Artificial Neural Networks Analysis of Surface-Enhanced Laser Desorption/Ionization Mass Spectra of Serum Protein Pattern Distinguishes Colorectal Cancer from Healthy Population

Yi-ding Chen,1 Shu Zheng,2 Jie-kai Yu,2 and Xun Hu2

1Department of Oncology, Second Affiliated Hospital, and 2Cancer Institute, College of Medicine, Zhejiang University, HangZhou, Zhejiang, Peoples Republic of China

ABSTRACT

Purpose: The low specificity and sensitivity of the carcinoembryonic antigen test makes it not an ideal biomarker for the detection of colorectal cancer. We developed and evaluated a proteomic approach for the simultaneous detection and analysis of multiple proteins for distinguishing individuals with colorectal cancer from healthy individuals.

Experimental Design: We subjected serum samples (including 55 colorectal cancer patients and 92 age- and sex-matched healthy individuals) from 147 individuals, for analysis by surface-enhanced laser desorption/ionization (SELDI) mass spectrometry. Peaks were detected with Ciphergen SELDI software version 3.0. Using a multilayer artificial neural network with a back propagation algorithm, we developed a classifier for separating the colorectal cancer groups from the healthy groups.

Results: The artificial neural network classifier separated the colorectal cancer from the healthy samples, with a sensitivity of 91% and specificity of 93%. Four top-scored peaks, at m/z of 5,911, 8,930, 8,817, and 4,476, were finally selected as the potential “fingerprints” for detection of colorectal cancer.

Conclusions: The combination of SELDI-TOF mass spectrometry with the artificial neural networks in the analysis of serum protein yields significantly higher sensitivity and specificity values for the detection and diagnosis of colorectal cancer.

INTRODUCTION

Colorectal cancer is ranked the fifth most frequent cause in cancer-related deaths in China (1). The high mortality rate caused by this disease is, to a significant extent, associated with the poor diagnosis at its presymptomatic stages. When colorectal cancer is detected early, more than 90% of persons with the disease live at least 5 years beyond diagnosis. Unfortunately, only 37% of colorectal cancers are diagnosed before they have spread (2). It has been demonstrated that the prognosis is inversely correlated with the stages of this malignant disease at the time diagnosed, with the significantly higher 5-year survival rates for those diagnosed at the early stages Dukes A and B than at the later stage Dukes C and D. Conceivably, the accurate “picking-up” of colorectal cancer at its presymptomatic stage is critical for reducing the death rate.

Currently, the noninvasive detection techniques used for clinical and screening purposes include (a) immunochemical detection of carcinoembryonic antigen (CEA; ref. 3), a colorectal cancer-associated biomarker, which yields the detection sensitivity (17–80%) and specificity (34–91%); (b) fecal occult blood test (FOBT), with very low sensitivity and relatively high specificity, which inevitably causes many unnecessary endoscopic examination (4); and (c) detection of other colorectal cancer-associated markers such as K-RAS, TP53, and BAT26 in feces (5, 6), all of which are plagued by unsatisfactory diagnostic power. Obviously, all of these techniques lack adequate sensitivity and specificity. Therefore, it is essential that more accurate and sophisticated methods be developed for this purpose.

Surface-enhanced laser desorption/ionization time-of-flight mass spectrometry (SELDI-TOF MS) is one of the recently developed sophisticated technologies that, based on capturing proteins/peptides by chemically modified surfaces, is specifically powerful for analyzing the complex biological samples (7, 8). This technology, combined with bioinformatics, has been successfully used to analyze the complex serum proteins to explore the cancer-specific “fingerprints” or “patterns.” These patterns are of far-reaching significance in the detection of the early-stage cancers, because they have demonstrated extremely high specificity and sensitivity in segregating cancers from noncancers (9–11). Besides the merits stated above, this technology is also high throughput, as compared with two-dimensional gel electrophoresis, and is cost-effective, because it can substantially reduce the expenditures on unnecessary medical tomographic examination, commonly caused by poor diagnostic accuracy by cancer biomarker testing.

Artificial neural networks (ANNs) are computer-based algorithms that are modeled on the structure and behavior of neurons in the human brain and can be trained to recognize and categorize complex patterns. Pattern recognition is achieved by adjusting parameters of the ANN by a process of error minimi-
zation through learning from experience. They can be calibrated by using any type of input data, such as proteomics data generated by SELDI-TOF MS, and the output can be grouped into any given number of categories. The pattern recognition techniques have been applied to diverse areas including gene microarray (12) and MS (13).

We report here the extraction and validation of a colorectal cancer-specific pattern, by analyzing 55 colorectal cancer serum samples and 92 healthy individuals by using SELDI-TOF; the samples were subsequently stringently processed by ANNs. The pattern presents the specificity of 93%, sensitivity 91%, in separating colorectal cancer from healthy groups.

