Anal Bioanal Chem (2008) 391:1163–1173
DOI 10.1007/s00216-008-1856-8

ORIGINAL PAPER

A new method of comparing 2D-PAGE maps based on the computation of Zernike moments and multivariate statistical tools

Emilio Marengo · Elisa Robotti · Marco Bobba · Marco Demartini · Pier Giorgio Righetti

Received: 15 October 2007 / Revised: 17 December 2007 / Accepted: 9 January 2008 / Published online: 7 February 2008

Abstract The aim of this work was to obtain the correct classification of a set of two-dimensional polyacrylamide gel electrophoresis map images using the Zernike moments as discriminant variables. For each 2D-PAGE image, the Zernike moments were computed up to a maximum *p* order of 100. Partial least squares discriminant analysis with variable selection, based on a backward elimination algorithm, was applied to the moments calculated in order to select those that provided the lowest error in cross-validation. The new method was tested on four datasets: (1) samples belonging to neuroblastoma; (2) samples of human lymphoma; (3) samples from pancreatic cancer cells (two cell lines of control and drug-treated cancer cells); (4) samples from colon cancer cells (total lysates and nuclei treated or untreated with a histone deacetylase inhibitor). The results demonstrate that the Zernike moments can be successfully applied for fast classification purposes. The final aim is to build models that can be used to achieve rapid diagnosis of these illnesses.

Keywords Zernike moments · Partial least squares discriminant analysis · 2D-PAGE maps · Classification · Multivariate analysis

E. Marengo (✉) · E. Robotti · M. Bobba · M. Demartini
Department of Environmental and Life Sciences,
University of Eastern Piedmont,
Via Bellini 25/G,
15100 Alessandria, Italy
e-mail: marengoe@tin.it

P. G. Righetti
Department of Chemistry, Materials and Engineering Chemistry
"Giulio Natta", Polytechnic of Milano,
Via Mancinelli 7,
20131 Milan, Italy

Introduction

Since every cell or biological fluid is rich in proteins, an efficient method for achieving their separation and successive determination is necessary. This problem was partially solved by the development of two-dimensional electrophoretic separation, which is certainly the most widely used analytical method in proteomics. This technique allows the efficient separation of the protein content of a particular cell or biological fluid, producing a two-dimensional image of the proteins present in the sample under investigation. Two-dimensional polyacrylamide gel electrophoresis [1, 2] has a unique capacity to resolve complex mixtures, permitting the simultaneous analysis of hundreds or even thousands of proteins. The separation is achieved by two successive electrophoretic runs: the first (through a pH gradient) separates the proteins according to their isoelectric points, while the second (through a porosity gradient) separates them according to their molecular masses. The result is a two-dimensional map with spots (proteins) spread all over the gel surface.

2D-PAGE (from 2-Dimensional PolyAcrylamide Gel Electrophoresis) maps can therefore be used for both diagnostic and prognostic purposes: by investigating the differences between the 2D-PAGE gels of control and pathological individuals, it is possible to classify the patients accordingly or even to capture the evolution of the disease [3–7].

The problem of how best to compare maps belonging to different individuals thus becomes the fundamental issue in the application of this technique for diagnostic/prognostic purposes.

In the field of drug development, especially for cancer, the 2D-PAGE technique is also widely applied [8, 9]. The study of two-dimensional maps can provide useful infor-

mation about the effectiveness of a drug treatment; that is, it can be performed to investigate whether the drug has had the expected effect on the protein contents of the pathological cells.

Unfortunately, the comparison of different 2D-PAGE maps is not a trivial process to perform [10, 11]. The main difficulty that arises during such comparisons is the high complexity of the specimen, which can result in maps with thousands of spots; this complexity is also increased by the complicated sample pretreatment, which is often characterized by many purification/extraction steps. These experimental steps may cause the appearance of spurious spots due to impurities in the final 2-D maps. Moreover, the differences between the treated and reference samples can be very small, thus complicating their identification in a real complex map.

In the classical approach, the comparison is performed by specific software, such as Melanie III or PD-Quest [12, 13]. In this case, each 2-D slab gel is analyzed by a densitometer that provides the optical density at each point on each map. The analysis performed by this software consists of the following different steps:

- Spot detection: the identification of protein spots in the gel image.
- Spot revelation: the software reveals the spots independently for each map.
- Matching the maps: the 2-D maps are matched to reveal common features (spots present in all maps) and those that differ between maps.

This procedure is usually time-consuming and affected by the particular ability of the operator, which tends to determine the final quality of the results.

Here, the classification of 2D-PAGE maps was performed in a completely different way: by decomposing the map images in terms of Zernike moments and applying multivariate tools usually adopted in image analysis problems. The moments calculated are then coupled to multivariate classification techniques; here we use partial least squares discriminant analysis (PLS–DA). The procedure proposed here allows us to bypass all of the steps listed previously, in particular the critical action of aligning the maps, which is not necessary here since Zernike moments are invariant with respect to map translations.

The Zernike moments of the map images were calculated using software written in Visual Basic and developed in-house, and then PLS–DA [14–18] together with variable selection procedures [14–18] were applied to classify the samples considered. This procedure was applied to four different datasets (six cases were studied overall) to check the general validity of the procedure.

Theory

Moment functions have a broad spectrum of application in image analysis, such as for invariant pattern recognition, object classification, pose estimation, image coding and reconstruction. A set of moments computed from a digital image generally represents global characteristics of the image shape and provides a lot of information about the different types of geometrical features of the image. The ability of image moments to represent features has been widely exploited in object identification techniques in several areas of computer vision and robotics [19–25]. Geometric moments were the first to be applied to images, as they are computationally very simple. As research into image processing has progressed, many new types of moment functions have been recently introduced, each with its own advantages for specific applications.

