of Variance (ANOVA) and related methods (*e.g.* General Linear Models) generally work under the assumption that the group sample distributions display certain properties (*e.g.* that they are normal and that their variances are homogeneous). While these methods are particularly sensitive, they also tend to perform less well in the presence of outliers and/or when significant deviations from the set assumptions exist.

Non-parametric approaches, on the other hand, generally avoid making assumptions about the underlying sample distributions. A commonly used example is the Wilcoxon-Mann-Whitney U-test, which displays a higher robustness in the presence of outliers compared to parametric approaches, besides having softer requirements (group distributions should be similar, but not necessarily normal). Another class of possible non-parametric approaches (based on resampling methods) explores the fact that, in a typical gel-based proteomic experiment and almost regardless of the type of test applied, the overwhelming majority of protein spots tend to display effects sizes that are considered "not significant" (*i.e.* the data does not support rejection of the null hypothesis of "no effect"), which provides an empirical basis for explicit estimation of the distribution of some test statistic or pivotal quantity (*e.g.* Cohen's d statistic, Behrens–Fisher's test) under the null hypothesis. Almeida *et al.* provide an example of these types of nonparametric approaches within the context of gel-based proteomics [23]. Fig. (**13**) also illustrates another possible iterative scheme based on the use of random permutations and bootstrap resampling to explicitly estimate the distributions of a test under both the null hypothesis and alternate hypothesis, rather than relying on some (possibly asymptotic) approximations that assume certain group distribution shape properties. It starts by applying some arbitrary preliminary test and considering its usual approximated distribution (in the case of Student's test, the t-distribution) as basis for setting an appropriate threshold. This allows an initial estimation of which spots are considered "significant" or not. Estimation of the distribution of the chosen pivotal quantity (in this case, Behrens–Fisher's test) under the null hypothesis can be performed using random permutations, as sample labels should be interchangeable in the case of "no effect". In the case of spots for which the null hypothesis was rejected (*i.e.* "significant" spots), sample labels are not interchangeable, so estimation of the distribution of the test under the alternate hypothesis has to be performed in some different way (*e.g.* bootstrapping samples with replacement). In both cases, it can be helpful to add a small amount of multivariate Gaussian noise (*e.g.* scaled so that its standard deviation is on the level of technical noise) to smooth out the discrete/sparse nature of the sampling. After having appropriate estimations of the distribution of the test under both the null and alternate hypothesis, it becomes possible to set an appropriate threshold for the test that achieves a balance between a low estimated false positive rate and a low estimated false negative rate. The whole process is then repeated, but using now the chosen pivotal quantity/statistical test and an appropriate threshold for segregation of "significant" from "non-significant" spots, being iterated several times. Due to the stochastic nature of these tests, there is no guarantee of convergence, so it is advisable to simply log the number of times each spot is selected as "significant" and stop iterating once it becomes clear which spots are consistently considered as such.

One important issue, regardless of the type of univariate approach employed, is how to set an appropriate threshold for a test statistic (or, alternatively, how to set an appropriate p-value threshold) to separate "significant" *vs.* "non-significant" spots, in a way that there is some degree of control over the overall number of false positives. This issue is particularly relevant within the context of proteomics (and other –omics), where univariate hypothesis tests are often performed over several hundred (or thousand) variables. In these cases, setting an arbitrary low threshold (like $p = 0.05$) is not sufficient to prevent the occurrence of false positives, which is why it is necessary to directly address this using "multiple testing correction" methods.

