Chemometrics of differentially expressed proteins from colorectal cancer patients

Lay-Chin Yeoh, Saravanan Dharmaraj, Boon-Hui Gooi, Manjit Singh, Lay-Harn Gam

Lay-Chin Yeoh, Lay-Harn Gam, School of Pharmaceutical Sciences, Universiti Sains Malaysia, Penang, 11800, Malaysia
Saravanan Dharmaraj, Centre for Drug Research, Universiti Sains Malaysia, Penang, 11800, Malaysia
Boon-Hui Gooi, Manjit Singh, Department of Surgery, Penang General Hospital, Penang, 10990, Malaysia

Author contributions: Gam LH conceived the design of the study and edited the manuscript; Yeoh LC carried out the experimental work and manuscript writing; Dharmaraj S carried out the statistical analyses; Gooi BH and Singh M provided the colorectal cancer specimens and patient information.

Supported by Research Universiti Grant, Grant No. 1001/PFAR MASI/815007

Correspondence to: Lay-Harn Gam, PhD, School of Pharmaceutical Sciences, Universiti Sains Malaysia, Penang, 11800, Malaysia. layharn@usm.my
Telephone: +60-4-6533888 Fax: +60-4-6570017
Received: August 13, 2010 Revised: September 18, 2010 Accepted: September 25, 2010 Published online: April 28, 2011

Abstract

AIM: To evaluate the usefulness of differentially expressed proteins from colorectal cancer (CRC) tissues for differentiating cancer and normal tissues.

METHODS: A Proteomic approach was used to identify the differentially expressed proteins between CRC and normal tissues. The proteins were extracted using Tris buffer and thiourea lysis buffer (TLB) for extraction of aqueous soluble and membrane-associated proteins, respectively. Chemometrics, namely principal component analysis (PCA) and linear discriminant analysis (LDA), were used to assess the usefulness of these proteins for identifying the cancerous state of tissues.

RESULTS: Differentially expressed proteins identified were 37 aqueous soluble proteins in Tris extracts and 24 membrane-associated proteins in TLB extracts. Based on the protein spots intensity on 2D-gel images, PCA by applying an eigenvalue > 1 was successfully used to reduce the number of principal components (PCs) into 12 and seven PCs for Tris and TLB extracts, respectively, and subsequently six PCs, respectively from both the extracts were used for LDA. The LDA classification for Tris extract showed 82.7% of original samples were correctly classified, whereas 82.7% were correctly classified for the cross-validated samples. The LDA for TLB extract showed that 78.8% of original samples and 71.2% of the cross-validated samples were correctly classified.

CONCLUSION: The classification of CRC tissues by PCA and LDA provided a promising distinction between normal and cancer types. These methods can possibly be used for identification of potential biomarkers among the differentially expressed proteins identified.

© 2011 Baishideng. All rights reserved.

Key words: Colorectal cancer; Proteomics; Marker protein; Principal component analysis; Linear discriminant analysis

Peer reviewer: Ki-Baik Hahm, MD, PhD, Professor, Gachon Graduate School of Medicine, Department of Gastroenterology, Lee Gil Ya Cancer and Diabetes Institute, Lab of Translational Medicine, 7-45 Songdo-dong, Yeonsu-gu, Incheon, 406-840, South Korea

Yeoh LC, Dharmaraj S, Gooi BH, Singh M, Gam LH. Chemometrics of differentially expressed proteins from colorectal cancer patients. World J Gastroenterol 2011; 17(16): 2096-2103 Available from: URL: http://www.wjgnet.com/1007-9327/full/v17/i16/2096.htm DOI: http://dx.doi.org/10.3748/wjg.v17.i16.2096

INTRODUCTION

Proteomic research has made great achievements in biomarker discovery, especially when incorporated with high-
throughput analytical tools and technology, for example 2D-PAGE and LC-MS/MS[1]. Two-dimensional gel electrophoresis is a fundamental tool for protein analysis to detect alterations in protein expression between control and disease states of cells, which can lead to the discovery of various biomarkers that contribute to pathogenesis or carcinogenesis[2]. Biomarkers can be used to discriminate variables for subsequent classification of normal and diseased groups[3]. The complexity of variables generated by mass spectra, microarray and immunohistochemistry often requires advanced statistical techniques or chemometrics to evaluate their clinical value.

