Available online at www.sciencedirect.com
SCIENCE DIRECT
Journal of Chromatography A, 1096 (2005) 86–91
JOURNAL OF CHROMATOGRAPHY A
www.elsevier.com/locate/chroma

Use of Legendre moments for the fast comparison of two-dimensional polyacrylamide gel electrophoresis maps images

Emilio Marengo [a,*], Marco Bobba [a], Maria Cristina Liparota [a], Elisa Robotti [a], Pier Giorgio Righetti [b]

[a] *Department of Life and Environmental Science, University of Eastern Piedmont, Via Bellini 25/G, 15100 Alessandria, Italy*
[b] *Department of Industrial Biotechnologies, University of Verona, Strada delle Grazie 15, 37134 Verona, Italy*

Available online 19 August 2005

Abstract

In this paper, Legendre moments are calculated to extract the global information from a set of two-dimensional polyacrylamide gel electrophoresis map images. The dataset contains 18 samples belonging to two different cell lines (PACA44 and T3M4) of control (untreated) and drug-treated pancreatic ductal carcinoma cells. The aim of this work was to obtain the correct classification of the 18 samples, using the Legendre moments as discriminant variables. For each image the Legendre moments up to a maximum order of 100 were computed. The stepwise linear discriminant analysis (LDA) was performed in order to select the moments with the highest discriminating power. The results demonstrate that the Legendre moments can be successfully applied for fast classification purposes and similarity analysis.

Keywords: Orthogonal moments; Legendre moments; Proteomics; Chemometrics; Stepwise linear discriminant analysis; Cluster analysis

1. Introduction

Since each cell or biological fluid has a rich protein content, an efficient method for achieving their separation and successive determination is necessary. Two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) [1,2] has a unique capacity for the resolution of complex mixtures, permitting the simultaneous analysis of hundreds or even thousands of proteins. The separation is achieved by two successive electrophoretic runs: the first run (through a pH gradient) separates the proteins according to their isoelectric point, while the second run (through a porosity gradient) separates them according to their molecular mass. The result of this technique is a two-dimensional map with spots spread all over the gel surface; each spot represents a particular protein and for this reason the 2D-PAGE maps may be considered as a "snapshot" of the protein content of the investigated cell at a given point of its life cycle.

The physiological state of a particular cell or tissue is related to its protein content, and a particular disease may cause differences in the proteins contained in the pathological tissue. These differences may consist of changes in the relative abundance or in the appearance/disappearance of some proteins [3–7]. Thus, 2D-PAGE maps are fundamental tools for both diagnostic and prognostic purposes. In fact it is possible to compare maps belonging to healthy subjects with maps belonging to subjects affected by any pathology, in order to point out the differences, which may permit to identify new powerful markers for diagnostic and prognostic purposes, as well as to understand the disease biochemical mechanism, through the identification of the proteins involved.

In the field of drug development, especially for cancer, the 2D-PAGE technique is widely applied [8,9]. The study of two-dimensional maps can give information about the effectiveness of a drug treatment, that is to investigate if this has played the expected role on the protein content of the pathological cell.

Unfortunately, the comparison of different 2D-PAGE maps is not a trivial process to achieve [10,11]. The difficulty which arises during the comparison is above all due

* Corresponding author. Tel.: +39 0131 360272; fax: +39 0131 287426.
E-mail address: marengoe@tin.it (E. Marengo).

0021-9673/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.chroma.2005.06.100

to the high complexity of the specimen, which can produce maps with thousands of spots; the complexity is also increased by the highly articulated sample pretreatment, often characterised by many purification/extraction steps. These experimental steps may cause the appearance of spurious spots due to impurities in the final 2D-maps. Moreover, the differences occurring between treated and reference samples can be very small, thus complicating their identification in a real complex map.

Usually, the differential analysis between classes of different 2D-PAGE maps is achieved by means of commercial software (i.e., Melanie III or PDQuest) [12,13], which exploit the following three-step method: (i) the 2D-PAGE images to be compared are aligned, so that all gels are reduced to the same size. The maps are then matched to each other; (ii) The spots present on each map are independently revealed; (iii) the maps are matched to each other in order to identify the common information and the differences.

