SPANDIDOS PUBLICATIONS

Plasma proteomic profiling: Search for lung cancer diagnostic and early detection markers

IQBAL UNNISA ALI[1,2], ZHEN XIAO[1], WINFRED MALONE[1], MYLINH SMITH[5], THOMAS P. CONRADS[3], TIMOTHY D. VEENSTRA[3], PETER GREENWALD[1], BRIAN T. LUKE[4] and JERRY W. McLARTY[5]

[1]Division of Cancer Prevention, National Cancer Institute, Bethesda, MD 20892; [2]Laboratory of Cancer Prevention, Center for Cancer Research, [3]Laboratory of Proteomics and Analytical Technologies and [4]Advanced Biomedical Computing Center, SAIC-Frederick Inc., National Cancer Institute at Frederick, Frederick, MD 21702; [5]Feist-Weiller Cancer Center, Louisiana State University Health Sciences Center, Shreveport, LA, USA

Received August 18, 2005; Accepted October 24, 2005

Abstract. Environmental and occupational exposure to asbestos is among the established risk factors for lung cancer, the leading cause of cancer-related deaths in the United States. This link between exposure to asbestos and the excessive death rate from lung cancer was evident in a study of former workers of an asbestos pipe insulation manufacturing plant in Tyler, TX. We performed comparative proteomic profiling of plasma samples that were collected from nine patients within 12 months before death and their age-, race- and exposure-matched disease-free controls on strong anion exchange chips using surface-enhanced laser desorption ionization time-of-flight mass spectrometry. A distance-dependent K-nearest neighbor (KNN) classification algorithm identified spectral features of *m/z* values 7558.9 and 15103.0 that were able to distinguish lung cancer patients from disease-free individuals with high sensitivity and specificity. The high correlation between the intensities of these two peaks (r=0.987) strongly suggests that they are the doubly and singly charged ions of the same protein product. Examination of these proteomic markers in the plasma samples of subjects from >5 years before death from lung cancer suggested that they are related to the early development of lung cancer. Validation of these biomarkers would have significant implications for the early detection of lung cancer and better management of high-risk patients.

Introduction

Lung cancer is the leading cause of cancer deaths in the United States and worldwide (1). It is often diagnosed in advanced stages of the disease and has one of the lowest 5-year survival rates of <15% among all cancers (2). Based on clinical and histopathological features, lung cancer is comprised of a broad spectrum of tumors of two main categories, non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC), accounting for about 80% and 20% of the disease, respectively. NSCLC is further classified into squamous cell carcinoma, adenocarcinoma, and large-cell carcinoma, with adenocarcinoma being the most frequent form. Another rare form of lung cancer, mesothelioma (3), a highly malignant tumor of the pleura, is predominantly associated with exposure to asbestos (4,5). Irrespective of several distinct histopathologies, all lung cancers are often detected in late stages and have a high fatality rate. Among the major risk factors for lung cancer are tobacco smoke, radon, asbestos, and heavy metals such as arsenic, chromium, and nickel (6). Although approximately 10% of smokers are estimated to eventually develop lung cancer, the highest risk of lung cancer is attributed to tobacco smoking, accounting for almost 90% of lung cancer deaths.

Occupational exposure to asbestos is also an established and second most common risk factor for lung cancer, with the severity of disease varying with the extent and length of exposure and the size of inhaled fibers (7,8). Furthermore, concomitant asbestos and tobacco exposure have a multiplicative effect on lung cancer development, translating to as much as a 50-fold higher lung cancer risk in comparison with non-smoking individuals in the general population (8-10). This association between exposure to asbestos and excess death from lung cancer was documented in a study on former workers of a plant in Tyler, TX that used amosite to manufacture pipe insulation material (11). In this study, we performed surface-enhanced laser desorption ionization time-of-flight mass spectrometry (SELDI-TOF MS) (12,13) on the stored plasma samples of lung cancer patients and disease-free controls from the Tyler, TX cohort to identify lung cancer-specific biomarkers. The availability of serial samples from the same cancer patients, collected from an extended period of time prior to patients' death, allowed us to search for potential early detection markers for lung cancer.