Multivariate analyses including the dimension reduction method known as principal component analysis (PCA), and classification methods such as linear discriminant analysis (LDA) are often employed in proteomic studies. PCA reduces the number of variables for further data analysis and interpretation while identifying the variables that retain most of the data variance[4]. A principal component (PC) is defined as a new variable to explain the maximum amount of variance in the original data and corresponds to a linear combination of the original variables. PCs are presented orthogonally to each other, which provides a more effective representation of the data than the original variables[5]. LDA is a multivariate technique to classify observations into groups or categories. LDA forms new variables from the original data and identifies the variables that provide the best discrimination between the groups[6].

Djidja et al[6] have used a novel approach that combines matrix-assisted laser desorption ionization-mobility separation-mass spectrometry (MALDI-IMS-MS) and PCA-discriminant analysis (PCA-DA) to generate tumor classification models based on pancreatic cancer protein patterns. Furthermore, Kamath et al[7] have used PCA-based k-nearest neighbor analysis to classify normal and cancerous autofluorescence spectra of colonic mucosal tissues. Zwijelen et al[8] have investigated the use of Fourier transform infrared microscopy for colon cancer diagnosis. Their model uses PCA to define spectral changes among normal and cancerous human biopsied colon tissues. Ragazzi et al[9] have reported the use of multivariate techniques on plasma proteins to diagnose colorectal cancer (CRC). The plasma protein profile generated by MALDI-MS is analyzed by PCA and LDA to discriminate ionic species from normal subjects and CRC patients.

In this study, we carried out the comparison of 2-D images of cancerous and normal colorectal tissues. The differentially expressed proteins from Tris and thiourea lysis buffer (TLB) extractions were respectively tested on a PCA-LDA model to find out the possibility of using protein expression to classify the disease and non-disease tissues of CRC.

MATERIALS AND METHODS

Tissue specimen collection

Matching pairs of normal colonic mucosa and cancerous colonic tissue (located 10 cm from each other) from 26 CRC patients were collected after surgery at the Penang General Hospital, Penang, Malaysia. The study was approved by the Human Ethical Committee of Universiti Sains Malaysia. Informed written consent was received from all patients before the study was conducted. Prior to surgery, the patients did not receive preoperative neoadjuvant chemotherapy and radiotherapy. The tissues were confirmed as cancerous and normal, respectively, by the hospital's pathologist. The cancerous tissues were classified using the TNM system. Surgically removed samples were stored at -80°C until use.

Protein analysis

The method of protein analysis was as described in Yeoh et al[10]. Frozen tissue (250 mg) was rinsed in distilled water to remove cell debris and excess blood. The tissues were homogenized in ice-cold Tris buffer (0.5 g tissue/mL buffer) [40 mmol/L, Tris and 1 X Protease Inhibitor Cocktail (Sigma, St Louis, MO, USA)] and centrifuged at 12000 rpm for 15 min at 18°C. The supernatant was recovered and labeled as Tris extract. The pellet was subjected to further extraction using TLB (1 g tissue/1 mL buffer) [8 mol/L urea, 2 mol/L thiourea, 4% (w/v) CHAPS, 0.4% (w/v) carrier ampholytes and 50 mmol/L dithiothreitol] and centrifuged at 12000 rpm for 15 min at 18°C. The supernatant was recovered and labeled as TLB extract. The extracts were subjected to 2D gel separation on 11 cm ReadyStrip™ IGP strip (linear pH 4-7, Bio-Rad, USA) followed by separation on 10% (w/v) PAGE at a constant voltage of 200 V. The gels were stained with Coomassie Blue. The images obtained were analyzed by PDQuest version 7.3 (Bio-Rad). Comparison of the protein expression levels was carried out between cancerous and normal tissues. Differentially expressed proteins were defined as proteins with a spot intensity that was 1.5-fold higher or lower in cancerous tissues when compared to that in the corresponding normal tissues. A differentially expressed protein was defined as upregulated when it was found at greater intensity in cancerous tissue than in the corresponding normal tissue. The downregulated proteins were detected at greater intensity in normal tissues than in the corresponding CRC cancerous tissues.

