Proteomics & Bioinformatics

Mazzara et al. J Proteomics Bioinform 2011, 4:1
http://dx.doi.org/10.4172/jpb.1000162

Research Article

Open Access

Application of Multivariate Data Analysis for the Classification of Two Dimensional Gel Images in Neuroproteomics

Saveria Mazzara[1]*, Sergio Cerutti[1], Sandro Iannaccone[2], Antonio Conti[3], Stefano Olivieri[3], Massimo Alessio[3] and Linda Pattini[1]

[1]*Department of Bioengineering, IIT Unit, Politecnico di Milano, Piazza Leonardo da Vinci, 32, 20133, Milan, Italy*
[2]*Department of Neurology, San Raffaele Scientific Institute, Via Olgettina, 58, 20132, Milan, Italy*
[3]*Proteome Biochemistry, San Raffaele Scientific Institute, Via Olgettina 58, 20132, Milan, Italy*

Abstract

Two-dimensional gel electrophoresis (2DE) still plays a key role in proteomics for exploring the protein content of complex biological mixtures. However, the development of fully automatic strategies in extracting interpretable information from gel images is still a challenging task. In this work, we present a computational strategy aiming at an automatic classification of the discriminant patterns emerging from separation images intended as fingerprints of the correspondent biological conditions. The method was applied to gel images acquired in a study on motor neuron diseases: 33 2DE maps generated from samples of cerebrospinal fluid were processed (26 pathologic and 7 control subjects). Quantitative image descriptors were extracted and fitted to a partial least squares-discriminant analysis (PLS-DA) assessing the chance to classify the samples. Moreover, the model was able to identify gel areas that most differ through the clinical categories. Combining multivariate statistical techniques with 2DE may represent a valid tool to extract informative protein patterns. This kind of approach can contribute to the development of a system of screening to discriminate different clinical conditions on the basis of the overall patterns emerging from the maps, representing a useful complementary analysis in the routine of a proteomic laboratory.

Keywords: Two-dimensional gel electrophoresis (2DE); Partial least squares discriminant analysis (PLS-DA); Motor neuron diseases

Abbreviations: 2DE: Two-Dimensional gel Electrophoresis; PLS-DA: Partial Least Squares Discriminant Analysis; CSF: Cerebrospinal Fluid; ALS: Amyotrophic Lateral Sclerosis; SMA: Spinal Muscular Atrophy

Introduction

Two-dimensional gel electrophoresis (2DE) is still a wide spread technique for the separation of proteins in biological samples, allowing the analysis of a large number of proteins through only one experiment (Rabilloud, 2002). 2DE provides a proteome mapping of the sample via orthogonal mass/charge analysis. The method is based on the combination of two single-dimension electrophoretic runs: the first run, via a pH gradient, separates the proteins according to their isoelectric point (pI), whereas the second run separates them according to their molecular mass. The result is a two-dimensional map where the proteins appear as spots spread all over the gel surface. Once the maps, obtained from protein migration, have been acquired as gray level images, the biological information embedded in the 2D maps is processed and quantified to perform a differential analysis between the single protein spots of different samples. Unfortunately, the comparison and the classification of gel images is a hard and time consuming process due to the complexity and low reproducibility of the maps. The computational aspects of image processing play a central role in the analysis of 2DE gels (Dowsey et al., 2003). This is a very labour intensive step and involves a considerable expertise to properly extract information (Fey and Larsen, 2001; Marengo et al., 2005). The employment of 2D gel electrophoresis in combination with multivariate data analysis (Grove et al., 2008), such as principal component analysis (PCA) and partial least squares discriminant analysis (PLS-DA), may represent a complementary approach to the classical differential analysis, based on univariate statistical analysis, providing the bases for an automatic classification protocol. PLS-DA is highly suited for the analysis of high dimensional data set characterized by few samples and many variables; moreover, it takes into account the noise in the system and multicollinearity. As a result of these properties, in recent years projection methods are being successfully applied to biological data such as DNA microarrays (Pérez-Enciso and Tenenhaus, 2003; Nguyen and Rocke, 2002a; Nguyen and Rocke, 2002b) and proteomic data (Verhoeckx et al., 2005; Jessen et al., 2002; Lee et al., 2003; Whelehan et al., 2006; Klenø et al., 2004; Gotffries et al., 2004). In particular, Gottfries et al. (1995) used PLS-DA to diagnose different types of dementia while Karp et al. (2005) utilized PLS-DA combined with an iterative threshold process to measure changes in protein expression but many solutions are still to be investigated. In this work, an innovative computational approach, based on this kind of technique, was applied for implementing an automatic identification of discriminant patterns emerging from separation images considered as fingerprints of the correspondent clinical conditions. To illustrate the strategy, we have considered a set of cerebrospinal fluid (CSF) samples from subjects who were diagnosed with amyotrophic lateral sclerosis (ALS) or spinal muscular atrophy (SMA) in comparison to samples from healthy subjects. Image descriptors were derived, on the basis of the extracted quantitative parameters, to be used in the successive exploratory analysis; it is worth noting that the use of descriptors coming from a spot quantification expressed in the space of the experimental coordinates may represent a significant improvement in the descriptors power. The extracted features were subsequently analyzed by PCA (Jain et al., 2000; Wold, 1987) and PLS-DA (Wold et al., 2001; Eriksson et al., 2006); the former was used as explorative tool to overview groupings and trends in the data while the latter was used to the data modelization and the discriminant analysis.