Correspondence to: Dr Iqbal Unnisa Ali, Division of Cancer Prevention, National Cancer Institute, Bethesda, MD 20892, USA
E-mail: alii@mail.nih.gov

Key words: lung cancer, surface-enhanced laser desorption ionization time-of-flight mass spectrometry, plasma proteomic profiling, diagnostic/early detection markers

Table I. Patient characteristics and dates of sample collection.

Patient ID	Sample ID	Date of collection	Sample collection sequence	Race[a]	Age	Smoking status[b]	Date of death	Months survived after last sample collection
Case								
1	C1-1	9/26/1989	Last	W	57	C	5/26/1990	8
	C1-2	3/11/1988	Mid-point					
	C1-3	8/27/1986	First					
2	C2-1	5/29/1985	Last	B	66	C	8/25/1985	3
	C2-2	4/16/1985	First					
3	C3-1	8/31/1992	Last	W	70	C	8/10/1993	12
	C3-2	11/14/1989	Mid-point					
	C3-3	1/11/1986	First					
4	C4-1	1/30/1991	Last	W	71	C	12/12/1991	11
	C4-2	7/14/1988	Mid-point					
	C4-3	12/12/1985	First					
5	C5-1	3/3/1987	Last	W	74	C	3/12/1988	12
	C5-2	8/25/1986	Mid-point					
	C5-3	7/14/1986	First					
6	C6-1	2/6/1990	Last	W	77	C	5/20/1990	3
	C6-2	8/19/1987	Mid-point					
	C6-3	1/23/1986	First					
7	C7-1	3/11/1985	Last	W	81	C	8/14/1985	5
	C7-2	1/28/1985	First					
8	C8-1	8/12/1986	Last	W	88	C	5/28/1987	9
	C8-2	6/25/1986	First					
9	C9-1	7/24/1992	Last	W	73	F	2/14/1993	5
	C9-2	11/14/1989	Mid-point					
	C9-3	1/11/1986	First					
Control								
1	N1	6/10/1989	Last	W	57	C		
2	N2	6/12/1985	Last	B	63	C		
3	N3	6/1/1992	Last	W	70	C		
4	N4	4/29/1991	Last	W	69	C		
5	N5	7/15/1987	Last	W	74	C		
6	N6	5/9/1990	Last	W	75	C		
7	N7	9/5/1985	Last	W	82	C		
8	N8	8/25/1986	Last	W	85	C		
9	N9	7/9/1992	Last	W	73	Q		

[a]W, White; B, Black. [b]C, current smoker; F, former smoker.

Materials and methods

Study population and plasma samples. Former workers (n=1095) of an asbestos manufacturing plant, which operated in Tyler, TX from 1954 to 1972, were exposed to high levels of amosite dust. These workers were enrolled in a follow-up study in 1978 that included periodic collection of sputum and plasma samples (14). The lung cancer mortality in this cohort has been described (11). Plasma samples from 50 workers, who died of lung cancer, and 50 age-, race-, and exposure-

Sensitivity, specificity, and positive and negative values for the spectral features used in the DD-KNN classifiers.

m/z	No. of neighbors	Sensitivity	Specificity	PPV	NPV
3316.2	4	83.3	100.0	100.0	84.6
15148.5					
3503.9	5	91.7	91.7	91.7	91.7
15148.5					
3505.9	6	83.3	84.6	83.3	84.6
15103.0					
6787.3	4	91.7	92.3	91.7	92.3
7558.9					
36866.0					
6787.3	5	91.7	92.3	91.7	92.3
7558.9					
36866.0					
5408.3	6	91.7	100.0	100.0	92.9
6787.3					
7558.9					