Protein identification

The differentially expressed proteins were excised from the gel and subjected to in-gel digestion using trypsin and the tryptic peptides were analyzed by LC/MS/MS using an electrospray ionization ion trap mass analyzer (Agilent Technologies, Santa Clara, CA, USA). The MS/MS data were subjected to the MASCOT protein database search engine for protein identification. The identities of a few proteins (dependent on the availability of antibodies) were further confirmed using western blotting.

Statistical analysis

The differential expression of the proteins was tested by the paired Student's t test that is included in PDQuest, to determine their statistical significance (P < 0.05). For
PCA and LDA, the protein spot intensities were exported out from PDQuest and imported into SPSS version 15.0 (Chicago, IL, USA) to perform multivariate analyses. Protein spot intensities were used as variables.

RESULTS

The tissues specimens from each patient were collected in pairs of cancerous and normal tissues. Table 1 shows the details of the tissues used in the analysis. The tissues were subjected to a sequential extraction method to extract aqueous soluble proteins and membrane-associated proteins in two different fractions using Tris and TLB, respectively. Tables 2 and 3 show the 37 and 24 differentially expressed proteins identified in Tris and TLB extracts, respectively. The average fold change indicates the degree of differentiation in expression levels of the protein in cancerous tissues compared to normal tissues in all the patients tested, where a positive sign indicates a greater expression level in cancerous tissues, whereas a negative sign indicates a greater expression level in normal tissues. The MOWSE score refers to the score values given by the MASCOT search. Tables 4 and 5 show the mean intensity of spots and SD, and percentage coefficient of variation (%CV) of spot intensity of differentially expressed proteins in all patients for Tris and TLB extracts, respectively. An example of the differentially expressed protein, as represented by different intensities of protein spots between normal and cancerous tissues for glutathione S-transferase P (GST-P), is shown in Figure 1; the bar chart was plotted according to the intensity of the respective protein spots. GST-P was detected as upregulated in cancerous tissues.

Data analysis

The significance of the expression levels of the differentially expressed proteins in both Tris and TLB extracts was analyzed by Student’s t test. After univariate analysis was performed, the normalized intensities of 37 differentially expressed protein spots in Tris extracts were subjected to PCA. The PCA reduced the original data to 12 PCs based on an eigenvalue of > 1, and these 12 PCs contributed 76.43% of the total data variance of the Tris extract data. Figure 2 shows the 3D PC plot with the x-y- and z-axes representing the first, second and third PC number. The variables that had the highest loadings were those that contributed most to the differentiation of the disease state. Figure 3 shows the scree plot of Tris extracts. Six PCs were chosen and these components contributed 53.97% of the total variance of the Tris extract.

Table 1  Clinicopathological features of 26 colorectal cancer patients involved in study

| Patient No. | Age (yr) | Race | Sex  | pTNM  | Stage | Degree of differentiation | Tumor location |
|-------------|----------|------|------|-------|-------|--------------------------|---------------|
| 1           | 62       | Malay| Male | pT3N1Mx | III   | B                        | Sigmoid colon |
| 2           | 79       | Malay| Male | pT2N0M0 | I     | MD                       | Descending colon |
| 3           | 74       | Malay| Male | pT3N0M0 | II A  | MD                       | Ascending colon |
| 4           | -        | Malay| Male | pT3N2Mx | III C | MD                       | Rectum |
| 5           | 37       | Malay| Male | pT3N0M0 | III A  | MD                       | Transverse colon |
| 6           | 58       | Malay| Female | pT4N2Mx | III C | MD                       | Recto-sigmoid |
| 7           | 59       | Malay| Female | pT4N2Mx | III C | MD                       | Ileoceleal |
| 8           | 69       | Malay| Male | pT3N0Mx | III A  | MD                       | Sigmoid colon |
| 9           | 63       | Malay| Female | pT3N0Mx | III A  | MD                       | Recto-sigmoid |
| 10          | 84       | Chinese| Female | pT4N0Mx | II B   | MD                       | Rectum |
| 11          | 58       | Chinese| Male | pT3N0Mx | III A  | MD                       | Recto-sigmoid |

MD: Moderately differentiated adenocarcinoma.

Figure 1  Comparison of protein spot intensity between normal and colorectal cancer tissues for glutathione S-transferase P.