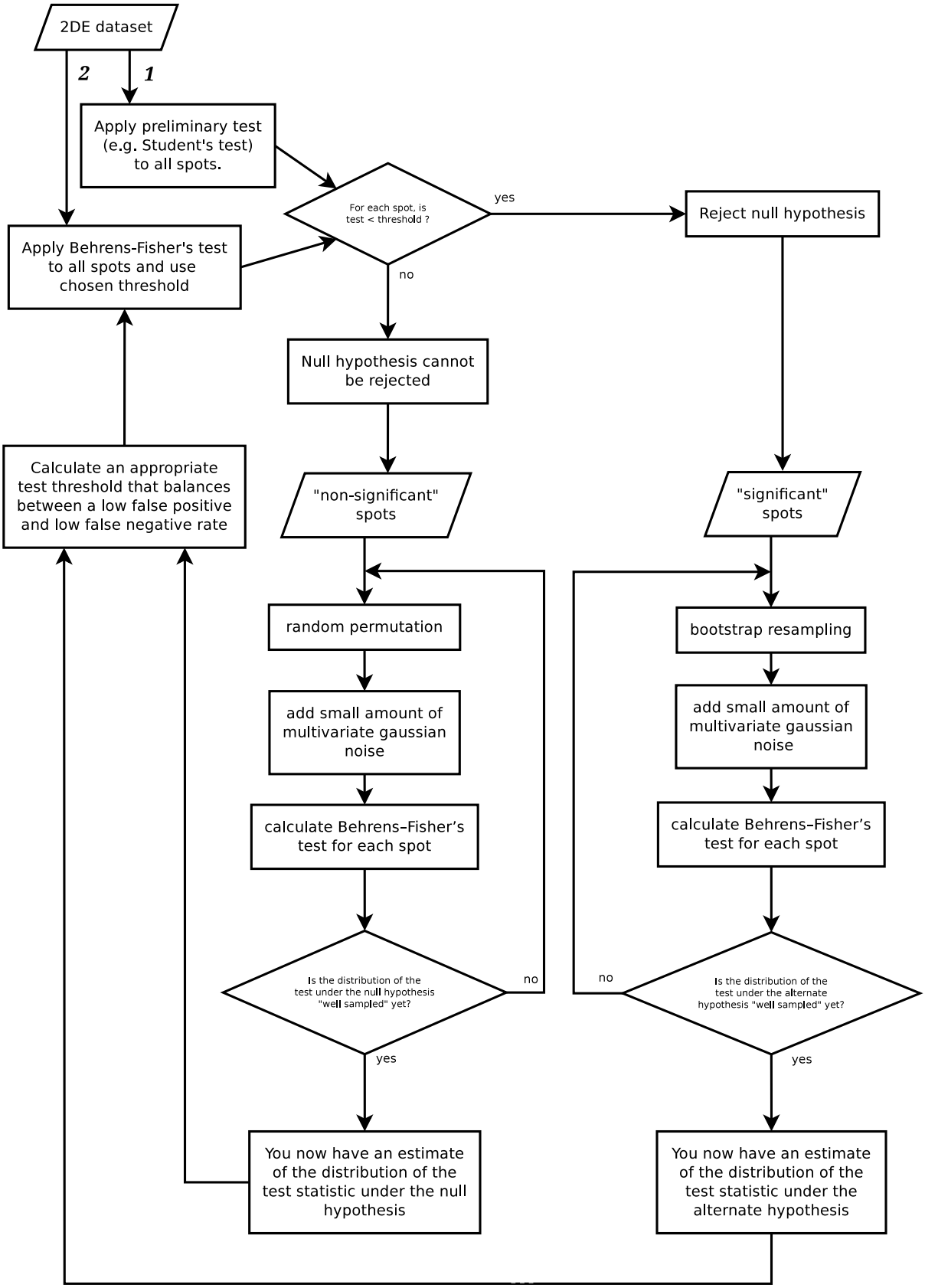


Fig. (13). Flowchart showing a generic non-parametric resampling approach for feature selection allowing the use of an arbitrary pivotal quantity/statistical test (in this case, Behrens-Fisher's test is used). These types of approaches avoid as much as possible making any predetermined assumptions regarding distribution of the test under the null hypothesis or the underlying distribution of the samples.

The simplest type of multiple testing correction is Bonferroni correction which consists in applying a p-value threshold of $p = \alpha/n$ (where n is the number of performed comparisons/spots and α is the overall probability that, over all the n comparisons, one or more constitute false positives). So, for example, if we are performing univariate hypothesis tests over 600 spots, while setting the p-value threshold to α should theoretically ensure that the probability of false positive for each of the n comparisons is only 5%, setting it to α/600 will ensure that the probability of any false positives over all the comparisons is only 5% (*i.e.* the family-wise error rate, or FWER, is 5%). Other types of multiple testing correction methods are known, some of which exploit knowledge on the distribution of p-values and the correlation between the different tests to achieve a better performance over the classic Bonferroni correction, while still effectively allowing some control over the FWER (*e.g.* Šidák correction, Holm-Bonferroni method, Simes procedure). It is also possible to use a less stringent criterion for multiple testing correction, such as the false discovery rate (FDR), rather than the FWER. The use of such FDR methods (such as Benjamini-Hochberg's or Storey's procedures) provides an increase in power (*i.e.* less false negatives) at the expense of a small (but quantifiable) increase in false positives. In the context of proteomics, some publications provide interesting discussions in this area (*e.g.* [1, 5, 49, 50]).

