Molecular classification of liver cirrhosis in a rat model by proteomics and bioinformatics

Xiu-Qin Xu[1], Chon K. Leow[1, 2], Xin Lu[3], Xuegong Zhang[4], Jun S. Liu[3], Wing-Hung Wong[3], Arndt Asperger[5], Sören Deininger[5] and Hon-chiu Eastwood Leung[1]

[1]Medical Proteomics and Bioanalysis Section, Genome Institute of Singapore, Singapore
[2]Department of Surgery, National University of Singapore, Singapore
[3]Harvard University, Cambridge, MA, USA
[4]Department of Automation and MOE Key Laboratory of Bioinformatics, Tsinghua University, Beijing, China
[5]Bruker Saxonia Analytik GmbH, Leipzig, Germany

Liver cirrhosis is a worldwide health problem. Reliable, noninvasive methods for early detection of liver cirrhosis are not availabe. Using a three-step approach, we classified sera from rats with liver cirrhosis following different treatment insults. The approach consisted of: (i) protein profiling using surface-enhanced laser desorption/ionization (SELDI) technology; (ii) selection of a statistically significant serum biomarker set using machine learning algorithms; and (iii) identification of selected serum biomarkers by peptide sequencing. We generated serum protein profiles from three groups of rats: (i) normal (n = 8), (ii) thioacetamide-induced liver cirrhosis (n = 22), and (iii) bile duct ligation-induced liver fibrosis (n = 5) using a weak cation exchanger surface. Profiling data were further analyzed by a recursive support vector machine algorithm to select a panel of statistically significant biomarkers for class prediction. Sensitivity and specificity of classification using the selected protein marker set were higher than 92%. A consistently down-regulated 3495 Da protein in cirrhosis samples was one of the selected significant biomarkers. This 3495 Da protein was purified on-chip and trypsin digested. Further structural characterization of this biomarkers candidate was done by using cross-platform matrix-assisted laser desorption/ionization mass spectrometry (MALDI-MS) peptide mass fingerprinting (PMF) and matrix-assisted laser desorption/ionization time of flight/time of flight (MALDI-TOF/TOF) tandem mass spectrometry (MS/MS). Combined data from PMF and MS/MS spectra of two tryptic peptides suggested that this 3495 Da protein shared homology to a histidine-rich glycoprotein. These results demonstrated a novel approach to discovery of new biomarkers for early detection of liver cirrhosis and classification of liver diseases.

Received	20/1/04
Revised	29/3/04
Accepted	30/3/04

Keywords: Cirrhosis / Recursive support vector machine / SELDI/MALDI-MS/MS

1 Introduction

Liver cirrhosis is the most common complication of chronic liver disease, secondary to chronic alcohol ingestion, or viral infection with hepatitis B or C virus. The gold standard for diagnosing cirrhosis is by histological examination of the liver obtained by liver biopsy. However, this is an invasive procedure associated with potential risk of internal bleeding following biopsy. While gross cirrhosis can be detected by computed tomography scanning, it is not able to detect early cirrhosis accurately. At present, there are no sensitive and specific serum or plasma markers available for the detection of cirrhosis. Recently, genomic analyses using RT-PCR or cDNA microarray were being used to classify diseased tissues from normal tissues. Mechanism of toxicity of hepatotoxins was revealed by using cDNA microarray [1]. Application of these technologies in a clinical setting is limited by the need to obtain liver tissue by an invasive procedure. Con-

Correspondence: Dr. Hon-chiu Eastwood Leung, Medical Proteomics and Bioanalysis Section, Genome Institute of Singapore, 60 Biopolis Street, Singapore 138672, Singapore
E-mail: leunge@gis.a-star.edu.sg
Fax: +65-6468-9060

Abbreviations: RSVM, recursive support vector machine; **SVM**, support vector machine; **WCX**, weak cation exchanger

versely, the utility of serum instead of tissue to classify diseased state would have great advantages in that serum is easy to collect, the procedure is minimally invasive and samples can be collected repeatedly with ease.

Proteomic approaches have been used to understand the pathogenesis of liver disease and to identify new biomarkers. Early proteomic approaches used 2-DE [2]. Differentially expressed proteins can then be identified by downstream mass spectrometry analyses. Although 2-DE is unchallenged in its ability to separate thousands of proteins, it resolves hydrophobic proteins poorly, low-abundant proteins are often not detected, and low molecular weight proteins are not resolved. It is not suitable for large-scale screening or clinical testing because it is laborious, and consumes large quantities of proteins.

Recent advances in mass spectrometry have led to it emerging as a powerful technology for proteomic profiling and characterization. SELDI-TOF-MS, an exension of MALDI-MS, offers a sensitive and high-throughput technology for protein profiling and subsequent biomarker discovery. This technology uses different surface chemistries for affinity capture of proteins from complex biological samples and is followed by mass spectrometric analysis. The SELDI approach has been successfully used to identify serum biomarkers in bladder carcinoma [3, 4], lung cancer [5], and ovarian cancer [6]. In the ovarian cancer serum biomarker discovery study, a bioinformatics algorithm was the key element that led to a high level of sensitivity and specificity ($>$95%) [6].

In order to identify the most relevant proteins from large numbers of proteins being monitored in experiments, and build a model using these identified markers to predict sample status, a cohort of machine learning algorithms could be used. These algorithms include genetic algorithms (GA), nearest-neighbor, decision trees, neural networks, and support vector machines (SVM). SVM is a new machine learning method developed by Vapnik and his colleagues in the mid-1990s [7, 8]. The key idea of SVM is generalization: a classifier needs not only to work well on training samples, but also to work equally well on previously unseen samples or test samples. Although this philosophy had long been recognized before the appearance of SVM, it is SVM that gives it a good implementation. SVM is built upon a theory about learning with limited samples, called statistical learning theory [7, 9]. The standard theory and algorithm of SVM has been described many times in the literature [7–10], and SVM codes are publicly available such as SVMTorch [11]. Considering the intrinsic complexity of biological objects, and their high-dimensionality in contrast to small sample sets for high-throughput biological techniques, SVM becomes an ideal selection of data analysis tool in these studies.

We report here classification of liver cirrhosis/fibrosis in rat models by using serum protein profiling together with identification of significant serum biomarkers. This pioneering study demonstrates that an algorithm model could be developed to identify a cluster pattern that segregates chemicaly-induced cirrhosis and bile duct ligation liver fibrosis from normal controls. We further analyzed the primary structure of a statistically significant protein peak by on-chip purification and peptide sequencing. Our results show that this approach may lead to identification of new markers for diagnosis and prognosis purposes for patients with liver cirrhosis.