PPV, positive predictive value; NPV, negative predictive value. The major discriminators in each model are bolded. Ten cancer and 2 control spectra have intensities of the 7558.9 peak below 2.95, while 2 cancer and 11 control spectra have intensities above this value. This peak alone therefore yields a sensitivity and PPV of 83.3% and a specificity and NPV of 84.6%; the other two peaks in the classifiers simply correct 2 or 3 of the errors. For the 15103.0 peak, 11 cancer and 3 control spectra have intensities below 3.10, while 1 cancer and 10 control spectra have intensities above this value, yielding a sensitivity of 91.7%, specificity of 76.9%, PPV of 78.6% and an NPV of 90.9%. Including the peak at 3505.9 simply corrects the prediction of one control spectrum and misclassifies an additional cancer spectrum. The 15148.5 intensities have comparable discriminating ability, but is a shoulder of the 15013.0 peak and therefore does not represent a biomarker.

matched cancer-free controls were used in this study. The samples were stored at -80˚C before analysis. Unfortunately, information on the date of diagnosis of lung cancer and histopathology of the disease were not available, except that 4 of the 50 cancers were mesotheliomas.

SELDI protein profiling. Strong anionic exchange (SAX) protein chips were used in combination with SELDI-TOF MS to generate plasma protein profiles. The chips were pre-equilibrated twice with binding buffer (20 mM Tris, pH 9.0) and agitated for 5 min each time. Plasma samples in 30 μl aliquots were thawed on ice, diluted with 30 μl 8 M urea, 1% CHAPS, pH 7.4, and mixed well with vortex for 10 min. The samples were further diluted with 20 mM Tris, pH 9.0 and applied in a randomized order on the pre-equilibrated chip arrays in duplicate at 100 μl/well. The arrays were incubated with agitation for 90 min. Subsequently, the protein chips were washed 3 times with binding buffer for 5 min each. After the final wash, chips were rinsed twice with deionized H_2O for 30 sec each, and air-dried.

Saturated sinapinic acid (Fluka, Milwaukee, WI) in 50% (v/v) acetonitrile, and 0.5% trifluoroacetic acid was applied to each spot twice, at 0.5 μl/time, and air-dried between applications. The chips were read using a PBS-II SELDI-TOF mass spectrometer (Ciphergen Biosystems, Inc.). Protein spectra were generated by averaging 104 laser shots collected on each spot with a laser intensity setting of 205, detector sensitivity of 8, high mass of 100000 Dalton, and optimized mass range from 1000 to 19000 Daltons. The spectra were calibrated using the All-in-1 protein molecular mass standard (Ciphergen Biosystems, Inc.). The raw spectral data, which contained about 35000 *m/z* values, and the corresponding peak intensities per spectrum were exported into Microsoft Excel for bioinformatics analysis.

Bioinformatics analysis. To remove any possible spectra generated by the energy-absorbing matrix, all intensities below 1500 *m/z* were removed. This filtering resulted in approximately 22300 *m/z* values per spectrum. For peak identification, each spectrum was then scaled to a constant total ion current, and the scaled spectra were summed to identify regions with sufficient intensity. The putative peaks were identified from these regions with the criteria that they had an intensity of at least 15% the average intensity in the summed spectrum, and not within 0.3% of *m/z* of any other peaks. A window of width of 0.3% of *m/z* was centered on each peak in each spectrum, and the maximum intensity within this window was used as the intensity for this peak. This analysis produced a set of 792 isolated peaks in each spectrum.

Since each sample was run in duplicate, the Euclidean distance was used across all 792 peak intensities to count the number of times the distance between duplicates was larger than a sample-to-sample distance (15). If this number was ≥2, the duplicate spectra were sufficiently different and kept separate; otherwise, they were averaged. If the nearest-neighbor distance or number of peaks with an intensity value within 5% of the overall minimum or maximum was more than two standard deviations above the mean value, the spectrum was considered an outlier and removed from the dataset.

