Evaluation of Two-Dimensional Gel Electrophoresis Maps by Local Tangent Space Alignment: an Application to Neuroproteomics

Saveria Mazzara, Antonio Conti, Stefano Olivieri, Sandro Iannaccone, Massimo Alessio, Sergio Cerutti, *Fellow, IEEE,* and Linda Pattini

***Abstract*—Proteomic analysis may be useful to investigate disorders of the central nervous system, in order to explore the protein content of cells and of biological fluids in respect of the onset and evolution of diseases. Today, one of the most used proteomic approach includes the separation and visualization of proteins by means of two-dimensional gel electrophoresis (2DE). However the development of fully automatic strategies in extracting information from gel images is still a challenging task. In this paper we applied a computational strategy to the aim of obtaining a compact representation of the original data. This method was applied to an experimental protocol including two different clinical groups of amyotrophic lateral sclerosis (ALS) and peripheral neuropathy patients : 32 2DE maps generated from cerebrospinal fluid (24 pathologic and 8 control subjects) were processed. Quantitative features were extracted to describe each image and dealt with the dimension reduction technique of local tangent space alignment (LTSA). The discovered low-dimensional structure allows to see the gel of the dataset as separable, according to their clinical conditions, showing the informativeness of the adopted descriptors and providing the bases for classification of this kind of samples.**

I. Introduction

TWO-DIMENSIONAL GEL ELECTROPHORESIS is still the most wide spread technique for the separation of proteins in biological samples, allowing the analysis of a large number of proteins through only one experiment [1], [2]. 2DE provides a proteome mapping of the sample by means of the orthogonal combination of two 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. The maps obtained from proteins migration are acquired as grey level images, which can be processed and quantified to perform a differential analysis between the single protein spots of the different samples. Unfortunately, the comparison of different gel images is a difficult and time consuming process, because of the complexity and low reproducibility of the maps. The computational aspects of image processing play a central role in the analysis of 2DE gels [3]. This is a very labour intensive step and involves a considerable expertise to properly extract information. Usually, the differential analysis is achieved by means of dedicated software packages but none is yet fully automatic and all of them still require a large amount of user interaction to complete the analysis [4]. Beside these tools it can be useful to develop automatic strategies based on the assessment of the complete ensembles of spots shown in the maps [5], [6], [7]. In this work, we propose a complementary approach for the evaluation of the sample categories involved, avoiding the steps of registration and matching. This approach was applied to the gel images set obtained from cerebrospinal fluid (CSF) acquired in studies on neurodisorders. To this aim, image descriptors are derived, on the basis of the extracted quantitative parameters, to be used in the successive exploratory analysis. Each gel image, represented as a vector of quantitative features, can be investigated by dimension reduction methods, as LTSA. Dimensionality reduction is the transformation of high-dimensional data into a meaningful representation of reduced dimensionality [8]. Its goal is to discover the hidden structure from the raw data automatically managing the curse of the dimensionality. As a result, dimensionality reduction facilitates visualization, classification and compression of high-dimensional data.

II. Methods

We analyzed 32 2DE maps generated from CSF of four groups: ALS patients (n=8), neuropathic patients with pain (PN, n=8), neuropathic patients without pain (NPN, n=8) and control subjects (CN, n=8); patient features and 2DE gels generation were already published in [9] and [10]. Gel images were acquired using a Molecular Dynamics Personal SI Laser Densitometer (Molecular Dynamics, Sunnyvale, CA) and saved as grey level images in .tif format. Fig. 1 represents an example, for each class, of the experimental 2D maps considered. The 2DE protein patterns covered the ranges from 3.2 to 10.4 in *pI* and from 5 kDa to 200 kDa in relative molecular mass (Mr). The image analysis was performed using Progenesis PG240 v2006 software (Nonlinear Dynamics, Newcastle, UK) [11]. A spot