In this paper, complex Zernike moments have been implemented as feature descriptors for 2D-PAGE map classification. Zernike moments were first introduced by Teague [26] based on orthogonal functions called Zernike polynomials [27]. Though computationally very complex, compared to geometric and Legendre moments [28–30], Zernike moments have been shown to provide superior feature representation and low noise sensitivity [31, 32]. Moreover, the orthogonal basis for Zernike moments means that a value of zero can be attained for the redundancy measure in a set of moment functions, such that these orthogonal moments correspond to independent characteristics of the image. In other words, moments with orthogonal basis functions can be used to represent the image with a set of mutually independent descriptors, yielding a minimum amount of information redundancy. Therefore, orthogonal moments are more robust than nonorthogonal moments in the presence of image noise.

For the specific application we are interested in, two of the various important features of Zernike moments, invariance to rotation and translations of the image, are particularly relevant.

Zernike moments

Zernike moments are based on orthogonal Zernike polynomials defined using the polar coordinates inside a unit circle. The two-dimensional Zernike moments of order p with repetition q for an image intensity function $f(r, \vartheta)$ are defined as:

$$Z_{pq} = \frac{p+1}{\pi} \int_{\vartheta=0}^{2\pi} \int_{r=0}^{1} V_{pq}^{*}(r,\vartheta) f(r,\vartheta) r \, \mathrm{d}r \, \mathrm{d}\vartheta, \qquad |r| \leq 1,$$

where the Zernike polynomials of order p with repetition q, $V_{pq}(r,\vartheta)$, are defined as:

$$V_{pq}(r,\vartheta) = R_{pq}(r)\mathrm{e}^{iq\vartheta},$$

and the real-value radial polynomial, $R_{pq}(r)$, is given as follows:

$$R_{pq}(r) = \sum_{k=0}^{(p-|q|)/2} (-1)^k \frac{(p-k)!}{k!((p+|q|)/2-k)!((p-|q|)/2-k)!} r^{p-2k},$$

$$0 \le |\mathrm{q}| \le p \text{ and } \mathrm{p} - |q| \text{ is even}$$

Since Zernike moments are defined in terms of polar coordinates (r,ϑ) with $|r| \le 1$, their computation requires a linear transformation of the image coordinates (i, j) (with i, j=0,1,2,...,N−1) to a suitable domain $(x,y) \in R^2$ inside a unit circle. In this way we can express the discrete approximation of the continuous integral of the moments as:

$$Z_{pq} = \frac{2(p+1)}{\pi(N-1)^2} \sum_{i=0}^{N-1}\sum_{j=0}^{N-1} R_{pq}(r_{ij}) e^{-iq\vartheta_{ij}} f(i,j)$$

where the general image coordinate transformation to the interior of the unit circle is given by:

$$r_{ij} = \sqrt{x_i^2 + y_j^2},\ \vartheta_{ij} = \tan^{-1}\left(\frac{y_i}{x_i}\right),\ \text{where}$$

$$x_i = \frac{\sqrt{2}}{N-1} i - \frac{1}{\sqrt{2}},\ y_j = \frac{\sqrt{2}}{N-1} j - \frac{1}{\sqrt{2}}$$

The image intensity function $f(i, j)$ can be reconstructed from a finite number n of Zernike moments, using the following equation:

$$f(r,\vartheta) = \sum_{p=0}^{n}\sum_{q} Z_{pq} V_{pq}(r,\vartheta),\ \text{defined for: } |q| \le p \text{ and } p-|q| = \text{even}$$

Zernike moments were calculated here by exploiting the so-called Q-recursive method, developed by Mukundan, Raveendran and Chong [33], which allows them to be calculated in a reduced computational time.

In this method, $R_{pq}(r)$ (with p=q−4) is calculated by:

$$R_{p(q-4)}(r) = H_1 R_{pp}(r) + \left(H_2 + \frac{H_3}{r^2}\right) R_{p(q-2)}(r)$$

where:

$$R_{pp} = r^p$$

$$R_{p(q-2)}(r) = pR_{pp}(r) - (p-1)R_{(p-2)(p-2)}(r)$$

$$H_1 = \frac{q(q-1)}{2} - qH_2 + \frac{H_3(p+q+2)(p-q)}{8}$$

$$H_2 = \frac{H_3(p+q)(p-q+2)}{4(q-1)} + (q-2)$$

$$H_3 = \frac{-4(q-2)(q-3)}{(p+q-2)(p-q+4)}$$

Partial least squares discriminant analysis (PLS–DA)

Partial least squares (PLS) [14–18] is a multivariate regression method that allows the relationship between one or more dependent variables (Y) and a group of descriptors (X) to be established. The X and Y variables are modeled simultaneously to find the latent variables (LVs) in X that will predict the latent variables in Y. These latent variables (also called the PLS components) are similar to the principal components calculated from principal component analysis [14-18]: they are extracted such that each successive latent variable accounts for the largest possible amount of variation that is not accounted for by the previous variables (they are orthogonal to each other), in both the descriptor space (X) and the response space (Y). The LVs are computed hierarchically so that the last LVs are mostly responsible for random variations and experimental error.

The optimal number of LVs (i.e., a model that uses the information in X to predict the response Y while avoiding overfitting) is determined by the residual variance in prediction. Here, leave-one-out cross-validation is applied to evaluate the predictive ability and to select the optimal number of latent variables in the final model.