Yeoh LC et al. Chemometrics of colorectal cancer biomarkers
| Spot No. | Protein name | Swissprot No. | MOWSE score | MW (Da) | pl | Sequence coverage (%) | GRAVY | Average fold change |
|----------|--------------|---------------|-------------|---------|----|----------------------|-------|---------------------|
| 1 | Proteasome subunit β type 6 | P28072 | 134 | 25573 | 4.80 | 16 | 0.034 | -2.967 |
| 2 | 14-3-3 protein ζ | P63104 | 336 | 35567 | 6.97 | 40 | -0.744 | 11.659 |
| 3 | Tropomyosin α-3C-like protein | A6NL28 | 127 | 27407 | 4.71 | 31 | -0.992 | 44.183 |
| 4 | Rho GDP-dissociation inhibitor 1 | P52565 | 167 | 23120 | 5.03 | 29 | -0.700 | -7.607 |
| 5 | 14-3-3 protein ζ | P63104 | 282 | 27919 | 4.73 | 16 | -0.621 | 4.127 |
| 6 | Tubulin β-2C chain | P60371 | 524 | 50304 | 4.83 | 40 | -0.362 | -52.184 |
| 7 | Cathepsin B | P02758 | 74 | 22981 | 5.20 | 18 | -0.433 | 33.149 |
| 8 | Rho GDP-dissociation inhibitor 2 | P52566 | 48 | 22920 | 5.10 | 18 | -0.799 | -10.625 |
| 9 | SEC13 homolog | P55735 | 78 | 36400 | 5.22 | 9 | -0.372 | 6.873 |
| 10 | Hsc70-interacting protein | P56502 | 164 | 28464 | 8.92 | 21 | -0.653 | 20.959 |
| 11 | Apolipoprotein A-1 | P02647 | 143 | 30777 | 5.56 | 26 | -0.717 | -4.478 |
| 12 | Proteasome subunit α type 3 | P25788 | 201 | 15958 | 6.82 | 41 | 0.008 | 4.249 |
| 13 | Actin, cytoplasmic 2 | P63261 | 105 | 26169 | 5.65 | 14 | -0.156 | 28.601 |
| 14 | 60 kDa heat shock protein | P10809 | 151 | 61304 | 5.70 | 14 | -0.074 | 131.219 |
| 15 | Peroxiredoxin-2 | P32119 | 283 | 21995 | 5.67 | 42 | -0.210 | 1.250 |
| 16 | Guanine nucleotide binding protein subunit β 2 | P62879 | 112 | 37954 | 5.60 | 11 | -0.183 | -14.442 |
| 17 | F-actin-capping protein subunit β | P47756 | 259 | 34187 | 6.02 | 37 | -0.574 | 33.554 |
| 18 | GST-P | P09211 | 730 | 23442 | 5.44 | 60 | -0.131 | 4.834 |
| 19 | Haptoglobin-related protein | P00739 | 49 | 39529 | 6.42 | 3 | -0.308 | 56.209 |
| 20 | Cathepsin Z | Q9LBR2 | 100 | 27780 | 5.48 | 15 | -0.545 | -60.766 |
| 21 | F-actin-capping protein subunit β | P47756 | 245 | 21280 | 7.93 | 34 | -0.540 | 13.278 |
| 22 | Actin-related protein 3 | K11581 | 148 | 47704 | 5.61 | 27 | -0.271 | 15.881 |
| 23 | Abhydrolase domain-containing protein 14B | Q961U4 | 200 | 25429 | 6.82 | 26 | -0.023 | 0.765 |
| 24 | Nucleoside diphosphate kinase A | P15531 | 87 | 19873 | 5.42 | 36 | -0.075 | 73.120 |
| 25 | L-lactate dehydrogenase B chain | P07195 | 228 | 36928 | 5.71 | 14 | 0.056 | 3.513 |
| 26 | Fibrinogen β chain | P02675 | 151 | 56624 | 8.54 | 22 | -0.758 | 41.329 |
| 27 | Leucokine elastase inhibitor | P60740 | 170 | 42857 | 5.90 | 11 | -0.249 | 10.458 |
| 28 | PDI A3 | P5101 | 674 | 57202 | 5.98 | 35 | -0.506 | 7.579 |
| 29 | Gelsolin | P63996 | 238 | 80103 | 5.90 | 20 | -0.415 | -11.917 |
| 30 | Heat shock 27 kDa protein | P04792 | 256 | 22840 | 5.98 | 47 | -0.567 | -1.508 |
| 31 | DI-1 protein | Q94997 | 122 | 20879 | 6.33 | 54 | 0.004 | 4.981 |
| 32 | Fibrinogen β chain | P02675 | 75 | 56624 | 8.54 | 22 | -0.758 | -72.722 |
| 33 | Selenium-binding protein 1 | Q13228 | 502 | 52971 | 5.93 | 21 | -0.254 | -26.544 |
| 34 | Selenium-binding protein 1 | Q13228 | 592 | 52938 | 5.93 | 30 | -0.254 | 27.403 |
| 35 | Selenium-binding protein 1 | Q13228 | 979 | 52938 | 5.93 | 37 | -0.254 | -1.887 |
| 36 | Lysokine A4 hydrolase | P09860 | 215 | 69792 | 5.80 | 22 | -0.259 | 29.759 |
| 37 | Proteasome subunit β type 6 | P69001 | 71 | 29888 | 8.57 | 39 | -0.247 | 0.768 |

