Available online at www.sciencedirect.com
ScienceDirect
Journal of Chromatography B, 849 (2007) 251–260
JOURNAL OF CHROMATOGRAPHY B
www.elsevier.com/locate/chromb

Improvements in the search for potential biomarkers by proteomics: Application of principal component and discriminant analyses for two-dimensional maps evaluation[☆]

Ana María Rodríguez-Piñeiro[*], Francisco Javier Rodríguez-Berrocal, María Páez de la Cadena

Departamento de Bioquímica, Genética e Inmunología, Universidad de Vigo, 36310 Vigo, Spain

Received 31 May 2006; accepted 8 September 2006
Available online 30 October 2006

Abstract

In this study, we evaluated if the application of multivariate analysis on the data obtained from two-dimensional protein maps could mean an improvement in the search for protein markers. First, we performed a classical proteomic study of the differential expression of serum *N*-glycoproteins in colorectal cancer patients. Then, applying principal component analysis (PCA) we assessed the utility of the 2-D protein pattern and certain subsets of spots as a tool to distinguish control and case samples, and tested the accuracy of the classification model by linear discriminant analysis (LDA). On the other hand we looked for altered spots by univariate statistics and then analysed them as a cluster by PCA and LDA. We found that those proteins combined presented a theoretical sensitivity and specificity of 100%. Finally, the spots with known protein identity were analysed by multivariate methods, finding a subgroup that behaved as the most obvious candidates for further validation trials.

© 2006 Elsevier B.V. All rights reserved.

Keywords: Two-dimensional gel electrophoresis; 2D-PAGE; Data evaluation; Multivariate statistics; Principal component analysis; Discriminant analysis

1. Introduction

Nowadays, the utility of proteomics in the search for new tumour markers is unquestionable and, as more analytical tools are described in the proteomics area, its weight is increasing in the clinical field. On the one hand, every new technique is obviously welcome to overcome the lack of clinically useful tumour markers. Precisely, proteomic techniques are especially suitable for developing blind searches without previous ideas, and therefore they allow the detection of alterations in proteins that before were not thought to be related to carcinogenesis. On the other hand, it is becoming increasingly evident that it will not be possible to find a single biomarker for a given pathology fulfilling all the requirements of a useful clinical marker. Any marker will always show a lack of sensitivity due to the simple fact that some tumours will be too small to release detectable amounts of protein. Furthermore, other diseases than cancer will lead to the appearance of various markers in blood (i.e. liver disease, infections, etc.) and therefore specificity will never be 100% [1,2].

Hence, proteomics would be a suitable tool to aid in the search for a panel of candidate biomarkers. In proteomic studies of differential expression in serum, a large number of protein candidates for biomarkers can be found; among them, there are both serum proteins that are differentially expressed or modified in cancer patients, and proteins that are secreted by tumour cells into the circulation or intracellular tumour proteins that are released when tumour cells die, which would increase the specificity of a panel of biomarkers. Besides, a proper panel could also include proteins altered in processes concomitant to tumour development (as inflammation and immune response) even if they are not directly related to carcinogenesis. These proteins would increase the sensitivity of the test [3].

The value of an altered molecule as a tumour marker has to be validated *a posteriori* by other specific techniques. This assessment requires a simple method that could be implemented in the clinical routine (as ELISA assays), but it also means to test

[☆] This paper is part of a special volume entitled “Analytical Tools for Proteomics”, guest edited by Erich Heftmann.

[*] Corresponding author. Tel.: +34 986 813841; fax: +34 986 812556.
E-mail address: anarp@uvigo.es (A.M. Rodríguez-Piñeiro).

1570-0232/$ – see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.jchromb.2006.09.021

a large number of samples that permit the pertinent statistical analyses (as ROC or survival curves) [4]. On the other hand, to check the specificity of a marker for diagnosis of a certain disease, it is necessary to study samples from related diseases. Therefore, it is unthinkable to perform all the required tests with each one of the proteins that could be found altered by proteomic techniques. In fact, there is some dismay about the results eventually applicable in health care with regard to proteomics [3]. From the many studies of differential expression between pathological and normal states, only a few ulterior works have confirmed the alterations described, and practically none have corroborated the clinical value of the potential markers. Thence, it would be greatly helpful to find methods to refine the conclusions of proteomics studies and select the altered proteins that are more prone to display a real utility.

Multivariate analyses have been proposed before to aid in the evaluation of proteomic experiments. Among them, there are methods for dimension reduction as principal component analysis (PCA) or partial least squares (PLS); methods for classification as linear discriminant analysis (LDA); clustering methods as hierarchical cluster analysis; and multivariate analyses of variance [5]. Besides, pattern recognition techniques and machine learning methods could also help in complex proteomic analyses [5].

PCA is a multivariate statistical method that allows the representation of the original dataset in a new reference system formed by new variables called factors or principal components (PCs). When these PCs effectively account for the variability of the different sample populations analysed, it is possible to cluster them into the correct groups. The LDA is another type of multivariate statistical method that builds one or more functions (called ‘discriminant’, ‘diagnostic’ or ‘classification’ functions) and uses them to characterise the groups, assigning or classifying the cases into them and measuring the degree of success of the classification model.

Some types of multivariate analyses have been already applied to proteomics. As an example, de Noo et al. [6] recently used a classical LDA to validate MALDI-TOF serum profiles in the detection of colorectal cancer (CRC). Regarding 2-DE, Kovarova et al. [7] applied 2-DE followed by PCA to characterise changes in the proteome of a leukemia cell line after a treatment, finding that the spot pattern after 6 h of treatment was similar to that of control cells, whereas longer times of exposure produced a different protein complement. These authors also attempted to find the proteins that contributed the most to the significant PCs, although they could just ascertain the ones that were significant by univariate analysis. Roblick et al. [8] applied PCA after 2-D of different tissues regarding the progression of CRC (normal tissue, polyps and adenomas, tumour tissues and metastasis) and showed that the proteome of benign lesions was similar to that of the normal tissues and differed from malignant lesions. Recently, Verhoeckx et al. [9] also tried the combination of the DIGE technology with PCA as an explorative data analysis tool, using the test to corroborate the separation of different groups of samples on the basis of their proteome, but eventually studying differences in protein expression levels by a univariate method. Another example is the report of Jia et al. [10], who proposed an approach similar to the one described in this work, using PCA to plot the samples although further analyses required computational knowledge to perform rotation tests and other calculations [11]. Karp et al. [5] recently reported an elegant combination of mean centering scaling with PCA to demonstrate clustering of the samples and absence of outliers, followed by PLS-DA to classify the samples through a regression model, refined by an iterative process. Marengo et al. [12] presented a preliminary approach to the discrimination of 2D-PAGE maps called 3-way PCA that included information regarding both the intensity and the position of the spots, and involved transformation of the data through a maximum scaling technique and other refinements regarding the conventional PCA. In relation to the differential analysis of spot position, in a previous work we have proposed the application of a multivariate method called the relative warp analysis, which is commonly used in the studies based on geometric-morphometrics [13]. On the other hand, Marengo et al. have also developed methods based on fuzzy logic [14] and tested tools as the soft-independent model of class analogy [15].