In the case where a large number of descriptors (X variables) are present or a large experimental error is expected, it can be quite difficult to obtain a final model with a suitable predictive ability. In these cases, some techniques for variable selection can be exploited. Here, two subsequent strategies were applied: an initial simplification of the model achieved by eliminating groups of nonsignificant X variables up to a maximum of 200 variables, based on the minimum error obtained in cross-validation; and a second phase where variables were eliminated one at a time to provide a final model with an overall minimum error in cross-validation.

PLS was created to model continuous responses, but it can also be applied for classification purposes by establishing an appropriate Y that is related to the membership of each sample to a class. When only two classes are present

(for example control and treated samples), a binary Y variable is added to the dataset, which is coded so that −1 is attributed to one class (control samples) and +1 to the other one (treated samples). When more than two classes are present, the Y matrix contains one column for each class, and the sample is coded to +1 for the column corresponding to the class it belongs to, and −1 for the other classes.

The regression is then carried out between the X-block variables (Zernike moments) and the Y variables just established. This process of classification is called PLS discriminant analysis (PLS–DA).

Model evaluation

The coefficient of multiple determination, R^2, for PLS was calculated as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \left(\hat{y}_i - y_i\right)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

where the two sums run on the samples used for calibration (R^2), or for validation (R^2_{cv}); $\hat{y}_i$ is the predicted value of the response for the i-th experiment; y_i is the experimental predicted value for the response for the i-th experiment; y is the average response of the samples used for calibration (R^2), or for validation (R^2_{cv}).

The root mean square error (RMSE) is calculated as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} \left(\hat{y}_i - y_i\right)^2}{n}}$$

where the sum runs on the samples used for calibration (RMSEC), and for validation (RMSECV).

In this case, the best model complexity in terms of both the number of X variables present and the number of latent variables in the model was selected by the minimum value of RMSECV (leave-one-out procedure).

Experimental

Datasets

The procedure was applied to six groups of 2D-PAGE maps belonging to four different pathologies: human lymphoma, neuroblastoma, human colon cancer, human pancreatic cancer. For lymphoma and neuroblastoma, the comparison involved:

- *Lymphoma*: Four samples from the GRANTA519 cell line of human lymphoma (control) and four samples from the MAVER-1 cell line
- *Neuroblastoma*: Four samples from control adrenal mouse glands (control) and four samples from adrenal mouse glands affected by neuroblastoma.

The other two cases under investigation were more complex:

- *Colon cancer* exposed to a histone deacetylase (HDAC) inhibitor. Nuclei and total cell lysates were investigated from colon cancer cell line HCT116. The nuclei dataset comprised six control (diseased) and five samples treated with a HDAC inhibitor. The lysates dataset instead comprised five control and five HDAC inhibitor-treated samples.
- *Pancreatic cancer*. Two human pancreatic cancer cell lines were investigated: the PACA44 and T3M4 cell lines, both treated or untreated with trichostatin A. For the PACA 44 cell line, four control and four drug-treated samples were investigated, while for the T3M4 cell line the dataset consisted of five control and five drug-treated samples.

The experimental protocols followed in order to obtain the several 2D-PAGE maps used in this study are not reported here since they are described elsewhere [10, 34, 35] and represent standard practice in proteomics.

Software

PLS–DA with variable selection was performed by PARVUS (M. Forina, University of Genova, Italy, http://www.parvus.unige.it/). Zernike moments were computed by software developed in-house in Visual Basic (Microsoft Visual Studio 6.0). Data pretreatment and graphical representations were performed by Visual Basic, Parvus and Microsoft Excel 2003.

Results and discussion

Image pretreatment

Each map was automatically digitalized to provide a grid of 100×100 pixels, where each pixel contained the gray-scale intensity at the corresponding position in the image. These values therefore ranged from 0 (black) to 255 (white).

Each image map was then pretreated to eliminate the contribution from the background to the signal. This correction exploits two threshold values that must be fixed: the value of the first derivative of each image (the *slope*), indicating the presence of an actual spot rather than noise, and the value of the pixel (the *cut*), indicating the threshold value corresponding to the background.

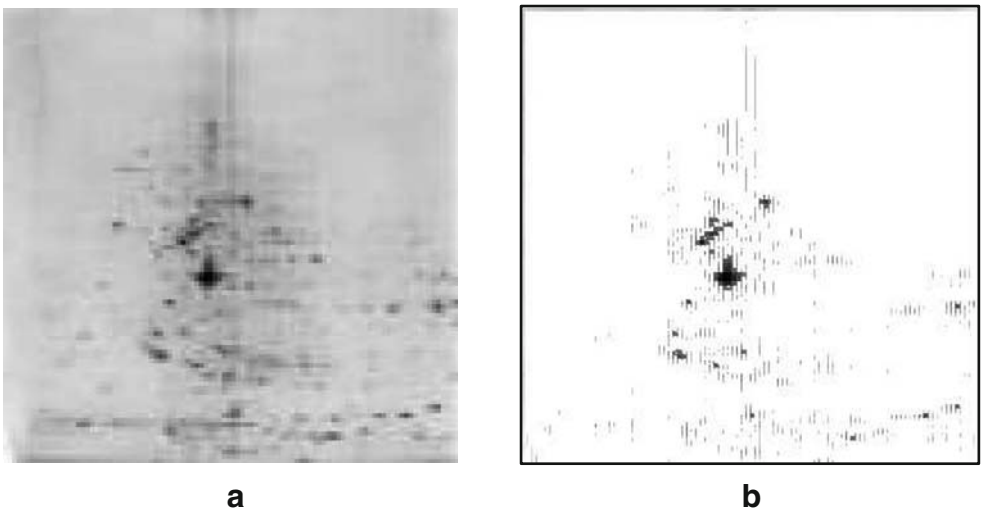

Fig. 1 Example of map pre-treatment: sample CTR1 from the Lymphoma dataset before (**a**) and after (**b**) background correction with cut=100 and slope=15

For each image, the first derivative is calculated as difference between the values of two adjacent pixels (calculations are performed row-wise). Each image is then corrected for the background by considering the first derivative and the value of each pixel simultaneously: if the first derivative is less than the slope and the pixel is larger than the cut, the value of the pixel is set to 255 (white). Good results can be obtained with cut values of 100–150 and slope values ranging from 10 to 20. Figure 1 shows an example of a sample corrected using cut=100 and slope=15. All of the maps were treated using these two values for cut and slope.