In this work the procedure of classification of the 2D-PAGE maps has been performed without the use of commercial software and the steps listed above were by-passed; in other words the original maps, without pre-treatment, have been used for classification. It is important to emphasize that the phase of alignment of the maps is not necessary because the Legendre moments are invariant with respect to translation. This is the main advantage of this new technique of clustering of the maps, based on the Legendre moments.

The Legendre moments were calculated by a Matlab algorithm and then the stepwise linear discriminat analysis [14–16] was executed to select the moments that allowed the best separation of the classes of maps. The dendrogram was also computed to verify the separation of the samples [17–19].

In this paper, the Legendre moments technique is applied to a dataset comprising 18 samples belonging to two different cell lines (PACA44 and T3M4) of pancreatic human cancer before and after the treatment with Trichostatin A. The aim of this work is to assign the 18 samples to the correct class.

If proved to be effective, this tool could permit to perform fast search and comparisons in 2D-PAGE maps data bases for diagnostic purposes.

2. Theory

Moments function of image intensity values can be used to capture the global features of the image. Various types of moments have been widely used for pattern recognition, image analysis and object classification [20–27]. Moment functions such as geometric moments, orthogonal moments, rotational moments and complex moments are all useful tools in the field of pattern recognition and can be used to describe the features of objects such as the shape, area, border, location and orientation.

In this paper orthogonal Legendre moments [28–30] have been implemented as feature descriptors in the application of 2D-PAGE maps classification. Orthogonal basis of the moments can attain a zero value of redundancy measure in a set of moment functions, so 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 by a set of mutually independent descriptors, with a minimum amount of information redundancy. Therefore, orthogonal moments have additional properties of being more robust with respect to the non-orthogonal moments, in the presence of image noise.

Orthogonal moments also permit the analytical reconstruction of an image intensity function from a finite set of moments, using the inverse moment transform.

2.1. Legendre moments

The two-dimensional Legendre moments of order $(p+q)$, given an image intensity map $f(x,y)$, are defined as:

$$L_{pq} = \frac{(2p+1)(2q+1)}{4} \int_{-1}^{1} \times \int_{-1}^{1} P_p(x) P_q(y) f(x,y)\,\mathrm{d}x\,\mathrm{d}y; \quad x, y \in [-1,1], \tag{1}$$

where Legendre polynomial, $P_p(x)$, of order p is given by:

$$P_p(x) = \sum_{k=0}^{p} \left\{ (-1)^{p-k/2} \frac{1}{2^p} \frac{(p+k)!x^k}{((p-k)/2)!((p+k)/2)!k!} \right\}_{p-k=\text{even}} \tag{2}$$

The recurrence relation of Legendre polynomials, $P_p(x)$, is gives as follows:

$$P_p(x) = \frac{(2p-1)xP_{p-1}(x) - (p-1)P_{p-2}(x)}{p}, \tag{3}$$

where $P_0(x) = 1$, $P_1(x) = x$ and $p > 1$. Since the region of definition of Legendre polynomials is the interior of $[-1,1]$, a square image of $N \times N$ pixels with intensity function $f(i,j)$, $0 \le i, j \le (N-1)$, is scaled in the region $-1 < x, y < 1$.

Eq. (1) can then be expressed in discrete form as:

$$L_{pq} = \lambda_{pq} \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} P_p(x_i) P_q(y_j) f(i,j), \tag{4}$$

where the normalizing constant is:

$$\lambda_{pq} = \frac{(2p+1)(2q+1)}{N^2} \tag{5}$$

x_i and y_j denote the normalized pixel coordinates in the range $[-1,1]$, which are given by:

$$x_i = \frac{2i}{N-1} - 1 \quad \text{and} \quad y_j = \frac{2j}{N-1} - 1 \tag{6}$$