In this work, we performed a classical proteomics comparison of serum *N*-glycoproteins from CRC patients and healthy donors. Then, we applied different approaches to analyse the data obtained from the 2-D comparison. First, we used a typical univariate test to look for differences of expression of individual proteins. Then, we used the multivariate methods PCA and LDA to find whether the global protein expression or a certain pattern would allow the differentiation of control and case 2-D maps. Furthermore, we took advantage of the combination of both univariate and multivariate techniques in order to aid in the selection of the proteins that could be grouped to discriminate the disease studied. Finally, we tried to select the minimum group of known proteins which alteration allows the separation of samples from patients and healthy individuals, proteins that could be more easily and effectively set up in the clinical routine as a panel of markers. All these different approaches demonstrate the potential of exploring multivariate analyses to complement the conventional tools used in the search of biomarkers through proteomic studies.

2. Experimental

2.1. Chemicals and reagents

Con A-Sepharose 4B was purchased from Sigma–Aldrich Chemie (Steinheim, Germany). Analytical grade of sodium di-hydrogen phosphate2-hydrate, di-sodium hydrogen phosphate anhydrous (PANREAC Quimica, Barcelona, Spain), and methyl-α-D-mannopyranoside (Sigma, Steinheim, Germany) were used as reagents.

ReadyStrip^{TM} IPG Strips (4% T; 3% C) were purchased from Bio-Rad (Hercules, California, USA). Lysis buffer was prepared with 7 M urea, 2 M thiourea and 4% (w/v) 3-[(3-cholamidopropyl)dimethylamonio]-1-propanesulfonate (CHAPS). Rehydration buffer contained 7 M urea, 2 M thiourea, 4% (w/v) CHAPS, 0.3% (w/v) dithiothreitol (DTT), and 0.5% (v/v) Bio-Lytes 3/10 ampholytes. SDS-PAGE equilibration buffer was prepared with 6 M urea, 50 mM Tris pH 8.8, 2% (w/v) SDS, and 30% (v/v) glycerol, and afterwards 1% (w/v) DTT, and 2.5% (w/v) iodoacetamide (IAA) were added. Thiourea was purchased from Sigma (Steinheim, Germany) and IAA was obtained from Merck (Schuchardt, Germany). Silver nitrate was obtained from Sigma (Steinheim, Germany) and the remaining chemicals were purchased from Bio-Rad (Hercules, California, USA).

2.2. *Equipment*

Chromatographic separation was carried out on an Econo Column from Bio-Rad (Hercules, California, USA). Fractions were collected with a Microfraction Collector 203 from Gilson (Middleton, Wisconsin, USA), and optical density was measured in an UVIKON Spectrophotometre 930 from Kontron Instruments (Milan, Italia). Samples were lyophilised in a Christ Alpha 2–4 freeze drier and resuspended in a Sanyo Gallenkamp orbital shaker incubator, both purchased from B. Braun Biotech (Leicester, UK). The Protean IEF Cell for isoelectric focusing and Protean II xi Cell for electrophoresis, the power supply Power PAC 1000, the calibrated densitometre GS-800, and the PDQuest software package were all purchased from Bio-Rad (Hercules, California, USA). MS was performed in a M@LDI-HT^{TM} with the MassLynx software (Micromass-Waters, Saint-Quentin, France). MASCOT Daemon search engine was from Matrix Science (London, UK).

2.3. *Sample preparation*

Blood samples were obtained by venipuncture from five patients operated on for CRC at Complejo Hospitalario Universitario de Vigo (Spain). Blood samples of the control group (five healthy and habitual blood donors) were provided by the Galician Transfusion Center. Drawn blood was allowed to coagulate at room temperature and centrifuged at 2000 × *g* for 15 min. Sera were stored at −85 °C. All procedures involving human samples were performed according to the clinical ethical practices of the Spanish Government and followed the tenets of the Helsinki Declaration. Informed consent was obtained from each subject's guardian, and anonymity was warranted tracing the patients through their clinical history number.

2.4. *Concanavalin A-Sepharose affinity chromatography*

The chromatographic method was performed as published elsewhere [16]. Briefly, 1 mL of filtered serum was applied to a Concanavalin A-Sepharose (Con A) column (0.8 cm × 7 cm), equilibrated in 10 mM sodium/disodium phosphate buffer pH 6.0. Flow was stated at 0.3 mL/min. The column was first washed with 30 mL of equilibrating buffer, releasing the flow-through fraction enriched in non-glycosylated proteins and *O*-glycoproteins. The fraction of interest, constituted mainly by *N*-glycosylated proteins, was selectively eluted with 0.5 M methyl-α-D-mannopyranoside. Optical density (OD) at 280 nm was measured along the chromatographic process. Chromatographic fractions were dialysed against milliQ water at 4 °C overnight, frozen and lyophilised, and then stored at −85 °C until used.

2.5. *Two-dimensional electrophoresis*

The detailed procedure and its reproducibility have been described before [16]. Briefly, the lyophilised eluate corresponding to the *N*-glycoprotein fraction was solubilised by suspension in lysis buffer, and protein concentration was measured according to the modified method of Bradford [17]. Then, 150 μg of protein were mixed with rehydration buffer and separated by isoelectric focusing (IEF) in 17 cm, pH 4–7, linear ReadyStrip^{TM} IPG Strips (4% T; 3% C) in the Protean IEF Cell focusing tray. After incubation with DTT and IAA equilibration buffers, the IPG gel was transferred onto a 9–16% gradient polyacrylamide gel (30% T; 2.6% C) and SDS-PAGE was performed in a Protean II xi Cell. Gels were stained with ammoniacal silver (modified from [18]).