2 Materials and methods

2.1 Induction of cirrhosis

Nine week old, male Wistar-Furth rats were obtained from Sembawang Animal Centre, Singapore. Following one week of acclimatization, the animals were treated as follows: rats (n = 22) received intraperitoneal thioacetamide injections thrice weekly at a dose of 300 mg/kg body weight for 10 weeks to induce cirrhosis. At the end of 10 weeks, these animals were rested for one week without injections prior to being sacrificed. Bile doct ligation rats (n = 5) underwent laparotomy under inhalational ether anaesthesia and their bile ducts were doubly ligated with silk suture. The animals were recovered, kept for 10 weeks and then sacrificed. Control animals (n = 8) were not treated and kept for 10 weeks prior to sacrifice. All animals were kept according to the institutional guidelines on animal experimentation. All animals were anaesthetized with ether. Through a laparotomy wound, blood was taken from the inferior vena cava (IV) prior to exsanguinations by transection of IVC just beneath the diaphragm. Livers were rapidly removed. Blood samples were spun; serum removed and snap frozen in liquid nitrogen for analysis at a later date. The left lobe of the liver was placed in 10% w/v phosphate buffered formalin for histological processing while the right lobe was snap frozen in liquid nitrogen and stored.

2.2 Liver histology and estimation of cirrhosis

Following fixation, the left lobe of liver was embedded in paraffin. Four micron sections were prepared, mounted and stained with Masson Trichrome stain. Each section was examined for the degree of fibrosis or cirrhosis. The degree of hepatic fibrosis was scored with a modified scoring system based on that described by Ruwart and coworkers [12]. The amount of positively stained collagen was graded as follows: 0, no collagen or fibrosis; 1, slightly increased collagen; 2, definite increase, without septa, generally seen as small stellate expansions of collagen

from the central zones or pericentrally in the lobules (septa were identified as linear collagenous extensions from microscopic landmarks, usually terminal hepatic venules); 3, definite increase with incomplete septa (those septa which did not interconnect with each other so as to divide the parenchyma into separate fragments); 4, definite increase with complete septa but thin septa (those septa which interconnected with each other so as to divide the parenchyma into separate fragments). Established cirrhosis scored 4 based on this scoring system.

2.3 SELDI analyses of serum samples

Four types of chip (Ciphergen Biosystems, Fremont, CA) with surface chemistry of hydrophobic, ionic, cationic, and metal binding were initially evaluated to determine which affinity chemistry provied the best serum profiles. The weak cation exchanger (WCX) type chip was used throughout this study because of the presence of differential peaks. Serum samples were diluted 1:5 in WCX chip binding buffer containing 0.1 M sodium acetate, 0.02% w/v Triton X-100, and pH 5.5. An aliquot, 2 μL, of diluted sample was applied to each spot on an eight-spot WCX chip. Sample preparation and SELDI analysis were performed according to the recommendations of the manufacturer. Mass accuracy was calibrated daily through the use of standard human insulin (5733 Da) before chip processing. CHA saturated in 50% acetonitrile, 0.1% v/v TFA, was added, 0.5 μL, twice before analyses of low molecular weight proteins. Samples from three groups were run concurrently, and repeated on intra-chip spots and on inter-chip spots to test reproducibility. Captured proteins were detected using the PBS-II mass reader. Data were collected using the automatic chip protocol. For detection of low molecular mass proteins, the protocol setting was: high mass 20 kDa, with optimized mass from 1.0 kDa to 16 kDa; laser intensity, 180; detector sensitivity, 9; 61 shots on average per sample. Peak intensities were normalized according to total ion current after background subtraction. Mass accuracy was normalized to calibrated internal standard peaks. For Biomarker Wizard setting, the signal-to-noise ratio was set between 2 and 5. Peak clusters were generated by allowing a mass difference of 0.02%.

2.4 Recursive support vector machine (RSVM) analyses and feature selection

We used a simple recursive support vector machine (RSVM) strategy to search for a suboptimal combination of predictive biomarkers [13]. The basic procedure was: step 0, predefine a decreasing series of feature numbers $d_0 > d_1 > d_2 > \ldots > d_k$ to be selected in the recursive procedure, where $d_0 = d$ is the number of all features available in the initial data, set $i = 0$; step 1; build the SVM decision function with current d_i features; step 2, rank the features according to their contribution in the decision function and select top d_{i+1} feature (or remove the bottom, $d_1 - d_{i+1}$ features as stated before [14]); step 3, $i = i+1$, repeat from step 1 until $i = k$.

In order to get an unbiased estimation of error rate, we followed a corrected cross-validation scheme. This was done by leaving samples out followed by recursively selecting feature selection on the training subset. Different lists of important features could be generated from different training subsets. From another point of view, these training subsets could be treated as resampled subsets from a potential population. Due to the intrinsic complexity of biological problems, we targeted for, not only an optimal subset of features that could best classify the samples, but also a stable list of features that were consistent among different available sample sets. Therefore, a frequency-based selection method was adopted to generate the final feature list [15]. After applying the recursive feature selection procedure to each of the training subsets, we counted the frequency of each feature being slected in each of the d_i levels, and the top d_i high frequency features were reported as the final d_i list.

2.5 SELDI-guided protein purification

Guided purification of the 3.5 kDa protein was performed as follows: normal rat serum was diluted ten-fold in 0.2 M ammonium acetate, pH 7. An aliquot, 2 μL, of diluted serum was loaded onto eight spots of a WCX chip. Each spot was preloaded with 4 μL of different binding buffers such as 0.2 M ammonium acetate, pH 7; 0.2 M ethanolamine-HCl, pH 9; 0.2 M ethaloamine-HCl, pH 10; 0.2 M ethanolamine-HCl, pH 10, with 0.1 M KCl; 0.2 M ethanolamine-HCl, pH 10, with 0.2 M KCl; 0.2 M ethanolamine-HCl, pH 10, with 0.5 M KCl; and 0.2 M ethanolamine-HCl, pH 10, with 1 M KCl. Proteins were allowed to bind at room remperature for 30 min in a humid chamber. Non-bound proteins were soaked away using Kimwipe paper. The reason to use Kimwipe paper was to ensure a slow and constant solution removal speed so that loosely bound peptides could still be retained on the chips's surface. An aliquot, 5 μL, of binding buffer was added to each spot followed by 2×5 μL HPLC water wash. Saturated CHCA was added, 0.5 μL, twice before mass reading.