PATIENTS AND METHODS

Patients. A total of 147 serum samples were obtained from the serum bank of Zhejiang University Cancer Institute. The cancer group consisted of 55 serum samples from colorectal cancer patients at different clinical stages: Dukes A (n = 8), Dukes B (n = 22), Dukes C (n = 13), and Dukes D (n = 12). These samples were collected during the period of January 2001 through June 2001. All of the serum samples were obtained before clinical treatment. Diagnosis was pathologically confirmed. The median age of the colorectal cancer patients was 57 (range, 28–78), 60 male and 32 female. The control group included sera from 92 healthy people, with the median age of 56 (range, 28–78), 60 male and 32 female. The control samples was obtained from screening clinics that were open to the general public during March 2001. All of the samples were obtained in the morning before food intake and were stored at −80°C until use. The intervals between collection and analysis were 1 to 7 months for the patient specimens, and 4 months for control specimens.

Serum Carcinoembryonic Antigen Determination. Serum CEA levels were determined with a Wallac DELFIA CEA kit (time-resolved fluoroimmunoassay, Perkin-Elmer) with a cutoff value of 5 ng/mL.

Proteinchip Array Analysis. Ten μL of each serum sample and 90 μL of a solution containing 0.5% CHAPS in phosphate-buffered saline (pH 7.4) were added to each well of 96-well plate. The mixture was vortex-mixed at 4°C for 15 minutes, followed by the addition of 100 μL of Cibacron Blue 3GA (Sigma, Inc., St. Louis, MO; prepared balanced 0.5% CHAPS three times). Plates were placed on a platform shaker at 4°C for 60 minutes. The supernatant (40 μL) was then transferred onto the H4 chips so that each chip (8-spot format) held four cancer and four healthy samples to rule out the systematic error. The chips were held by a bioprocessor (Ciphergen Biosystems, Freemont, CA), a device that holds 12 chips and allows application of larger volumes of serum to each chip array. The samples were allowed to react with the surface of the H4 chip for 60 minutes at room temperature. The chips then were washed three times by gently shaking on a platform shaker at a speed of 700 rpm for 5 minutes with 200 μL of 20 mmol/L HEPES (pH 7.4), were air dried, and were crystallized by the addition of α-cyano-4-hydroxycimamic acid (CHCA; Ciphergen Biosystems).

Chips were detected on the protein biological system II (PBS-II) plus mass spectrometer reader (Ciphergen Biosystems). Data were collected by averaging 140 laser shots with an intensity of 150, a detector sensitivity of 6, a highest mass of 30,000 Da and an optimized range of 2,000–20,000 Da. Mass accuracy was calibrated to <0.1% with the All-in-1 peptide molecular mass standard (Ciphergen Bios). The upper limit was 22,000. Under the SELDI-TOF MS, and the output can be grouped into any given number of categories. The pattern recognition techniques have been applied to diverse areas including gene microarray (12) and MS (13).

RESULTS

After cluster analysis by Biomarker Wizard (Ciphergen Inc., Fremont, CA; Proteinchip Software versions 3.0), there were 54 peaks detected for discriminating colorectal cancer patients from healthy individuals. The peaks were between 2 and 30 kDa. Peaks with a m/z <2 kDa were mainly ion noise from the matrix and were, therefore, excluded (15). The m/z of the four candidate biomarkers were 5911, 8930, 8817, and 4476. The normalized intensities of the four peaks were all significantly elevated in colorectal cancer patients compared with healthy controls (Fig. 1), with the P values of t tests <10−9 and the area under the receiver operating characteristic (ROC) curve >8.0.

Evaluation of Artificial Neural Networks. The four peaks were combined and evaluated by integrated ANNs. The hidden layer was composed of two hidden nodes. The architecture of ANNs is plotted in Fig. 2. The training data were performed 3,000 epochs in ANN. In the procedure of training ANNs, we used a cross-validation approach to reduce the risk of “over fit” (16). We had done a 5-fold cross-validation approach. In each round of cross-validation, 80% of each group was assigned to be the training set and the remaining 20% were held out as the test set. There were 74 healthy samples and 44 colorectal cancer samples in the training set, and 18 healthy samples and 11 colorectal cancer samples in the test set. The total sample size is the sum in the 5-fold cross-validation. Table 1 shows the result for this classifier. For the integrated ANNs classifier, the estimated specificity in the test sets was 93%, the estimated sensitivity was 91%, the estimated positive predictive value was 89%, the area under the ROC curve was 0.967 (Fig. 3). It is generally.