Classification of the samples used a distance-dependent K-nearest neighbor (DD-KNN) algorithm with a Euclidean distance metric. For each analysis, intensities of 2 or 3 peaks were used to place the samples in 2- or 3-dimensional space, allowing the identification of 4 to 6 nearest-neighbors for each sample. The un-normalized probability of belonging to the same class as a given neighbor was proportional to the inverse of their distance. If the probability of being in the same class was <0.8, a probability of being undetermined was assigned as 0.1. As the probability of being in the same class increased to 1.0, the undetermined probability decreased to 0.0. After normalizing the probabilities across all neighbors, the cost of this set of peaks was taken as the sum of one minus the probability of a correct assignment across all samples. A total of six DD-KNN analysis runs (each with 2-3 peaks and 4-6 neighbors) were examined for putative biomarkers.

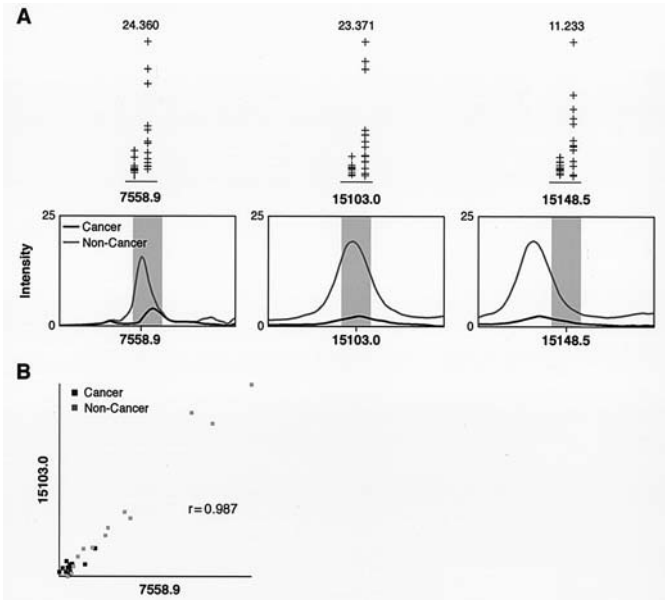

Figure 1. (A) Spectra at *m/z* 7558.9, 15103.0, and 15148.5 from lung cancer patients and non-cancer controls. Top panel, the intensities of all three spectral features, *m/z* 7558.9, 15103.0, and 15148.5, from all nine lung cancer patients (left column) and matched non-cancer controls (right column). The number at the top indicates the maximum intensities. Bottom panel, the representative examples of individual spectral features. (B) Scatter plot of the peak intensities at *m/z* 7558.9 and 15103.0.