2.6. *Computer analysis of electrophoretic patterns*

Gels were digitised with the GS-800 calibrated densitometre and protein patterns were compared using the PDQuest 7.1.1 software package. Protein spots were detected by the software based on the spot parameters chosen by the user through selection of the biggest, smallest and least intense spot. After subtraction of background, the resulting filtered images were edited to correct inaccuracies (smeared, streaked and overlapped spots were manually cancelled for ulterior comparisons). The intensity levels of the spots were normalised by expressing the intensity of each spot in a gel as a proportion of the total protein intensity detected for the entire gel (relative volume), in order to correct for differences in protein loading and gel staining [19].

For comparison, gels belonging to control and patient samples were matched independently in two sets, and a representative standard gel was obtained from each comparison. Then, these analyses within groups (control group; patient group) were matched between groups (control group standard versus patient group standard). Only spots that were consistently found in all the samples analysed (matched in the 10 individuals) were selected for further statistical comparisons. This selection procedure was applied to avoid null values in the dataset.

2.7. *Data evaluation by univariate and multivariate statistical analyses*

Relative volumes of the spots matched between control and patient samples were imported into the SPSS software package (release 11.5). For univariate analyses, differences in the relative levels of each spot were assessed with the non-parametric Mann–Whitney *U*-test. *P* values ≤0.05 were considered statistically significant.

Multivariate analyses were performed on datasets constituted by the relative levels of all the spots that were consistently matched between the 10 individuals analysed, in order to avoid the replacement of null values by inference. PCA is based on the variability of all the information contained in the complete dataset. This method calculates the eigenvalues and eigenvectors of the correlation matrix generated from the original matrix of ten rows (gels) and as many columns as spots considered in each analysis. Each of the PCs is calculated by the software so as to explain the maximum amount of variance contained in the original dataset and corresponds to a linear combination of the original variables (i.e. the relative volume of each spot in this study). Moreover, the PCs are defined as orthogonal to each other and therefore allow a more effective representation of the system than the original variables. The relevant PCs obtained (those with eigenvalue >1) were considered statistically significant if $P \leq 0.05$ by the Mann–Whitney U-test.

The LDA was performed in order to test the utility of a spot pattern to correctly predict the classification of the samples in their original groups. The test was used to build one or more discriminant functions based on the linear combination of the predictive variables that performed the better discrimination between the two groups considered. Then, the actual and the predicted membership were compared, and the degree of success of the classification model was measured. The accuracy of the classification was assessed through leave-one-out cross-validation, removing each individual in turn from the dataset, recalibrating the discriminant rule and predicting the group for the leftover data. This allowed the correction for over-estimates, so when the percentage of correct classification was much lower for the cross-validated samples, the model was discarded.

2.8. Mass spectrometric protein identification

Individual spots found altered in 2-D gels were identified by MS as described elsewhere [16]. Briefly, spots were excised from Coomassie-stained gels, and destained with 50 mM ammonium bicarbonate and 50% (v/v) ACN. Then, gel pieces were dried and digested with 10 μg/mL trypsin in 25 mM ammonium bicarbonate at 37 °C overnight. Peptides were eluted with 5% (v/v) TFA and 75% (v/v) ACN. Finally, samples were mixed with an α-cyano-4-hydroxycinnamic acid matrix and analysed on a M@LDI-HT. Data processing was performed with MassLynx and peptide fingerprints were searched against nrNCBI and SwissProt databases with the MASCOT Daemon search engine.

3. Results and discussion

In this work, we have applied different statistical approaches to analyse the results of a conventional 2-D experiment comparing two sample conditions. Fig. 1 shows the procedure followed in this study.

3.1. Comparison of two-dimensional protein patterns

Serum samples from five healthy donors and five CRC patients were processed through Con A chromatography [16] in order to obtain a serum fraction enriched in N-glycoproteins. Then, these fractions were separated by 2D-PAGE and silver-stained. Images of the 2-D maps were acquired and analysed

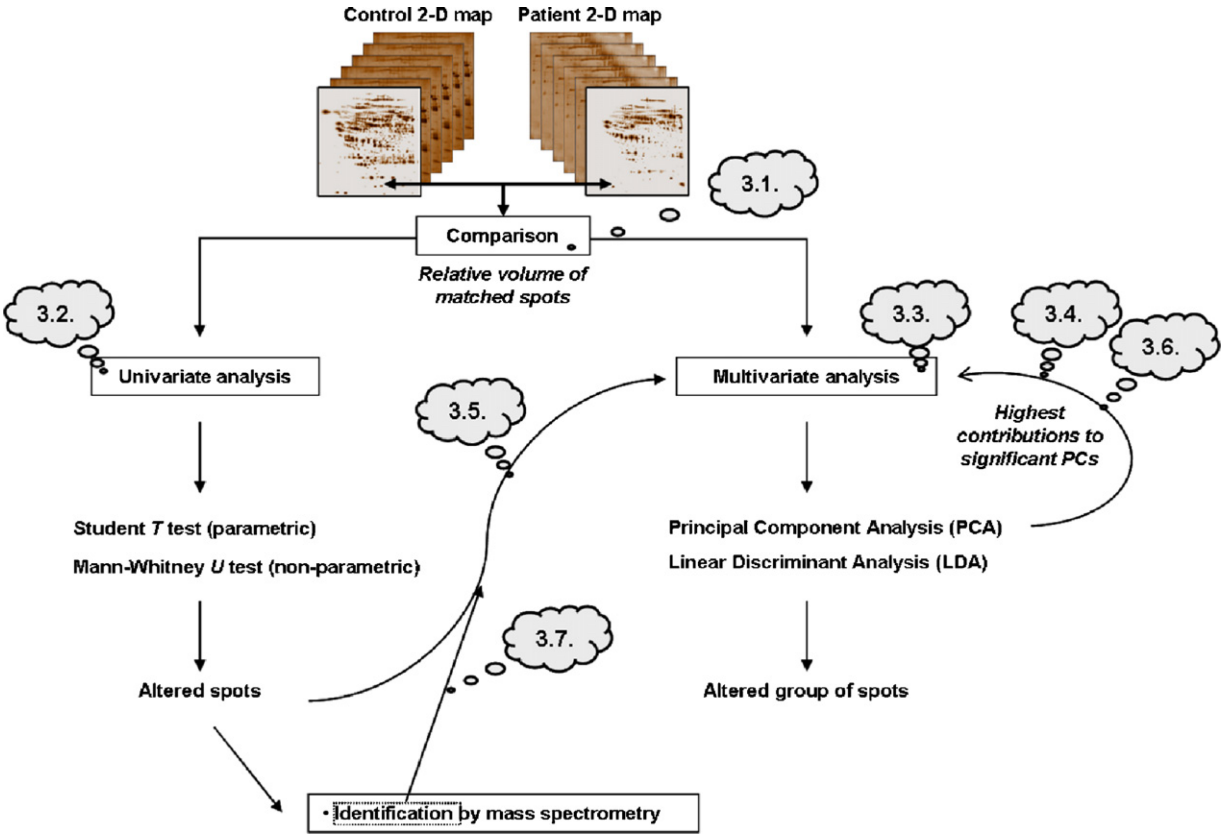
Control 2-D map
Patient 2-D map
3.1.
Comparison
Relative volume of matched spots
3.2.
3.3.
3.4.
3.6.
Univariate analysis
Multivariate analysis
3.5.
Highest contributions to significant PCs
Student T test (parametric)
Mann-Whitney U test (non-parametric)
Principal Component Analysis (PCA)
Linear Discriminant Analysis (LDA)
3.7.
Altered spots
Altered group of spots
Identification by mass spectrometry