Finally, many authors have expressed concerns regarding the overly high reliance on univariate hypothesis testing and reporting p-values in most scientific fields [51-53]. One of the problems is that the interpretation of a p-value is not immediately intuitive, since it doesn't usually provide information on the probability of the effect being real (*i.e.* what the researcher usually would like to know), but on the probability of a non-existent effect attaining the same level of significance, which is not exactly equivalent. Another problem is that simply reporting p-values, although it allows the reader to infer on the likeliness of a certain effect being a false positive, does not really provide information on the magnitude of that effect. Despite these shortcomings, univariate approaches are likely to remain a staple for feature selection (both in gel-based proteomics and other scientific areas) due to their simplicity and power. Particularly in the –omics fields, the application of these methods within multiple testing correction frameworks allows a formal approach to feature selection that explicitly addresses and limits the occurrence of false positives, which probably justifies their central role in feature selection. Nevertheless, it is recommended that the reporting of results of univariate hypothesis testing as p-values is supplemented by some estimate (or confidence intervals) for the effect size (as assessed by, *e.g.* Cohen's d), providing the reader with information on both the likeliness of an observed effect being a false positive and on the apparent magnitude of said effect.

6.2. Multivariate methods

Besides the classical univariate approaches for feature/variable selection, it is increasingly clear that (also unsupervised, but particularly supervised) multivariate approaches are essential tools that complement univariate approaches in the context of both data visualization and feature selection, also providing important insights on variable correlation that are otherwise impossible to obtain through univariate approaches [1, 5, 7, 11, 14, 17, 31-36].

By far, the most used method for multivariate feature selection in gel-based proteomics is PLS regression (and variants), which was already mentioned in the context of data visualization. The advantage of this method is flexibility, since it allows one to look for protein spots that "explain" a certain set of experimental factors or protein spots that correlate to some other co-measured parameters. PLS regression allows, for example, the development of novel and useful "data fusion" approaches (*e.g.* [54]) that exploit knowledge on the joint distribution of proteomic and non-proteomic data. On the other hand, a disadvantage is that PLS regression tends to assign nonzero coefficients to (almost) every spot, for every latent variable, which is an issue most PLS variants try to address. One such method is called "PLS with jack-knifing", which basically consists in applying jack-knifing cross-validation (*i.e.* removing one sample at a time from the training set) and assessing the stability of the coefficients for each latent variable. In the case of a coefficient sign inversion, for instance, we can assume that its loading in the latent variable is not significant, and its coefficient can therefore be set to zero (effectively ignoring variables which do not have a consistent correlation with the factors/co-measurements of the Y-matrix). Many other variants exist, and the performance of some of these (PLS with jack-knifing, CovProc, PLS with cross-model validation and Power-PLS) has already been compared and discussed in the context of gel-based proteomics [55]. There are also other approaches (generically called "sparse PLS") which use regularization methods (*e.g.* LASSO penalization, elastic net regularization) to minimize the number of nonzero coefficients in the latent variable loadings (*i.e.* make the loading matrix sparse), easing the distinction between significant and non-significant loadings [56-58].

Besides PLS regression methods, other supervised classification and machine learning methods also have potential for application as multivariate feature selection approaches, such as decision tree learning methods [48], soft independent modeling by class analogy [59-61] (SIMCA), support vector machines [62, 63] (SVM) and artificial neural networks [23, 48] (ANN). These methods are very general and tend to be applicable to all sorts of datasets, including the ones from gel-based proteomics. It is important to note that some of these also tend to display the non-sparsity problem (*i.e.* rely on many/all spots for classification, even the ones that do not provide much information), and often maintain complex internal representations that prevent a straightforward interpretation in terms of "interesting protein spots" (*e.g.* multilayer ANN, kernel SVM). On the other extreme, some of these methods also suffer from the opposite problem, which is to simply focus on a minimal set of non-redundant spots that enable sample classification (*e.g.* heavily pruned decision trees). The issue with these approaches (outside of the context of classification) is that they tend to ignore spots which might be relevant, but that display a high degree of correlation with some other, even more relevant, protein spot. This issue can be partially mitigated either with filtering strategies (*e.g.* previous removal of highly relevant and/or generally irrelevant spots) or through the application of sequential methods (*e.g.* a classification strategy is applied to determine

