Available online at www.sciencedirect.com
ScienceDirect
Clinical Biochemistry 40 (2007) 1266–1271
CLINICAL BIOCHEMISTRY

An approach to the characterization of serum low-molecular weight proteins/peptides in liver injury using SELDI–TOF MS and factor analysis

Xiaoli Liu [a], Lanjuan Li [a,*], Guoyi Zhang [b], Guoping Sheng [a], Wei Xu [a]

[a] *State Key Laboratory for Infectious Diseases Diagnosis and Treatment, The First Affiliated Hospital, College of Medicine, Zhejiang University, 79 QingChun Road, Hangzhou, 310003, PR China*
[b] *Department of Mathematics and Statistics, Arizona State University, USA*

Received 3 December 2006; received in revised form 25 April 2007; accepted 24 May 2007
Available online 3 July 2007

Abstract

Objectives: The objective of this work was to investigate if the application of factor analysis to the SELDI–TOF MS data could contribute to the characterization of the serum low-molecular weight proteins/peptides in liver injury.

Design and methods: Distinguished SELDI–TOF mass spectral peaks of the liver injury group were identified by comparing with those of the control group. Factor analysis was introduced to classify these peaks into different groups, assignable to the possible underlying factors. Based on original mass spectral plot, loading and current medical knowledge, the common characteristics of the peaks in same group were revealed and the underlying factors were tentatively defined.

Results: The SELDI profiles of liver injury group exhibit 43 discriminating peaks from the control group. Factor analysis extracted 4 common factors, which were the cholestasis, coagulation, attenuation and 9292 factors. And a plausible interpretation for some undetermined peaks was proposed.

Conclusion: The application of factor analysis to SELDI–TOF MS data extracted valuable information out of complex and high-dimensional mass spectra data.

Keywords: Hepatitis; Proteomics; Mass spectrum; Factor analysis; Low molecular weight

Introduction

Liver injury leads to distinctive alterations in its protein metabolism and synthesis. The changes in proteins such as albumin, alpha-foetal protein (AFP), fibrinogen, and transferrin are observed in patients with the hepatocellular dysfunction. Plasma amino acid levels are severely deranged in acute liver failure patients. A decreasing ratio of branched-chain amino acids to aromatic amino acids is implicated in the pathogenesis of hepatic encephalopathy. Although the changes of large proteins and amino acids are fairly well known, the understanding of the changes of low-molecular weight proteins/peptides in liver injury is poor.

Mass spectrometry (MS) has been a powerful platform for the identification and characterization of proteins. SELDI–TOF MS (surface-enhanced laser desorption ionization–time of flight–mass spectrometry), a MS-based technique, is characterized by the investigation of proteins less than 20 kDa [1,2]. Application of this technique has led to the discovery of many serum low-molecular weight proteins/peptides biomarkers for various diseases, and attempts to sequence and identify these molecules are currently underway [3–8]. However, the characteristics of obtained mass spectral data are complexity and high dimensionality. It is important to address the correlations among peaks and mine information as much as possible from mass spectral data before the further sequence identification of the biomarkers [9].

As a model-based method, factor analysis is often applied as a data reduction or structural detection method. It had been widely used in disciplines as diverse as chemistry, sociology,

* Corresponding author. Fax: +1 86 0571 87236755.
E-mail address: ljli@zjwst.gov.cn (L. Li).

0009-9120/$ - see front matter © 2007 The Canadian Society of Clinical Chemists. Published by Elsevier Inc. All rights reserved.
doi:10.1016/j.clinbiochem.2007.05.019

economics, psychology, education and biological sciences [10–13]. Commonly, the main applications of factor analysis were: (1) to reduce the number of variables and (2) to detect structure in the relationships between variables, which was to classify variables.

In this paper, SELDI–TOF MS was used to detect the serum low-molecular weight proteins/peptides in hepatitis B patients. The peaks associated with liver function impairment were identified by comparing mass spectra of hepatitis group with those of the control group. Factor analysis was then introduced to investigate whether it could contribute to extracting information out of complex and high-dimensional mass spectra data.