Zernike moment calculation and dataset preparation

For each image, Zernike moments were calculated with a maximum p order of 100. This procedure provides a total of 2601 moments. The algorithm allows the separate calculation of the real and imaginary parts of the moments, providing a total of 5202 descriptors for each image: 2601 corresponding to the real parts of the moments and 2601 to the imaginary parts. Since the X matrix can only contain real numbers, only the coefficient of the imaginary part of the Zernike moment was considered (i.e., the numerical coefficient multiplying the i character). For example, the complex number $(-5.34 - 0.0478i)$ can be separated into two parts: −5.34 is the real part and −0.0478 is the imaginary part.

The samples in the four different datasets were coded as follows:

- Lymphoma dataset: GRANTA519 cell line (CTR1-4) and MAVER-1 cell line (MAV1-4)
- Neuroblastoma dataset: adrenal mouse glands (CTR1-4) and adrenal mouse glands affected by neuroblastoma (ILL1-4);
- Colon cancer dataset: nuclei from colon cancer cells (CTR1-6) and nuclei from colon cancer cells treated by a HDAC inhibitor (NHD1-5); lysates from colon cancer cells (CTR1-5) and lysates from colon cancer cells treated with the inhibitor (LHD1-5)
- Pancreatic cancer dataset: Control (PACA1-4) and drug-treated (PTSA1-4) PACA 44 cell line; control (T3M41-5) and drug-treated (TTSA1-5) T3M4 cell line.

PLS–DA

Each dataset was autoscaled before performing PLS–DA. PLS–DA was applied, as specified in the "Theory" section, with variable selection, exploiting a backward elimination algorithm. This procedure enables only the most relevant moments—those that allow the correct classification of the samples (minimum RMSECV)—to be identified.

Due to the large number of variables present (5202), the backward elimination procedure was applied in two consecutive steps:

- A first selection was made where groups of nonsignificant variables were eliminated, providing a final

Table 1 Variance (%) explained by the first LV for the X and Y variables for each case study

	LV_1	
	% Expl. Var. X	% Expl. Var. Y
Lymphoma	88.90	99.90
Neuroblastoma	74.65	99.68
Pancreas: PACA cell line	94.20	99.43
Pancreas: T3M4 cell line	79.77	99.35
Colon cancer: Lysates	94.58	99.22
Colon cancer: Nuclei	65.16	96.66