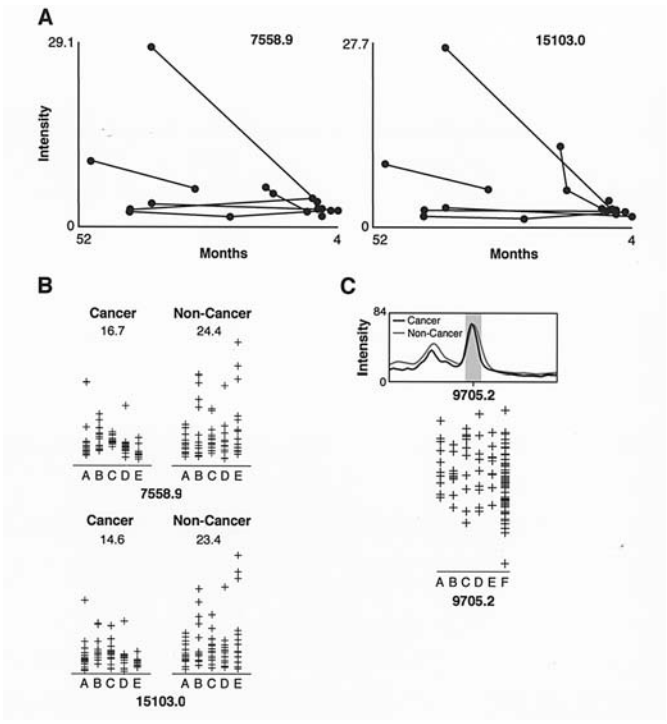

Figure 2. (A) Intensities of the peaks at *m/z* 7558.9 and 15103.0 for seven lung cancer patients over time. Intensities of both spectral features are from the first and midpoint blood draws, as shown in Table I, ranging between 4 and 52 months before death. Intensities of the markers are either low or decrease sharply during the time of progression to death. (B) Intensities of the peaks at *m/z* 7558.9 and 15103.0 in lung cancer patients and non-cancer controls over time. Cancer samples were collected at different time points before death. A, >5 years; B, between 3 and 5 years; C, between 2 and 3 years; D, between 1 and 2 years; E, within 1 year. The numbers at the top indicate the maximum intensities. (C) Intensities of the peak at *m/z* 9705.2 in lung cancer patients and non-cancer controls over time. Samples were collected at different time points before death. A, >5 years; B, between 3 and 5 years: C, between 2 and 3 years; D, between 1 and 2 years; E, within 1 year; F, matched non-cancer controls. Representative examples of the spectral region with *m/z* 9705.2 in a lung cancer patient and non-cancer control are also shown.

The best set of peaks was determined using a modified evolutionary programming algorithm (16,17). This population-based feature selection algorithm included an operator that ensured the uniqueness of peaks in the parent and offspring populations. This function reduced the probability that the algorithm prematurely converged on a suboptimal set of markers.

Results

Diagnostic markers for lung cancer. Our first set of analyses to identify lung cancer-specific markers by comparing proteomic profiles of plasma samples from 50 lung cancer patients and 50 controls did not yield meaningful results. This may have been because the last available plasma samples from lung cancer patients before death were from a period ranging between 3 months to >5 years. Therefore, the individuals whose plasma samples were collected several years before death may not have had discernable disease at the time of sample collection and could be technically classified as 'disease-free controls.' A vast majority of lung cancers are diagnosed in the late stage of disease and approximately 60% of these patients die within 1 year of diagnosis according to the available estimates from the American Cancer Society. Thus, to increase the probability of finding true lung cancer-specific proteomic markers, we focused on a spectra of nine patients, whose plasma samples were available within 3-12 months before death (Table I). Nine age-, race-, and exposure-matched disease-free individuals from the same cohort were used as controls. Since each sample was examined twice, the analysis started with 18 cancer and 18 control spectra. The duplicate spectra in three cancer cases were averaged to produce 15 spectra, and 3 of these were removed as outliers, yielding 12 spectra with 792 peaks. For the nine control samples, four duplicate spectra were averaged and one was identified as an outlier; producing a dataset of 13 spectra with 792 peak intensities.

The optimum classifier obtained in each of the six DD-KNN runs, along with its quality, is shown in Table II. Each classifier consists of a major distinguishing feature shown in bold, and one or two other features that reduce the misclassification. The intensities of each of the distinguishing features and the regions in cancerous and control spectra are shown in Fig. 1A. While all three features have low intensities in the cancer spectra and significantly higher intensities in many of the control spectra, only the features at *m/z* values of 7558.9 and 15103.0 correspond to peaks; the feature at 15148.5 was a shoulder of the 15103.0 peak. The strong correlation between the intensity in the markers 7558.9 and 15103.0 in each of the 25 spectra (r=0.987, Fig. 1B) suggests they are the doubly and singly charged ions of the same protein product.

Evaluation of m/z 7668.9 and 15103.0 markers for early detection of lung cancer. We attempted to investigate if the

s of *m/z* 7668.9 and 15103 markers could be used ble identification of lung cancer patients in this cohort several years before their death from lung cancer. We first analyzed the intensities of *m/z* 7558.9 and 15103.0 markers in the available plasma samples of 7 of the 9 cancer patients that were drawn at earlier time points. The time range of these samples varied between 3 and 52 months. Fig. 2A displays the intensities of the two markers plotted as a function of time; it is obvious that either the intensities of both markers were always low or decreased during the time until death. We then examined the intensities of the *m/z* 7558.9 and 15103.0 markers in the plasma samples of all 50 lung cancer patients and their matched controls. The cancer samples were divided into five groups according to the time of blood draw until death. The results displayed in Fig. 2B demonstrate that the intensities of both proteomic markers decreased from >5 years (group A) to <1 year (group E) before death. Some of the controls, however, were also found to have low intensities of both markers.