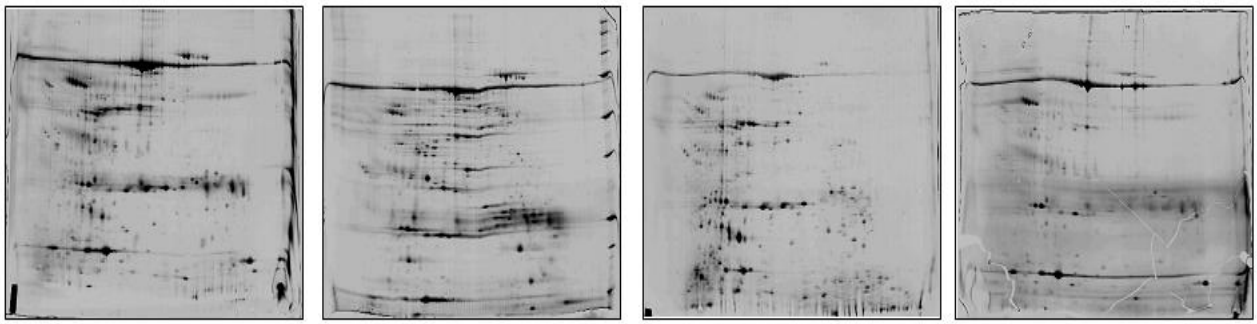
Manuscript submitted July 5, 2008. This work was supported by Fondazione Cariplo (NOBEL GuARD Project; 2006.0537/105411 project; and 2006.0538 project).

S. Mazzara, S. Cerutti and L. Pattini are with the Department of Bioengineering, IIT Unit, Politecnico di Milano, Milan, Italy (e-mail: saveria.mazzara@mail.polimi.it).

A. Conti, S. Olivieri and M. Alessio are with the Proteome Biochemistry Unit, San Raffaele Scientific Institute, Milan, Italy.

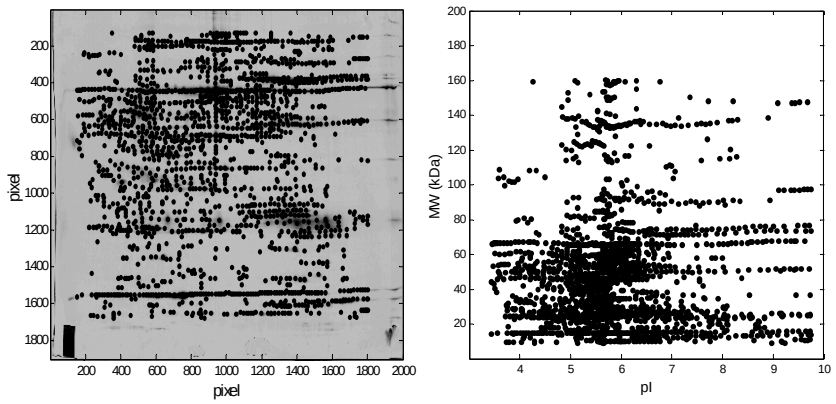
S. Iannaccone is with the Department of Neurology, San Raffaele Scientific Institute, Milan, Italy.

detection phase was included in the developed strategy to improve the signal to noise ratio and let emerge only the position or even exactly the same shape or same spatial distributions. In general, the positions in pixel, Fig. 2(a),



(a) (b) (c) (d)

Fig. 1. 2DE maps of CSF of the four groups investigated: control subjects (a), patients with ALS (b) and neuropathic subjects without pain (c) or with pain (d). Proteins were separated by *pI* in the first dimension and molecular mass in the second dimension.



(a) (b)

Fig. 2. Images of protein migration in the different samples do not cover equivalent pixel areas; to bypass this problem, after ad hoc calibration, the positions of the detected spots were converted from pixel (a) in biochemical coordinates, *pI* and MW, allowing the compilation of a linear virtual map of the protein migration (b) .

useful signal; the
idea was that the features that are then derived refer to areas of the image segmented as spots excluding from the quantitative description artifacts and background signal. Conversely, the extraction of descriptors directly from the images, meant as matrices of pixel intensities, without the step of spot detection, could include information from areas not correspondent to real spots. As a final result, the automated spot detection gives back for each analyzed image the collection of identified protein spots with their set of quantitative parameters as volume, maximal intensity, area. A normalization step was included to compensate non-expression related variation in spot intensity between gel images, caused by experimental variations.

What makes critical and time consuming the spot matching is that proteins may not necessarily be in the same physical from a gel to another one are not equivalent in respect to the separation and also proportions are not conserved through the collection of samples. To tackle this problem, gels were calibrated, as reported in Fig. 2(b), to obtain the position of each identified spot in terms of the biochemical coordinates: apparent Mr were estimated by comparison with molecular weight (MW) reference markers (Precision, Bio-Rad, Hercules, CA) and *pI* values assigned to detected spots by calibration as described in the GE-Healthcare guidelines. This step makes the samples comparable, without the accomplishment of a canonic image matching by means of registration techniques. The new space (*pI*,Mr) is, in principle, invariant to the alterations of protein migration, allowing the inclusion in the analysis of the gel images that otherwise had to be excluded because of the lack of the necessary homogeneity. Only at this point it is possible to