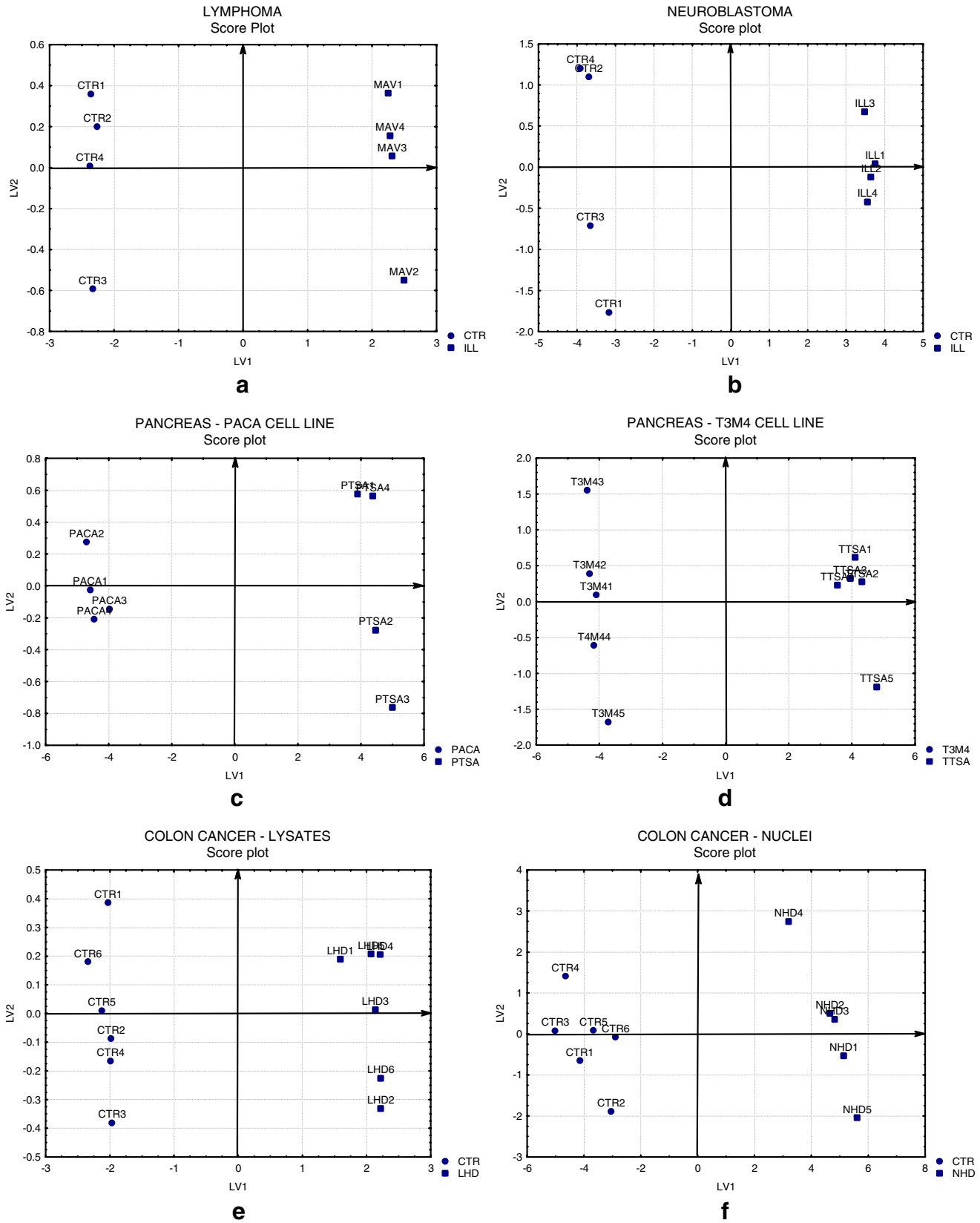

Fig. 2 Score plots of the first two LVs for the four datasets investigated: Lymphoma (**a**); Neuroblastoma (**b**); Pancreas (PACA cell line: **c**; T3M4 cell line: **d**); Colon cancer (Lysates: **e**; Nuclei: **f**). Control samples are represented as *circles*; samples belonging to the second class are represented as *squares*

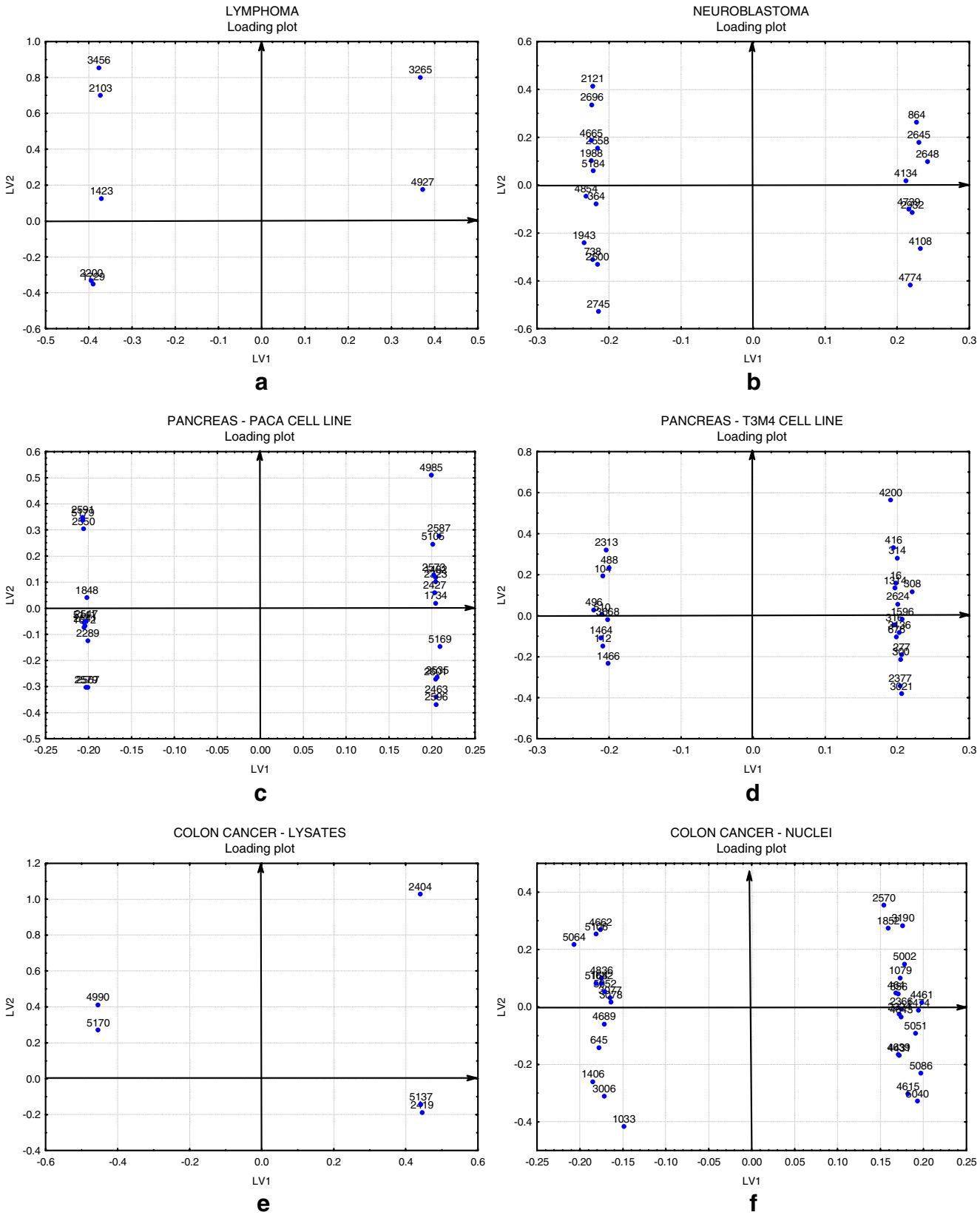

Fig. 3 Loading plots of the first two LVs for the four datasets investigated: Lymphoma (**a**); Neuroblastoma (**b**); Pancreas (PACA cell line: **c**; T3M4 cell line: **d**); Colon cancer (Lysates: **e**; Nuclei: **f**). Zernike moments are represented by numbers