Table 2  List of proteins found in 2D gel of Tris extracts

1Protein accession number at SwissProt at http://www.expasy.org/uniprot; 2MOWSE score from MASCOT protein database search at http://www.matrixscience.com, where score > 41 is statistically significant (P < 0.05); 3Average ratio of the spot intensity in normal mucosa over tumor tissue (negative variation or decrease) or tumor tissue over normal tissue (positive variation or increase). GST-P: Glutathione S-transferase P; PDE: Protein disulfide isomerase.

data. Table 6 shows the LDA results for Tris extract proteins, where 22 out of 26 original normal tissues, and 21 out of 26 original cancer tissues were correctly classified. In cross-validated samples, 22 out of 26 normal tissues and 21 out of 26 cancer tissues were correctly classified.

Figure 2 Principal component plot of Tris proteins.

Figure 3 Scree plot showing principal components and their eigenvalues in Tris extracts.
Both original and cross-validation samples had an average 82.7% correct classification.

Figure 4 shows the 3D view of the PCs plot for the TLB extract. PCA reduced the original data of the TLB extract to seven PCs based on an eigenvalue one of > 1, and the seven PCs accounted for 72.46% of the total data variance. The 3D view indicates that tissues can be grouped according to CRC disease state. Figure 5 shows the scree plot of the TLB extracts. Six PCs were chosen based on the slope of scree plot, which contributed 67.61% of the total data variance of TLB extracts. Table 7 shows the LDA results of TLB extracts, where 22 out of 26 original normal tissues, and 19 out of 26 original cancerous tissues were correctly classified. In cross-validated samples, 21 out of 26 normal tissues and 16 out of 26 cancerous tissues were correctly classified. The average percentages of correct classification for original and cross-validation samples were 78.8% and 71.2%, respectively.

**DISCUSSION**

The expression levels of the differentially expressed protein between colorectal cancerous and normal tissues were
analyzed using PCA based on a multivariate analysis approach, to assess their usefulness in classifying colorectal tissues as cancerous or normal. The differentially expressed proteins identified showed good consistency in their expression levels in cancerous and normal tissues. The proteins were extracted in two fractions according to their polarities. In the PCA-LDA model, the selected proteins from the first few PCs were able to discriminate colorectal tissues with and without CRC.

A scree plot was derived by plotting the eigenvalues against the PC number. The shape of the plot was used to evaluate the number of PCs to be retained. In general, the point at which the scree plot straightens out indicates the number of PCs to be extracted \(^{[11]}\). Cross-validation is a method to estimate the accuracy of a predicted classification model if performed using new future data sets (samples); this is because a classification model is considered incomplete until the prediction error is estimated \(^{[12]}\). One method of cross validation is leave-one-out cross-validation, where one sample from the data set of N samples is removed, the discriminant rule is recalibrated, and a classification model is built based on the remaining N - 1 data. The one sample that is left out is classified in this model and the process repeated N times \(^{[13]}\).