Bioinformatics and Biostatistics. Using Biomarker Wizard software (Ciphergen Inc., Fremont, CA), we compiled all spectra. Qualified mass peaks (signal-to-noise ratio >5) with mass-to-charge ratios (m/z) between 2000 and 20,000 were automatically detected. Peak clusters were completed with second-pass peak selection (signal-to-noise ratio >2, within a 0.3% mass window), and estimated peaks were added. The peak intensities were normalized to the total ion current of m/z between 2000 and 20,000. All of these were done with ProteinChip Software 3.0 (Ciphergen Bios).
accepted that the area under the ROC curve \( \geq 90\% \) is of satisfactory power in diagnosis.

The same data were used by discriminant analysis, in original grouped data; the estimated specificity in the test sets was 87\%, the estimated sensitivity was 79\%, and the estimated positive predictive value was 85.7\%. In cross-validated grouped data, the estimated specificity in the test sets was 85\%, the estimated sensitivity was 81\%, the estimated positive predictive value was 83.7\% cases correctly classified.

**Serum Carcinoembryonic Antigen Levels.** Serum CEA levels were determined with a Wallac DELFIA CEA kit, and a cutoff value of 5 ng/mL was used. In 92 healthy individuals, mean serum CEA level was 1.60 ng/mL (range, 0.36~7.49). In the CEA-positive group (>5 ng/mL), the number of healthy individuals was 6, and in the CEA-negative group \( \leq 5 \text{ ng/mL} \) the number was 86 (Table 2). In the colorectal cancer group \((n = 55)\), mean CEA level was 25.54 ng/mL (range, 0.60~442.0). In the CEA-positive group (>5 ng/mL), the number of colorectal cancer was 26 and in the CEA-negative group \( \leq 5 \text{ ng/mL} \), 29 (Table 2). With the use of the cutoff value of the CEA level of 5 ng/mL, this testing has the detection sensitivity of 47.3\% (26 of 55) and the specificity of 93.5\% (86 of 92) in separating colorectal cancer samples from healthy samples.

To exclude the possibility that the four top-scored proteins could be the multicharged forms of CEA, we investigated the correlation between serum CEA level and each of the four potential biomarkers from the 55 colorectal cancer persons and the 92 healthy persons. As shown in Fig. 4, linear regression analysis indicated that there was no statistically significant correlation \(( P > 0.05)\).

**DISCUSSION**

In screening for colorectal cancer, the aim should be to detect disease at either Dukes A stage or Dukes B stage. The available biomarkers for detection of colorectal cancer, such as CEA and FOBT, are at a disadvantage in sensitivity, or specificity, or both. Using an upper limit of normal of 2.5 ng/mL, Fletcher (3) calculated that CEA has a sensitivity of 36\% and a specificity of 87\% in screening for Dukes A and B colorectal cancer. All of the existing biomarkers in serum lack sufficient sensitivity for screening and diagnosing colorectal cancer (17–20). These findings, combined with the low prevalence of this malignant disease in unselected populations, render the positive
predictive value of CEA unacceptably low and, thus, of little value in screening healthy subjects. Therefore, additional FOBTs and endoscopy were required to achieve adequate specificity and sensitivity in screening colorectal cancer (21). But almost one half of the colorectal cancer cases were FOBT negative, especially in early stages (4). In our result, we also find that CEA has a sensitivity of 47.3% and a specificity of 93.5% in screening for colorectal cancer with an upper limit of 5.0 ng/mL. Because FOBT and CEA examination lack the adequate accuracy in detecting colorectal cancer, especially the early-stage disease, exploring a new method to improve sensitivity and specificity in the screening or detection for colorectal cancer is urgently needed.

Because of the multifactorial nature of cancer, it is very likely that a combination of several markers will be necessary to effectively detect and diagnose cancer. To look for such fingerprints of cancer apparently requires not only high-throughput genomic or proteomic profiling (22, 23) but also sophisticated bioinformatics tools for complex data analysis and pattern recognition. Taking advantage of the recently developed SELDI-TOF and the ProteinChip technologies, we were able to simultaneously analyze the protein profiles of 147 serum samples from patients and healthy individuals. The software package Biomarker Wizard allows evaluation of each mass peak according to its collective contribution toward the maximal separation of the cancer patients from the healthy controls. The ROC curve, t test were applied to rank and to select the peaks according to their contribution to the separation of two groups.

On acquiring the serum protein profiles from the training sets by SELDI-TOF, the colorectal cancer diagnostic pattern has been stringently extracted and tested for its specificity and sensitivity in segregating the cancers from the validation set. Our data showed that the pattern effectively discriminated the colorectal cancers from healthy controls, with superb diagnostic power as compared with CEA. The diagnostic pattern presents the detection specificity of 93% and sensitivity of 91%. Of particular importance is that all cancers at Dukes A and B stages were precisely recognized by the pattern in the masked set, with the detection specificity of 100% and sensitivity of 100%. In contrast, serum CEA elevation occurred mostly in the patients with Dukes stages C and D. Conceivably, this SELDI diagnostic pattern is of important potential for detecting the early-stage cancers from presymptomatic high-risk populations. Because of its high diagnostic accuracy, it may substantially reduce the numbers of unnecessary endoscopic examination and relieve the patients of unnecessary anxiety caused by high false-positive/negative frequencies by CEA or FOBT testing.