Methods

Study population and sample collection

After approval by the local ethics committee, the patients and control cases were recruited at the First Affiliated Hospital of Medical College, Zhejiang University with written consent form. Liver injury group consisted of 52 chronic hepatitis B patients who were hospitalized for acute deterioration in liver function: model for End-Stage Liver Disease (MELD) score <20 (n=23), 20<MELD score<40 (n=24), MELD score>40 (n=5). The control group consisted of blood samples from 22 healthy individuals, taken during physical check-up. All of them had no evidence of the diseases. Demographics of both groups were summarized in Table 1. The gender of the liver injury group was similar to that of the control group.

Fasting blood samples were drawn into standard and citrated tubes for sera and plasma, respectively. The samples were stored at room temperature for 2 h. After centrifugation at 3000×g for 10 min at 4 °C, the serum/plasma was stored at −70 °C prior to the analysis.

SELDI–TOF MS analysis

Various chip chemistries (hydrophobic, anionic, cationic, and metal binding) were initially evaluated. The Q10 chip (a strong anion exchange chip; Ciphergen Biosystems, Freemont, CA, USA) was used throughout the study because it produced

Table 1
Demographic and laboratory data for control group and liver injury B group

	Control group (n=22)	Liver injury group (n=52)
Male, n (%)	16 (73)	40 (77)
Age, years	38 (30, 45)	36 (31, 43)
ALT, U/L	26 (23, 31)	125 (57, 391)
AST, U/L	22 (19, 25)	127 (63, 237)
TB, μmol/L	15 (12, 16)	226 (70, 426)
PT, s	12 (11, 13)	19 (15, 26)
MELD score	/	22 (16, 29)

ALT, alanine aminotransferase; AST, aspartic aminotransferase; TB, total bilirubin; PT, prothrombin time; MELD, Model for End-Stage Liver Disease.
Data are expressed as the number of subjects (percentage) or median values (25, 75 percentile).
MELD score was calculated as $9.57\times\log_e(\text{Cr mg/dL})+3.78\times\log_e(\text{TB mg/dL})+11.20\times\log_e(\text{international normalized ratio})+6.43$.

