An automated procedure for the statistical analysis of two-dimensional electrophoresis gels for biomarkers discovery

Mauro Fasano (✉ mauro.fasano@uninsubria.it)

Department of Theoretical and Applied Sciences, University of Insubria

Tiziana Alberio

Department of Theoretical and Applied Sciences, University of Insubria

---

Method Article

Keywords: Two-dimensional gel electrophoresis, biomarker discovery, statistics, linear discriminant analysis, power analysis, Pearson correlation.

Posted Date: January 2nd, 2013

DOI: <https://doi.org/10.1038/protex.2013.001>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

A protocol for the analysis of two-dimensional electrophoresis images has been written in R, an open-source environment for statistical computing, with the purpose of identifying and verifying potential disease biomarkers. In a stepwise procedure, spot volumes are normalized and missing values are replaced. A quality control procedure permits to efficiently discard gels that should not be included in a comparative statistics. Spots associated with confounding factors are identified and discarded from any further analysis. Comparisons between groups are performed with both parametric and non parametric tests, thus allowing researchers to choose the most appropriate condition. Hereafter, a classification model is built by linear discriminant analysis. The model is then verified so to obtain expected sensibility and specificity values.

Introduction

Proteomics is a fundamental tool for the discovery of disease biomarkers and to find new therapeutic targets ¹. Originally proposed as the protein complement of the genome, the proteome recently assumed a more functional significance ^{2 3 4}. In integration with genomics, transcriptomics, bioinformatics and biostatistics, proteomics became an indispensable instrument for translational aspects of modern molecular medicine ⁵. Proteomics is an intrinsically unbiased approach, aimed at generating a list of candidate proteins deserving further targeted studies ⁶. Such a global approach to biology and medicine allows to manage with the great complexity of the system being investigated, thus overcoming limits of conventional biochemistry or molecular biology tools ⁷. In order to be unbiased, proteome should be approached globally. Despite the protein separation technique being employed (either gel-based or gel-free) ^{8 9 10}, it is possible to observe protein levels spanning over three order of magnitude even if they cover almost six orders of magnitude in a cell and up to ten in plasma ¹¹. Although there is a strong technological push towards gel-free techniques, 2-DE continues to be preferred for its capacity to resolve thousands of spots simultaneously and represents the only technique that can be routinely applied for parallel quantitative expression profiling of large sets of complex protein mixtures such as whole cell lysates ^{12 13}. 2-DE gels are at the same time an image of protein distribution (in terms of molecular weight and pI) and levels, and a container of separated proteins available for further characterization ¹³. Nevertheless, there are intrinsic weaknesses in 2-DE. Due to the reduced dynamic range of the detection technique, less represented proteins are lost, thus limiting the global characteristic of the proteomic approach. Moreover, proteins of extreme hydrophobicity and those of extreme basic or acidic pI are not considered in the majority of published studies ¹⁴. In this protocol we propose a suite of standardized computational procedures to analyze densitometry data from 2-DE gel image analysis. Briefly, data are internally normalized and quality-checked to ensure that the right comparative statistics procedure is applied. Confounding factors are identified in order to exclude spots significantly associated to them from any further analysis. Given the log-normal distribution of magnitude values, a non-parametric test is performed to identify a set of spots that best contribute to a discriminant function, the latter being identified by linear discriminant analysis. Furthermore, selected spots are ranked in terms of their

contribution to the predictive model. Eventually, the efficiency of the set of spots is verified by power analysis, predictions are calculated by the leave-one-out cross validation and performance parameters are calculated in terms of receiver operating characteristics (ROC) analysis.

Equipment

2-DE gel images were acquired (12 bit grayscale) with the GelDoc-It Imaging System (UVP, Upland, CA, U.S.A.) and analyzed with ImageMaster 2D Platinum (GE Healthcare, Uppsala, Sweden). All procedures for data analysis and graphics were written in R, an open-source environment for statistical computing¹⁵.