2.6 On-chip protein purification and tryptic digestion

Each WCX chip was prepared according to vendor's instructions except that the binding buffer used was 4 μL 0.2 M ammonium acetate, pH 7.0. Normal serum samples

were diluted tenfold in 50 mM sodium acetate, pH 5.0. Two μL diluted serum was spotted onto a WCX chip in triplicate. Proteins were allowed to bind for 30 min in a humid chamber at room temperature. Non-bound proteins were gently soaked from chip onto Kimwipe papers without touching the surface. Addition of 5 μL binding buffer was used as a wash step. Again, solution on spots was removed by Kimwipe paper and then two rounds of 5 μL water were applied to samples as a final wash. One spot was subjected to two rounds of loading 0.5 μL saturated CHCA solution. The other spots were saved for later on-chip trypsin digestion and peptide sequencing. The presence of a 3495 Da peptide peak was checked in the spectrum generated by the PBS-II mass reader.

Unread spots on a chip containing a purfied 3495 Da peptide peak were subjected to on-chip trypsin digestion. Sequencing grade modified trypsin (Roche; Indianapolis, IN, USA) was resuspended in 25 mM ammonium bicarbonate at final concentration of 20 ng/μL. 5 μL diluted trypsin was loaded onto the unread spots. The chip was put into a humidified 15 mL blue cap tube and incubated at 37°C for 3 h. The tryptic digest was removed with a Kimwipe. Two rounds of 5 μL HPLC-water wash were performed before addition of two rounds of 0.5 μL 20% w/v CHCA solution. Digested peptides were resolved in PBS-II mass reader by reading the second spot.

2.7 Peptide sequencing using MS/MS

Digested peptides on the third unread spot were extracted with 5 μL 50% v/v acetonitrile, 0.1% v/v TFA. Extracted peptides were loaded onto one spot of AnchorChip (Bruker Daltonics, Billerica, MA, USA). The function of AnchorChip was to condense each tryptic digest in a very small area so that higher signal intensity could be obtained even with low concentration of analytes on the target plate. MS spectra of peptides were generated by using an Ultraflex MALDI TOF/TOF mass spectrometer (Bruker Daltonics). The most prominent tryptic peptides were subject to MS/MS analysis.

3 Results

3.1 Gross liver appearance and liver histology

At the time of sacrifice, livers of control animals showed normal gross morphology. On histological examination, normal liver had no evidence of increased fibrosis (Fig. 1a). All thioacetamide treated animals had livers with gross nodules typical of liver cirrhosis. Livers of bile duct ligated animals were grossly enlarged with multiple cystic structures on the surface. Cysts contained bile color fluid. While liver sections of both treatment groups showed marked fibrous deposition, the patterns of deposition were different between two treatment groups. Thioacetamide treated liver sections showed a classical appearance of established cirrhosis with typical cirrhotic nodules (Fig. 1b). All twenty-two specimens scored 4, based on the scoring system described above. Histological sections of bile duct ligated liver showed marked ductal proliferation and deposition of collagen (Fig. 1c). However the histological pattern of fibrosis was not typical of that seen in cirrhosis. We used histological analyses as a bench mark that these three groups of rats were showing different liver histological characteristics. The next question was whether serum protein profiles could distinguish different groups of samples.

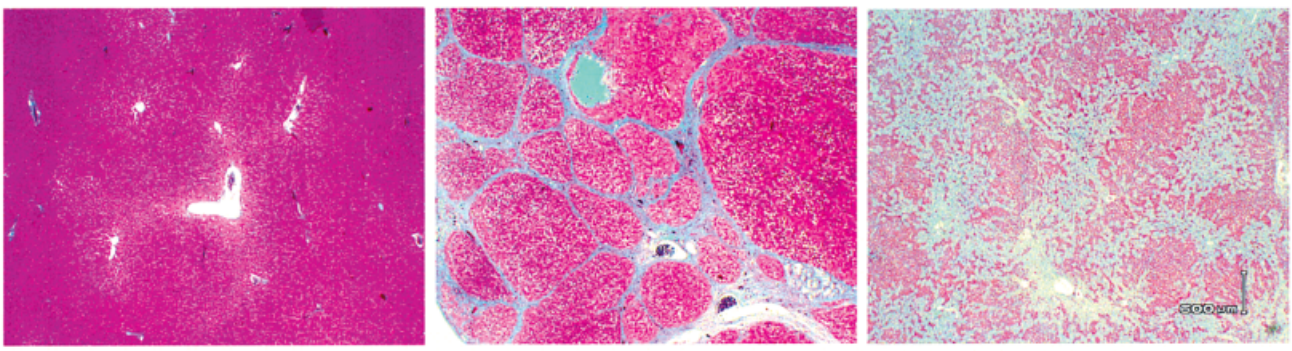

Figure 1. Liver sections stained with Masson Trichrome. (A) normal liver, (B) cirrhotic liver, and (C) liver after bile duct ligation. The fold magnification in 1a) and 1b) was × 20; while 1c) was × 40.

3.2 Scanning of rat serum protein profiles

Each serum sample was processed at least in triplicate to confirm reproducibility in resolving the proteins. The averaged coefficient of variation in this study was 12%. Figure 2a depicts a representative protein spectrum showing the proteins in low molecular mass between 1500 Da and 9000 Da of a single serum sample. Cirrhosis-related down-regulated serum biomarkers were 3495.38 and 8201.04 Da peaks (Figs. 2b and e, respectively). Cirrhosis-related up-regulated serum biomarkers were 4186.07, and 4902.63 Da peaks (Figs. 2c and d, respectively).