Table 2
Summary of factor loadings and communalities

m/z	Factor 1	Factor 2	Factor 3	Factor 4	Communalities
The variables for factor 1 (ordered by the loadings on factor 1)					
1698	0.93777	−0.18524	−0.19894	−0.13948	0.97276135
1714	0.93461	−0.19087	−0.20404	−0.14252	0.97186940
2174	0.92230	−0.17397	−0.15937	−0.09244	0.91485676
1890	0.91796	−0.12900	−0.16552	−0.07740	0.89268117
1731	0.91679	−0.22578	−0.13005	−0.12318	0.92357337
2459	0.91597	−0.12507	−0.12798	−0.10381	0.88179147
1874	0.91079	−0.13522	−0.19773	−0.07953	0.89325551
1522	0.86737	−0.18543	−0.19754	−0.16314	0.85236207
1537	0.86484	−0.20889	−0.24917	−0.16470	0.88080158
2475	0.84780	0.01589	−0.15970	−0.16365	0.77129584
1221	0.81102	−0.25975	−0.19120	−0.12264	0.77681340
TB	0.73699	−0.26706	−0.39122	−0.14423	0.78833500
DB	0.70499	−0.26097	−0.42121	−0.11462	0.75566600
3089	0.59186	−0.21764	−0.04938	−0.36868	0.53602868
1065	0.51263	−0.10858	−0.42162	−0.01181	0.45247926
The variables for factor 2 (ordered by the loadings on factor 2)					
4075	−0.10661	0.91778	0.19088	0.11341	0.90297528
4051	−0.02431	0.89752	0.18898	0.19560	0.88009786
3954	−0.10126	0.88615	0.13147	0.09868	0.82253881
3937	−0.02813	0.88168	0.13695	0.14312	0.81739092
5882	−0.28111	0.85299	0.11982	0.14050	0.84071355
5948	−0.33111	0.83938	0.07504	0.23747	0.87621566
2955 (5908, 2H+)	−0.25237	0.83895	0.13857	0.11898	0.80088850
5908	−0.35763	0.80350	0.09090	0.24307	0.84084938
5924	−0.34800	0.79529	0.03113	0.19323	0.79189516
4239	−0.11266	0.74041	0.32722	0.15234	0.69117595
4175	−0.24977	0.69683	0.27689	0.42248	0.80311942
6053	−0.40666	0.69426	0.15375	0.16148	0.69707641
8570	−0.22020	0.66078	0.58889	0.01269	0.83206775
4060	−0.12124	0.62431	0.27157	0.48940	0.71771768
4187	0.01498	0.61284	0.35223	0.33645	0.61306169
4206	−0.20479	0.51648	0.50790	0.39675	0.72406527
The variables for factor 3 (ordered by the loadings on factor 3)					
3318 (6635, 2H+)	−0.24352	0.17284	0.83086	0.06901	0.78426312
6635	−0.34983	0.25142	0.81551	0.14516	0.87172757
4715 (9431, 2H+)	−0.13142	0.06619	0.81040	0.43897	0.87110133
6436	−0.30367	0.17767	0.79499	0.14420	0.77659173
8695	−0.34655	0.45808	0.72484	−0.00609	0.85537095
8924	−0.23852	0.18447	0.70403	0.34340	0.70450967
4348 (8695, 2H+)	−0.34136	0.43120	0.69905	−0.03790	0.79256238
9431	−0.24083	0.06042	0.66564	0.55117	0.80851414
6894	−0.36827	0.30641	0.62158	0.13978	0.63540513
The variables for factor 4 (ordered by the loadings on factor 4)					
9292	−0.07818	0.33827	0.41809	0.68356	0.76259536
9195	−0.17363	0.08952	0.38194	0.68030	0.64684070
4618	−0.12939	0.11368	0.54544	0.65942	0.76200732
4598 (9195, 2H+)	−0.15066	0.09567	0.38903	0.62117	0.56904262
6314	−0.18766	0.18166	0.26453	0.57670	0.47077481
PT	0.26739	−0.34573	−0.04623	−0.60085	0.55418251
AST	0.08776	−0.26147	0.07203	−0.62112	0.46705157
ALT	0.08165	−0.23262	0.13761	−0.62646	0.47216214

more discriminating peaks than the other chips. The Q10 chips were installed in 8-well bioprocessors to allow for a larger volume of serum for the chip array and were equilibrated twice with the binding buffer (100 mmol/L Tris–HCl, pH=9.0). Samples (10 μL) were denatured in U9 solution (20 μL, 9 mmol/L urea, 20 g/L 3-[(3-cholamidopropyl)dimethylammonio]-1-propanesulfonate, and 1% dithiothreitol) at 4 °C for 30 min and were then diluted by 360 μL binding buffer. The diluted samples (100 μL) were added to each well on the chip surfaces. After gentle agitation for 1 h at the room temperature, the chips were washed with binding buffer (3μtimes, 5 min) and 1 mmol/L 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES), pH=9 (2 times, 30 s). The chips were then removed from the processors. CHCA solution (1 μL, 50%) was applied to each spot of the chips twice. The chips were then air-dried at room temperature.

The arrays were analyzed by using a SELDI–TOF mass spectrometer (PBS IIc, Ciphergen Biosystems). The instrument was set as follows: laser intensity, 170; detector sensitivity, 7; mass deflector, 500 Da. The mass accuracy was calibrated externally using the All-in-1 peptide molecular mass standard (Ciphergen Biosystems).

Statistical analysis

The peak intensities of all spectra between masses ranging from 1 kDa to 10 kDa, were normalized according to the total ion current. 237 common peaks among the SELDI–TOF mass spectra were identified, and the peak intensities were compared between the control and liver injury groups by a Ciphergen Biomarker Wizard software. The liver function-associated peaks were obtained with the *p*-value less than 0.00001. These significantly different peaks plus 5 traditional markers of biochemical liver tests, i.e. total bilirubin (TB), direct bilirubin (DB), alanine aminotransferase (ALT), aspartic aminotransferase (AST), and prothrombin time (PT), were considered as variables for the following data analysis.