accomplish the ideal partition in subquadrants of the migration area, expressed in (*pI*,MW), at a resolution of 0.3 in *pI* and 3 kDa in MW. The subdivision is linear in the experimental coordinates but does not correspond to a regular grid on the gel image. The subquadrants are identified consistently, and track the virtual separation area in *pI* and MW space in each gel image, in spite of the presence of deformations in the electrophoretic diffusion process. The collection of spots was determined for each subquadrant and the integral of the correspondent spot volumes was obtained and considered as a quantitative feature to be used in the successive exploratory data analysis [5]. In this way, the samples were described as vectors of sorted features and could be investigated by means of a nonlinear dimensionality reduction technique.

The purpose of dimension reduction is to find a manifold (coordinate system) of smaller dimension that will allow to project the data vectors on it obtaining a low-dimensional compact representation of the data. Traditional dimension reduction techniques such as principal component analysis (PCA) and factor analysis usually work well when the data lie on or near a linear subspace of the input space [12]. Unfortunately, they fail to discover nonlinear structures embedded in the set of data points [13]. In contrast to the linear techniques, the nonlinear techniques have the ability to deal with complex nonlinear data such as biological data. In this application we used the LTSA on handling the problem of high dimensionality; this technique describes local properties of high-dimensional data using the tangent space in the neighborhood of a data point to represent the local geometry, and then align these local tangent spaces to construct the global coordinate system for the nonlinear manifold. LTSA is based on the observation that, if local linearity of the manifold is assumed, there exists a linear mapping from high-dimensional data point to its local tangent space, and that there exists a linear mapping from the corresponding low-dimensional data point to the same local tangent space. LTSA attempts to align these linear mappings in such a way, that they construct the local tangent space of the manifold from the low-dimensional representation. In practice, LTSA simultaneously searches for the coordinates of the low-dimensional data representations, and for the linear mappings of the low-dimensional data points to the local tangent space of high-dimensional data. LTSA starts with computing bases for the local tangent spaces at the data points x_i. This is done by applying PCA on the k data points x_{ij} that are neighbors of data point x_i. This results in a mapping M_i from the neighborhood of x_i to the local tangent space Θ_i. A property of the local tangent space Θ_i is that there exists a linear mapping L_i from the local tangent space coordinates θ_{ij} to the low-dimensional representations y_{ij}. Using this property of the local tangent space, LTSA performs the following minimization

$$\min_{Y_i, L_i} \sum_i \left\| Y_i J_k - L_i \Theta_i \right\|^2, \qquad (1)$$

where J_k is the centering matrix of size k. Zhang and Zha [12] have shown that the solution of the minimization is formed by the eigenvectors of an alignment matrix B, that correspond to the d smallest nonzero eigenvalues of B. The entries of the alignment matrix B are obtained by iterative summation (for all matrices V_i and starting from $b_{ij} = 0$ for $\forall\, ij$)

$$B_{N_i, N_i} = B_{N_i, N_i} + J_k (I - V_i V_i^T) J_k, \qquad (2)$$

where N_i is a selection matrix that contains the indices of the nearest neighbors of data point x_i. Subsequently, the low-dimensional representation Y is obtained by computation of the eigenvectors corresponding to the d smallest nonzero eigenvectors of the symmetric matrix

$$\frac{1}{2}(B + B^T). \qquad (3)$$

To assess the accuracy of the discrimination of the samples belonging to different groups obtained with the proposed representation, a leave one out cross-validation through linear discriminant analysis (LDA) was accomplished.

III. Results and Discussion

The procedure was applied on the considered data set to the aim of assessing the chance to discriminate patients affected by ALS from control subjects and neuropathic patients (either with or without pain) and to verify whether the method would be able to discern between subjects with or without algic symptomatology in the peripheral neuropathy groups (these latter samples were already processed with the linear method of dimensionality reduction of PCA, as reported in [5]). The four possible pairwise comparisons between the considered clinical groups were accomplished.