Procedure

1. Import spot volumes and clinical/demographic data Spot volumes, exported from the ImageMaster software or other image analysis softwares in .csv (comma separated values) format, are read as a data matrix. Each row represents one of the L gels, whereas columns from 3 to N+2 are spot volumes for each of the N detected spots. Column 1 contains the gel labels, whereas column 2 contains the group factor (e.g., control vs. affected). A second matrix is read in .csv format, containing again one row per gel. Columns represent clinical and demographic data (age, gender, years from onset, disease staging, (any) drug daily dose, treated or not with a given drug, etc). Columns are read as independent vectors for further analysis. ****2. Normalize 2-DE spot volumes**** In each gel, a set of spots common to all gels (if L is small) or to 75% of gels (if L is large) is identified for normalization purposes. The sum of these spot volumes is considered as a loading/staining reference, thus providing L normalization factors. Each line in the spot volume matrix is then divided by the appropriate normalization factor to obtain a new matrix of normalized spot volumes. ****3. Replace missing values**** After normalization, missing spot values are replaced by the mean value of the spot volume of the group (i.e., control or affected) or, if the mean was lower than the 98th percentile, by the minimum value observed in the group¹². This ensures to safely prevent the artefactual assignment of missing spots (qualitative differences) as false negatives. To this purpose, the data matrix is split into separate submatrices for distinct groups, i.e., controls (co), early-onset patients (eo) and late-onset patients (lo). ****4. Check the quality of data sets**** Like many other biological observables, 2-DE spot volumes show a log-normal distribution, with less-intense spots more frequent than more-intense ones. To simply verify this, spot volumes are transformed logarithmically and the standard deviation is calculated within the group. "Figure 1":<http://www.nature.com/protocolexchange/system/uploads/2372/original/Figure1.jpg?1355321820> shows almost constant values of standard deviations as a function of log(volume), as expected for a log-normal distribution. This means that, in principle, a parametric test is suitable for comparison between groups after logarithmic transformation of spot volumes¹⁶. Then, the biological variability of the subjects in each group is evaluated. Pairs of 2-DE gels of specimens from different subjects in the same group are compared by Pearson linear correlation of corresponding spot volumes after logarithmic transformation. If the number of specimens (i.e., L) is large, it is preferable to take a single subject per