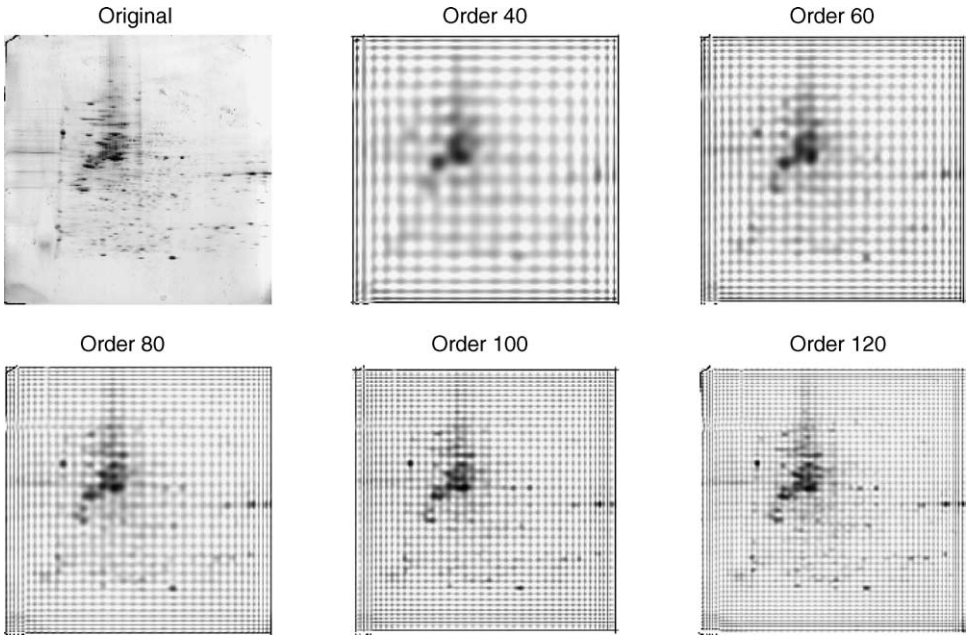

Fig. 1. Image reconstruction using Legendre moments of increasing order.

The reconstruction of the image function from the calculated moments can be expressed as:

$$f(i,\, j) = \sum_{p=0}^{p_{\max}} \sum_{q=0}^{q_{\max}} \lambda_{pq} P_p(x_i) P_q(x_j) \tag{7}$$

Fig. 1 shows the results of the reconstruction of a 2D-PAGE map with Legendre moments of increasing order:

3. Cluster analysis

Cluster analysis technique is a multivariate statistical tool to aggregate the objects according to their similarity. As a result hierarchically or non-hierarchical ordered clusters are formed. The similarity of objects is described by an appropriate similarity measure.

One possibility for clustering objects is their hierarchical aggregation. Here the objects are combined according to their distances or similarities. Two different procedures exist: agglomerative and divisive. Divisive cluster formation is based on splitting the whole set of objects into individual clusters. With the most frequently used agglomerative clustering one starts with single objects and merges them to larger objects groups.

The result of such clustering is a graph, called dendrogram, in which the objects (x-axis) are connected at decreasing levels of similarity (y-axis). The results of hierarchical clustering methods depend on the specific linking method and on the specific measure of similarity.

3.1. Linear discriminant analysis (LDA)

LDA is a Bayesian classification method that allows the discrimination of the samples present in a dataset considering its multivariate structure. The assignment of a sample, $\boldsymbol{x}$, characterized by p features, to a class j of all classes g is based on maximizing the *posterior probability*

$$P(j|\boldsymbol{x}) \text{ for } j = 1, \ldots, g \tag{8}$$

Application of Bayes's theorem for calculation of the posterior probability gives:

$$P(j|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|j)P(j)}{p(\boldsymbol{x})} \tag{9}$$

According to Eq. (9), the posterior probability is computed from the probability density function for the considered class, $P(j|\boldsymbol{x})$, the prior probability for that class $P(j)$ and the probability density function over all classes $p(\boldsymbol{x})$. A sample $\boldsymbol{x}$ is then assigned to that class j, for which the largest posterior probability is found.