In Fig. 3(a) it is reported the result of the first comparison, between ALS and CN samples. After dimension reduction, the two groups are disposed in separable regions. The output may reveal the structure and the distribution of the input data set. Leave one out cross-validation provided an accuracy of 68.75%. The coordinates of the low-dimensional space were able to account for the differences between ALS and NPN subjects as shown in the Fig. 3(b). The accuracy estimated by leave one out cross-validation was 81.25%. The result of LTSA for the comparison of the data relative to ALS and PN samples is shown in the Fig. 3(c): the projection of the data vectors on a low-dimensional manifold allows to see the two groups of samples in

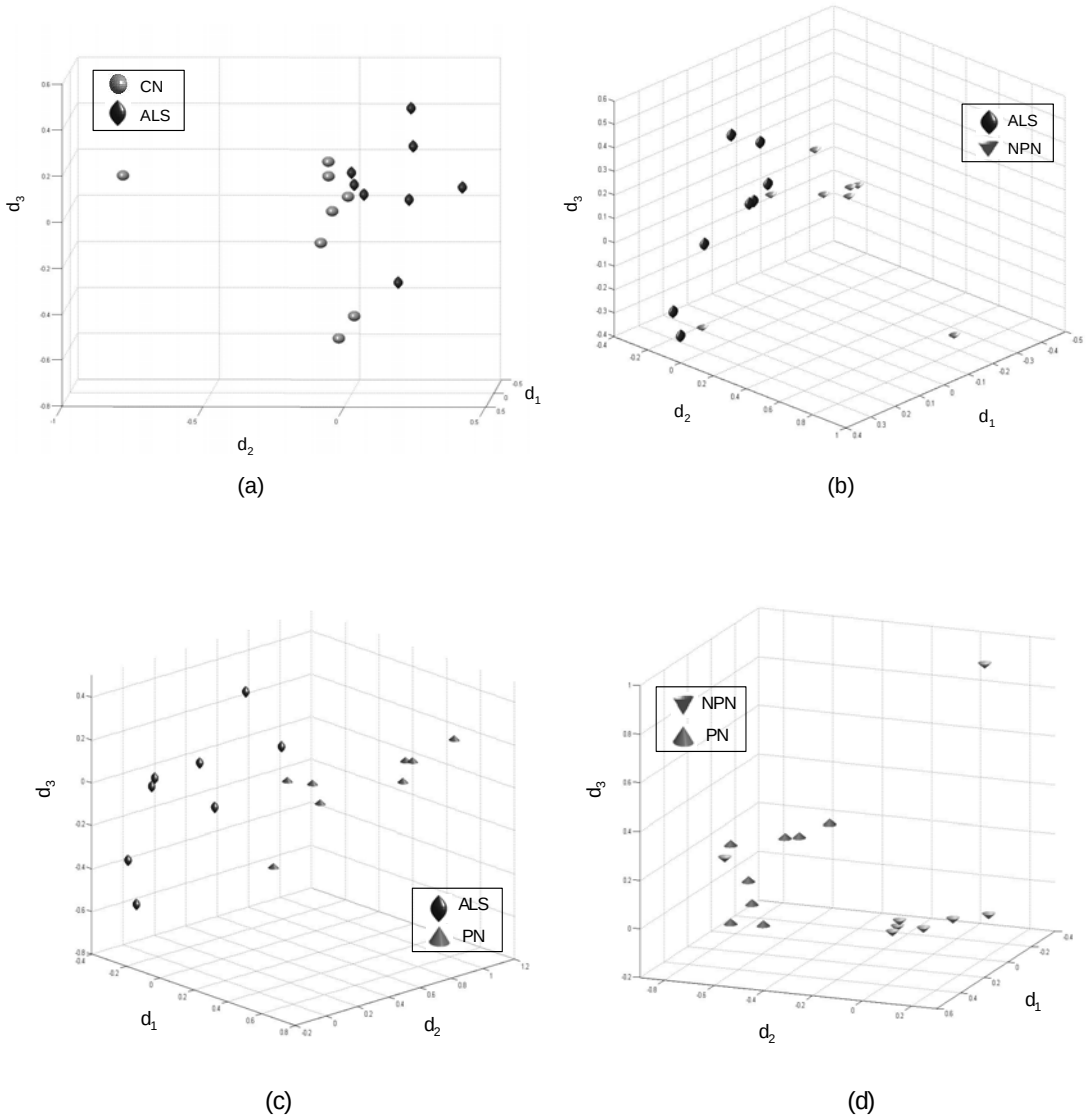


Fig. 3. Visualization of the investigated data set after feature extraction and dimension reduction via LTSA. The projection on 3D manifold shows that patients affected by Amyotrophic Lateral Sclerosis (ALS, dark diamonds) are separated from control subjects (CN, light grey spheres) (a); the compact representation on the low-dimensional space evidences the differences between ALS and neuropathic patients without pain (NPN, dark grey cones); the items correspondent to the two categories examined, ALS samples vs neuropathic patients with pain (PN, dark grey cones), are positioned in different regions of the space (c) and even subjects without pain and with algic symptomatology are compared, the corresponding positions are clearly clustered according their clinical conditions, except for a sample of the NP group, that, although, initially classified as NPN, has successively developed algic symptomatology (d).