the most relevant spot, that spot is removed from the dataset and the strategy is repeated to determine the second most relevant spot, etc. until all spots are ranked according to relevance).

To conclude, though these multivariate approaches do not replace the use of univariate hypothesis testing in gel-based proteomics, for the purpose of feature selection, they do provide an important complement, being particularly relevant in experimental contexts with more complex designs (*e.g.* fractional factorial designs) and/or when extensive co-measurement of non-proteomic data is performed. While results obtained through different feature selection approaches do not perfectly overlap, protein spots that display a very strong and consistent effect regarding a particular factor do tend to be selected, regardless of the type of approach undertaken, which can be exploited as a way of separating strongly affected spots from the ones that are only marginally so.

7. FINAL CONSIDERATIONS

Throughout this paper, several ways of approaching data visualization and feature selection within the context of gel-based proteomics were discussed. One important fact to have in mind is that all these methods depend on appropriate data quality, regardless of the chosen approach, which is why the experimenter must follow through the upstream steps (experimental design, experimental practice, sampling, sample processing and instrumental analysis) with all the required care in order to ensure both meaningful and statistically-sound results (this is commonly known as the "garbage in, garbage out" principle).

After image analysis, it is essential to understand which type of data is provided by the analysis software and, if needed, pre-process it. The experimenter should then proceed to exploratory data visualization (in order to assess data quality and determine the general trends) and feature selection (in order to obtain the subset of protein spots that are affected by the experimental conditions), both of which depend on the type of dataset and the type of biological problem involved (*i.e.* experimental design factors).

Looking at all assessed data visualization methods, PCA does seem like a good "default" choice, as it provides a fairly good and unbiased view of a dataset along the subspace where most variation occurs. It is therefore not surprising its standard use in 2DE-based publications, as most readers are probably already familiarized with it, providing them with a global overview of sample distribution/group homogeneity. Within this context, although it does not yet seem to be a standard, the use of heatmaps for 2DE-based publications also seems to be a very good idea, as it provides very detailed information about spot intensity across all gels, in a dense (but still visually informative) way. Other types of unsupervised approaches also seem to be relevant, at least for data exploration purposes, with ICA, MDS (with Kendall correlation-based distance) and SOM being suggested as particularly interesting. While there are similarities in how these methods approach data visualization, observing the differences in how they organize samples can provide additional insight about the groupings and heterogeneity in the dataset.

Regarding feature selection methods, it is important to take into account that what distinguishes a "relevant" from a "non-relevant" feature depends on the underlying biological question that is being addressed, as well as on the specific purpose of the experiment. As such, it is important to make the distinction between strict situations, where no false positives should occur at all (*e.g.* if one is looking for consistent and reliable molecular markers of some pathological state), and less strict situations, where a low (but quantifiable) amount of false positives is acceptable in order to reduce the number of false negatives (*e.g.* if one is simply interested in generally knowing which major pathways are affected by a particular stimulus). In the first case, it is highly advisable to restrict oneself to the use of conservative approaches that explicitly attempt to control the occurrence of false positives (*e.g.* univariate test with multiple testing correcting). Also, in this case, unless one is sure that the dataset follows the usual parametric assumptions (normality and homoscedasticity), it is best to use non-parametric approaches (*e.g.* estimation of the distribution of a test under the null hypothesis using random permutations), as the computational overhead associated with them is negligible nowadays. Regardless, it is highly advisable to present some estimate of the effect size (*e.g.* Cohen's d, Hedges's g) along with reported p-values (among other things, for the purpose of meta-analysis). In the second case, the use of alternative feature selection methods (particularly PLS regression-based methods) have been shown to provide a useful complement to the classical approaches based on univariate hypothesis testing. The use of these tools also allows the possibility of directly addressing intra-group heterogeneity, by enabling the integration of co-measured parameters in the feature selection process, opening up the possibility of novel "data fusion" methodological approaches for a better understanding of the sources of biological variability.