Yeoh LC et al. Chemometrics of colorectal cancer biomarkers
PCA and LDA results from Tris extract indicated that six out of 37 proteins were reliable to determine the tissues with CRC. The proteins comprised five upregulated proteins, namely GST-P, tropomyosin α-3C-like protein, F-actin capping protein subunit β, selenium binding protein 1 and DJ-1 protein, and one downregulated protein, namely, proteasome subunit β type 6. DJ-1 protein and GST-P contributed the most to the first PC based on the weight of their loadings. This was followed by the tropomyosin α-3C-like protein and proteasome subunit β type 6 that contributed to the second PC, while F-actin capping protein subunit β and selenium binding protein 1 contributed to the third PC. The initial PCA reduced the original data and therefore enabled LDA to be carried out because LDA is sensitive to the number of variables. In LDA, the six PCs chosen were shown to be capable of predicting whether the tissues were with or without CRC. Two-way validation by using original and cross-validation analyses was applied to validate the state of the tissues, where the cancerous and normal tissues were classified correctly at 82.7% for both original and cross-validation samples.

Two proteins that contributed most to PC1 in Tris extract were DJ-1 and GST. DJ-1 is a putative oncoprotein that is able to transform cells with H-Ras[14]. Overexpression of DJ-1 activates protein kinase B, which subsequently increases cell survival. Furthermore, increased DJ-1 expression also activates Nrf2 (nuclear factor erythroid 2-related factor), which in turn increases expression of antioxidant enzymes that confer a survival advantage to tumor cells[15]. Upregulation of DJ-1 protein in esophageal squamous cell carcinoma is correlated with lymph node metastasis[15]. Although there is no reported role of DJ-1 in CRC, its upregulation in CRC is undeniable, and we have shown that its expression can be used to discriminate between CRC cancerous and normal tissues.

GST catalyzes the conjugation of reduced glutathione to electrophiles[16]. GST functions to remove peroxides from endogenous compounds such as lipids and DNA[17]. Overexpression of GST-P1 in CRC may be involved in cell proliferation, differentiation and apoptosis[18]. GST-P1 is overexpressed in liver cancer cells[19].

In TLB extract, six of the 24 differentially expressed proteins identified were found to be useful in discriminating CRC cancerous from normal tissues. The leave-one-out cross-validation obtained 71.2% correct classification of normal and cancerous tissues. The value for original grouped samples was higher with 78.8% correct classification.

Two proteins that contributed most to PC1 in TLB extracts were PDI and GC1q-R. PDI catalyzes the formation and breakage of disulfide bonds between two cysteine residues[20]. PDI regulates cell transformation and intracellular and extracellular redox activities via its redox activity[21]. PDI regulates STAT3 signaling and proliferation, which is thought to induce malignancy[22]. PDI is upregulated in CRC cell lines and its upregulation is correlated with cancer cell differentiation[21,22].

GC1q-R is a cell surface glycoprotein, which binds to the globular heads of Clq molecules[23]. Clq molecules bind to a variety of cells such as B cells, monocytes, macrophages, endothelial and smooth muscle cells[23]. Clq elicits responses such as phagocytosis in monocytes and activation of tumor cytotoxicity of macrophages[24-26]. GC1q-R is overexpressed in colon cancer cells and may be involved in tumor metastasis. However, PDI and GC1q-R were downregulated when using average fold change to determine their expression levels.

Proteins are the expression components that regulate cell activity. Differential expression of proteins is expected upon transformation of normal cells to cancerous cells. These differentially expressed proteins are useful in diagnosis and prognosis of the disease. In the present study, the specimens used in the analysis comprised tissues from female and male patients who were diagnosed with various stages, grades and locations of CRC. Regardless of the sex of the patients and pathological specification of the tissues, we showed that the differentially expressed protein identified from 2D protein profiles of cancerous and normal tissues could be used to separate and classify normal and cancerous tissues by combining PCA and LDA. The data reduction technique of PCA was sufficient to provide a classification of tissues according to CRC disease state. These statistical models simplify the data management through the reduced dimensionality of protein spots from the 2D gel images. Therefore, multivariate analysis of differentially expressed proteins identified from cancerous and normal tissues may be used as a tool for diagnosis and prognosis of CRC disease state.