For computation of the class probability density, $p(\boldsymbol{x}|j)$, the multidimensional normal distribution is assumed:

$$p(\boldsymbol{x}|j) = (2\pi)^{-d/2} |\boldsymbol{S}_j|^{-0.5} \times \exp[-0.5(\boldsymbol{x} - \bar{\boldsymbol{x}}_j)\boldsymbol{S}_j^{-1}(\boldsymbol{x} - \bar{\boldsymbol{x}}_j)^{\mathrm{T}}] \tag{10}$$

where the covariance matrix $\boldsymbol{S}_j$ based on the class centroid $\bar{\boldsymbol{x}}_j$ is obtained from:

$$\boldsymbol{S}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_j)^{\mathrm{T}} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_j) \tag{11}$$

$$\bar{\boldsymbol{x}}_j = \frac{1}{n}\sum_{i=1}^{n_j} \boldsymbol{x}_i^{(j)} \quad (12)$$

n_j describes the number of samples in class j.

Maximizing the posterior probability is related to minimizing the discriminant scores obtained from:

$$\mathrm{d}_j(\boldsymbol{x}) = (\boldsymbol{x} - \bar{\boldsymbol{x}}_j)^{\mathrm{T}} \boldsymbol{S}_j^{-1} (\boldsymbol{x} - \bar{\boldsymbol{x}}_j) + \ln |\boldsymbol{S}_j| - 2 \ln P(j) \quad (13)$$

An unknown sample is assigned to the class j for which the distance to its class centroid is shortest. The first term of Eq. (13) ($(\boldsymbol{x} - \bar{\boldsymbol{x}}_j)^{\mathrm{T}} \boldsymbol{S}_j^{-1} (\boldsymbol{x} - \bar{\boldsymbol{x}}_j)$) represents the Mahalanobis distance between the sample $\boldsymbol{x}$ and the class centroid $\bar{\boldsymbol{x}}_j$.

In LDA it is assumed that the class covariance matrices are equal, i.e. $\boldsymbol{S}_j = \boldsymbol{S}$ for all $j = 1$ to g.

4. Experimental

4.1. Dataset

Each cell line pool, T3M4 and PACA44, was split so that half of the pool was treated with TSA and the other was not. Therefore, the dataset, containing 18 2D maps of pancreatic human cancer, was divided into four classes:

- four replicate 2D maps of a PACA44 cell line pool;
- four replicate 2D maps of a PACA44 cell line treated for 48 h with TSA;
- five replicate 2D maps of a T3M4 cell line pool;
- five replicate 2D maps of a T3M4 cell pool treated for 48 h with TSA.

Fig. 2 shows an example of the experimental 2D PAGE maps obtained for each class.

4.2. Chemicals and materials

Urea, thiourea, 3-[(cholamidopropyl)dimethylammonium]-1-propane-sulfonate (CHAPS), iodoacetamide (IAA), tributylphosphine (TBP) and sodium dodecyl sulfate (SDS) were obtained from Fluka (Buchs, Switzerland). Bromophenol blu and agarose were from Pharmacia-LKB (Uppsala, Sweden). Acrylamide, *N*′,*N*′-methylenebisacrylamide, ammonium persulfate, TE-MED, the Protean IEF Cell, the GS-710 Densitometer and the 17 cm long, immobilised pH 3–10 linear gradient strips were from Bio-Rad Labs. (Hercules, CA, USA). Ethanol, methanol and acetic acid were from Merck (Darmstadt, Germany). Trichostatin A (TSA) was obtained from Sigma–Aldrich (St. Louis, MO, USA). A 2.2 mM solution of TSA in absolute ethanol was prepared and stored at −80 °C until use.

4.3. Cell treatment with TSA

PACA44 and T3M4 cell were grown in RPMI 1640 supplemnted with 20 mM glutamine and 10% (v/v) FBS (BioWhittaker, Italy) and were incubated at 37 °C with 5% (v/v) CO_2. Subconfluent cells were treated with 0.2 mM TSA for 48 h.

4.4. Cell lysis

Protein extraction from cells was performed with lysis buffer (40 mM Tris, 1% (v/v) NP40, 1 mM Na_2VO_4, 1 mM NaF, 1 mM PMSF, protease inhibitor cocktail). Cells were left in lysis buffer for 30 min in ice. After centrifugation at 14,000 × *g* at 4 °C for removal of particulate material, the protein solution was collected and stored at −80 °C until used.