Concluding, this paper has shown the usefulness of several multivariate tools in the context of gel-based proteomics. The importance and relevance of these types of integrative approaches will have a tendency to increase, as technical and technological advances increasingly push forward both the quantity and complexity of available data within the context of proteomics (and molecular biology in general).

CONFLICT OF INTEREST

The authors confirm that this article content has no conflicts of interest.

ACKNOWLEDGEMENTS

This work is part of project 21595-INUTR, co-financed by FEDER through PO Algarve 21 in the framework of QREN 2007-2013. Tomé S. Silva was supported by a research grant within the context of project 21595-INUTR. Nadège Richard was supported by grant SFRH/BPD/65578/2009 (FCT, Portugal). The authors would like to thank Bo Jørgensen, Flemming Jessen, Mahaut de Vareilles and Tune Wulff for helpful discussions. The authors would also like to thank Marek Misztal for corrections of the manuscript.

REFERENCES

[1] Chich, J.F.; David, O.; Villers, F.; Schaeffer, B.; Lutomski, D.; Huet, S. Statistics for proteomics: experimental design and 2-DE differential analysis. *J. Chromatogr. B.*, **2007**, *849*, 261-272.

[2] Hunt, S.M.N.; Thomas, M.R.; Sebastian, L.T.; Pedersen, S.K.; Rebecca, L.; Sloane, A.J.; Wilkins, M.R. Optimal replication and the importance of experimental design for gel-based quantitative proteomics. *J. Proteome Res.*, **2005**, *4*, 809-819.

[3] Karp, N.A.; Lilley, K.S. Design and analysis issues in quantitative proteomics studies. *Proteomics*, **2007**, *7*, 42-50.

[4] Oberg, A.L.; Vitek, O. Statistical design of quantitative mass spectrometry-based proteomic experiments. *J. Proteome Res.*, **2009**, *8*, 2144-2156.

[5] Valledor, L.; Jorrín, J. Back to the basics: maximizing the information obtained by quantitative two dimensional gel electrophoresis analyses by an appropriate experimental design and statistical analyses. *J. Proteomics*, **2011**, *74*, 1-18.

[6] Westermeier, R.; Naven, T.; Höpker, H. Proteomics in Pratice: A guide to Successful Experimental Design. Second Edition. Wiley-VCH, Germany: 2008.

[7] Marengo, E.; Robotti, E.; Bobba, M.; Demartini, M.; Righetti, P. G. A new method of comparing 2D-PAGE maps based on the computation of Zernike moments and multivariate statistical tools. *Anal. Bioanal. Chem.*, **2008**, *391*, 1163-1173.

[8] Marengo, E.; Robotti, E.; Antonucci, F.; Cecconi, D.; Campostrini, N.; Righetti, P. G. Numerical approaches for quantitative analysis of two-dimensional maps: A review of commercial software and home-made systems. *Proteomics*, **2005**, *5*, 654-666.

[9] Rye, M.B.; Færgestad, E.M.; Martens, H.; Wold, J.P.; Alsberg, B.K. An improved pixel-based approach for analyzing images in two-dimensional gel electrophoresis. *Electrophoresis*, **2008**, *29*, 1382-1393.

[10] Berth, M.; Moser, F.M.; Kolbe, M.; Bernhardt, J. The state of the art in the analysis of two-dimensional gel electrophoresis images. *Appl. Microbiol. Biotechnol.*, **2007**, *76*, 1223-1243.

[11] Daszykowski, M.; Stanimirova, I.; Bodzon-Kulakowska, A.; Silberring, J.; Lubec, G.; Walczak, B. Start-to-end processing of two-dimensional gel electrophoretic images. *J. Chromatogr. A*, **2007**, *1158*, 306-317.

[12] Dowsey, A.W.; English, J. A.; Lisacek, F.; Morris, J.S.; Yang, G. Z.; Dunn, M.J. Image analysis tools and emerging algorithms for expression proteomics. *Proteomics*, **2010**, *10*, 4226-4257.