ACKNOWLEDGMENTS

We want to express our appreciation to Universiti Sains Malaysia for provided the grant under RU funding to conduct this project. We would also like to thank National Institute of Pharmaceutical and Neutracetical for their generosity in allowing us to conduct the LC/MS/MS experiments.

COMMENTS

Background

Colorectal cancer (CRC) is one of the leading causes of death worldwide. Diff-
ferentially expressed proteins between cancerous and normal colonic tissues were identified using 2D gel separation followed by LC/MS/MS analysis. The protein spot intensities of the 2D gel images were analyzed using principal component analysis (PCA) and linear discriminant analysis (LDA) for their possible use in classification of disease state.

**Research frontiers**

Multivariate analyses, including the dimension reduction method known as PCA and classification methods such as LDA, are used in cancer proteomic studies to identify the protein variables that provide the best discrimination between the cancerous and normal tissues.

**Innovations and breakthroughs**

The authors used sequential protein extraction to extract aqueous soluble and membrane-associated proteins from colorectal tissues. Differentially expressed proteins were analyzed using a combination of PCA and LDA to determine their usability in differentiating normal and cancerous colonic tissues. Using this method, the authors successfully classified the tissues according to their respective types. DJ-1 protein and glutathione S-transferase P1 of the aqueous soluble proteins, protein disulfide isomerase and complement component 1 Q subcomponent-binding protein of the membrane-associated proteins gave the best classification of the tissues.

**Applications**

The identified biomarkers may be used for the diagnosis and prognosis of CRC.

**Terminology**

Chemometrics is defined as the information aspects of complex biological and chemical systems. Chemometrics utilize mathematical, statistical or formal logic-based methods to extract chemical information, which in this case, is for biomarker discovery.

**Peer review**

This study investigated the use of PCA and LDA of differential protein expression between normal and cancerous tissues for classification of disease state. The method gave good classification of cancerous and normal colonic tissues.