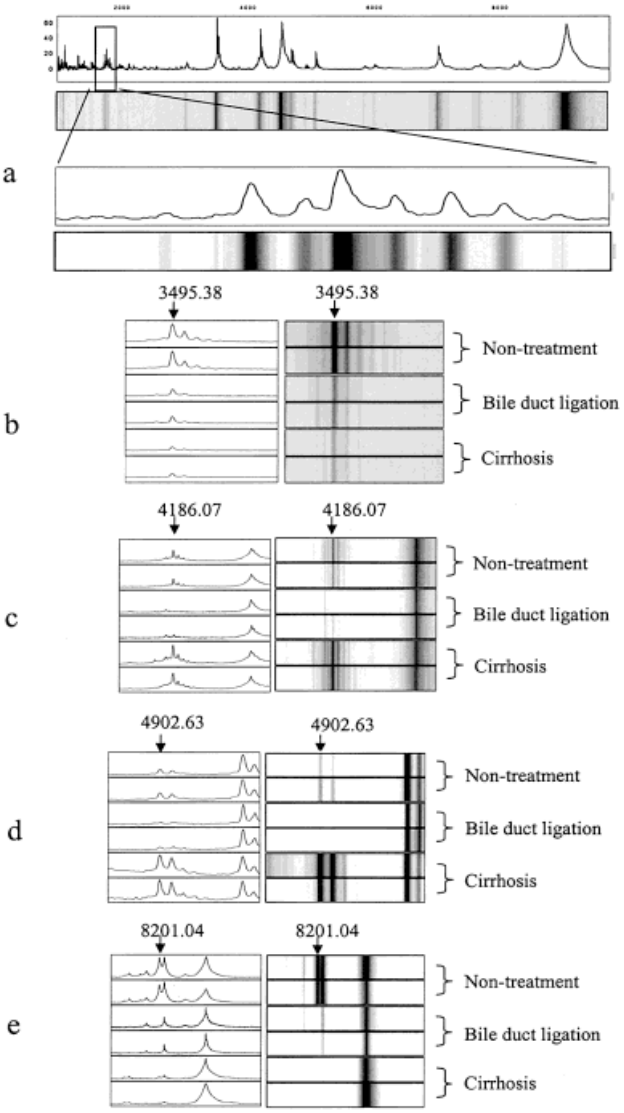

Figure 2. Protein profiling on WCX chips. Representative overview of protein profiling on WCX chips showing spectral map (upper panel) and gel view (lower panel) of serum from one sample (a). X-axis represents the molecular mass calculation (*m/z* values), Y-axis represents relative intensity. A boxed region of the spectrum is zoomed to reveal the resolution of the spectrum. SELDI analysis of rat serum for proteomic pattern in normal control, bile duct ligation, and cirrhotic samples with mass spectra (left) and gel view (right) (2b to 2e). Differential expressed proteins in non-treated control, bile duct ligation, and cirrhosis samples with *m/z* values of (B) 3495.38 Da, (C) 4186.07 Da, (D) 4902.63 Da and (E) 8201.04 in each panel were from different individuals.

We used the Biomarker Wizard function of the SELDI software to identify clusters of peaks differentially expressed in cirrhotic, bile duct ligation and control non-treated samples (Fig. 3). Among 94 clusters, five peaks with molecular masses 1743.12, 3495.38, 4186.07, 4902.63, and 8201.04 Da displayed differences in distribution of intensities of peaks in three groups. Three peaks with masses 1743.12, 3495.38 and 8201.04 Da were lower in bile duct ligation samples and even lower in the cirrhotic rats. The 1743.12 Da peak was considered to be a doubly-charged component of 3495.38 Da, as its molecular mass was almost exactly one-half of the latter protein. Sodium adducts of 3495 Da peptide were also observed as 3515 and 3537 Da peaks with gradient decrease in intensity. Two proteins with masses 4186.07 and 4902.63 Da were detected highly expressed in cirrhotic samples when compared with the normal controls, and interestingly, these two proteins were also down-regulated in bile-duct ligation samples (Fig. 2c, 2d, and Fig. 3).

3.3 Selection of significant biomarkers and classification of proteomic pattern using the RSVM algorithm

We used a machine learning algorithm to select a panel of statistically significant biomarkers and to segregate sample classes. There were in total 15422 data points in the whole spectrum. Since there was hardly any signal in the region above 10 kDa and too much noise below 1 kDa, only the region between 1 and 10 kDa was used in the following analysis, where there were 7457 data points. As each spectrum had a slightly different mass axis, we carried out an interpolation smoothing before performing a point-wise comparison and obtained 4607 points for each spectrum.

The biomarker detection function of Ciphergen software V3.0 detected 78 biomarkers from the region between 1 and 10 kDa. From the comparison between cirrhosis and normal samples, RSVM feature selection algorithm identified 6 important markers: [1743.12, 3515.68, 3537.26, 4186.07, 4902.63, and 8201.04] (in Da).

As a comparison, we also tried two other methods to find important markers. The first one was to use all points, instead of only biomarkers detected, in RSVM feature selection. While using this method, we looked at the "important regions" (the region where important data

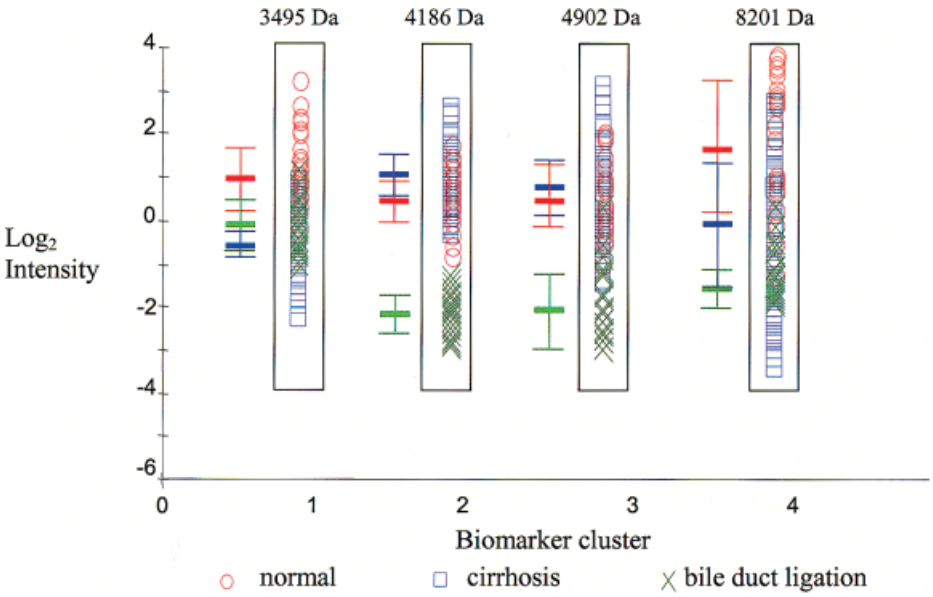

Figure 3. Distribution of signal intensity of four selected biomarkers generated by SELDI Biomarker Wizard software. Four boxed clusters with molecular masses of 3495.38, 4186.07, 4902.63, and 8201.04 Da are illustrated among normal (circle), bile duct ligation (cross) and cirrhotic rats (square). Thick bars represent averaged value and thin bar represents standard deviation.

points were highly condensed) instead of individual points, otherwise some data points that reside at the shoulder or valley of peaks along the spectrum were hard to explain and validate biologically (see [10] and following discussion). The "important regions" included exactly the same 6 markers as selected by RSVM (1743.12, 3515.68, 3537.26, 4186.07, 4902.63, 8201.04 Da), but the final 7 top points selected from all 4607 points were not exactly these 6 points. Instead, there were some points beside them (1744.56, 3513.31, 3515.07, 3518.60, 3520.36, 4187.13, and 8209.99 Da).