In this study, four top-ranked peaks with m/z values (5,911 Da, 8,930 Da, 8,817 Da, 4,476 Da) were discovered to be of Table 1 Classification of training and test sets in 5-fold cross-validation

<table>
<thead>
<tr>
<th></th>
<th>Training set</th>
<th></th>
<th>Test set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total HP CRC</td>
<td>Total HP CRC</td>
<td></td>
</tr>
<tr>
<td>HP</td>
<td>370</td>
<td>359</td>
<td>11</td>
</tr>
<tr>
<td>CRC</td>
<td>220</td>
<td>11</td>
<td>209</td>
</tr>
<tr>
<td>Specificity</td>
<td>97% (359 of 370)</td>
<td>93% (84 of 90)</td>
<td></td>
</tr>
<tr>
<td>Sensitivity</td>
<td>95% (209 of 220)</td>
<td>91% (50 of 55)</td>
<td></td>
</tr>
<tr>
<td>Positive value</td>
<td>95% [209 of (209 + 11)]</td>
<td>89% [50 of (50 + 6)]</td>
<td></td>
</tr>
</tbody>
</table>

NOTE. In each round of cross-validation, 80% of each group were assigned to be the training set and the remaining 20% were held out as the test set. Specifically, there were 74 HP and 44 CRC samples in the training set, and 18 healthy samples and 11 CRC samples in the test set. The total sample size is the sum in the 5-fold cross-validation.

Abbreviations: HP, healthy people; CRC, colorectal cancer.

Table 2 CEA levels in CRC and healthy samples

<table>
<thead>
<tr>
<th>Sample</th>
<th>n</th>
<th>CEA ≤5.0 ng/ml</th>
<th>CEA &gt;5.0 ng/ml</th>
<th>Correct rate %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Healthy</td>
<td>92</td>
<td>86</td>
<td>6</td>
<td>93.5</td>
</tr>
<tr>
<td>CRC of Dukes A</td>
<td>55</td>
<td>29</td>
<td>26</td>
<td>47.3</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>6</td>
<td>2</td>
<td>25.0</td>
</tr>
<tr>
<td>B</td>
<td>22</td>
<td>13</td>
<td>9</td>
<td>40.9</td>
</tr>
<tr>
<td>C</td>
<td>13</td>
<td>6</td>
<td>7</td>
<td>53.8</td>
</tr>
<tr>
<td>D</td>
<td>12</td>
<td>4</td>
<td>8</td>
<td>66.7</td>
</tr>
<tr>
<td>Total</td>
<td>76</td>
<td>76</td>
<td>6</td>
<td>76.2</td>
</tr>
</tbody>
</table>

Fig. 3 Receiver operating characteristic (ROC) curve for validation cases from the neural network. For the integrated ANNs classifier, the area under the ROC curve was 0.967.
high value in the colorectal cancer diagnosis. Could they be the multiple charged forms of CEA? We have checked the intensity of the markers relative to CEA level. The results showed that there were no significant correlations between these proteins, which strongly implied that these four proteins/peptides were unknown potential biomarkers. Therefore, it is particularly important to identify these proteins. If the molecular nature of these proteins/peptides were known, the application potential of these proteins in the detection of colorectal cancer would expand greatly. At this moment, the detection of colorectal cancer by the pattern of these four proteins is solely depending on SELDI-TOF MS. However, when these molecules are identified, antibodies of high specificity and high affinity to these molecules could be generated and could be used separately or in combination for the development of all sorts of immunochemical methods, which ultimately would truly benefit those patients with colorectal cancer at the presymptomatic stages.

REFERENCES


Artificial Neural Networks Analysis of Surface-Enhanced Laser Desorption/Ionization Mass Spectra of Serum Protein Pattern Distinguishes Colorectal Cancer from Healthy Population

Yi-ding Chen, Shu Zheng, Jie-kai Yu, et al.


Updated version
Access the most recent version of this article at:
http://clincancerres.aacrjournals.org/content/10/24/8380

Cited articles
This article cites 20 articles, 8 of which you can access for free at:
http://clincancerres.aacrjournals.org/content/10/24/8380.full#ref-list-1

Citing articles
This article has been cited by 12 HighWire-hosted articles. Access the articles at:
http://clincancerres.aacrjournals.org/content/10/24/8380.full#related-urls

E-mail alerts
Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions
To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions
To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org.