***Corresponding author:** Saveria Mazzara, Department of Bioengineering, IIT Unit, Politecnico di Milano, Piazza Leonardo da Vinci, 32, 20133, Milan, Italy, Tel: + 39 02 2399 3302; Fax: + 39 02 2399 3360; E-mail: saveria.mazzara@mail.polimi.it

Received November 10, 2010; **Accepted** January 03, 2011; **Published** January 05, 2011

Citation: Mazzara S, Cerutti S, Iannaccone S, Conti A, Olivieri S, et al. (2011) Application of Multivariate Data Analysis for the Classification of Two Dimensional Gel Images in Neuroproteomics. J Proteomics Bioinform 4: 016-021. doi:10.4172/jpb.1000162

Materials and Methods

CSF samples

A total of twenty-six CSF samples taken from patients with motor neuron disease (MND) were analyzed in this study. According to the clinical diagnosis, these patients were classified into two groups: seven were diagnosed as SMA and nineteen as ALS. In spite of the serious difficulty in collecting CSF from control subjects, the experimental protocol also included seven healthy individuals. For each sample, a 2D electrophoresis map was generated as described in Conti et al. (2008) and corresponding images were processed (see Figure 1). All samples used in the study were obtained having secured approval from the San Raffaele Hospital's Institutional review board (Milan, Italy) and informed consent from patients according to the Helsinki Declaration and local legislation.

Image analysis

The gel images were visualized by scanning the gels at a resolution of 100 µm using the ProXPRESS 2D Proteomic Imaging System (PerkinElmer) and saved as 16 bit gray scale images in ".tif" format. The protein migration area covered pI values ranging from 3.2 to 10.4 and molecular weight (MW) values ranging from 5kDa to 250 kDa. Scanned images were, then, imported in Progenesis PG240 v2006 software (Nonlinear Dynamics, Newcastle, UK) (Rosengren et al., 2003) for the successive quantitative image processing. After background subtraction, detection and quantification of the protein spots were automatically performed with default settings. Spot intensities were normalized as percentage of the total spots optical density to accurately compare the quantification results despite non-expression related variation in spot intensity between gel images. Subsequently, gels were calibrated, still in the Progenesis environment, to obtain the position of each identified spot in terms of the biochemical coordinates: apparent relative molecular mass (Mr) was estimated by comparison with MW reference markers (Precision, Bio-Rad, Hercules, CA), and pI values were assigned to detected spots by calibration as described in the GE-Healthcare guidelines. The position of the identified spots was expressed in terms of pI and MW by using some reference markers and interpolating (by means of a cubic spline) these values to obtain the calibration curve that empirically defines the relation between geometrical and physicochemical coordinates for every gel image. Including the calibration step has been an essential requisite because a reasonable correspondence between different gels was restored. This ad hoc calibration made possible the virtual partition in subquadrants of the migration area, expressed in (pI, MW), at a chosen resolution of pI and MW. The subdivision was linear in the experimental coordinates but did not correspond to a regular grid on the gel image. The collection of spots was determined for each subquadrant, so we were able to extract an overall feature, as the summation of the correspondent spot optical densities, to be used in the successive exploratory data analysis (Pattini et al., 2008). In this way, the two-dimensional electrophoretic maps were described as vectors of sorted features (protein abundances) and could be processed through the application of the considered multivariate projection method.