Table 2 R^2 and RMSE values calculated for fitting (R^2, RMSEC) and cross-validation (R^2_{cv}, RMSECV) for all of the datasets investigated

	R^2	R^2_{cv}	RMSEC	RMSECV
Lymphoma	0.9990	0.9980	0.0367	0.0508
Neuroblastoma	0.9968	0.9951	0.0653	0.0846
Pancreas: PACA cell line	0.9943	0.9894	0.0875	0.1045
Pancreas: T3M4 cell line	0.9935	0.9890	0.0904	0.1093
Colon cancer: Lysates	0.9922	0.9891	0.0965	0.1046
Colon cancer: Nuclei	0.9666	0.9471	0.2012	0.2331

dataset containing a maximum of 200 moments, based on finding the smallest error in cross-validation

– A second refinement was performed by eliminating the variables one at a time in order to select the actual number of moments that provide the smallest error in cross-validation.

Table 1 reports the amount of variance explained by the first LV for each case study for both X and Y variables. The first LV was considered the significant one for all of the cases under investigation (leave-one-out cross-validation). The first LV in fact explains more than 99% of the total amount of information contained in the Y variable; the only exception is the Nuclei dataset, for which it explains about 96%. The use of one LV in each classification model allows the correct classification of all of the samples in each dataset with a final NER% (non-error rate) of 100%.

Figures 2 and 3 report scores and loadings plots, respectively, for all of the investigated datasets. The control sample class is plotted as circles and the other class as squares in the score plots. For all cases, the control samples are located at large negative values along the first LV, while the other class is located at large positive scores. This behavior confirms the ability of the first LV to separate the samples in the two classes present for each dataset.

The analysis of the corresponding loading plots allows the identification of the Zernike moments responsible for the differences between the classes of samples. In each loading plot, each moment is represented by a number from 1 to 5202. Moments located at large negative values along LV_1 show large positive values for control samples and large negative ones for the other class; the moments located at large positive values along LV_1 show the opposite behavior: large negative values for control samples and large positive ones for the other class in each dataset. The loading plots report only the moments found to be significant by the backward elimination procedure applied to each dataset.

Table 2 reports the R^2 and RMSE values obtained during fitting and cross-validation for all of the datasets investigated. The R^2 values show that the models provide very

Table 3 Zernike moments considered to be significant for the six cases under investigation; real and imaginary parts are reported separately

Lymphoma		Neuroblastoma		Pancreas: PACA cell line		Pancreas: T3M4 cell line		Colon cancer: Lysates		Colon cancer: Nuclei	
Real	Imaginary	Real	Imaginary	Real	Imaginary	Real	Imaginary	Real	Imaginary	Real	Imaginary
p67q45R	p87q3I	p93q57R	p100q14I	p65q61R	p28q14I	p100q70R	p100q56I	p27q3R	p28q4I	p91q43R	p92q68I
P58q4R	p82q48I	p85q29R	p100q8I	p61q57R	p18q14I	p98q94R	p92q38I	p26q0R	p15q3I	p89q41R	p91q19I
p43q29R	p32q6I	p82q30R	p99q13I	p57q53R	p14q10I	p98q78R	p89q83I		p10q4I	p87q41R	p91q17I
p39q3R		p50q16R	p98q14I	p53q49R	p10q6I	p95q41R	p77q27I			p78q16R	p88q64I
		p48q26R	p94q28I	p34q12R	p8q6I	p94q90R	p62q20I			p77q3R	p54q30I
		p42q36R	p65q11I	p32q12R		p94q74R				p68q10R	p53q25I
		p19q3R	p64q24I	p25q11R		p94q62R				p66q10R	p47q23I
		p12q2R	p45q17I	p22q12R		p94q58R				p59q39R	p46q22I
			p42q2I	p15q5R		p92q46R				p53q41R	p46q14I
			p40q16I	p14q8R		p90q86R				p32q10R	p45q23I
			p36q12I	p13q11R		p90q70R				p29q21R	p44q14I
			p7q5I	p13q5R		p90q42R				p10q2R	p37q11I
				p10q8R		p86q66R					p27q9I
				p9q7R		p70q54R					p24q12I
				p8q4R		p66q30R					p23q15I
				p6q4R		p66q26R					p23q13I
				p5q3R		p62q26R					p22q12I
				p3q3R		p32q32R					p20q12I
				p0q0R		p28q28R					p18q12I
											p11q5I