group as a reference, so that all gels in the group are compared to this one. For each comparison, the Pearson correlation coefficient is shown below the scatter graph. Dispersions of the studentized residuals against ranked magnitudes (QQ) plots are shown on the left of each correlation ("Figure 2":<http://www.nature.com/protocolexchange/system/uploads/2373/original/Figure2.jpg?1355322300>). Spots deviating from linearity indicate non-normal distribution of residuals. A linear behavior indicates linear proportionality between gels. Dashed lines indicate 95% CI. On this basis, a direct proportionality between gels from different subjects may be assumed ¹⁶. Otherwise, the gel shall be discarded. **5. Identify spots associated to confounding factors** Confounding factors may be either grouping classes (gender, assuming or not a given treatment, familiarity) or real numbers (age, daily drug dose, age at onset, disease duration, disease staging). In the first case, a nonparametric Wilcoxon test is performed on spot volumes grouped according to the classes, so to determine which spots are significantly different in a class with respect to the other one ("Figure 3":<http://www.nature.com/protocolexchange/system/uploads/2374/original/Figure3.jpg?1355322720>). Otherwise, a linear correlation analysis is performed to determine which spots are significantly correlated with the magnitude of the confounding factor (e.g., daily drug dose or age). Important to notice, the analysis of correlation with age should be performed on control subjects only ("Figure 4":<http://www.nature.com/protocolexchange/system/uploads/2374/original/Figure3.jpg?1355322720>). Similarly, in the drug dose correlation only patients should be included, treated or not with that drug. Eventually, spots that significantly correlate with one or more confounding factor(s) are removed from the dataset. **6. Calculate p values for all comparisons** This is the most critical step of the procedure. If data are log-normally distributed, a parametric test may be applied after logarithmic transformation of the dataset. A Welch test is preferred, since distinct spots may have different variance. Otherwise, a non-parametric test such as the Wilcoxon test may be a safer choice ¹⁶. Working with three groups, five contrast are suggested, in principle. Three are binary contrast (one group vs. another one), a fourth compares patients (both eo and lo) vs. control subjects and the last contrast compares a peculiar group (e.g., eo) vs. all other subjects. In this way, ten p values are computed for each spot (five Welch test and five Wilcoxon test). If the purpose is to identify a biomarker able to discriminate patients from controls, then the threshold for significance should be $0.05/N$, where N is the number of spots, and a non-parametric variant of the classic ANOVA test might be preferable ¹⁶. However, our procedure might be employed to identify all those spots that are different in the contrast, in order to build a classification function as a linear combination of all the selected spots or of a subset of them. **7. Build a linear discriminant function to classify the individuals and verify the predictive model** Predictive models for the classification of patients with respect to control subjects, or of patients with a particular disease subtype with respect to idiopathic disease patients (e.g., early-onset vs. late-onset) are built by linear discriminant analysis (LDA) ¹⁶. Volumes of spots selected as described above are linearly combined in a discriminating function, thus providing a likelihood score for the disease. The predictive model is verified by leave-one-out cross-validation. Each subject is iteratively removed from the dataset and classified on the basis of the spot volumes observed in the remaining individuals. This procedure permits to obtain reliable estimates of sensitivity and specificity. A disease likelihood score is attributed to each subject

and a ROC curve¹⁶ is calculated by plotting sensitivity and specificity as a function of the cutoff value. The contribution of each spot may be of different relevance, depending on the separation of mean volume values and their LDA coefficient. The product of these two parameters indicates the intrinsic contribution of every spot to the discriminating model, so to permit the selective removal of worst-contributing spots. Eventually, a report table is produced to summarize the descriptive statistics results, the fold of change in log2 units, the LDA coefficient and the weighted contribution of each spot to the model \(_i.e._\), the product of the difference of mean values times the LDA coefficient) \("Table 1":<http://www.nature.com/protocolexchange/system/uploads/2376/original/Table1.jpg?1355330640>). On this basis, a subset of most relevant spots may be selected instead of the list including all identified spots. **8. Make a power analysis** Spots to be included in the classification function have been selected on the basis of a non-parametric test. However, linear discriminant analysis takes into account the separation of two normal distributions. It may be useful, at this point, to check which one would be the appropriate number of subjects to be enrolled in a verification study, assuming a normal distribution of each spot volume across the individuals. **Note:** the complete source is available "here":<http://www.nature.com/protocolexchange/system/uploads/2394/original/protocol.R?1356083285> as a text file.

Timing

The whole procedure usually requires less than five minutes to be completed, depending on the size of the dataset and on the CPU clock of the computer being used.

Troubleshooting

The appropriate choice of the comparative statistics test is a critical issue. As we discussed above, proteomics data - including 2-DE spot volumes - are expected to show a log-normal distribution. In this case, the researcher faces two possibilities: either data are transformed on a logarithmic scale and are analyzed by parametric tests, possibly considering a different variance for the compared groups, or data are compared on a linear scale after ranking, i.e. with a non-parametric test. Our procedure offers both possibilities, with a parametric Welch test or a non-parametric Wilcoxon test being performed on all spots. Related to this issue is the choice of the p value used as a threshold for statistical significance. When dealing with multiple comparisons, the p value should be corrected for the number of observables being observed \(_Bonferroni\ correction\)^{16}. Nevertheless, it is up to the investigator to apply any correction of the p value. Actually, identified spots should not be considered _per se_ as being significant in discriminating a patient group from a control group. Rather, a larger number of spots should be included in the LDA procedure to further refine them in a separate step. LDA is a parametric approach, therefore it may not be effective in the classification of subjects based on multiple comparisons obtained with a non-parametric test. For this reason, an accurate refinement of the spots to be included in the model is advisable. We tackled this problem by progressively refining the list of candidate spots on the basis of their contribution to the model. When the reported procedure was applied¹⁷, we started from a 20-spot