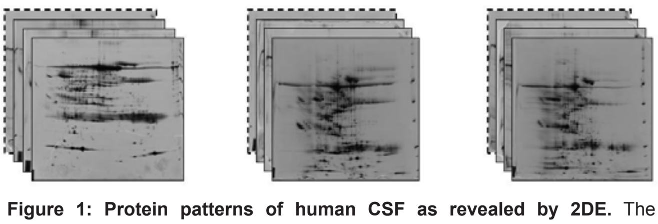

Figure 1: Protein patterns of human CSF as revealed by 2DE. The experimental 2D maps were generated from samples of CSF obtained from three groups: control subjects (n=7), patients with spinal muscular atrophy (n=7) and patients affected by amyotrophic lateral sclerosis (n=19).

LV	R^2X(%)	R^2X(cum)(%)	R^2Y(%)	R^2Y(cum)(%)	Q^2Y(%)	Q^2Y(cum)(%)
1	19.32	19.32	93.58	93.58	71.78	71.78
2	11.68	31.01	4.58	98.43	-8.26	69.45
3	14.86	45.86	1.19	99.62	23.94	76.77

R^2X, R^2Y: fraction of the variance of descriptor matrix (X) and class response (Y) explained by each latent variable (LV) in %, Q^2Y: fraction of the variance predicted (cross-validated) in %; R^2X(cum), R^2Y(cum): cumulative explained variation and Q^2Y(cum) predicted variation in %.

Table 1: Summary statistics of PLS-DA model for the SMA vs CN comparison. A three component model utilizes 45.86% of X for modelling 99.62% and predicting 76.77% of the response variation.

Multivariate statistical analysis

For the multivariate statistical analysis, the quantification parameters, generated by the commercial software, were organized in the form of a matrix, X (descriptor matrix), where the rows represent gel samples and each column (variable) represents one of the subquadrants in which each gel image was partitioned and the integral of the intensities of the protein spots in a given subquadrant the X-value. Initially, the data matrix X was processed by PCA with the aim of exploring the dataset and identifying outliers to be excluded from the modeling. By applying PLS regression, it is possible to connect the information in two blocks of variables, X (descriptor matrix) and Y (response matrix), to each other via a linear multivariate model. In this case, the response matrix, Y, is a matrix of dummy variables describing the class membership of each observation in the descriptor matrix X. PLS calculates latent variables (LV) as linear combination of X in such way that they well approximate X and Y and maximize the covariance between X and Y. The application of PLS (Boulesteix and Strimmer, 2006) as a classification method is indicated as partial least squares discriminant analysis (PLS-DA) (Barker and Rayens, 2003; Wold et al., 2001). Furthermore, the model allows to identify which descriptors explain most of the differences in the two groups by means of the variable influence in projection (VIP). The VIP is a weighted sum of squares of the PLS loading weights taking into account the amount of explained Y-variation in each dimension. The rule "greater than one" is used for detecting the descriptors with the greatest importance in the projection (Eriksson et al., 2006).

Model validation

A reliable way for predictive validation of a model is given by external validation, which consists of precisely predicting the Y-values of observations with new X-values. Unfortunately, an independent and representative validation set is a critical point in clinical/proteomics studies because of difficult availability of samples, especially for control subjects. In absence of a validation set, cross-validation (CV) provides a fast first clue of the predictive power of a model. CV is performed by dividing the data set in a number of groups and then developing a number of parallel models from reduced data with one of the groups deleted. After developing the reduced model, the deleted data are used as a test set, and the differences between actual and predicted responses are calculated from all the parallel models to form PRESS (predictive residual sum of squares). This is a measure of the predictive ability of the model and generally it is re-expressed as Q^2 (the cross-validated R^2) which is 1-PRESS/SS where SS is the sum of squares of the response, corrected for the mean (Eriksson et al., 2006). It is relevant to point out that cross-validation may provide over-optimistic results, so in order to test the statistical significance of the Q^2Y value a permutation test was performed. Thus, only the Y-block (class membership) was randomly reordered while the X-block (protein spot volumes) was left intact. For every permutation of the Y-block, a PLS-DA model was fitted to the permuted version of the Y-matrix and the new estimates of R^2Y and Q^2Y values were computed. The distribution of the R^2Y and Q^2Y parameters obtained by fit to random data is useful for estimating their statistical significance; if the "real" values are found outside such distribution this is a sign of high validity of the model (Eriksson et al., 2003).

Data analysis was performed using Matlab v. 7.0 (The Mathworks, Inc., Natick, MA) and the R plsm package (http://cran.r-project.org/web/packages/plspm/index.html).