[13] Millioni, R.; Puricelli, L.; Sbrignadello, S.; Iori, E.; Murphy, E.; Tessari, P. Operator-and software-related post-experimental variability and source of error in 2-DE analysis. *Amino Acids*, **2012**, 1-8.

[14] Marengo, E.; Robotti, E.; Bobba, M. Multivariate statistical tools for the evaluation of proteomic 2D-maps: recent achievements and applications. *Curr. Proteomics*, **2007**, *4*, 53-66.

[15] Marengo, E.; Robotti, E.; Bobba, M. 2D-PAGE maps analysis. *Methods Mol. Biol.*, **2008**, *428*, 291.

[16] Albrecht, D.; Kniemeyer, O.; Brakhage, A.A.; Guthke, R. Missing values in gel-based proteomics. *Proteomics*, **2010**, *10*, 1202-1211.

[17] Grove, H.; Hollung, K.; Uhlen, A.K.; Martens, H.; Faergestad, E.M. Challenges related to analysis of protein spot volumes from two-dimensional gel electrophoresis as revealed by replicate gels. *J. Proteome Res.*, **2006**, *5*, 3399-3410.

[18] Miecznikowski, J.C.; Damodaran, S.; Sellers, K.F.; Rabin, R.A. A comparison of imputation procedures and statistical tests for the analysis of two-dimensional electrophoresis data. *Proteome Sci.*, **2010**, *8*, 66.

[19] Ahmad, N.; Zhang, J.; Brown, P.J.; James, D.C.; Birch, J.R.; Racher, A.J.; Smales, C.M. On the statistical analysis of the GS-NS0 cell proteome: Imputation, clustering and variability testing. *Biochim. Biophys. Acta*, **2006**, *1764*, 1179-1187.

[20] Krogh, M.; Liu, Y.; Waldemarson, S.; Valastro, B.; James, P. Analysis of DIGE data using a linear mixed model allowing for protein-specific dye effects. *Proteomics*, **2007**, *7*, 4235-4244.

[21] Corzett, T.H.; Fodor, I.K.; Choi, M.W.; Walsworth, V.L.; Chromy, B.A.; Turteltaub, K.W.; McCutchen-Maloney, S.L. Statistical analysis of the experimental variation in the proteomic characterization of human plasma by two-dimensional difference gel electrophoresis. *J. Proteome Res.*, **2006**, *5*, 2611-2619.

[22] Engelen, K.; Sifrim, A.; Van de Plas, B.; Laukens, K.; Arckens, L.; Marchal, K. Alternative experimental design with an applied normalization scheme can improve statistical power in 2D-DIGE experiments. *J. Proteome Res.*, **2010**, *9*, 4919-4926.

[23] Almeida, J.S.; Stanislaus, R.; Krug, E.; Arthur, J.M. Normalization and analysis of residual variation in two-dimensional gel electrophoresis for quantitative differential proteomics. *Proteomics*, **2005**, *5*, 1242-1249.

[24] Chang, J.; Remmen, H.V.; Ward, W.F.; Regnier, F.E.; Richardson, A.; Cornell, J. Processing of data generated by 2-dimensional gel electrophoresis for statistical analysis: missing data, normalization, and statistics. *J. Proteome Res.*, **2004**, *3*, 1210-1218.

[25] Kultima, K.; Scholz, B.; Alm, H.; Sköld, K.; Svensson, M.; Crossman, A.; Bezard, E.; Andrén, P.; Lönnstedt, I. Normalization and expression changes in predefined sets of proteins using 2D gel electrophoresis: a proteomic study of L-DOPA induced dyskinesia in an animal model of Parkinson's disease using DIGE. *BMC Bioinformatics*, **2006**, *7*, 475.

[26] Huber, W.; Von Heydebreck, A.; Sültmann, H.; Poustka, A.; Vingron, M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, **2002**, *18*, S96-S104.

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

[28] Rowell, C.; Carpenter, M.; Lamartiniere, C.A. Modeling biological variability in 2-D gel proteomic carcinogenesis experiments. *J. Proteome Res.*, **2005**, *4*, 1619-1627.

[29] Somorjai, R.L.; Dolenko, B.; Baumgartner, R. Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions. *Bioinformatics*, **2003**, *19*, 1484-1491.