model and progressively discarded the six spots showing the worst contribution ($\text{weight} < 0.3$) and the LDA was performed on the remaining 14 spots. In a further refinement, we discarded additional spots with $\text{weight} < 1$, thus obtaining a 9-spot model. A comparison of the performance of the three predictive models is shown in "Table

2":<http://www.nature.com/protocolexchange/system/uploads/2377/original/Table2.jpg?1355330700> ¹⁷.

Anticipated Results

A correct analysis of proteomic data is fundamental for the identification of candidate biomarkers. First, an adequate assessment of the gel quality grants a remarkable spare of time. Indeed, the inclusion of poorly resolved gels or bad quality gel images in the dataset leads to loss of significance of potentially interesting results. Any foreseeable correlation with confounding factors has to be taken into account, age and gender being just an example. Using this approach, we identified several spots that are correlated with dopaminergic therapies ¹⁸ and excluded them from the further analysis. Eventually, LDA allowed us to set up a discriminant function based on a linear combination of a set of independent observations, that dramatically improves sensitivity and specificity of the proposed diagnostic tool. We anticipate that this procedure might permit a rapid screening of 2-DE spots to be efficiently included in a panel of protein biomarkers for molecular diagnostics.

References

1. Alberio, T. & Fasano, M. Proteomics in Parkinson's disease: An unbiased approach towards peripheral biomarkers and new therapies. *J Biotechnol.* 156, 325-337 (2011).
2. Blackstock, W.P. & Weir, M.P. Proteomics: quantitative and physical mapping of cellular proteins. *Trends Biotechnol.* 17, 121-127 (1999).
3. Pandey, A. & Mann, M. Proteomics to study genes and genomes. *Nature* 405, 837-846 (2000).
4. Tyers, M. & Mann, M. From genomics to proteomics. *Nature* 422, 193-197 (2003).
5. Ostrowski, J. & Wyrwicz, L.S. Integrating genomics, proteomics and bioinformatics in translational studies of molecular medicine. *Expert Rev. Mol. Diagn.* 9, 623-630 (2009).
6. Caudle, W.M. et al. Using 'omics' to define pathogenesis and biomarkers of Parkinson's disease. *Expert Rev. Neurother.* 10, 925-942 (2010).
7. Villoslada, P. et al. Systems biology and its application to the understanding of neurological diseases. *Ann. Neurol.* 65, 124-139 (2009).
8. Jacobs, J.M. et al. Utilizing human blood plasma for proteomic biomarker discovery. *J. Proteome Res.* 4, 1073-1085 (2005).
9. Miller, I. et al. Protein stains for proteomic applications: which, when, why? *Proteomics* 6, 5385-5408 (2006).
10. Schulze, W.X. & Usadel, B. Quantitation in mass-spectrometry-based proteomics. *Annu. Rev. Plant Biol.* 61, 491-516 (2010).
11. Anderson, N.L. & Anderson, N.G. The human plasma proteome: history, character, and diagnostic prospects. *Mol. Cell. Proteomics* 1, 845-867 (2002).
12. Albrecht, D. et al. Missing values in gel-based proteomics. *Proteomics* 10, 1202-1211 (2010).
13. Görg, A. et al. Current two-dimensional electrophoresis technology for proteomics. *Proteomics* 4, 3665-3685 (2004).
14. Wilkins, M.R. Hares and tortoises: the high- versus low-throughput proteomic race. *Electrophoresis* 30 Suppl 1, S150-S155 (2009).
15. R Development Core Team. R: A language and environment for statistical computing. Vienna:

R Foundation for Statistical Computing \ (2009). 16. McDonald, J.H. Handbook of Biological Statistics. 2nd ed. Baltimore: Sparky House Publishing \ (2009). 17. Alberio, T. et al. Discovery and verification of panels of T-lymphocyte proteins as biomarkers of Parkinson's disease. Sci. Rep. 2, 953; DOI:10.1038/srep00953 \ (2012). 18. Alberio, T. et al. Dopaminergic therapies modulate the T-cell proteome of patients with Parkinson's disease. IUBMB Life 64, 846-852 \ (2012).