Results and Discussion

We have evaluated the 2DE maps of CSF samples using the outlined strategy in order to discriminate between different groups of MND subjects and healthy (not affected by neurological disorders) individuals. According to the clinical diagnosis, subjects with MND were further classified as SMA and ALS; so, considering the control group (CN), three categories were considered and the three possible pairwise comparisons were accomplished. Firstly, SMA and CN subjects were considered; in the first stage of the data analysis, an unsupervised approach by means of PCA was applied. The investigation by PCA was done in order to get an overview of the data and also to identify possible outliers and trends in the data. The features extracted from the 2DE maps were organized in the X matrix, where the samples are represented as vectors of descriptor variables (protein expression levels in the different quadrants); using all 14 observations of the set, a PCA overview of the data was obtained. In order to investigate if the 2D electrophoregrams class discriminating information could be modeled, the three CN samples that were not segregating consistently with their clinical conditions in the PCA space were omitted from the data set. The PLS-DA modeling has, therefore, been based on the 11 selected samples. We created a dummy matrix of two Y-variables expressing class membership (diagnosis) of the CSF samples. Data were standardized

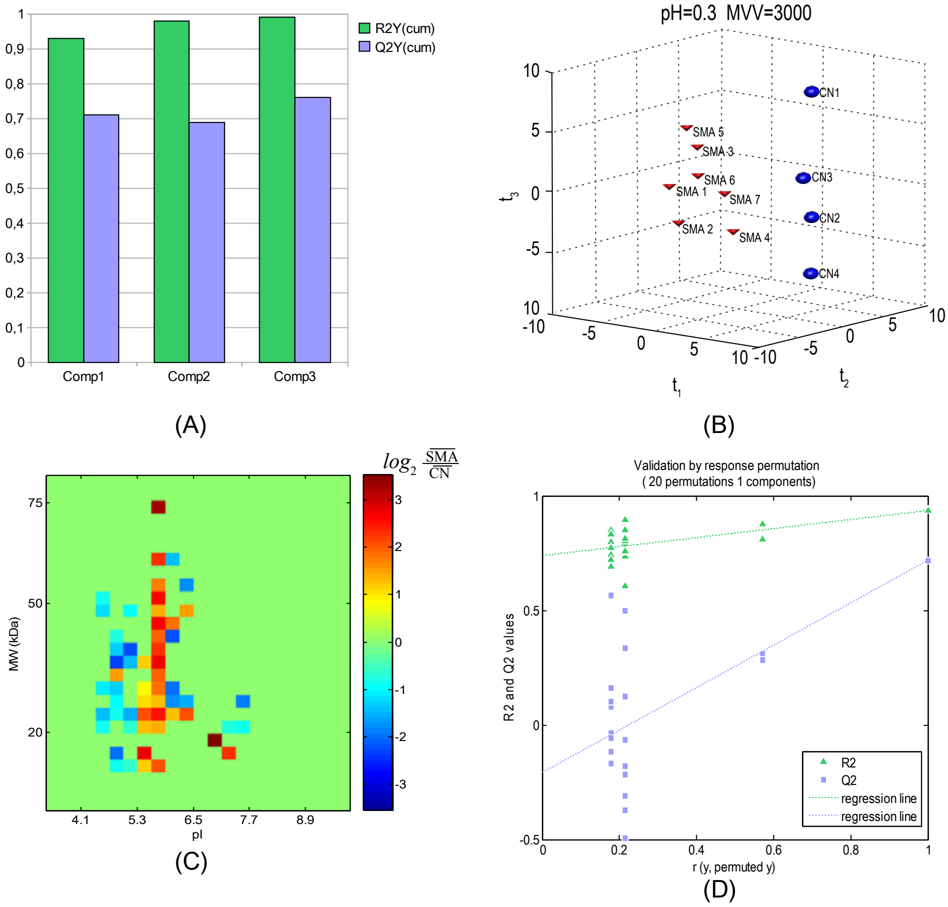

Figure 2: Overview of the PLS-DA analysis for the SMA vs CN comparison. (A) Plot of R^2Y (explained variation) and Q^2Y (predicted variation); it shows how the considered parameters change as a function of increasing model complexity. Note that the PLS-DA modeling resulted in one significant component; however, three components were generated to enable the construction of the 3-dimensional score plot. (B) PLS-DA score plot reveals grouping according to class membership: all patients affected by Spinal Muscular Atrophy (SMA, red cones) are positioned in the left part of the space, whereas all control subjects (CN, blue spheres) are found in the right part of the space. (The axes of the plot indicate PLS-DA components 1-3). (C) Identification of the subquadrants, in terms of pI and MW, with the highest VIP-values. Protein expression changes in the most influential subquadrants, involved in the discrimination of the disease from the healthy group, are plotted as a heat map. In this heat map, a red color reflects expression greater in SMA than CN patients, a blue color less in SMA than CN and a light green color reflects a similar expression in the two groups. (D) Validation plot by permutation test. The X-axis denotes the correlation coefficient between original and permuted data response, whereas the Y-axis shows the R^2Y (triangles) and Q^2Y (squares) values of all models. The two last points in the plot correspond to the values of R^2Y and Q^2Y for the original model. Two regression lines have been fitted, one among the R^2Y points and another one among the Q^2Y points. The two intercepts can be considered as measures of degrees of overfit and overprediciton.