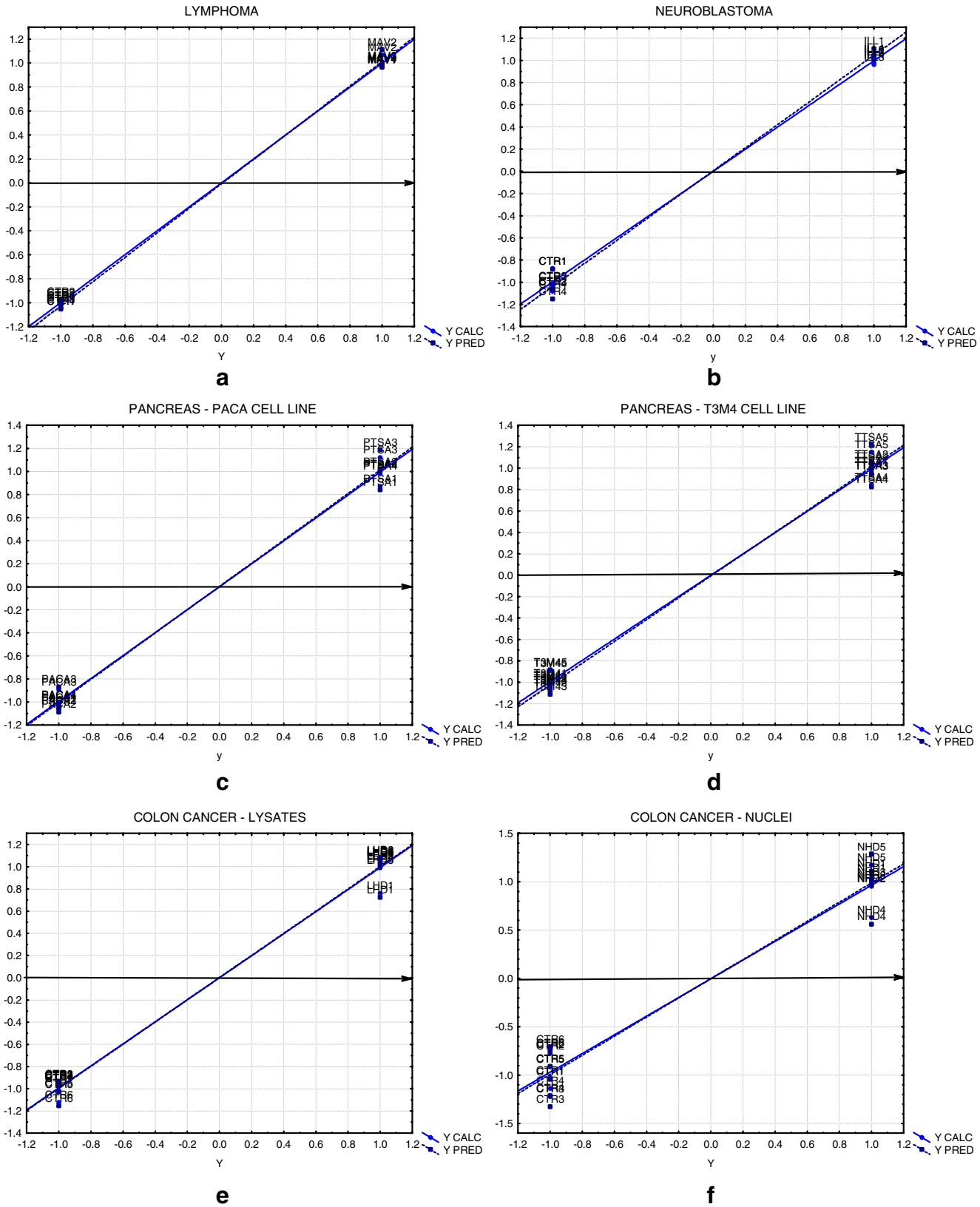

Fig. 4 Calculated and predicted *Y* values vs. reference *Y* values for the four datasets investigated: Lymphoma (**a**); Neuroblastoma (**b**); Pancreas (PACA cell line: **c**; T3M4 cell line: **d**); Colon cancer (Lysates: **e**; Nuclei: **f**). Calculated values are represented as *circles*, while predicted values are shown as *squares*. *Solid regression lines* correspond to calculated values, while *dotted regression lines* correspond to predicted values

good performances in terms of both fitting and validation. The worst results were obtained for the Nuclei dataset, which still however presents R^2 and R^2_{cv} values of above 0.94. The good abilities of the derived classification models to describe the information provided (fitting) and to predict new values (validation) are also demonstrated by the RMSE values calculated: the fitting errors (RMSEC) are almost all below 0.1 (the only exception is for the Nuclei dataset), while validation errors (RMSECV) are almost all below 0.11 (again, the only exception is for the Nuclei dataset).

These conclusions are also confirmed by Fig. 4, which reports, for each case study, the calculated and predicted *Y* values vs. the actual *Y* values. In all cases there is good agreement between the actual and the calculated or predicted values. Since in this case PLS is used as a classification tool, the most important information provided by these diagrams is represented by the variations in the calculated and predicted values along the *Y* axis: the positions of both the fitted and the validated values at negative values for control samples and at positive ones for the other class in each dataset prove that the models derived here provide 100% NER%. This is also true for the nuclei dataset, even if the variations along the *Y* axis appear to be the largest in this case.

Table 3 reports the number of significant moments selected for each dataset; for each case studied, the real and imaginary parts of the moments are reported separately. Moments are represented by an alphanumeric string reporting the values of the orders *p* and *q* followed by *R* if the moment represents the real part or by *I* if it represents the imaginary part. The number of significant moments (ranging from five for the Lysates dataset to 32 for the Nuclei dataset) shows the importance of the selection procedure, which eliminates information present in the maps that is not directly related to the classification of the samples (i.e., redundant information). The analysis of the *p, q* orders found to be significant then shows that the significant moments do not show recursive *p, q* values; in other words, for the different cases studied, different moments are significant. This is logical, since different classes of maps will show differences in different areas of the maps themselves. Unfortunately, it is not a trivial task to directly identify the features in each group of images that determine the differences between the two classes investigated: this is due to the particular nature of Zernike moments (and other image moment functions), which capture global independent aspects of each image.

Conclusions

A new method for the fast comparison of proteomic 2-D maps is presented here. The method exploits Zernike moment functions coupled to classification tools. Zernike moments were calculated for four different datasets (six case studies in total) of varying complexity, all characterized by the presence of two classes: control samples and diseased or treated samples.