4.5. Two-dimensional electrophoresis

Seventeen centimetre long, pH 3–10 immobilized pH gradient strips (IPG; Bio-Rad Labs., Hercules, CA, USA) were rehydrated for 8 h with 450 μL of 2D solubilizing solution (7 M urea, 2 M thiourea, 5 mM tributylphosphine, 40 mM Tris and 20 mM iodoacetamide) containing 2 mg mL^{-1} of total reduced/alkylated protein from sample cells. Isoelectric focussing (IEF) was carried out with Protean IEF Cell (Bio-Rad Labs.) with a low initial voltage and then by applying a voltage gradient up to 10,000 V with a limiting current of 50 μA. The total product time × voltage applied was 70,000 Vh for each strip, and the temperature was set at 20 °C. For the second dimension, the IPGs strips were equilibrated for 26 min by rocking in a solution of 6 M urea, 2% (w/v) SDS, 20% (v/v) glycerol, 275 mM Tris–HCl, pH 8.8. The IPG strips were then laid on a 8–18% T gradient SDS-PAGE with 0.5% (w/v) agarose in the cathode buffer (192 mM glycine, 0.1% (w/v) SDS and Tris to pH 8.3). The anodic buffer was a solution of 375 mM Tris–HCl, pH 8.8. The electrophoretic run was performed

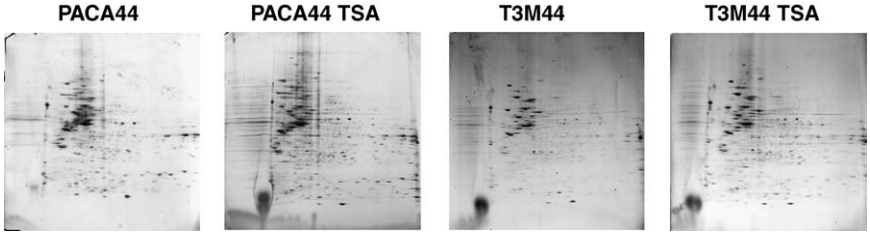

Fig. 2. 2D-PAGE maps of the real sample of pancreatic human cancer.

by setting a current of 2 mA for each gel for 2 h, then 5 mA/gel for 1 h, 10 mA/gel for 20 h and 20 mA/gel until the end of the run. During the whole run the temperature was set at 11 °C. Gels were stained overnight with colloidal Coomassie blu [0.1% (w/v) Comassie Brilliant Blue G, 34% (v/v) methanol, 3% (v/v) phosphoric acid and 17% (w/v) ammonium sulphate]; destaining was performed with a solution of 5% (v/v) acetic acid until a clear background was achieved.

4.6. Software

Stepwise LDA was performed with STATISTICA (Statsoft, ver. 6.1, USA). Legendre moments were computed with MATLAB (The Mathworks, ver. 6.5, USA); this software was also used for data treatment and for graphical representations.

5. Results and discussion

The 2D gels of all the samples (PACA44 control and treated with TSA, T3M4 control and treated with TSA) were scanned with a GS-710 densitometer (Bio-Rad Labs.).

Each 2D-PAGE, which was automatically digitalised, is described by a matrix of dimension 200 pixels × 200 pixels; the value of each pixel varying from 0 to 1 indicate the staining intensity in the given position.

The Legendre moments of the 18 digitalised images were calculated. Moments up to a maximum order of 100 where computed from the images and therefore a matrix of dimension 101 × 101 is obtained for each image. This matrix holds the global information of the 2D-PAGE map.

The final dataset contains 18 samples and 10,201 variables. The number of variables is very large and many of them do not contain information related to the specific target of correctly classifying the samples, for this reason a method for selecting the variables having the highest power of discrimination was applied (stepwise LDA).

5.1. Stepwise linear discriminant analysis

LDA was performed on the final dataset. The most discriminant moments were selected by a stepwise procedure in forward search with $F_{\text{to-enter}} = 4.0$.