The second method used to find important markers was scanning spectra by using a sliding window with a width of 41, and then calculating the ratio of mean intra-class distance over mean inter-class distance within the sliding window. The regions with a ratio of mean distance below a threshold (0.75 here) were important regions. Twenty-one markers were detected in these important regions, and the RSVM further selected 6 markers from these 21 markers (1743.12, 1787.89, 3515.68, 3537.26, 6207.55, and 8201.04 Da).

The CV2 error rate (external CV; leave – one – out first and select feature from training set left), and CV1 error rate (internal CV; feature selection from whole data set and then calculating CV, the error rate was underestimated) of these 3 groups of biomarkers selected were listed in Table 1. The type specific CV2 error rates were about 7.7% for negative group (normal rat), 2.9% for positive

Table 1. Overall error rates of different statistical biomarkers selection approaches

Statisical Method	CV2 (external CV) errors		CV1 (internal CV) errors	
	False positive (Type 1) count	False negative (Type 2) count	False positive (Type 1) count	False negative (Type 2) count
Point-to-point RSVM	2	2	1	1
Sliding window selection	2	0	2	0
Biomarker Wizard RSVM	2	2	1	1
Type Specific Error Rate	7.7%	0 to 2.9%	3.8 to 7.7%	0 to 1.4%
Overall Error Rate	2.1 to 4.2%		2.1%	
Overall sensitivity	97.1 to 100%			
Overall specificity	92.3%			

group (cirrhosis), and the CV1 error rates were about 3.8 to 7.7% for negative group and 0 to 1.4% for positive group. The overall error rate for CV2 was about 2.1 to 4.2% and that for CV1 is about 2.1%. This translated to sensitivity and specificity of 97.1 to 100%, and 92.3%, respectively (Table 1). These subsets of markers were able to segregate 3 classes sufficiently well when subject to multidimensional scaling analyses (Fig. 4). These six markers also comprised the minimal panel of the most significantly differential markers with higest specificity and sensitivity. Accuracy of class prediction tended to decrease if less than six markers were chosen (data not shown).

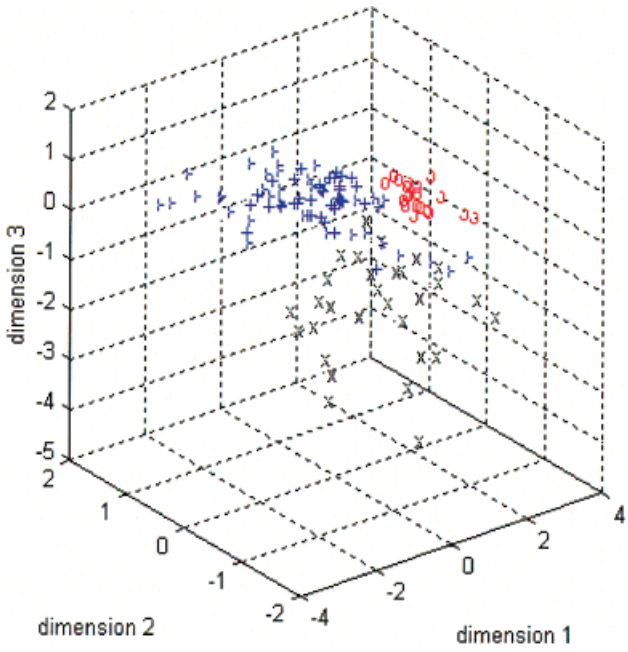

Figure 4. 3-D multidimensional scaling of three groups of serum samples. The most important biomarkers identified by various feature selection methods were used in this figure. Red circles (o): bile duct ligation samples; blue cross (+): cirrhotic samples; black cross (x): normal samples.

We also evaluated each marker separately by single marker statistics including Student's *t*-test and receiver operating characteristic (ROC) analyses. Most of the selected markers were significant by these tests, but not necessarily the ones at the highest ranking. While combining the 3515 Da peak with other selected markers such as 1743, 3537, 4186, 4902, and 8201 Da peaks, they can separate the classes better than other combinations in SVM classifier. We also tried to build an SVM model on those top scoring Student's *t*-test or ROC markers, but accuracy of class prediction was not as good as these six markers according to CV error (data not shown). Therefore, multivariable methods were more suitable to select important markers, not only based on performance of single maker, but more importantly, based on performance of combination of multiple markers.

3.4 SELDI-guided on-chip purification of the 3495 Da peptide

The 1743, 3515 and 3537 Da peaks were selected as important markers by three statistical tests, and these peptides were mostly dicharged peptide or sodium adducts of the 3495 Da peak. As an attempt to purify these protein peak series rapidly, we decided to purify these peptides on-chip. Protein profiles were generated after diluted normal serum was loaded onto spots with binding and washing buffers from pH 7 to 10 with increasing concentration of potassium chloride. In a preliminary trial, 1743, 3495, 3515, and 3537 Da peaks condensed to one single 3495 Da peak as the most prominent peak. This 3495 Da peak could not be eluted completely even in 0.2 M ethanolamine-HCl, pH 10, buffer with 0.5 M NaCl. This implied that this protein was highly positively charged. We then replaced NaCl with KCl, as shown in Fig. 5a, the signal intensity of the peak at 3495 Da was 55.6 in pH 7 buffer. This was the most prominent peak in the spectrum. In buffer of pH 9, a 53% decrease of signal intensity was observed (signal intensity 26). In buffer of pH 10 with increasing concentration of KCl (0.1 to 0.5 M), the signal intensity dropped to 20 or 14% of that at pH 7. The intensity decreased to an undetectable level in buffer of pH 10 with 1.0 M KCl. As a conclusion, the 3495 Da peptide was a highly positively charged peptide.