to have mean 0 and standard deviation 1. This type of preprocessing is equivalent to giving all of the variables equal importance, hence this procedure is particularly useful in proteomic data sets where variables are characterized by large scale effects. The result of PLS-DA modeling is summarized in Table 1 and Figure 2A. According to cross-validation, even only one component was sufficient to model the relationship between the inputs (X variables) and the outputs (Y-variables). Nevertheless, a three component model was generated (R^2Y=99.62% and Q^2Y=76.77%) to allow score plotting (Eriksson et al., 2006). In PLS-DA modeling, it is interesting to analyze the t score plot for investigating the class discriminating ability of a developed model: the 3-dimendional X-score plot is shown in Figure 2B. There is clear discrimination between the two groups; it is mainly the first component that is responsible for separating the two groups of subjects from each other. The model also gives the possibility to obtain a quantitative measure of the discriminating power of each gel descriptors by means of VIP. It provides a synthetic information for model interpretation because VIP is a weighted sum of squares of the loading weights taking into account the amount of explained Y-variance in each dimension; for a given model there will always be one VIP-vector, summarizing all components and Y-variables. X-variables characterized by VIP values larger than 1 have major importance for modeling the responses whereas variables with VIP values smaller than 0.8 have a minor influence on the model. Thus, using the VIP score, a number of 57 protein descriptors were found to have a major discriminatory power between different clinical conditions. At this point, it may be interesting to return to the gel images and visualize the regions of the gel (in terms of pI and MW) that carried most class discriminating information. To overview the protein expression level and to give a rapid comparison between the two clinical groups a heat map was used (Figure 2C). Columns represent pI values, rows represent MW and each square indicate the log ratio of averaged abundances between the SMA and the CN group of that particular subquadrant. The color code of the heat map ranges from red to blue indicating increases or decreases in protein abundance for the SMA group, light green indicates no change. This kind of information may support the identification of proteins or groups of