[30] Simon, R. Supervised analysis when the number of candidate features (p) greatly exceeds the number of cases (n). *SIGKDD Explor.*, **2003**, *5*, 31-36.

[31] Kjærsgård, I.V.H.; Nørrelykke, M.R.; Jessen, F. Changes in cod muscle proteins during frozen storage revealed by proteome analysis and multivariate data analysis. *Proteomics*, **2006**, *6*, 1606-1618.

[32] Nedenskov Jensen, K.; Jessen, F.; Jørgensen, B.M. Multivariate Data Analysis of Two-Dimensional Gel Electrophoresis Protein Patterns from Few Samples†. *J. Proteome Res.*, **2008**, *7*, 1288-1296.

[33] Engkilde, K.; Jacobsen, S.; Sondergaard, I. Multivariate data analysis of proteome data. *Methods Mol. Biol.*, **2006**, *355*, 195.

[34] Jacobsen, S.; Grove, H.; Nedenskov Jensen, K.; Sørensen, H.A.; Jessen, F.; Hollung, K.; Uhlen, A.K.; Jørgensen, B.M.; Færgestad, E.M.; Søndergaard, I. Multivariate analysis of 2-DE protein patterns–Practical approaches. *Electrophoresis*, **2007**, *28*, 1289-1299.

[35] Mazzara, S.; Cerutti, S.; Iannaccone, S.; Conti, A.; Olivieri, S.; Alessio, M.; Pattini, L. Application of multivariate data analysis for the classification of two dimensional gel images in neuroproteomics. *J. Proteomics Bioinform.*, **2011**, *4*, 016-021.

[36] Rye, M.B. Image segmentation and multivariate analysis in two-dimensional gel electrophoresis. Norwegian University of Science and Technology, **2007**.

[37] Abdi, H.; Williams, L.J. Principal component analysis. *WIREs Comp. Stat.*, **2010**, *2*, 433-459.

[38] Safavi, H.; Correa, N.; Xiong, W.; Roy, A.; Adali, T.; Korostyshevskiy, V.R.; Whisnant, C.C.; Seillier-Moiseiwitsch, F. Independent component analysis of 2-D electrophoresis gels. *Electrophoresis*, **2008**, *29*, 4017-4026.

[39] Jessen, F.; Lametsch, R.; Bendixen, E.; Kjærsgård, I.V.; Jørgensen, B.M. Extractin information from two-dimensional electrophoresis gels by partial least squares regression. *Proteomics*, **2002**, *2*, 32-35.

[40] Beyer, K.; Goldstein, J.; Ramakrishnan, R.; Shaft, U. When is "nearest neighbor" meaningful? In: *Database Theory—ICDT'99*, Springer: 1999; pp 217-235.

[41] Marengo, E.; Robotti, E.; Gianotti, V.; Righetti, P.G.; Cecconi, D.; Domenici, E. A new integrated statistical approach to the diagnostic use of two-dimensional maps. *Electrophoresis*, **2003**, *24*, 225-236.

[42] Hu, Y.; Malone, J.P.; Fagan, A.M.; Townsend, R.R.; Holtzman, D.M. Comparative proteomic analysis of intra-and interindividual variation in human cerebrospinal fluid. *Mol. Cell. Proteomics*, **2005**, *4*, 2000-2009.

[43] Hilario, M.; Kalousis, A. Approaches to dimensionality reduction in proteomic biomarker studies. *Brief. Bioinform.*, **2008**, *9*, 102-118.

[44] Bylesjö, M.; Rantalainen, M.; Nicholson, J.K.; Holmes, E.; Trygg, J. K-OPLS package: Kernel-based orthogonal projections to latent structures for prediction and interpretation in feature space. *BMC Bioinformatics*, **2008**, *9*, 106.

[45] Kohonen, T. The self-organizing map. *Neurocomputing*, **1998**, *21*, 1-6.

[46] Lee, J.A.; Lendasse, A.; Verleysen, M. Nonlinear projection with curvilinear distances: Isomap versus curvilinear distance analysis. *Neurocomputing*, **2004**, *57*, 49-76.

[47] Weinberger, K.Q.; Saul, L.K. In *An introduction to nonlinear dimensionality reduction by maximum variance unfolding*, Proceedings of the National Conference on Artificial Intelligence, Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999: 2006; p 1683.