Factor analysis (FA) was employed to classify the biomarkers into several groups, which belonged to the possible underlying factors. The factor analysis was conducted using SAS 8.2. Three steps were taken: (1) extraction of the initial factors using PCA; (2) orthogonal factor rotation to transform the extracted factors into the interpretable factors; and (3) interpretation based on the rotated factor loadings. In the factor analysis, it is important to decide how many factors should be kept. By comparing the results from 3, 4, and 5 factors, four factors were selected. The final choice was based on the proportions of the sample variance explained (about 80%), the subject matter knowledge, and the "reasonableness" of the results.

Results

The SELDI profiles of liver injury groups exhibited 43 discriminating peaks from the control group. The intensities of 13 peaks were significantly higher and 30 peaks were significantly lower in the liver injury group than that in the control group. These 43 peaks and TB, DB, ALT, AST, and PT were included in the following factor analysis. The factor loadings and communalities of the variables were listed in Table 2.

The four factors explained 76.57% of the total variance. Communalities (most of them were greater than 70%) indicated that the four factors could explain the data without losing too

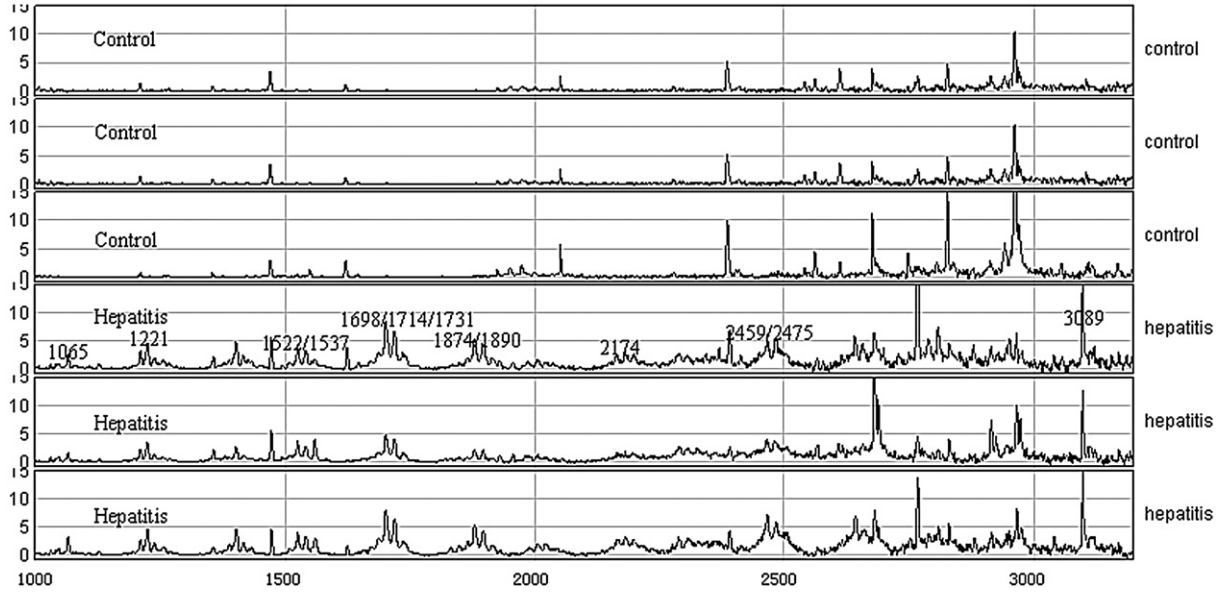

Fig. 1. SELDI–TOF mass spectra of healthy control and hepatitis serum samples (*m/z* 1000–3200). The discriminating peaks were labeled. The distance between clusters of *m/z* 1522/1537 and 1698/1714/1731 is 176 Da. So is the difference between 1698/1714/1731 and 1874/1890. 176 Da is the molecular weight of glucuronide group. Furthermore, *m/z* 1522, 1698, 1874, and 2459 are exactly the molecular weights of different combinations of monoglucuronidated and diglucuronidated bilirubin. These characteristic features strongly suggest that the peaks of *m/z* 1522/1537, 1698/1714/1731, 1874/1890, and 2459/2475 are the micropolymers of bilirubin.