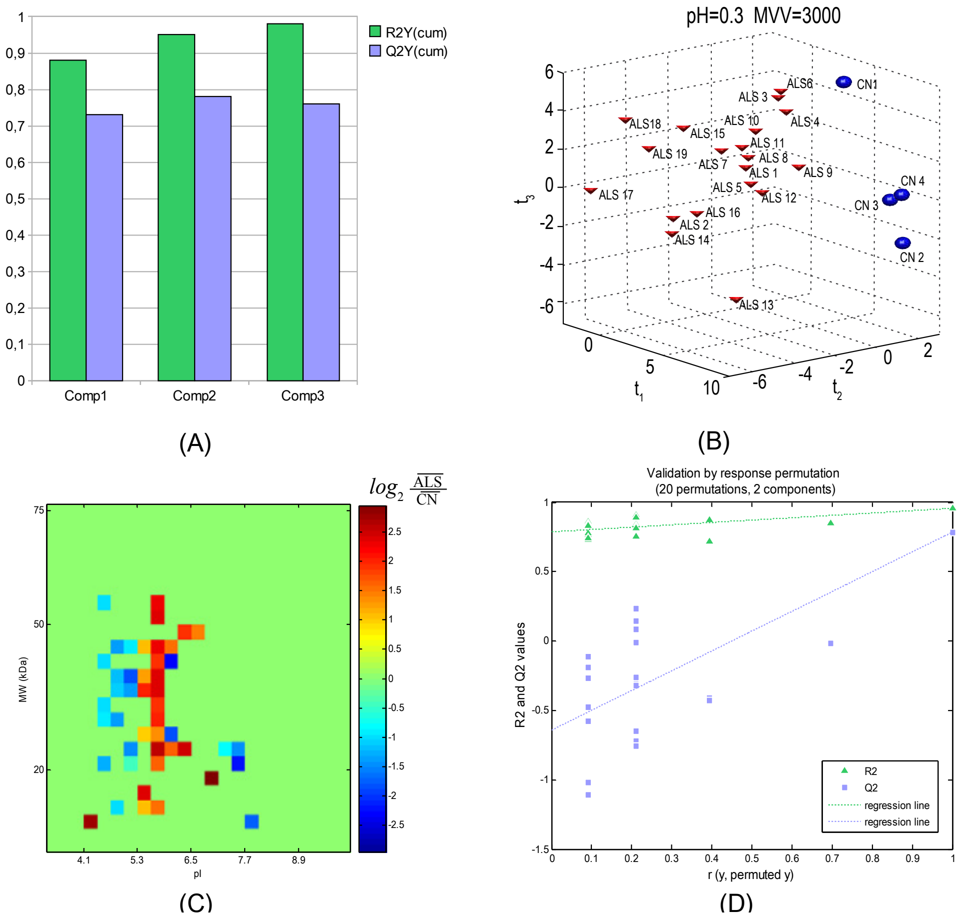

Figure 3: Overview of the PLS-DA analysis for the ALS vs CN comparison. (A) Plot of R^2Y (explained variation) and Q^2Y (predicted variation); it shows how the considered parameters change as a function of increasing model complexity. According to the cross-validation, two components resulted significant in order to explain the relationship between the descriptor matrix and the class response; nevertheless, three components were considered to allow score plotting. (B) PLS-DA score plot reveals no overlap between the two clouds of items correspondent to the two categories examined, Amyotrophic Lateral Sclerosis (ALS, red cones) vs control samples (CN, blue spheres). (The axes of the plot indicate PLS-DA components 1-3). (C) Identification of the subquadrants, in terms of pI and MW, with the highest VIP-values. Protein expression changes in the most influential subquadrants, involved in the discrimination of the disease from the healthy group, are plotted as a heat map. In this heat map, a red color reflects expression greater in ALS than CN patients, a blue color less in ALS than CN and a light green color reflects a similar expression in the two groups. (D) Validation plot by permutation test. The X-axis denotes the correlation coefficient between original and permuted data response, whereas the Y-axis shows the R^2Y (triangles) and Q^2Y (squares) values of all models. The two last points in the plot correspond to the values of R^2Y and Q^2Y for the original model. Two regression lines have been fitted, one among the R^2Y points and another one among the Q^2Y points. The two intercepts can be considered as measures of degrees of overfit and overprediciton.

LV	R^2X(%)	R^2X(cum)(%)	R^2Y(%)	R^2Y(cum)(%)	Q^2Y(%)	Q^2Y(cum)(%)
1	14.24	14.24	88.95	88.95	73.56	73.56
2	11.49	25.72	6.73	95.68	18.37	78.42
3	8.40	34.12	2.38	98.06	-10.87	76.07

R^2X, R^2Y: fraction of the variance of descriptor matrix (X) and class response (Y) explained by each latent variable (LV) in %, Q^2Y: fraction of the variance predicted (cross-validated) in %; R^2X(cum), R^2Y(cum): cumulative explained variation and Q^2Y(cum) predicted variation in %.

Table 2: Summary statistics of PLS-DA model for the ALS vs CN comparison. A three component model utilizes 34.12% of X for modelling 98.06% and predicting 76.07% of the response variation.

proteins differentially expressed in different conditions. Furthermore, a more rigorous way to validate the model was considered using response permutation testing. Figure 2D reports the results of the permutation test after 20 permutations. The y axis represents the R^2Y/Q^2Y values for every model fitted with the permuted data, including the "real" one. The x axis reports the correlation coefficient between permuted and original response variables. The two points in the right side are the R^2Y and Q^2Y for the original model because 1.0 is the correlation coefficient obtained when correlating a variable with itself. Recently, a quantitative way of synthesizing this plot has been put forward by Eriksson et al. (2003); that is, a regression analysis is carried out, one regression line is fitted among the R^2Y values and another line among the Q^2Y values. The two intercepts can be seen as sign of the degrees of overfit and overprediciton of the model; valid models are characterized by R^2Y intercept below 0.3-0.4 and Q^2Y intercept below 0.05. Hence, a close examination of the Figure 2D reveals the over parameterization of the model indeed the R^2Y value (goodness of fit) alone is not very dependable: when permuting response variables at random, there is a high probability of yielding model with similar or higher R^2Y; that is, it is mathematically possible to reproduce any possible combination of the Y variables by means of X matrix. On the contrary, the distribution of the Q^2Y values offers a favorable picture; the response permutation plot shows the validity of the PLS-DA model because the distribution of Q^2Y values of reordered response data is always lower than the real one.