different areas of the space. For this comparison we obtained a leave one out cross-validation of 100%. At last we analyzed the NPN and PN groups to assess the chance to discriminate between subjects without pain and with algic symptomatology. The synthetic representation provided by LTSA for these two categories, observable in Fig. 3(d), allows to segregate the samples consistently with their clinical conditions. The capacity of identifying discriminative patterns between different clinical conditions through the application of the developed method has been confirmed, also in this case, as in [5], by the detection of a single outlier. As showed in Fig. 3(d), a sample of NPN is clearly positioned in the opposite category; in fact at the first diagnosis this patient was annotated as NPN but successively showed disease progression with pain appearance, as learnt in further clinical controls. A leave one out cross-validation reported an accuracy of 68.75%.

The results obtained are very significant and robust considering that the samples of the data set are not technical replicates, i.e. gels obtained from fractions of the

same biological sample. In this work, “biological” replicates were processed, i.e. each gel image is representative of a different human subject, so the gels are characterized by low homogeneity, however none of the gel was excluded from the analysis; this brings the tackled problem to a much higher level of variability and complexity. The method developed can be a useful complement in the routine of a proteomics laboratory, because it is highly repeatable and does not need any “a priori” information. It may provide an effective visualization tool and lead to the definition of a protocol of automatic classification, that may represent a complementary approach to the differential analysis aiming to perform rapid and systematic screening tests. 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.

REFERENCES

[1] T. Rabilloud, “Two-dimensional gel electrophoresis in proteomics: Old, old fashioned, but it still climbs up the mountains,” *Proteomics*, vol. 2, pp. 3–10, 2002.

[2] E. Marengo, E. Robotti, P.G. Righetti and F. Antonucci, “New approach based on fuzzy logic and principal component analysis for the classification of two-dimensional maps in health and disease. Application to lymphomas,” *J. Chromatogr. A*, vol. 1004, pp. 13-28, 2003.

[3] A.W. Dowsey, M.J. Dunn and G.Z. Yang, “The role of bioinformatics in two-dimensional gel electrophoresis,” *Proteomics*, vol. 3, pp. 1567–1596, 2003.

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

[5] L. Pattini, S. Mazzara, A. Conti, S. Iannaccone, S. Cerutti and M. Alessio, “An integrated strategy in two-dimensional electrophoresis analysis able to identify discriminants between different clinical conditions,” *Exp. Biol. Med.*, vol. 233(4), pp. 483-91, 2008.

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

[7] E. Marengo, R. Leardi, E. Robotti, P.G. Righetti, F. Antonucci and D. Cecconi, “Application of three-way principal component analysis to the evaluation of two-dimensional maps in proteomics,” *J. Proteome Res.*, vol. 2, pp. 351-360, 2003.

[8] J. Tenenbaum, V. de Silva, J. Langford, “A global geometric framework for nonlinear dimensionality reduction,” *Science*, vol. 290, pp. 2319-2323, 2000.

[9] A. Conti, P. Ricchiuto, S. Iannaccone, B. Sferrazza, A. Cattaneo et al., “Pigment epithelium-derived factor is differentially expressed in peripheral neuropathies,” *Proteomics*, vol.5, pp. 4558-4567, 2005.

[10] A. Conti, S. Iannaccone, B. Sferrazza, L. De Monte, S. Cappa et al., “Differential expression of ceruloplasmin isoforms in the cerebrospinal fluid of Amyotrophic Lateral Sclerosis patients,” *Proteomics Clin. Appl.*, to be published (in press).

[11] A. T. Rosengren, J. M. Salmi, T. Aittokallio, J. Westerholm, R. Lahesmaa et al., “Comparison of PDQuest and Progenesis software packages in the analysis of two-dimensional electrophoresis gels,” *Proteomics*, vol. 3, pp. 1936-1946, 2003.

[12] Z. Zhang and H. Zha, “Principal manifolds and nonlinear dimension reduction via local tangent space alignment,” *SIAM Journal of Scientific Computing*, vol. 26(1), pp. 313-338, 2004.

[13] T. Zhang, J. Yang, D. Zhao and X. Ge, “Linear local tangent space alignment and application to face recognition,” *Neurocomputing*, vol. 70, pp. 1547-1553, 2007.