The results of stepwise LDA procedure shows that only six different Legendre moments are necessary in order to discriminate the four classes of samples. In this case, the discriminant functions are a linear combination of the moments L_{pq} selected, which are reported in Table 1.

Using the selected variables all samples are correctly classified by LDA. In order to evaluate the classification models predictive ability leave-one-out cross-validation method [31] was used, since the number of samples available is not sufficient for performing a more severe validation. Also the leave-one-out validation provided a 100% of correct classification, showing that the six selected variables can be effectively used also for predictive purposes.

Table 1
Legendre moments selected by means of stepwise LDA

Order of moment	
p	*q*
2	0
2	11
3	10
5	5
86	8
96	0

5.2. Cluster analysis

The six moments selected with stepwise LDA are able to separate the four classes of samples and therefore they were used to perform a cluster analysis to verify how the samples are grouped. The dendrogram calculated using the Ward method [32] and the Euclidean distance to compute the samples similarity is reported in Fig. 3.

The ordinate labelled $(D_{\text{leg}}/D_{\text{max}}) \times 100$ is a percentage dissimilarity scale expressing the linking distance (D_{leg}) of the groups of objects as a fraction of the maximum possible distance (D_{max}). It is possible to observe that the samples are divided into two main groups, at a normalised distance of more than 80%; the first group contains the samples belonging to PACA44 cells while the second group contains the samples belonging to T3M4 cells.

At a normalised distance of more then 45% the two main groups are then separated into two sub-groups: both the considered cell lines are correctly separated in control and treated samples.

The dendrogram demonstrates and confirms that the six Legendre moments, selected in the stepwise LDA phase, are able to correctly separate the four classes of samples.

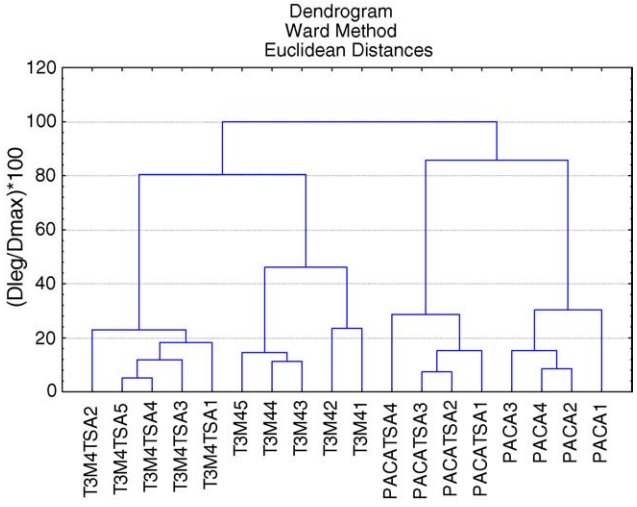

Fig. 3. Dendrogram calculated on the basis of the six Legendre moments.

6. Conclusions

In this work a new method, based on the Legendre moments, has been developed for classifying 2D-PAGE maps. Legendre moments were calculated here on a dataset comprising 18 samples that belong to two different cell lines (T3M4 and PACA44) of control (untreated) and drug-treated pancreatic human cancer cells.

Legendre moments proved to be a successful tool for extracting the global information present in the images of the 2D maps obtained from 2D gel-electrophoresis: stepwise-LDA provided the correct classification of the samples by mean of only six moments, thus greatly simplifying the system dimensionality.

The separation of the 18 samples into four groups by mean of the six Legendre moments was also confirmed by cluster analysis.

This method could be applied in principle for a fast comparison and/or search along the available proteomic databases. Its diagnostic/prognostic application is straightforward.

Work is in progress in our laboratory to compare the results obtained with Legendre moments to other image analysis tools, like Zernicke and Chebyschev moments.

Acknowledgements

Professor Emilio Marengo gratefully acknowledges financial support from Regione Piemonte, Ricerca Scientifica Applicata, D.D.R.R. no. 78-9416 of 19 May 2003 and no. 111-20277 of 1 August 2003.