Table 3
Relationship of 6 discriminating clusters and bilirubin

m/z	Comments
1522–1537	761*2
1698–1714–1731	761+937
1874–1890	937*2
2459–2475	761*2+937

The molecular weights of unconjugated bilirubin and bilirubin in the forms of monoglucuronides and diglucuronides are 585 Da, 761 Da, and 937 Da, respectively.

much information. But we should be careful about the variables, which had communalities less than 0.5, such as ALT, AST, *m/z* 1065 and *m/z* 6314 (Table 2). The result suggested that these variables were not suitable for the model assumptions.

The peaks, which had major loadings on factor 1, were shown in Fig. 1. Similar shapes were observed in the discriminating clusters of *m/z* 1522/1537, 1698/1714/1731, and 1874/1890. It was interesting to note that the mass difference between them was 176, which corresponds to the molecular weight of glucuronide group. Furthermore, *m/z* 1522, 1698, and 1874 were exactly the molecular weights of different combination of monoglucuronidated and diglucuronidated bilirubin, respectively (Table 3). The characteristics of *m/z* 1522/1537, 1698/1714/1731, and 1874/1890 peaks indicated that they were the micropolymers of bilirubin (the SELDI mass spectra could include a few components besides the proteins and peptides). Similarly, *m/z* 2459/2475 was assigned to the polymers of bilirubin (Table 3). Thus, we could define factor 1 as the cholestasis factor.

The results of factor analysis suggested that *m/z* 1221 and *m/z* 2174 were also produced by factor 1. As up-regulated peaks, the lower loadings of *m/z* 3089 and *m/z* 1065 from factor analysis suggested that the reason of the rising of *m/z* 3089 and *m/z* 1065 might be different from other variables on the factor 1.

The peaks other than *m/z* 8570, which had major loadings on factor 2, were shown in Fig. 2. The spectra of serum samples were compared with those of the plasma ones. It was noticed that all these peaks except for *m/z* 8570 (Fig. 3) were not observed in the plasma samples of both the liver injury and control groups. In addition, these peaks were down-regulated in the sera of the patients. It strongly indicated that these peaks should relate to the coagulation process. Accordingly, we defined factor 2 as the coagulation factor. *m/z* 8570 had moderate loadings on both factors 2 and 3. From the original plot in Fig. 3, we assume that the peak of *m/z* 8570 belonged to factor 3. The peak of *m/z* 4206 had moderate loadings on factors 2, 3 and 4. This variable was different from the other variables on factor 2.

The variables, which had major loadings on factor 3, were illustrated in Fig. 3. The peak intensities of *m/z* 6436, 6635, 6894, 8695, 8924, and 9431 were all down-regulated in hepatitis serum samples, and they were not related to the coagulation process according to Fig. 3. The factor analysis suggested that they were highly correlated and influenced by the same factor and could be classified as a group. However, their