One concern was that the low intensities of m/z 7558.9 and 15103.0 markers in cancer patients could be time-related, especially in the two patients with a high level of these markers, and may not be an indication of impending lung cancer. To address this question, we analyzed another randomly selected plasma proteomic marker at *m/z* 9705.2. As evident in Fig. 2c, the intensity of this marker was neither time- nor disease-related. This suggests that the intensity changes in *m/z* 7558.9 and 15103.0 were primarily associated with disease.

Discussion

The Tyler, TX cohort was unique in several respects; namely, the workers of the pipe insulation material manufacturing plant were exposed to high levels of pure amosite for an extended period of time resulting in excess death from lung cancer (11,14). Although not prospectively designed, the availability of sequential plasma samples in this cohort, especially in some lung cancer patients, allowed meaningful molecular studies to search for diagnostic and potential early detection markers. When plasma samples of the last blood draw from patients within 12 months before death were subjected to SELDI-TOF MS, the DD-KNN algorithm identified two proteomic markers that could distinguish lung cancer cases from matched disease-free controls. According to the mass values and a high correlation coefficient, these two markers likely represent the doubly and singly charged ions of the same protein product. In separate runs that constructed eight classifiers using either average-linkage (ALC) or complete-linkage clustering (CLC) (2 or 3 features, 2-5 clusters), the peak at *m/z* 15103.0 was used in one of the ALC classifiers and 6 of the 8 CLC classifiers with other highly correlated features (data not shown). The robustness of both markers also transcended several important and much criticized technological challenges of SELDI-TOF MS (18,19), including protein chip and instrument-related concerns and also potential time-dependent changes in the plasma proteomic profiles (unpublished data).

Several lines of evidence suggest that either of the two biomarkers at *m/z* 7558.9 and 15103.0 can serve as a diagnostic marker for lung cancer and possibly monitor the disease before clinical diagnosis. First, the intensity of these highly correlated biomarkers (r=0.987) was either low from the start or decreased sharply during the time of progression to death in the nine lung cancer patients, whose last blood draws were within 12 months before death (Fig. 2A). Second, all 50 lung cancer patients analyzed in different time periods starting >5 years before death showed a trend toward decreased intensities of both markers with the lowest intensity within 12 months prior to death (Fig. 2B). Although the pattern of intensities of the two markers in 50 disease-free controls is quite different from that of lung cancer cases, it is obvious that many of the 50 cancer-free controls also had low intensities of both markers. Since these 50 individuals of the Tyler, TX cohort, used as age- and race-matched controls, were also exposed to amosite for approximately the same length of time, it is possible that at least some of them may become cases, and the low intensities of *m/z* 7558.9 and 15103.0 markers are an indication of the onset of lung cancer. Third, the low intensity of the markers is not the result of a direct interaction with amosite since the controls from the same cohort generally have high intensity of the markers. Finally, it is possible that the low intensity of *m/z* 7558.9 and 15103.0 markers in sequential samples of the nine patients is not a time- or disease-dependent decrease, but merely a fortuitous event. Although this possibility cannot be ruled out, the measurement of another randomly selected plasma proteomic marker in all 50 lung cancer patients (Fig. 2C) suggests that the decrease in the intensity of *m/z* 7558.9 and 15103.0 proteomic markers is an indication of lung cancer development.

In conclusion, despite some limitations of our study (unavailability of the date of clinical diagnosis of lung cancer and histological classification), we were able to identify two robust proteomic markers with low intensities, which were diagnostic for lung cancer. These markers, identified from lung cancer patients with advanced disease (within 12 months before death), also surprisingly had low intensities in the plasma samples of patients several years before their death from lung cancer. Given the fact that only 10-15% of patients with all stages of lung cancer have a 5-year survival (2), our data suggest that low intensities of *m/z* 7558.9 and 15103.0 proteomic markers can serve as indicators for lung carcinogenesis; individuals with low intensities of these markers should therefore be monitored closely for lung cancer development.