3.5 On-chip purification and identification of the 3495 Da peptide

We took advantage of the high binding affinity to the WCX chip surface by the 3495 Da peptide. After binding and washing with 0.2 M ammonium acetate buffer, pH 7.0, the 3495 Da peptide became the most prominent peak (Fig. 6a) in the PBS-II mass reader. The peak intensity of the 3495 Da peak was reduced after three hours of on-chip trypsin degistion (Fig. 6b) whilst the daughter peaks intensities increased. The most prominent daughter peaks, in decreasing order of intensity, were 1000, 1228, 1509.9, 1790.5, and 881 Da peaks (Fig. 6b). These daughter peaks were not observed in spectra with trypsin alone (Fig. 6c). The PBS-II mass reader itself generated average masses. However, the determination of the exact monoisotopic mass of the most significant marker (3495 Da) was an essential precondition for further MS/MS analysis and subsequent sequencing on the basis of MS/MS data. Since the PBS-II mass reader was not

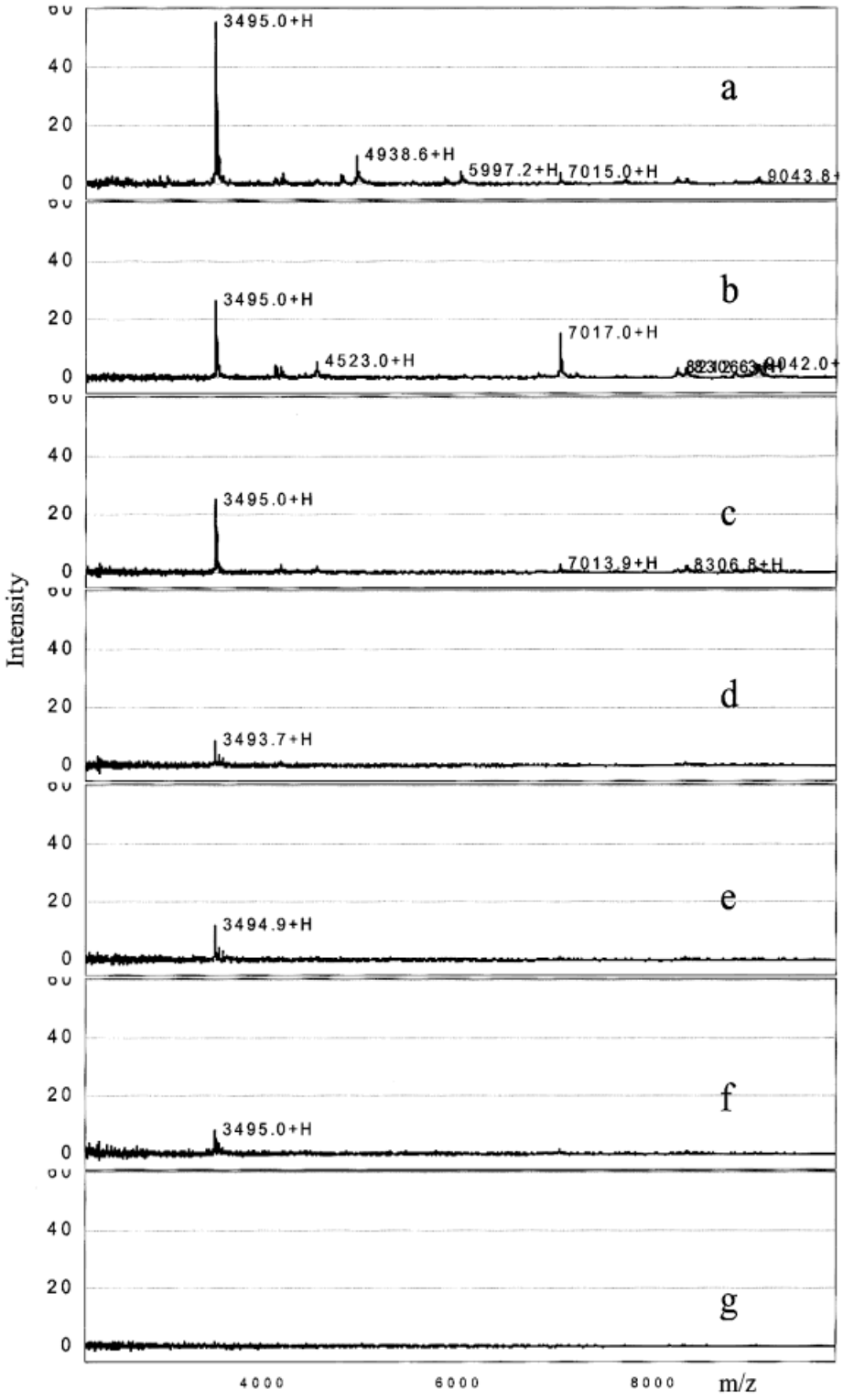

Figure 5. SELDI-guided purification of the 3495 Da protein on a WCX chip. Protein profiles were generated after binding and washing with different binding buffers. (A) 0.2 M ammonium acetate, pH 7; (B) 0.2 M ethanolamine-HCl, pH 9; (C) = 0.2 M ethanolamine-HCl, pH 10; (D) 0.2 M ethanolamine-HCl, pH 10 with 0.1 M KCl; (E) 0.2 M ethanolamine-HCl, pH 10, with 0.2 M KCl; (F) 0.2 M ethanolamine-HCl at pH 10 with 0.5 M KCl; (G) 0.2 M ethanolamine-HCl, pH 10, with 1 M KCl.

equipped with MS/MS capability, we transferred tryptic digests from WCX chips to AnchorChips which is one of the targets used by Ultraflex TOF/TOF mass spectrometer. After tranferring tryptic peptides from WCX chip to AnchorChip, we observed 1001, 1228, and 1791 Da as the most prominent peaks (Fig. 6d) in MS analysis by Ultraflex mass spectrometer. This peptide mass pattern was consistent with the peptide mass fingerprint on WCX chip read with the PBS-II mass reader. Therefore, we chose the three most prominent peptides (1001, 1228, and 1791 Da) for sequencing by MALDI TOF/TOF MS/MS in the Ultraflex mass spectrometer.