Fig. 1. Workflow of the analyses performed with the data obtained after a 2-DE experiment. Bubbles show the section of 'Results' where the approach is described.

with the PDQuest software. First, maps obtained from the control group were compared in a first-level matchset and a standard image was generated with all the spots shared by the five control samples. Likewise, maps from the patient group were compared and used to build a standard patient map. After removing background and noise signals (mainly streaks and areas of overlapping), we could detect 870 common spots in control samples whereas samples from patients showed a lower average, 675 shared spots, due to higher noise. After the denominated high-level matchset, consisting in pair alignment between control and patient standard maps, 363 spots were matched through the whole set of samples. In order to establish a semi-quantitative comparison, the intensity of each spot in a gel was normalised in relation to the total density detected in the gel, obtaining the relative volume, which is the most representative measure of the spot quantity considering the gel stain. Relative volumes from the 363 matched spots in the 10 gels were exported to the SPSS programme for statistical comparison.

3.2. Univariate analysis of the complete dataset

Following the traditional univariate approach employed with 2-D data [14], differences in spot relative volumes between control and patient samples were assessed with the non-parametric Mann–Whitney *U*-test. Considering a 95%-confidence level, we found 28 spots that were significantly altered (Fig. 2). All these spots were regarded as potential biomarkers and therefore they were excised and submitted to MS identification. Eventually, 13 out of the 28 spots were associated with significant hits (see identities in Table 1 and MS data in supplementary Table S1), remaining a 54% of the altered spots unidentified due to insufficient amount of peptides or low signal intensity during MS, and non-significant hits. Among the spots identified we found clusterin, haptoglobin, and β-2-glycoprotein I, which had also been detected in a previous study [16].

Fig. 2. Virtual standard map for the comparison of serum *N*-glycoproteins from healthy individuals and CRC patients. The spots found significantly altered by univariate analysis are shown in boxes if up-regulated or circled if down-regulated in CRC patients.

3.3. Multivariate analysis of the complete dataset

After the conventional univariate analysis, we considered of interest to test the role of the whole protein pattern, since a number of proteins that would be contributing to the pathological state could not have a significant alteration by themselves, but taken together could help more effectively than single proteins in the detection of the disease [3]. Multivariate data handling has been suggested before to have important utility in proteomic experiments, since they are usually characterised by a large number of variables (as spots in 2-D maps or mass-to-charge peaks in MS) and a lower number of observations (samples), and these types of datasets are most efficiently analysed by multivariate methods [5]. As other authors have stated, applying univariate tests to proteomic datasets increases the likelihood of false-positive results and does not permit the detection of trends [5,20,21]. Thus, we submitted a complex dataset, comprehending the relative volumes of the 363 valid spots matched through all the gels, to two multivariate analyses.

First, we applied a data reduction approach by means of the factorial analysis called PCA. Out of the potential 363 components extracted, the first nine PCs accounted for a 100% of the biological variability contained in the data with a low individual contribution (Table 2). When these PCs were tested for differences by the Mann–Whitney *U*-test, we found that only PC2 was significant with >99% confidence ($P = 0.006$). As it can be seen in Fig. 3A, PC2 allowed the effective separation of the samples in their original groups. This application of PCA had been successfully tested before by other authors as Kovarova et al. [7], Roblick et al. [8] or Verhoeckx et al. [9].

On the other hand, the second approach employed was the LDA. When the whole dataset was tested, we found a discriminant function for the control and patient groups that accounted for the total (100%) variability of the data, setting a punctuation for each sample that classified it in the original group with a confidence level higher than 99% ($P = 0.008$). As an example, the classification table shown in the output of the test is reproduced as Table 3. Therefore, in our work the 2-D protein pattern formed by the 363 spots matched through all the gels could be employed to classify the individuals by LDA, behaving as a diagnostic tool similar to that proposed by Karp et al. [5].

However, from Table 2 it is obvious that employing the 363 variables does not yield the best classification due to an excess of information in the explanation of the total variability (nine PCs for 100% variance) and a low individual contribution (i.e. percentage of variance explained) of each PC. Therefore, though the whole dataset allows a separation of the groups and could be used for diagnostic purposes, it seems reasonable to look for a simplified set of more relevant proteins to further improve the separation of the samples in order to find a panel of disease markers.

Table 1
Spots significantly altered in patients with CRC regarding healthy subjects