References

[1] P.G. Righetti, A. Stoyanov, M. Zhukov, The Proteome Revisited: Theory and Practice of all Relevant Electrophoretic Steps, Elsevier, Amsterdam, 2001.
[2] M.R. Wilkins, K.L. Williams, R.D. Appel, D.F. Hochstrasser, Proteome Research: New Frontiers in Functional Genomics, Springer, Berlin, 1997.
[3] M. Fountoulakis, E.J. Schlaeger, Electophoresis 24 (2003) 260.
[4] P.S. Gromov, M. Ostergaard, I. Gromova, J. Celis, Prog. Biophys. Mol. Biol. 80 (2002) 3.
[5] M.V. Dwek, S.L. Rawlings, Mol. Biotechnol. 22 (2002) 139.
[6] A. Castegna, M. Aksenov, V. Thongboonkerd, J.B. Klein, W.M. Pierce Jr., R. Booze, W.R. Markesbery, D.A. Butterfield, J. Neurochem. 82 (2002) 1524.
[7] P. Sinha, S. Kohl, J. Fisher, G. Htter, M. Kern, E. Kttgen, M. Dietel, H. Lage, M. Schnlzer, D. Schadendorf, Electrophoresis 21 (2000) 3048.
[8] T.E. Ryan, S.D. Patterson, Trends Biotechnol. 20 (Suppl. S) (2002) 45.
[9] S. Steiner, F.A. Witzmann, Electrophoresis 21 (2000) 2099.
[10] E. Marengo, E. Robotti, P.G. Righetti, N. Campostrini, J. Pascali, M. Ponzoni, M. Hamdan, H. Astener, Clin. Chim. Acta 1–2 (2004) 55.
[11] H.R. Schmid, D. Schmitter, O. Blum, M. Miller, D. Vonderschmitt, Electrophoresis 16 (1995) 1961.
[12] G. Westergren-Thorsson, J. Malmstrom, G. Marko-Varga, J. Pharm. Biomed. Anal. 24 (2001) 815.
[13] K. Bathia, R. Lord, P. Stanton, Eur. J. Cancer 28 (2002) S156.
[14] K. Jensen, C. Kesmir, I. Sondergaard, Electrophoresis 17 (1996) 694.
[15] F.H. Grus, A.J. Augustin, Ophthalmologie 97 (2000) 54.
[16] D. Swets, J. Weng, IEEE Trans. Pattern Anal. Mach. Intell. 18 (1996) 831.
[17] C. Henning, J. Multivariate Anal. 86 (2003) 183.
[18] S.S. Khan, A. Ahmad, Pattern Recogn. Lett. 25 (2004) 1293.
[19] M. James, Classification Algoritsms, Collins, London, 1985.
[20] C. Wee, R. Paramesran, F. Takeda, Inform. Sci. 159 (2004) 203.
[21] C. Kan, M.D. Srinath, Pattern Recog. 35 (2002) 143.
[22] H. Zenkouar, A. Nachit, Mater. Sci. Eng. B: Solid 49 (1997) 211.
[23] J. Yin, A. Rodolfo De Pierro, M. Wei, App. Math. Comput. 132 (2002) 249.
[24] M.K. Hu, IRE Trans. Inf. Theory 8 (1962) 179.
[25] M.R. Teague, J. Opt. Soc. Am. 70 (1980) 920.
[26] A. Khotanzad, Y.H. Hong, IEEE Trans. Pattern Anal. Mach. Intell. 12 (1990) 489.
[27] B.C. Li, J. Shen, Pattern Recog. 24 (1991) 807.
[28] C. Chong, P. Raveebdram, R. Mukundan, Pattern Recog. 37 (2004) 119.
[29] R. Mukundan, K.R. Ramakrishnan, Pattern Recog. 28 (1995) 1433.
[30] J.D. Zhou, H.Z. Shu, L.M. Luo, W.X. Yu, Pattern Recog. 35 (2002) 1143.
[31] S. Wold, Technometrics 20 (1978) 397.
[32] M. Otto, Chemometrics. Statistics and Computer Application in Analytical Chemistry, Wiley-VCH, Weinheim, 1999.