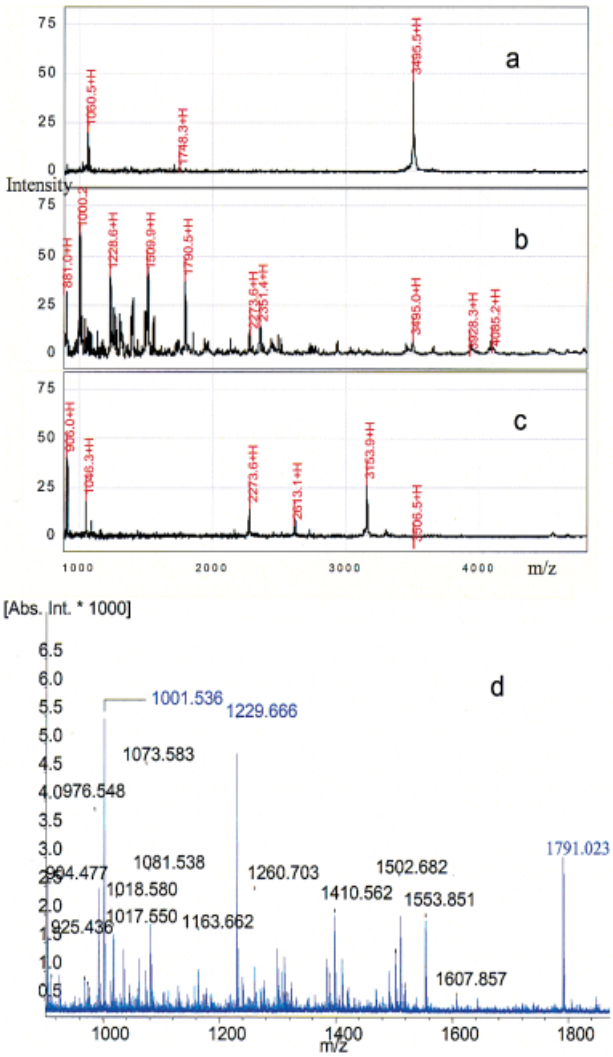

Figure 6. On-chip digestion of the purified 3495 Da protein. (A) Mass spectrum before trypsin digestion on WCX chip. (B) Mass spectrum after 3 h of on-chip trypsin digestion at 37°C. (C) Mass spectrum of trypsin alone after 3 h of on-chip digestion on WCX chip. PMF of tryptic digest after transferring to an AnchorChip and analyzed using an Ultraflex MALDI TOF/TOF mass spectrometer.

As shown in Fig. 7, MALDI-TOF/TOF analysis of the 1001 Da peak generated a short sequence tag GLFPFHQR. BLAST search showed that the sequence was from histidine-rich glycoprotein (accession no. NP_596919) as the top hit in the search of rat proteins. A short sequence tag from the 1791 Da peak also yielded histidine-rich glycoprotein (data not shown). The region around the sequence tag showed a high density of histidine. Furthermore, analysis of the region around the sequence tag under ExPASy-PeptideMass analysis (http://us.expasy.org/tools/peptide-mass.html) showed that the p*I* of this region was above 12. This was consistent with the observation from on-chip purification that this peptide was highly positively charged. However, sequencing of the 1229 Da peptide yielded a sequence tag of QINSQDQLK, which shared homology with glutaredoxin (accession no. XP_229437) with one mismatch. When the region around this sequence tag (total 30 amino acid residues) was analyzed under ExPASy-PeptideMass analysis, the p*I* was only 9.78. This could not account for the binding of the 3495 Da peptide on the chip surface even at pH 10. Taken together, sequencing of two out of the three most prominent peptides showed that the protein was histidine-rich glycoprotein, and the high density of positive charge residues around the sequence tag region suggested that the 3495 Da peak was a fragment of histidine-rich glycoprotein.

4 Discussion

We report here a three-step workflow from protein profiling to significant biomarker selection using machine learning algorithms followed by sequencing of a selected biomarker. We used this workflow to test the hypothesis that serum protein profiles can be utilized to segregate samples from animals with livers which were normal, cirrhotic and suffered prolonged biliary obstruction. Our approach is not only able to classify serum samples with different degrees of liver fibrosis/cirrhosis, but also to identify selected discriminating protein markers. The specificty and sensitivity are higher than 92%. The serum profile segregated the three groups of animals with high accuracy and this concurred with the underlying histological appearance. Most studies on classification of various human diseases [6, 16] use a combination of proteomics and bioinformatics without knowing the identity of significant biomarkers. We further extended the utility of the workflow by identifying each statistically significant biomarker. Clustering of hepatotoxins based on readout in gene expression profiles proved to be a successful tool to reveal mechanism of toxicity [1]. Although this study was conducted using sera from an homogenous group of animals rather than heterogeneous human serum samples,

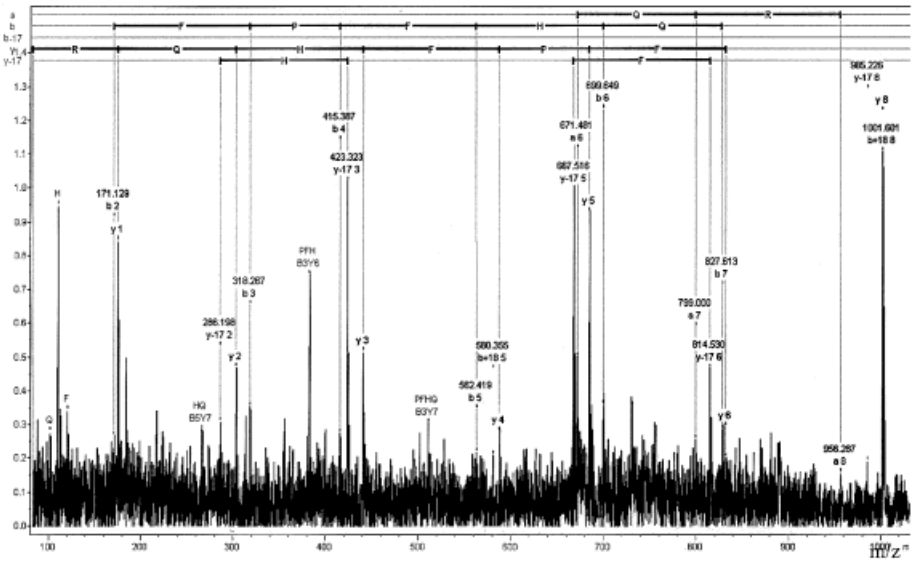

Figure 7. Peptide sequencing of the 1001 Da peak using Ultraflex MALDI TOF/TOF mass spectrometer. X-axis is the *m/z* value; the y-axis is the signal intensity.

with the progress of proteomics and bioinformatics, it may be possible to cluster human serum samples of different pathogenesis and stages of liver cirrhosis/fibrosis.

The specificity and sensitivity in this study are 92%, and 97 to 100%, respectively. It did not reach 100% in each parameter because of the intrinsic complexity of biological objects and the different ways of modeling the data by various machine learning algorithms. Although experiments on rat benefit from a more homogenous genetic background and well controlled experimental conditions, in comparison with using human subjects, there are still various sources of noise and variation. For example, the individual variations, and noise coming from sample preparation and instrument. The RSVM scheme distinguishes itself from another SVM based algorithm, the SVM-RFE method [14], in three aspects: (i) difference in the scheme of cross-validation, (ii) difference in criteria for ranking contributions of a feature to the decision function, and (iii) difference in final important gene list is based on frequency-based selection. These three features of RSVM enhance the accuracy of final marker selection. Other machine learning algorithms such as genetic selection or boosted decision tree have been used in analyses of SELDI data, however, no study has reported 100% in both sensitivity and specificity in spite of algorithm being used [6, 16]. The reason may stem from: (i) differences in protein composition of serum samples; (ii) sensitivity of detection methodology; or (iii) the machine learning algorithm. It will be interesting to see a systematic test of different machine learning algorithms in analyzing several sets of publically available SELDI data.