Spot	Controls relative volume (mean ± SD)	Patients relative volume (mean ± SD)	P value	Fold change	pI	M_r (kDa)	Identity	Accession no.
1	0.0539 ± 0.039	0.0080 ± 0.011	0.036	−6.78	4.9	38.7	Clusterin (CLU)	P10909
2	0.1027 ± 0.090	0.0150 ± 0.033	0.036	−6.85	4.8	62.0	–	–
3	0.2156 ± 0.126	0.0166 ± 0.037	0.016	−13.02	5.0	29.3	–	–
4	0.0353 ± 0.028	0.0025 ± 0.006	0.032	−13.91	5.0	37.4	Clusterin (CLU)	P10909
5	0.1385 ± 0.310	0.7294 ± 0.373	0.028	+5.27	5.0	46.0	Haptoglobin (HPT)	P00738
6	0.0256 ± 0.016	0.0043 ± 0.007	0.028	−5.92	5.1	38.1	Clusterin (CLU)	P10909
7	0.0288 ± 0.016	0.0001 ± 0.000	0.009	−205.86	5.1	36.1	Clusterin (CLU)	P10909
8	0.0257 ± 0.022	0.0049 ± 0.007	0.028	−5.24	5.1	73.2	–	–
9	0.0260 ± 0.015	0.0043 ± 0.008	0.028	−6.05	5.1	93.8	–	–
10	0.0408 ± 0.042	0.0013 ± 0.003	0.036	−30.42	5.1	94.9	–	–
11	0.0277 ± 0.024	0.0025 ± 0.005	0.021	−11.24	5.1	93.9	–	–
12	0.3245 ± 0.420	1.0915 ± 0.661	0.047	+3.36	5.3	43.5	Haptoglobin (HPT)	P00738
13	0.0742 ± 0.048	0.0226 ± 0.031	0.036	−3.28	5.2	71.1	–	–
14	0.0249 ± 0.025	0.0018 ± 0.004	0.036	−13.81	5.2	92.8	–	–
15	0.0545 ± 0.048	0.0042 ± 0.009	0.028	−12.91	5.3	93.0	–	–
16	0.0089 ± 0.011	0.0369 ± 0.024	0.044	+4.16	5.3	82.3	–	–
17	0.3214 ± 0.213	0.0223 ± 0.050	0.016	−14.44	5.4	16.8	Haptoglobin (HPT)	P00738
18	0.0296 ± 0.018	0.0740 ± 0.026	0.047	+2.50	5.5	27.2	Immunoglobulin, light chain (IGLC)	P99007
19	0.4289 ± 0.178	0.1179 ± 0.162	0.028	−3.64	5.7	17.1	Haptoglobin (HPT)	P00738
20	0.0693 ± 0.036	0.0139 ± 0.020	0.028	−4.99	5.6	54.9	Complement factor I (CFAI)	P05156
21	0.0837 ± 0.046	0.0014 ± 0.003	0.016	−58.15	6.2	17.1	Haptoglobin (HPT)	P00738
22	0.0285 ± 0.025	0.0677 ± 0.025	0.047	+2.37	6.1	58.0	–	–
23	0.0802 ± 0.053	0.1617 ± 0.034	0.028	+2.02	6.3	56.9	Immunoglobulin, heavy chain gamma (IGHG)	P99006
24	0.0781 ± 0.167	0.3302 ± 0.085	0.047	+4.23	6.3	56.9	–	–
25	0.1897 ± 0.059	0.0103 ± 0.014	0.009	−18.45	6.4	59.5	β-2-Glycoprotein I	P02749
26	0.0278 ± 0.021	0.0037 ± 0.008	0.028	−7.44	6.7	63.9	–	–
27	0.0152 ± 0.034	0.0674 ± 0.043	0.028	+4.43	6.5	87.2	–	–
28	0.0483 ± 0.090	0.1942 ± 0.119	0.028	+4.02	6.5	115.7	–	–

For each spot, it is shown the average relative volume, statistical significance, fold variation, experimental parameters, and identity found by MS.
SD: standard deviation; fold change is given in positive for relative volumes increased in CRC patients and in negative for decreased volumes; pI: experimental isoelectric point; M_r: experimental molecular mass; accession no. is given for the Swiss-Prot database.

3.4. Selection of the spots with highest contributions to the discrimination of different samples

When the aim of a 2-D study is to find a protein pattern that would discriminate different samples, it is desirable that the number of spots that should be localised and matched through gels is minimal. Therefore, in order to reduce the number of variables (spot relative volumes) that could be used as a differential pattern between healthy individuals and CRC patients, we looked for the spots with the highest loadings (contributions or correlations) to the significant factor found before (i.e. PC2).

Table 2
Principal components (PCs) calculated from the 363 spots matched between control and patient samples

Component	Eigenvalues			Significance ($P < 0.05$)
	Total	% of variance	Cumulative %	
PC1	118.096	32.533	32.533	0.889
PC2	53.628	14.773	47.307	0.006[a]
PC3	49.229	13.562	60.868	0.370
PC4	39.000	10.744	71.612	0.955
PC5	31.291	8.620	80.232	0.452
PC6	28.931	7.970	88.202	0.718
PC7	19.424	5.351	93.553	0.278
PC8	13.631	3.755	97.308	0.712
PC9	9.770	2.692	100.000	0.829

[a] Significant by the Mann–Whitney U-test.

Studying the factor matrix generated by the software during the analysis, we selected in the first place those spots with a correlation (either positive or negative) above 0.6 with PC2. These data formed a subset containing 45 variables that was reanalysed by PCA. Thus, we found seven PCs of which PC1, PC2 and PC3 explained 80% of the variance. PC1 was significantly different between groups ($P = 0.009$), accounted for a 46.5% of the variability and allowed the graphical discrimination of control and patient samples (Fig. 3B). When only the spots with correlations higher than 0.7 were selected, 16 variables were reanalysed; in this case, we found two relevant PCs that explained 85% of the variance contained in the data. In this set, PC1 was significantly different between groups ($P = 0.009$), explained a 59% of the total variability and permitted the graphical separation of the groups (Fig. 3C).

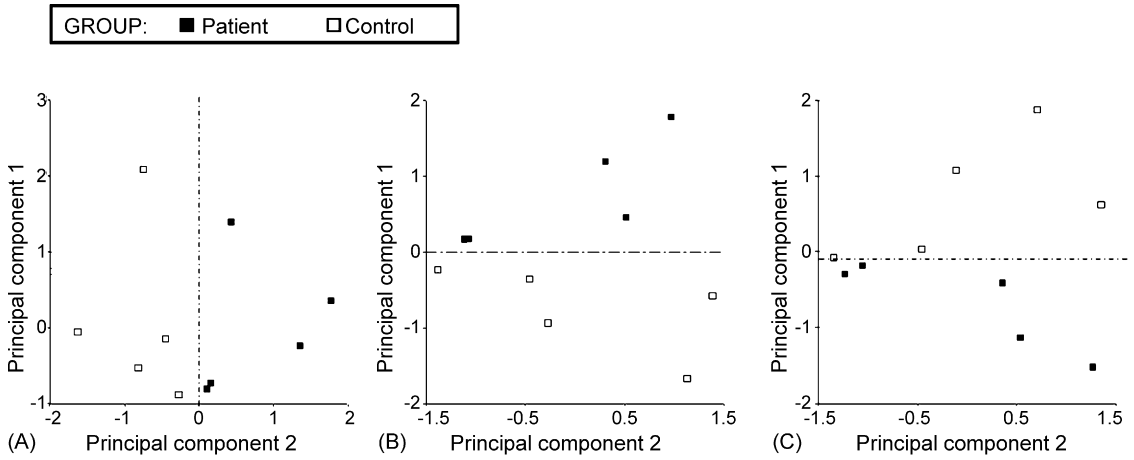

Fig. 3. Representation of the main principal components found after PCA of: (A) the 363 spots matched through all samples; (B) the spots with contributions above 0.6 in the 363-spot analysis; (C) the spots with contributions above 0.7 in the 363-spot analysis. Notice that the separation of the groups is given by PC2 in A, while in B and C the samples are clustered in relation to PC1.