In the second analysis, the data set including ALS and CN samples was investigated. In order to screen for outliers and to survey possible groupings, useful for directing further modeling efforts, a PCA model of the full data set was fitted; the analysis showed that the data were not strongly clustered owing to mispositioning of some samples (still the same three CN samples). Thus, also in this comparison, these samples were removed from the model due to their ambiguous behavior. A PLS-DA model was then generated from the reduced data set of 23 gels described through 108 variables (X matrix) using the knowledge related to class membership as the response variable (Y matrix). The number of significant components was determined using cross-validation; this yielded two components with an R^2Y of 0.95 and a cross-validated R^2 (Q^2) of 0.78 (Figure 3A). However, a three component model was generated to enable the construction of the 3-dimensional score plot, see Figure 3A and Table 2. From the t score plot, Figure 3B, it is clear that the two groups cluster according to their clinical conditions. To examine which protein descriptors contributed to this separation, the subquadrant with the highest VIP value are visualized as a heat map, see Figure 3C. The heat map representation gives information about the log ratio of the integral intensities of the protein spots in a given subquadrant between the two groups. The subquadrants are color coded according to the colorized scale with red, over-expression in ALS group, blue, under-expression in ALS group, and light green, no change between the two clinical groups. Next we applied the response permutation testing to provide an estimate of the significance of a Q^2Y value, we have permuted the response randomly 20 times and computed the new model with the original X-data matrix and reordered Y-data. For each derived model, both R^2Y and Q^2Y values were calculated and then compared with the estimates of the R^2Y and Q^2Y of the real model. As shown by Figure 3D, the Q^2Y distribution is sign of high predictive validity of the original model indeed it is impossible to obtain a model with the same predictive value by chance. To investigate our approach's ability to discriminate between subjects that have manifested degeneration of motor neurons, samples of CSF from patients diagnosed with ALS and patients diagnosed with SMA were processed through PLS-DA. Differently from the previous two comparisons, PLS-DA modeling of the complete set was not able to come out class separating information between the two diseased groups. The modeling, evidently, is not necessarily successful if the classes are not distinguishable according to the adopted descriptors. On the other hand, ALS and SMA may be considered as two forms of the same pathology, indeed ALS is characterized by dysfunction and loss of both upper and lower motor neurons, whereas SMA results from the loss of lower motor neurons (Bowser and Shneider, 2001). Indeed, by examining the regions of the 2D gel with the highest VIP-values (Figure 2C and Figure 3C) we can see a large number of common discriminant subquadrants (about 64%) between the two models.

Conclusions

In this work, we have proposed a computational strategy, based on multivariate data analysis techniques, for an automatic processing of two-dimensional gel electrophoresis maps, able to identify potential discriminant patterns between different clinical conditions. Using PLS-DA we were able to discriminate the two groups representing diseased and healthy subjects. Moreover, the analysis tell us which regions of the 2D-gels are responsible for class separation by means of VIP; these regions that include protein spots might be further investigate by means of mass spectrometry for proteins identification. It should also be pointed out that the samples of the data set are not technical replicates, i.e. gels obtained from fractions of the same biological sample. In this study, "biological" replicates were investigated, i.e. each gel image was representative of a different human subject, so the gels are characterized by low homogeneity, conferring to the protocol a very high level of variability and complexity. Another critical point is the difficulty of the availability of samples, in particular for control subjects. Notwithstanding these critical aspects, the strategy outlined, based on techniques of multivariate data analysis applied on the quantification parameters, obtained through the commercial software, gave promising and reliable results. This kind of study can contribute to the development of a system of screening to discriminate different clinical conditions on the basis of the overall patterns emerging from the maps representing a useful complementary analysis in the routine of a proteomic laboratory. Finally, it is worth remembering that no expert eye is able from a visual inspection of 2D electrophoretic gels to discriminate physiological from pathological conditions; thus, the information extraction in the processing of 2DE images is an important topic in computational biology and the proposed strategy may provide an interesting and fruitful point of view capturing the essential information from gel images.

Acknowledgements

This work was supported by a grant from Fondazione Cariplo 2006. MA is supported by MoH (RF07-ALS).

References

1. Barker M, Rayens W (2003) Partial least squares for discrimination. J Chemom 17: 166-173.

2. Boulesteix AL, Strimmer K (2006) Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. Briefings in Bioinformatics 8: 32-44.

3. Bowser LP, Shneider NA (2001) Amyotrophic Lateral Sclerosis. N Engl J Med 344: 1688-1700.

4. Conti A, Iannaccone S, Sferrazza B, De Monte L, Cappa S, et al. (2008) Differential expression of ceruloplasmin isoforms in the cerebrospinal fluid of amyotrophic lateral sclerosis patients. Proteomics Clinical Applications 2: 1628-1637.