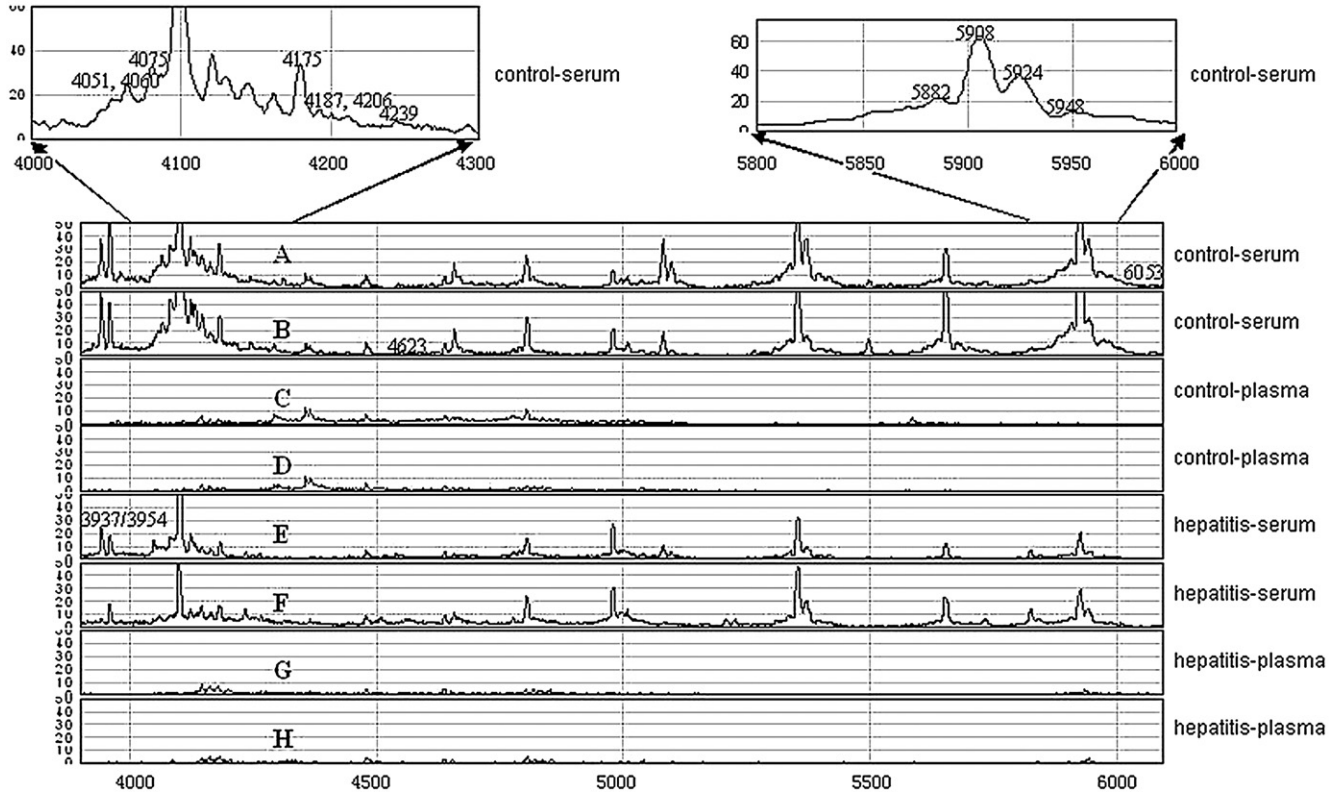

Fig. 2. Comparison of SELDI–TOF mass spectra of control and hepatitis samples (*m/z* 3900–6200): (A, B) serum of healthy controls; (C, D) plasma of healthy controls; (E, F) serum of hepatitis patients; (G, H) plasma of hepatitis patients.