Therefore, we were able to look for a reduced set of variables (spots) that could aid in the separation of the samples in their true group of origin, and improved the explanation of the variability contained in those groups. To corroborate if the protein pattern defined in 2-D maps by either set of spots could be considered as a tool to classify a blind sample after 2-DE processing, we performed LDA on both datasets (the 45 spots with loadings over 0.6 and the 16 spots with loadings over 0.7 on the PC2 of the 363-spot PCA). This analysis showed 100% correct classification of the original cases for both sets. However, after leave-one-out cross-validation we found a 30% and 50% of mismatches for the 45-spot and the 16-spot datasets, respectively. Therefore both clusters of spots were disregarded as clinically useful since results depended on the samples analysed. These results confirm the necessity of combining different statistical methodologies to assess the findings after data evaluation and the advantages of performing cross-validation tests.

Despite the negative results obtained in this particular case, this approach should be taken into consideration when planning to classify unknown samples on the basis of their proteome, since reducing the number of spots during the matching process reduces the occurrence of operator-derived errors.

Table 3
Classification results obtained after linear discriminant analysis of the 363 spots matched through all the samples

	Group	Predicted group membership		Total
		Control	Patient	
Original and cross-validation				
Count	Controls	5	0	5
	Patients	0	5	5
%	Controls	100.0	0.0	100.0
	Patients	0.0	100.0	100.0

The same results were found after cross-validation.

3.5. Multivariate analysis of the significantly altered spots found by univariate statistics

In the previous examples, multivariate data analysis was oriented to the assessment of the utility of the 2-D protein pattern as a tool to discriminate different types of samples regarding a disease condition.

Another approach that should be considered is the application of multivariate analyses on 2-D data to select the precise spots that would be the better markers before entering validation trials. As shown in a previous section, univariate methods allowed the detection of 28 spots that were significantly altered in the samples when individually tested. However, multivariate techniques allow clustering the differences given by each of those spots in PCs, and therefore would theoretically aid in the selection of the spots which could perform better in the clinical separation of the samples. Furthermore, it should be noticed that using univariate methods a number of proteins could appear altered due to chance. For instance, considering the 363 spots matched and setting a 95%-significance level, up to 18 of the 28 spots detected could not be truly altered.

When the 28 spots detected by univariate analysis were explored by PCA, the total variability of these spots was explained by nine PCs, the first three showing about a 78% of cumulative variance. This percentage shows how the separation achieved with the 28 spots was better than that obtained with the 363 spots (see Table 2, PC1 to PC3 accounted for 61% of cumulative variance), although similar to the separation given by the two subsets of proteins analysed above (see Section 3.4, 80% and 85% of variance explained by PC1 to PC3, respectively). Similarly to the PCA of those subsets, PC1 was significantly different ($P = 0.008$) between the control and patient groups. However, it could be graphically assessed (Fig. 4A and B) that the separation of the samples was neater in this analysis, reflecting that those spots included in the first PCs were altogether more relevant to the discrimination of the disease state. Noticeably, the distribution of the scores for the patient group seemed to display a high

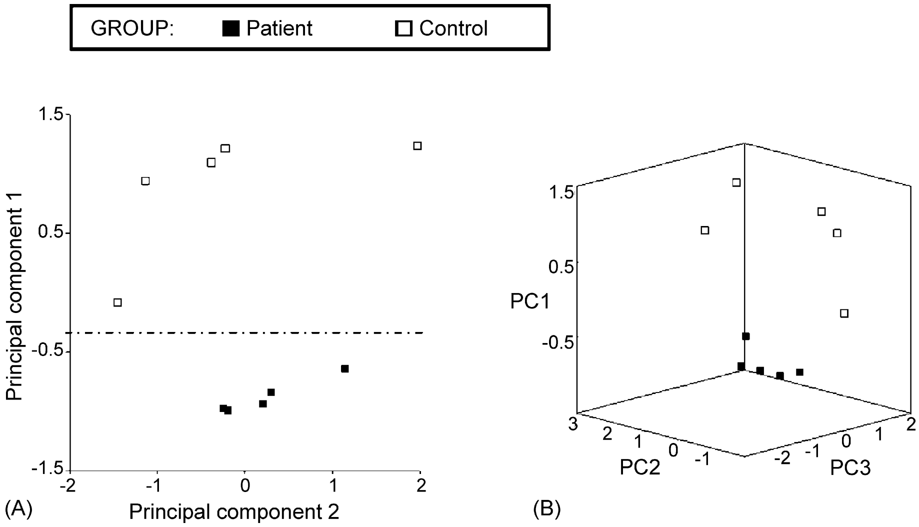

Fig. 4. Representation of the main principal components found after PCA of the 28 spots significantly altered by univariate statistics. (A) 2-D plot of PC1 and PC2; (B) 3-D plot of PC1, PC2 and PC3.

degree of homogeneity among tumour samples, as it was shown before by other authors in CRC and other cancers [8,22].

On the other hand, this set of variables was also challenged by LDA, finding that the group of 28 spots selected by univariate methods was able to correctly discriminate 100% of the samples tested, and this was corroborated through cross-validation. Therefore, the 28 spots together would present 100% specificity and 100% sensitivity if analysed from 2-D gels.

3.6. Searching for the best markers candidates among the altered spots

In the search for protein alterations with clinical utility in the diagnosis of a certain disease, the implementation of a panel of biomarkers requires a previous validation with specific assays including a high number of control and case samples, samples of related diseases, etc. [6]. Therefore, assaying all those samples for a large number of proteins (any of the successful sets tested in this study) could imply an elevated cost and extended time.

Therefore, we followed the same strategy described before to reduce the number of spots included in the search for a discriminant 2-D protein pattern, trying to select a subset of proteins that could yield a better separation of the samples. Hence, we sought for spots with contributions above 0.6, 0.7, 0.8 and 0.9 on the first factor (PC1) of the 28-spot PCA. Applying a new PCA on those sets of spots, the separation and explanation of the variance could not be improved (data not shown) with regard to the results obtained with the 28 spots.