[48] Luk, J.M.; Lam, B.Y.; Lee, N.P.; Ho, D.W.; Sham, P.C.; Chen, L.; Peng, J.; Leng, X.; Day, P.J.; Fan, S.T. Artificial neural networks and decision tree model analysis of liver cancer proteomes. *Biochem. Biophys. Res. Commun.*, **2007**, *361*, 68-73.

[49] Diz, A.P.; Carvajal-Rodríguez, A.; Skibinski, D.O. Multiple hypothesis testing in proteomics: a strategy for experimental work. *Mol. Cell. Proteomics*, **2011**, *10*.

[50] Karp, N.A.; McCormick, P.S.; Russell, M.R.; Lilley, K.S. Experimental and statistical considerations to avoid false conclusions in proteomics studies using differential in-gel electrophoresis. *Mol. Cell. Proteomics*, **2007**, *6*, 1354-1364.

[51] Cohen, J. The earth is round (p<. 05). *Am. Psychol.*, **1994**, *49*, 997-1003.

[52] Nickerson, R.S. Null hypothesis significance testing: a review of an old and continuing controversy. *Psychol. Methods*, **2000**, *5*, 241.

[53] Wagenmakers, E.J. A practical solution to the pervasive problems of p values. *Psychon. Bull. Rev.*, **2007**, *14*, 779-804.

[54] Olafsdottir, G.; Nesvadba, P.; Di Natale, C.; Careche, M.; Oehlenschläger, J.; Tryggvadottir, S.V.; Schubring, R.; Kroeger, M.; Heia, K.; Esaiassen, M. Multisensor for fish quality determination. *Trends Food Sci. Technol.*, **2004**, *15*, 86-93.

[55] Grove, H.; Jørgensen, B.M.; Jessen, F.; Søndergaard, I.; Jacobsen, S.; Hollung, K.; Indahl, U.; Færgestad, E. M. Combination of statistical approaches for analysis of 2-DE data gives complementary results. *J. Proteome Res.*, **2008**, *7*, 5119-5124.

[56] Rossouw, D.; Robert-Granié, C.; Besse, P. A sparse PLS for variable selection when integrating omics data. *Genet. Mol. Biol.*, **2008**, *7*, 35.

[57] Lê Cao, K.A.; Boitard, S.; Besse, P. Sparse PLS Discriminant Analysis: biologically relevant feature selection and graphical displays for multiclass problems. *BMC Bioinformatics*, **2011**, *12*, 253.

[58] Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Series B Stat. Methodol.*, **2005**, *67*, 301-320.

[59] Marengo, E.; Robotti, E.; Bobba, M.; Righetti, P.G. Evaluation of the variables characterized by significant discriminating power in the application of SIMCA classification method to proteomic studies. *J. Proteome Res.*, **2008**, *7*, 2789-2796.

[60] Marengo, E.; Robotti, E.; Bobba, M.; Liparota, M.C.; Rustichelli, C.; Zamò, A.; Chilosi, M.; Righetti, P.G. Multivariate statistical tools applied to the characterization of the proteomic profiles of two human lymphoma cell lines by two-dimensional gel electrophoresis. *Electrophoresis*, **2006**, *27*, 484-494.

[61] Marengo, E.; Robotti, E.; Righetti, P.G.; Campostrini, N.; Pascali, J.; Ponzoni, M.; Hamdan, M.; Astner, H. Study of proteomic changes associated with healthy and tumoral murine samples in neuroblastoma by principal component analysis and classification methods. *Clin. Chim. Acta*, **2004**, *345*, 55-67.

[62] Xu, X.Q.; Leow, C.K.; Lu, X.; Zhang, X.; Liu, J.S.; Wong, W.H.; Asperger, A.; Deininger, S.; Eastwood Leung, H.C. Molecular classification of liver cirrhosis in a rat model by proteomics and bioinformatics. *Proteomics*, **2004**, *4*, 3235-3245.

[63] Supek, F.; Peharec, P.; Krsnik-Rasol, M.; Šmuc, T. Enhanced analytical power of SDS-PAGE using machine learning algorithms. *Proteomics*, **2008**, *8*, 28-31.

Received: October 13, 2013 Revised: October 13, 2013 Accepted: November 19, 2013