Acknowledgements

Authors gratefully acknowledge Prof. Riccardo Fesce for helpful discussion and Mr. Simone Olgiati and Mrs. Agnese Pippione for testing the protocol.

Figures

Figure 1

Log-normal distribution of data Values of standard deviations as a function of $\log(\text{volume})$ are almost constant, as expected for a log-normal distribution. Open circles: control group. Filled circles: patients group.

Figure 2

Gel quality control Left panel: scatter plot of a pair of gels belonging to the same group. The Pearson correlation coefficient is reported below, together with the slope of the straight line. Right panel: quantile-quantile plot for the residuals of the linear fit in panel A. A linear correlation between t quantiles and Studentized residuals indicates that gels are comparable.

Spot No. 391 p = 0.016

Figure 3

Identification of spots associated with the gender Box and whiskers plot for the volume of Spot 391 in females (green) and males (red).

Spot No. 513 , $p = 0.007$

Figure 4

Identification of spots correlated with the age Linear correlation of the volume of Spot 513 _vs._ the age of control subjects.

	20-spot model	14-spot model	9-spot model
Cutoff	0.67	-0.31	-2.90
Sensitivity (%)	87	100	100
Specificity (%)	81	94	88
ROC AUC	0.906	0.996	0.992
p value	< 10 ⁻⁸	< 10 ⁻⁸	< 10 ⁻⁸

Figure 5

Table 1 Summary of performance parameters for the three predictive models Cutoff values are expressed in terms of the likelihood PD score obtained as a linear combination of selected spot relative volumes; ROC = receiver operating characteristic; AUC = area under curve; p values refer to Wilcoxon test on PD scores. Sensitivity, specificity and ROC AUC were obtained by leave-one-out cross-validation.

Spot		mean.co	mean.pd	SD.co	SD.pd	log2.fold	p.val.Wilcox	LDA.Coeff	Delta.norm
1	86	4,48E-03	1,56E-03	2,88E-03	9,72E-04	-1,521	3,32E-04	-425,75	1,244
2	87	9,34E-03	4,16E-03	7,19E-03	3,91E-03	-1,166	1,26E-02	-236,86	1,226
3	335	9,41E-03	2,73E-03	9,70E-03	1,83E-03	-1,786	1,96E-03	178,98	1,196
4	362	9,56E-03	4,38E-03	6,24E-03	3,48E-03	-1,126	2,72E-03	382,08	1,980
5	365	3,95E-02	1,40E-02	2,77E-02	1,01E-02	-1,493	8,10E-04	39,32	1,001
6	368	1,44E-02	6,63E-03	9,32E-03	4,55E-03	-1,122	7,82E-03	177,71	1,387
7	369	4,31E-02	1,47E-02	2,29E-02	1,02E-02	-1,557	6,06E-05	-165,94	4,724
8	382	1,86E-02	7,94E-03	1,64E-02	3,27E-03	-1,224	3,47E-04	-100,27	1,064
9	657	5,92E-03	2,10E-03	7,31E-03	1,37E-03	-1,497	9,97E-03	-299,77	1,146

Figure 6

Table 2 Statistics report A report table is produced to summarize the descriptive statistics results, the fold of change in log2 units, the LDA coefficient and the weighted contribution of each spot to the model (i.e., the product of the difference of mean values times the LDA coefficient).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [supplement0.r](#)