When the subsets chosen on the basis of the loadings on PC1 were tested by LDA, we found that they consistently misclassified at least one donor sample, although the patients were correctly discriminated by the subsets of spots with loadings above 0.6 and 0.7. Eventually they were regarded as non-useful for clinical management, due to a low specificity of the test notwithstanding a high sensitivity for the CRC condition.

3.7. Searching for the best markers candidates among the identified spots

When planning how to reduce the set of proteins for validation tests, it is reasonable to try to select candidates among the already identified spots. Hence, we applied multivariate analyses on the 13 spots identified by MS after the classical 2-D univariate analysis.

PCA determined that PC1 to PC3 obtained from these 13 spots accounted for 87% of the total variance (almost 10% more than the same PCs for the 28 significant spots) and allowed a graphical separation of the samples as good as in that test (Fig. 5A and B). Furthermore, an advantage of working with known proteins is that the researcher is able to remove from the analysis those proteins that have been related to common processes not specifically related to the disease studied. In this work, eliminating the spots identified as haptoglobins (Swiss-Prot accession number P00738) and immunoglobulin chains (heavy chain: P99006; light chain: P99007), widely known to be altered during acute phase and inflammation, we achieved a 96% cumulative variance and a similar graphical separation of the groups (Fig. 5C and D).

LDA showed that the 13 spots could classify correctly 100% of the original cases, although cross-validation showed a 40% failure-rate. Surprisingly, the spots identified as clusterin (P10909), complement factor I (P05156) and β-2-glycoprotein I (P02749) (that is, the 13 spots minus haptoglobins and immunoglobulins) correctly classified 100% of the original cases and 90% of the samples during cross-validation. Furthermore, when the variables were introduced in a step-by-step mode to allow the automatic refinement of the calculations, the set performed well (100% true assignments) both for the original cases and the cross-validated samples.

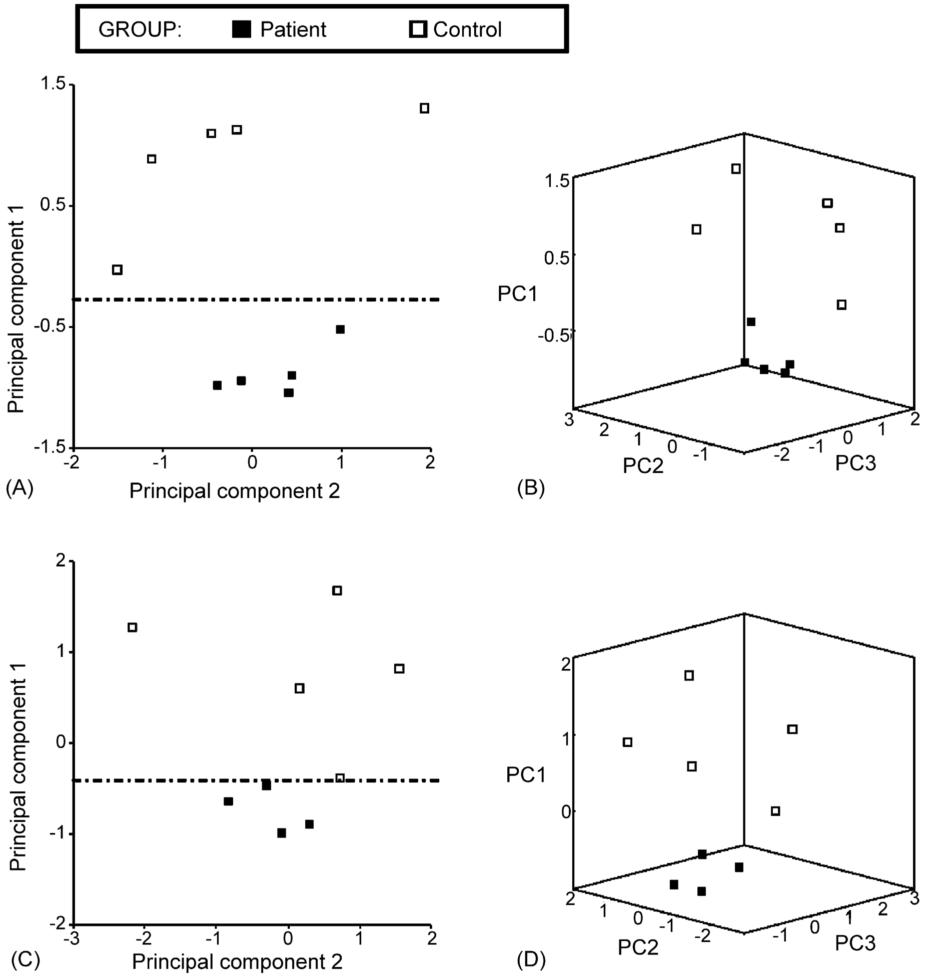

Fig. 5. 2-D and 3-D representation of the main principal components obtained after PCA on the 13 spots identified by MS (A, B), and PCA on the spots identified as clusterin, complement factor I and β-2-glycoprotein I (C, D). In C and D, two patient samples are superimposed.

Hence, those proteins (clusterin, complement factor I and β-2-glycoprotein I) could be regarded by themselves as a potential panel of biomarkers for CRC worth of further extensive testing. Interestingly, β-2-glycoprotein I has been suggested to bind to damaged cells [23], and both clusterin and complement factor I have been already related to carcinogenesis [24,25], supporting the interest of their validation.

4. Conclusions

To conclude, in this work we challenged a complex dataset obtained from a 2-D comparison of serum *N*-glycoproteins from healthy individuals and CRC patients, taking advantage both of univariate and multivariate approaches. Regarding the latter, we have found that both PCA and LDA are useful both to corroborate the utility of the whole proteome displayed by the 2-D maps as a tool to classify unknown samples and to verify the accuracy of the classification model. Moreover, these statistics offer parameters that allow the reduction of the number of spots considered, facilitating the management of the information contained in the maps. On the other hand, PCA and LDA were successfully applied to a group of spots previously detected as differentially expressed by way of univariate statistics, assessing the utility of the classification model provided by the whole group and by reduced subgroups of proteins. Finally, both methods were applied to the spots identified by MS, finding a combination of proteins with potential utility as markers for the disease studied. In any case, combination of univariate and multivariate techniques maximises the information obtained and improves the detection of true and relevant proteomic changes.