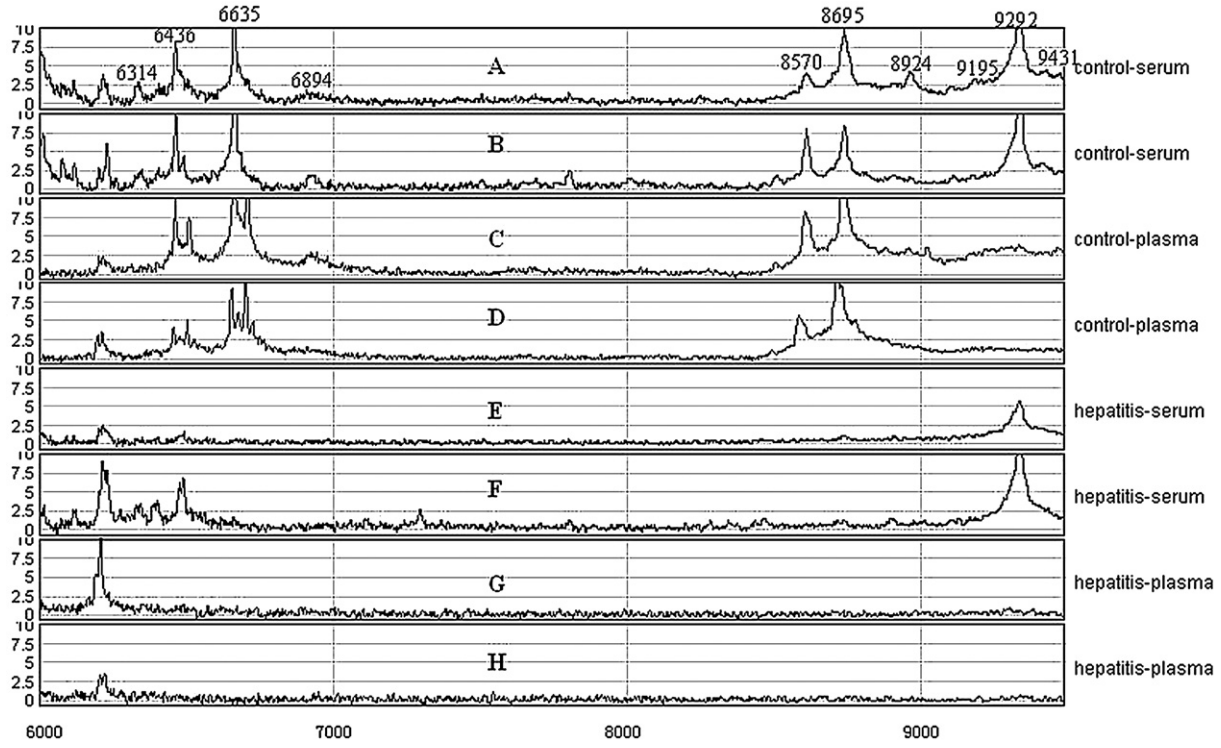

Fig. 3. Comparison of SELDI–TOF mass spectra of control and hepatitis samples (*m/z* 6000–9500): (A, B) serum of healthy controls; (C, D) plasma of healthy controls; (E, F) serum of hepatitis patients; (G, H) plasma of hepatitis patients.

chemical property was still unknown. Therefore, we tentatively define the factor 3 as the attenuation factor.

The variables, which had major loadings on factor 4, were shown in Fig. 2 (*m/z* 4618) and Fig. 3 (*m/z* 6314, 9195, and 9292). It can be seen that the peaks of *m/z* 4618, 6314, 9195, and 9292 were related to the coagulation process. PT, a traditional clotting parameter, had greatest loadings on factor 4. Therefore, similar to factor 2, factor 4 was also related to the coagulation process. Prior to a further exploration, we called factor 4 as 9292 factor since the *m/z* 9292 had a greater loading on factor 4.

Discussion

The liver performs a complex array of critical biochemical, synthetic, and excretory functions. It was understandable that the serum of liver injury patients exhibits 43 discriminating peaks in the SELDI–TOF MS spectra. Facing so many peaks, it was important to address the correlations among peaks and mine information as much as possible from mass spectral data before further sequence identification.

In our study, factor analysis was used to analyze the 43 peaks in SELDI–TOF MS spectra of compounds associated with the liver function impairment and 5 traditional markers of biochemical liver tests. The factor analysis model fit our data very well, and the discriminating peaks were classified into 4 groups belonging to possible underlying factors. Based on original mass spectral plot, loading and current medical knowledge, the common characteristics of the peaks in same group were revealed and 4 underlying factors were tentatively defined. Moreover, a plausible interpretation for some unknown peaks such as *m/z* 1221, *m/z* 2174, and *m/z* 3089 was proposed. Our study revealed that factor analysis was a very useful method to identify peaks produced by the same process in one spectrum and provided a valuable insight into the identity of the unknown peaks.

Recent studies revealed that most of the serum low-molecular weight proteins/peptides were fragments of circulating proteins that have been partially degraded by various enzymes [14–16]. Villanueva et al. identified 61 cancer-type-specific peptides, the results showed that the peptides fell into several tight clusters, and most of them were fragments generated ex vivo by proteinase-mediated enzymatic cleavage as part of the coagulation and complement activation pathways [17]. Our study suggested that many serum low-molecular weight proteins/peptides associated with liver injury were related to coagulation process, which was consistent with the current knowledge about low-molecular weight proteome and the coagulation disorders of liver diseases.

Coagulation is a very complicated process, in which prothrombin is cleaved to thrombin, fibrinogen is removed (to form the clot), and a limited series of other protein changes (mainly proteolytic cleavages) take place. Considering the complexity of the coagulation process, it was reasonable that both factors 2 and 4 (defined as coagulation factor and 9292 factor, respectively) were related to the coagulation disorders in liver injury. However, the difference between them was unclear, which would have to be further investigated.