The procedure developed proved to be a successful tool for extracting the global information present in the maps obtained from 2-D gel electrophoresis: PLS–DA provided the correct classification of all of the samples for all of the cases investigated. The application of backward elimination procedures enabled the most parsimonious set of moments that provided the best cross-validation results to be identified. For the cases investigated, final numbers of moments ranging from 5 to 32 were found to be significant for classification.

The method proposed could be applied in principle to perform rapid comparisons of 2-D proteomic maps; increasing the number of samples in each class could also lead to its use in diagnostic applications.

It is, however, important to point out that Zernike moments extract general independent aspects of an image, and so they do not easily and directly provide information about the differences that exist between the classes of maps investigated. At the moment, the reconstruction of the images based on the significant moments selected enables large areas of the image containing the most relevant differences to be identified. For all the cases studied, these areas contain relevant information (i.e., actual spots). This initial information is important, since we can state that Zernike moments classify the images based on the spots rather than differences in the background or image artefacts. However, this information is not sufficient if the purpose of the diagnostic (a role ably fulfilled by the proposed procedure) is connected to the identification of differences (functional proteomics). Work is in progress in our lab to solve this problem and thus to also make Zernike moments useful from this point of view.

References

1. Righetti PG, Stoyanov A, Zhukov M (2001) The proteome revisited: theory and practice of all relevant electrophoretic steps. Elsevier, Amsterdam
2. Wilkins MR, Williams KL, Appel RD, Hochstrasser DF (1997) Proteome research: new frontiers in functional genomics. Springer, Berlin
3. Fountoulakis M, Schlaeger EJ (2003) Electophoresis 24:260–275
4. Gromov PS, Ostergaard M, Gromova I, Celis JE (2002) Prog Biophys Mol Biol 80:3–22
5. Dwek MV, Rawlings SL (2002) Mol Biotechnol 22:139–152
6. Castegna A, Aksenov M, Thongboonkerd V, Klein JB, Pierce WM Jr, Booze R, Markesbery WR, Butterfield DA (2002) J Neurochem 82:1524–1532
7. Sinha P, Kohl S, Fisher J, Htter G, Kern M, Kttgen E, Dietel M, Lage H, Schnlzer M, Schadendorf D (2000) Electrophoresis 21:3048–3057

8. Ryan TE, Patterson SD (2002) Trends Biotechnol 20(Suppl):S45–S51
9. Steiner S, Witzmann FA (2000) Electrophoresis 21:2099–2104
10. Marengo E, Robotti E, Righetti PG, Campostrini N, Pascali J, Ponzoni M, Hamdan M, Astener H (2004) Clin Chim Acta 345:55–67
11. Schmid HR, Schmitter D, Blum O, Miller M, Vonderschmitt D (1995) Electrophoresis 16:1961–1968
12. Westergren-Thorsson G, Malmstrom J, Marko-Varga G (2001) J Pharm Biomed Anal 24:815–824
13. Bathia K, Lord R, Stanton P (2002) Eur J Cancer 28:S156
14. Massart DL, Vandeginste BGM, Deming SM, Michotte Y, Kaufman L (1988) Chemometrics: a textbook. Elsevier, Amsterdam
15. Vandeginste BGM, Massart DL, Buydens LMC, De Jong S, Lewi PJ, Smeyers-Verbeke J (1998) Handbook of chemometrics and qualimetrics: part B. Elsevier, Amsterdam
16. Eisenbeis RA (ed) (1972) Discriminant analysis and classification procedures: theory and applications. DC Heath and Co., Lexington, USA
17. Martens H, Naes T (1989) Multivariate calibration. Wiley, London
18. Kleinbaum D, Kupper L, Muller K (1988) Applied regression analysis and other multivariate methods, 2nd edn. PWS-Kent, Boston, MA
19. Wee C, Paramesran R, Takeda F (2004) Inform Sci 159:203–220
20. Kan C, Srinath MD (2002) Pattern Recogn 35:143–154
21. Zenkouar H, Nachit A (1997) Mat Sci Eng B–Solid 49:211–215
22. Yin J, Rodolfo De Pierro A, Wei M (2002) Appl Math Comput 132:249–263
23. Hu MK (1962) IRE Trans Inf Theory 8:179–187
24. Khotanzad A, Hong YH (1990) IEEE Trans Pattern Anal Mach Intell 12:489–497
25. Li BC, Shen J (1991) Pattern Recogn 24:807–813
26. Teague MR (1980) J Opt Soc Am 70:920–930
27. Zernike F (1934) Physica 1:689–704
28. Chong C, Raveebdram P, Mukundan R (2004) Pattern Recogn 37:119–129
29. Mukundan R, Ramakrishnan KR (1995) Pattern Recogn 28:1433–1442
30. Zhou JD, Shu HZ, Luo LM, Yu WX (2002) Pattern Recogn 35:1143–1152
31. Belkasim SO, Shridhar M, Ahmadi M (1991) Pattern Recogn 24:1117–1138
32. Teh CH, Chin RT (1988) IEEE Trans Pattern Anal Mach Intell 10:496–513
33. Chong CW, Raveendran P, Mukundan R (2003) Pattern Recogn 36:731–742
34. Marengo E, Robotti E, Cecconi D, Scarpa A, Righetti PG (2004) Anal Bioanal Chem 379(7–8):992–1003
35. Marengo E, Robotti E, Bobba M, Liparota MC, Antonucci F, Rustichelli C, Zamò A, Chilosi M, Hamdan M, Righetti PG (2006) Electrophoresis 27:484–494