5. Dowsey AW, Dunn MJ, Yang GZ (2003) The role of bioinformatics in two-dimensional gel electrophoresis. Proteomics 3: 1567-1596.

6. Eriksson L, Jaworska J, Worth AP, Cronin MTD, McDowell RM, et al. (2003) Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression- based QSARs. Environmental Health Perspectives 111: 1361-1375.

7. Eriksson L, Johansson E, Kettaneh-Wold N, Trygg J, Wikström C, et al. (2006) Multi- and megavariate data analysis. Basic Principles and Applications. Umetrics AB.

8. Fey SJ, Larsen PM (2001) 2D or not 2D. Two dimensional gel electrophoresis. Curr Opin Chem Biol 5: 26-33.

9. Gottfries J, Blennow K, Wallin A, Gottfries CG (1995) Diagnosis of dementias using partial least squares discriminant analysis. Dementia. 6: 83-88.

10. Gottfries J, Sjögren M, Holmberg B, Rosengren L, Davidsson P, et al. (2004) Proteomics for drug target discovery. Chem Intell Lab System 73: 47-53.

11. Grove H, Jorgensen BM, Jessen F, Sondergaard I, Jacobsen S, et al. (2008) Combination of statistical approaches for analysis of 2-DE data gives complementary results. J Proteome Res 7: 5119-5124.

12. Jain AK, Duin RPW, Mao J (2000) Statistical pattern recognition: a review. IEEE Trans Pattern Analysis Machine Intelligence. 22: 4-37.

13. Jessen F, Lametsch R, Bendixen E, Kjærsgard IVH, Jørgensen BM (2002) Extracting information from two-dimensional electrophoresis gels by partial least squares regression. Proteomics 2: 32-35.

14. Karp NA, Griffin JL, Lilley KS (2005) Application of partial least squares discriminant analysis to two-dimensional difference gel studies in expression proteomics. Proteomics 5: 81-90.

15. Klenø TG, Leonardsen LR, Kjeldal HØ, Laursen SM, Jensen ON, et al. (2004) Mechanisms of hydrazine toxicity in rat liver investigated by proteomics and multivariate data analysis. Proteomics 4: 868-880.

16. Lee KR, Lin X, Park DC, Eslava S (2003) Megavariate data analysis of mass spectrometric proteomics data using latent variable projection method. Proteomics 3: 1680-1686.

17. Marengo E, Robotti E, Antonucci F, Cecconi D, Campostrini N, et al. (2005) Numerical approaches for quantitative analysis of two-dimensional maps: a review of commercial software and home-made systems. Proteomics 5: 654-666.

18. Nguyen DV, Rocke DM (2002a) Tumor classification by partial least squares using microarray gene expression data. Bioinformatics 18: 39-50.

19. Nguyen DV, Rocke DM (2002b) Multi-class cancer classification via partial least squares with gene expression profiles. Bioinformatics 18: 1216-1226.

20. Pattini L, Mazzara S, Conti A, Iannaccone S, Cerutti S, et al. (2008) An integrated strategy in two-dimensional electrophoresis analysis able to identify discriminants between different clinical conditions. Exp Biol Med 233: 483-491.

21. Pérez-Enciso M, Tenenhaus M (2003) Prediction of clinical outcome with microarray data: a partial least squares discriminant analysis (PLS-DA) approach. Human Genet 112: 581-592.

22. Rabilloud T (2002) Two-dimensional gel electrophoresis in proteomics: Old, old fashioned, but it still climbs the mountains. Proteomics 2: 3-10.

23. Rosengren AT, Salmi JM, Attokallio T, Westerholm J, Lahesmaa R, et al. (2003) Comparison of PDQuest and Progenesis software packages in the analysis of two-dimensional electrophoresis gels. Proteomics 3: 1936-1946.

24. Verhoeckx KC, Gaspari M, Bijlsma S, van der Greef J, Witkamp RF, et al. (2005) In search of secreted protein biomarkers for the anti-inflammatory effect of $ß_2$-adrenergic receptor agonists: application of DIGE technology in combination with multivariate and univariate data analysis tools. J Proteome Res 4: 2015-2023.

25. Whelehan OP, Earll ME, Johansson E, Toft M, Eriksson L (2006) Detection of ovarian cancer using chemometric analysis of proteomic profiles. Chem Intell Lab System 84: 82-87.

26. Wold S, Sjöström M, Eriksson L (2001) PLS-regression: a basic tool of chemomterics. Chem Intell Lab System 58: 109-130.

27. Wold S, Esbensen K, Geladi P (1987) Principal component analsysis. Chem Intell Lab System 2: 37-52.