Thioacetamide is a soft nucleophile that induces up-regulation of tumor necrosis factor-alpha (TNF-α), interleukin 1 beta (IL-1β), superoxide dismutase, and glutathione peroxidase [17]. However, we did not find differentially expressed cytokines or panels of proteins related to oxidative stress (except for glutaredoxin) in this study. This can be accounted for by two reasons: (i) p*I* of these cytokines are below 5.2 and these proteins carry negative charge in pH 5.5 buffer. Binding of negatively charged molecules to the negatively charged surface of chips is inhibited in the assay condition. Secondly, these proteins are all larger than 20 kDa in molecular mass. To resolve these proteins in protein profiles, a strong anion exchange surface and screening at a higher molecular mass range (29 to 200 kDa) sould be used. cDNA microarray analyses of thioacetamide treated HepG2 cell line showed that protein transport machinery and oxidative stress response were down-regulated [19]. Effects of alteration of normal liver secretory pathways induced by thioacetamide on serum protein profiles remain to be studied systematically.

We tried to identify the 3495 Da peptide by using the mass tag for simple database search. We used the TagIdent tool from the ExPASy molecular biology (UTL: http://us.expasy.org/tools/tagident.html). In searches on the Swiss-Prot and TrEMBL protein databases, we found that only glucagon precursor (Swiss-Prot accession no. P06883) or glucagon-like peptide with a mass of 3482.79 Da that barely matched our query. However, when we performed immunoassay with glucagon antibody, the result was negative (data not shown), suggesting that this 3495.38 Da peptide could be a fragmental

peptide from an unkown protein. This down-regulated 3495 Da biomarker in cirrhotic liver is identified by MS/MA analysis of peptide sequence as a fragment of histidine-rich glycoprotein. This finding is consistent with two known facts of this protein: (i) histidine-rich glycoprotein is a plasma protein; and (ii) mRNA of murine histidine-rich glycoprotein is expressed in normal liver exclusively [18]. Thioacetamide has never been reported as a mutagenic agent, so disappearance of histidine-rich glycoprotein in cirrhotic samples should not relate to changes of DNA sequence. Analyses of gene expression changes of HepG2 cell line upon thioacetamide treatment showed overall down-regulation of mitochondrial energy production and other basic cellular functions [19]. Taking these accounts together, we speculate that down-regulation of histidine-rich glycoprotein in cirrhotic liver may be a manifestation of loss of normal liver function, including secretory pathways upon treatment with thioacetamide. Although it was hypothesized that histidine-rich peptide sequence may provide a highly charged area that interacts with complement components in plasma [20], the exact mechanism of origin and function of this 3495 Da fragment of histidine-rich glycoprotein remains to be explored.

The authors of this paper would like to acknowledge the financial support of Agency for Science Technology and Research of Singapore. We would also like to thank Ms. Lih-yin Lim and Mr. Teck Yew Low for technical assistance and Dr. M. Salto-Tellez, consultant pathologist, for examining the liver sections.

5 References

[1] Waring, J. F., Jolly, R. A., Ciurlionis, R., Lum, P. *et al.*, *Toxicol. Appl. Pharmacol.* 2001, *175,* 28–42.
[2] Bruck, R., Shirin, H., Aeed, H., Matas, Z. *et al.*, *J. Hepatol.* 2001, *35,* 457–464.
[3] Vlahou, A., Schellhammer, P. F., Mendrinos, S., Patel, K. *et al.*, *Am. J. Pathol.* 2001, *158,* 1491–1502.
[4] Adam, B. L., Vlahou, A., Semmes, O. J., Wright, G. L. Jr., *Proteomics* 2001, *1,* 1264–1270.
[5] Zhukov, T. A., Johanson, R. A., Cantor, A. B., Clark, R. A., Tockman, M. S., *Lung Cancer* 2003, *40,* 267–279.
[6] Petricoin, E. F., Ardekani, A. M., Hitt, B. A., Levine, P. J. *et al.*, *Lancet* 2002, *359,* 572–577.
[7] Vapnik, V. N., *The Nature of Statistical Learning Theory.* New York, Springer-Verlag, 1995.
[8] Cortes, C., Vapnik, V., *Machine Learing* 1995, *20,* 273–297.
[9] Vapnik, V. N., *Statistical Learning Theory. New York, Wiley, 1998.*
[10] Vapnik, V. N., *IEEE Trans Neutral Networks* 1999, *10,* 988–999.
[11] Collobert, R., *J. Machine Learning Res.* 2001, *1,* 143–160.
[12] Ruwart, M. J., Wilkinson, K. F., Rush, B. D., Vidmar, T. J. *et al.*, *Hepatol.* 1989, *10,* 801–806.
[13] Zhang, X., Wong, W. H., Technical Report, Department of Biostatistics, Harvard School of Public Health, 2001 (http://bisun1.harvard.edu/~xzhang/R-SVM/R-SVM.html)
[14] Guyon, I., Weston, J., Barnhill, S., Vapnik, V., *Machine Learning* 2002, *46,* 389–422.
[15] Li, L., Darden, T., Weinberg, C., Levine, A., Pederson, L., *Combinational Chemistry and High Throughput Screening* 2001, *4,* 727–739.
[16] Qu, Y., Adam, B. L., Yasui, Y., Ward, M. D. *et al.*, *Clin. Chem.* 2002, *48,* 1835–1843.
[17] Akbay, A., Cinar, K., Uzunalimoglu, O., Eranil, S. *et al.*, *Hum. Exp. Toxicol.* 1999, *18,* 669–676.
[18] Hulett, M. D., Parish, C. R., *Immunol. Cell Biol.* 2000, *78,* 280–287.
[19] Gore, M. A., Morshedi, M. M., Reidhaar.-Olson, J. F., *Funct., Integr. Genomics* 2000, *1,* 114–126.
[20] Chang, N. S., Leu, R. W., Rummage, J. A., Anderson, J. K., Mole, J. E., *Blood* 1992, *79,* 2973–2980.