Our study represents a proof-of-principle approach and demonstrates the feasibility of identifying biomarkers for the early detection of lung cancer using high-throughput proteomic profiling of serial plasma samples from lung cancer patients. Identification of the protein product that generates two spectral features of *m/z* 7558.9 and 15103.0 would provide insight into the biological function(s) and its relevance for lung carcinogenesis. While we hope to further elucidate the identity of these two closely associated markers, it is also important to validate these proteomic biomarkers in a larger prospective clinical trial, with the outcome having significant implications for the early detection of lung cancer and better surveillance and management of high risk patients well before the appearance of clinical symptoms of cancer.

Acknowledgements

We are indebted to Dr Eva Szabo for her very helpful discussions and critical comments on the manuscript. This work was funded in whole or in part with federal funds from the U.S. National Cancer Institute, National Institutes of Health, under contract no. NO1-CO-12400. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does any mention of trade names, commercial products or organizations imply endorsement by the U.S. Government.

References

1. Jemal A, Murray T, Ward E, *et al*: Cancer statistics. CA Cancer J Clin 55: 10-30, 2005.
2. Fry WA, Phillips JL and Menck HR: Ten-year survey of lung cancer treatment and survival in hospitals in the United States: a national cancer database report. Cancer 86: 1867-1876, 1999.
3. Corson JM: Pathology of mesothelioma. Thorac Surg Clin 14: 447-460, 2004.
4. Hughes RS: Malignant pleural mesothelioma. Am J Med Sci 329: 29-44, 2005.
5. Godleski JJ: Role of asbestos in etiology of malignant pleural mesothelioma. Thorac Surg Clin 14: 479-487, 2004.
6. Williams MD and Sandler AB: The epidemiology of lung cancer. Cancer Treat Res 105: 31-52, 2001.
7. Tweedale G: Asbestos and its lethal legacy. Nat Rev Cancer 2: 311-315, 2002.
8. Nelson HH and Kelsey KT: The molecular epidemiology of asbestos and tobacco in lung cancer. Oncogene 21: 7284-7288, 2002.
9. Lee DH and Selikoff IJ: Historical background to the asbestos problem. Environ Res 18: 300-314, 1979.
10. Lee PN: Relation between exposure to asbestos and smoking jointly and the risk of lung cancer. Occup Environ Med 58: 145-153, 2001.
11. Levin JL, McLarty JW, Hurst, GA, Smith AN and Frank AL: Tyler asbestos workers: mortality experience in a cohort exposed to amosite. Occup Environ Med 55: 155-160, 1998.
12. Von Eggeling F, Junker K, Fiedle W, *et al*: Mass spectrometry meets chip technology: a new proteomic tool in cancer research? Electrophoresis 22: 2898-2902, 2001.
13. Yip TT and Lomas L: SELDI ProteinChip array in onco-proteomic research. Technol Cancer Res Treat 1: 273-280, 2002.
14. Hurst GA, Spivey CG, Matlage, WT, *et al*: The Tyler Asbestos Workers Program. I. A medical surveillance model and method. Arch Environ Health 34: 432-439, 1979.
15. Slotta DJ, Heath LS, Ramakrishnan N, Helm R and Potts M: Clustering mass spectrometry data using order statistics. Proteomics 3: 1687-1691, 2003.
16. Luke BT: An overview of genetic methods. In: Genetic Algorithms in Molecular Modeling. Devillers J (ed). Academic Press, London, pp35-66, 1996.
17. Luke BT: Genetic algorithms and beyond. In: Nature-Inspired Methods in Chemometrics: Genetic Algorithms and Neural Networks. Leardi R (ed). Elsevier, Amsterdam, pp3-54, 2003.
18. Diamandis EP: Point: proteomic patterns in biological fluids: do they represent the future of cancer diagnostics? Clin Chem 49: 1272-1275, 2003.
19. Diamandis EP: Analysis of serum proteomic patterns for early cancer diagnosis: drawing attention to potential problems. J Natl Cancer Inst 96: 353-356, 2004.