**REFERENCES**

1. Cowan ML, Vera J. Proteomics; advances in biomarker discovery. Expert Rev Proteomics 2008; 5: 21-23
2. Rodriguez-Piñeiro AM, Rodriguez-Berrocal FJ, Páez de la Caden M. Improvements in the search for potential biomarkers by proteomics: application of principal component and discriminant analyses for two-dimensional maps evaluation. J Chromatogr B Analyt Technol Biomed Sci Life Sci 2007; 849: 251-260
3. Hilario M, Kalousis A. Approaches to dimensionality reduction in proteomic biomarker studies. Brief Bioinform 2008; 9: 102-118
4. Karson MJ. Multivariate statistical methods: An introduction. Iowa: Iowa State University Press, 1982: 159, 191
5. Giri NC. Multivariate statistical analysis. New York: Marcel Dekker, 1996: 293-294
6. Djidja MC, Claude E, Snel MF, Francese S, Scriven P, Caro-lan V, Clerch MR. Novel molecular tumour classification using MALDI-mass spectrometry imaging of tissue microarray. Anal Bioanal Chem 2010; 397: 597-601
7. Kamath SD, Mahato KK. Principal component analysis (PCA)-based k-nearest neighbor (k-NN) analysis of colonic mucosal tissue fluorescence spectra. Photomed Laser Surg 2009; 27: 659-668
8. Zwölly A, Mordechai S, Sinielniak I, Salaman A, Bogomolny E, Argov S. Advanced statistical techniques applied to comprehensive FTIR spectra on human colonic tissues. Med Phys 2010; 37: 1047-1055
9. Ragazzi E, Pucciarelli S, Saraglia R, Molin L, Agostini M, Lise M, Traldi P, Nitti D. Multivariate analysis approach to the plasma protein profile of patients with advanced colorectal cancer. J Mass Spectrom 2006; 41: 1546-1553
10. Yeoh LC, Loh CK, Gooi BH, Singh M, Gam LH. Hydrophobic protein in colorectal cancer in relation to tumor stages and grades. World J Gastroenterol 2010; 16: 2754-2763
11. McCarlage K, Cushman S, Staffor S. Ordination: Principal component analysis. In: McCarlage, editor. Multivariate statistics for wildlife and ecology research. New York: Springer-Verlag, 2000: 41-42
12. Dziuza DM. Biomarker discovery and classification. In: Dziuza, editor. Data mining for genomics and proteomics: Analysis of gene and protein expression data. Hoboken: John Wiley and Sons, 2010: 110-112
13. Nagakubo D, Taira T, Kitaura H, Ikeda M, Tamai K, Iuchi-Ariga SM, Ariga H. DJ-1, a novel oncogene which transforms mouse NIH3T3 cells in cooperation with ras. Biochem Biophys Res Commun 1997; 231: 509-513
14. Clements CM, McNally RS, Cotti BJ, Mak TW, Ting JP. DJ-1, a cancer- and Parkinson's disease-associated protein, stabilizes the antioxidant transcriptional master regulator Nrf2. Proc Natl Acad Sci USA 2006; 103: 15091-15096
15. Yuen HF, Chan YP, Law S, Srivastava G, El-Tanani M, Mak TW, Chan K. DJ-1 could predict worse prognosis in esophageal squamous cell carcinoma. Cancer Epidemiol Biomarkers Prev 2008; 17: 3993-3992
16. Mannervik B, Danielson UH. Glutathione transferases--structure and catalytic activity. CRC Crit Rev Biochem 1988; 23: 283-337
17. Park HJ, Pak SY, Jiang H, Hong KH, Song J. Functional studies of cysteine residues in human glutathione S-transferase P1-1 by site-directed mutagenesis. Bull Korean Chem Soc 2001; 22: 77-83
18. Lo HW, Antoun GR, Ali-Osman F. The human glutathione S-transferase P1 protein is phosphorylated and its metabolic function enhanced by the Ser/Thr protein kinases, cAMP-depen-dent protein kinase and protein kinase C, in glioblastoma cells. Cancer Res 2004; 64: 9131-9138
19. Tsuchida S, Sato K. Glutathione transferases and cancer. Crit Rev Biochem Mol Biol 1992; 27: 357-384
20. Wilkinson B, Gilbert HF. Protein disulfide isomerase. Biochin Biophys Acta 2004; 1699: 35-44
21. Hirano N, Shibasaki F, Sakai R, Tanaka T, Nishida J, Yazaki Y, Takenawa T, Hirai H. Molecular cloning of the human glucose-regulated protein ERp57/GRP58, a thiol-dependent reductase. Identification of its secretory form and inducible expression by the oncogenic transformation. Eur J Biochem 1995; 234: 336-342
22. Coe H, Jung J, Groenendyk J, Pints D, Michalak M. ERp57 modulates STAT3 signaling from the lumen of the endoplasmic reticulum. J Biol Chem 2010; 285: 6725-6738
23. Katayama M, Nakano H, Ishizuki A, Wu W, Oshima R, Sakurai J, Nishikawa H, Yamaguchi S, Otsubo T. Protein pattern difference in the colon cancer cell lines examined by two-dimensional differential in-gel electrophoresis and mass spectrometry. Surg Today 2006; 36: 1085-1093
24. Stierum R, Gaspari M, Dommels Y, Ouatas T, Phuk H, Jespersen S, Vogels J, Verhoeckx K, Groten J, van Ommen B. Proteome analysis reveals novel proteins associated with pti-tition and differentiation of the colorectal cancer cell line Caco-2. Biochim Biophys Acta 2003; 1659: 73-91
25. Ghebrehiwet B, Lim BL, Peerschke EI, Willis AC, Reid KB. Isolation, cDNA cloning, and overexpression of a 33-kD cell surface glycoprotein that binds to the globular “heads” of C1q. J Exp Med 1994; 179: 1809-1821
26. Ghebrehiwet B. Functions associated with the C1q receptor. Behring Inst Mitt 1989; 204-215
27. Bobak DA, Frank MM, Tenner AJ. C1q acts synergistically with phorbol dibutyrate to activate CR1-mediated phagocytosis by human mononuclear phagocytes. J Immunol 1988; 13: 2001-2007
28. Leu RW, Zhou AQ, Shannon B, Herriott MJ. Inhibitors of C1q biosynthesis suppress activation of murine macrophages for both antibody-independent and antibody-dependent tumor cytotoxicity. J Immunol 1990; 144: 2281-2286

S-Editor Sun H L-Editor Kerr C E-Editor Zheng XM