Up to date, the use of multivariate methods has been proposed by some researchers. However, those applications usually involved the knowledge of complex mathematical formulae or programming. The main advantage of the approach we employed is that both the PCA and the LDA can be simply applied through user-friendly statistical software as the SPSS used here, and therefore the researcher just needs to learn how to input the data and how to interpret the output of the programme. Notwithstanding, nowadays most of the 2-D dedicated software incorporates simple statistics as the Student *T* test, and thus it is possible that they could include this type of multivariate analysis, facilitating the extraction of useful trends from the complex datasets generated in 2-DE experiments.

Acknowledgements

The authors thank Dr. Emilio Rolán-Alvarez and Mr. Jorge Guerra-Varela for valuable technical comments and revision. We also thank 'Centro de Transfusión de Galicia' and 'Complejo Hospitalario Universitario de Vigo' for providing samples. This work was supported by grants from 'Xunta de Galicia' (PGIDIT02BTF30103PR) and 'Universidad de Vigo'.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jchromb.2006.09.021.

References

[1] A.S. Schrohl, M.N. Holten-Andersen, H.A. Peters, M.P. Look, M.E. Meijer-van Gelder, J.G. Klijn, N. Brunner, J.A. Foekens, Mol. Cell. Proteomics 2 (2003) 378.

[2] R. Aebersold, L. Anderson, R. Caprioli, B. Druker, L. Hartwell, R. Smith, J. Proteome Res. 4 (2005) 1104, doi:10.1021/pr050027n.

[3] J. LaBaer, J. Proteome Res. 4 (2005) 1053, doi:10.1021/pr0501259.

[4] F. Vitzthum, F. Behrens, N.L. Anderson, J.H. Shaw, J. Proteome Res. 4 (2005) 1086, doi:10.1021/pr050080b.

[5] N.A. Karp, J.L. Griffin, K.S. Lilley, Proteomics 5 (2005) 81, doi:10.1002/pmic.200400881.

[6] M.E. de Noo, B.J.A. Mertens, A. Özalp, M.R. Bladergroen, M.P.J. van der Werff, C.J.H. van de Velde, A.M. Deelder, R.A.E.M. Tollenaar, Eur. J. Cancer 42 (2006) 1068, doi:10.1016/j.ejca.2005.12.023.

[7] H. Kovarova, M. Hajduch, G. Korinkova, P. Halada, S. Krupickova, A. Gouldsworhty, N. Zhelev, M. Strnad, Electrophoresis 21 (2000) 3757, doi:10.1002/1522-2683(200011)21:17<3757::AID-ELPS3757>3.0.CO;2-X.

[8] U.J. Roblick, D. Hirschberg, J.K. Habermann, C. Palmberg, S. Becker, S. Krüger, M. Gustafsson, H.-P. Bruch, B. Franzén, T. Ried, T. Bergaman, G. Auer, H. Jörnvall, Cell. Mol. Life Sci. 61 (2004) 1246, doi:10.1007/s00018-004-4049-4.

[9] K.C.M. Verhoeckx, M. Gaspari, S. Bijlsma, J. van der Greef, R.F. Witkamp, R.P. Doornbos, R.J.T. Rodenburg, J. Proteome Res. 4 (2005) 2015, doi:10.1021/pr050183u.

[10] X. Jia, K. Hollung, M. Therkildsen, K.I. Hildrum, E. Bendixen, Proteomics 6 (2006) 936, doi:10.1002/pmic.200500249.

[11] O. Langsrud, Stat. Comput. 15 (2005) 53.

[12] E. Marengo, R. Leardi, E. Robotti, P.G. Righetti, F. Antonucci, D. Cecconi, J. Proteome Res. 2 (2003) 351, doi:10.1021/pr030002t.

[13] A.M. Rodríguez-Piñeiro, A. Carvajal-Rodríguez, E. Rolán-Álvarez, F.J. Rodríguez-Berrocal, M. Martínez-Fernández, M. Páez de la Cadena, J. Proteome Res. 4 (2005) 1318, doi:10.1021/pr0500307.

[14] E. Marengo, E. Robotti, F. Antonucci, D. Cecconi, N. Campostrini, P.G. Righetti, Proteomics 5 (2005) 654, doi:10.1002/pmic.200401015.

[15] E. Marengo, E. Robotti, M. Bobba, M.C. Liparota, C. Rustichelli, A. Zamo, M. Chilosi, P.G. Righetti, Electrophoresis 27 (2006) 484, doi:10.1002/elps.200500323.

[16] A.M. Rodríguez-Piñeiro, D. Ayude, F.J. Rodríguez-Berrocal, M. Páez de la Cadena, J. Chromatogr. B 803 (2004) 337, doi:10.1016/j.jchromb.2004.01.019.

[17] L.S. Ramagli, L.V. Rodríguez, Electrophoresis 6 (1985) 559.

[18] J. Heukeshoven, R. Dernick, Electrophoresis 6 (1985) 103.

[19] I. Byrjalsen, P. Mose Larsen, S.J. Fey, L. Nilas, M.R. Larsen, C. Christiansen, Mol. Hum. Reprod. 5 (1999) 748.

[20] S. Dudoit, J.P. Shaffer, J.C. Boldrick, Stat. Sci. 18 (2003) 71.

[21] M.R. Wilkins, R.D. Appel, J.E. Van Eyk, M.C.M. Chung, A. Görg, M. Hecker, L.A. Huber, H. Langen, A.J. Link, Y.-K. Paik, S.D. Patterson, S.R. Pennington, T. Rabilloud, R.J. Simpson, W. Weiss, M.J. Dunn, Proteomics 6 (2006) 4, doi:10.1002/pmic.200500856.

[22] B. Franzén, S. Linder, A.A. Alaiya, E. Eriksson, K. Uruy, T. Hirano, K. Okuzawa, G. Auer, Br. J. Cancer 74 (1996) 1632.

[23] B. Bouma, P.G. de Groot, J.M. van den Elsen, R.B. Ravelli, A. Schouten, M.J. Simmelink, R.H. Derksen, J. Kroon, P. Gros, EMBO J. 18 (1999) 5166, doi:10.1093/emboj/18.19.5166.

[24] X. Chen, R.B. Halberg, W.M. Ehrhardt, J. Torrealba, W.F. Dove, Proc. Natl. Acad. Sci. U.S.A. 100 (2003) 9530, doi:10.1073/pnas.1233633100.

[25] J.O. Minta, M. Fung, B. Paramaswara, Biochim. Biophys. Acta 1442 (1998) 286.