In conclusion, factor analysis method was introduced to analyze the SELDI–TOF MS data of liver injury samples. Based

on original mass spectral plot, loading and current medical knowledge, the peaks produced by the same process were identified and 4 common factors were extracted. Moreover, a plausible interpretation for some undetermined peaks was proposed. Our study revealed that factor analysis was capable of extracting valuable information out of complex and high-dimensional mass spectra data. Of course, the current work was largely descriptive; the application of factor analysis to the MS data could not substitute sequence identification. Without sequence identification of the proteins/peptides, definite conclusions regarding the nature of the peaks and factors are not possible.

Acknowledgments

This work was supported by Hi-Tech Research and Development Program of China (863 Program) No. 2003AA205150 and National Basic Research Program of China (973 Program) No. 2003CB515506.

References

[1] Tang N, Tornatore P, Weinberger SR. Current developments in SELDI affinity technology. Mass Spectrom Rev 2004;23:34–44.
[2] Rogers MA, Clarke P, Noble J, et al. Proteomic profiling of urinary proteins in renal cancer by surface enhanced baser desorption ionization and neural-network analysis: identification of key issues affecting potential clinical utility. Can Res 2003;63:6971–83.
[3] Petricoin EF, Ardekani AM, Hitt BA, et al. Use of proteomic patterns in serum to identify ovarian cancer. Lancet 2002;359:572–7.
[4] Adam BL, Qu Y, Davis JW, et al. Serum protein fingerprinting coupled with a pattern matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and health men. Cancer Res 2002;62: 3606–14.
[5] Zhang L, Yu W, He T, et al. Contribution of human alpha-defensin 1, 2, and 3 to the anti-HIV-1 activity of CD8 antiviral factor. Science 2002;298: 995– 1000.
[6] Marshell J, Kupchak P, Zhu W, et al. Processing of serum proteins underlies the mass spectral fingerprinting of myocardial infarction. J Proteome Res 2003;2:361–72.
[7] Schaub S, Rush D, Wilkins J, et al. Proteomic-based detection of urine proteins associated with acute renal allograft rejection. J Am Soc Nephrol 2004;15:219–27.
[8] Diamandis EP. Mass spectrometry as a diagnostic and a cancer biomarker discovery tool. Opportunities and potential limitations. Mol Cell Proteome 2004;3:367–78.
[9] White CN, Chan DW, Zhang Zhen. Bioinformatics strategies for proteomic profiling. Clin Biochem 2004;37:636–41.
[10] Goodman E, Dolan LM, Morrison JA, Daniels SR. Factor analysis of clustered cardiovascular risks in adolescence: obesity is the predominant correlate of risk among youth. Circulation 2005;111:1970–7.
[11] Statheropoulos M, Mikedi K, Tzamtzis N, Pappa A. Application of factor analysis for resolving thermogravimetric–mass spectrometric analysis spectra. Anal Chim Acta 2002;461:215–27.
[12] Holden RR, DeLisle MM. Factor analysis of the beck scale for suicide ideation with female suicide attempters. Assessment 2005;12: 231–8.
[13] Schatz M, Mosen D, Apter AJ, et al. Relationships among quality of life, severity, and control measures in asthma: an evaluation using factor analysis. J Allergy Clin Immunol 2005;115:1049–55.
[14] Richter R, Schulz-Knappe P, Schrader M, et al. Composition of the peptide fraction in human blood plasma: database of circulating human peptides. Chromatogr, B, Biomed Sci Appl 1999;726:25–35.
[15] Liotta LA, Petricoin EF. Serum peptidome for cancer detection: spinning biologic trash into diagnostic gold. J Clin Invest 2006;116:26–30.
[16] Tirumalai RS, Chan KC, Prieto DA, Issaq HJ, Conrads TP, Veenstra TD. Characterization of the low molecular weight human serum proteome. Mol Cell Proteome 2003;2:1096–103.
[17] Villanueva J, Shaffer DR, Philip J, Chaparro CA, ErdjumentBromage H, Olshen AB. Differential exoprotease activities confer tumor-specific serum peptidome patterns. J Clin Invest 2006;116:271–84.