Pacific Symposium on Biocomputing 13:216-227(2008)

ANALYSIS OF MALDI-TOF MASS SPECTROMETRY DATA FOR DETECTION OF GLYCAN BIOMARKERS

HABTOM W. RESSOM[1†], RENCY S VARGHESE[1], LENKA GOLDMAN[1], CHRISTOPHER A LOFFREDO[1], MOHAMED ABDEL-HAMID[2], ZUZANA KYSELOVA[3], YEHIA MECHREF[3], MILOS NOVOTNY[3], RADOSLAV GOLDMAN[1]

[1]*Georgetown University, Lombardi Comprehensive Cancer Center, Washington, DC*

[2]*Minia University and Viral Hepatitis Research Laboratory, NHTMRI, Cairo, Egypt*

[3]*National Center for Glycomics and Glycoproteomics, Department of Chemistry, Bloomington, IN*

We present a computational framework for analysis of MALDI-TOF mass spectrometry data to enable quantitative comparison of glycans in serum. The proposed framework enables a systematic selection of glycan structures that have good generalization capability in distinguishing subjects from two pre-labeled groups. We applied the proposed method for a biomarker discovery study that involves 203 participants from Cairo, Egypt; 73 hepatocellular carcinoma (HCC) cases, 52 patients with chronic liver disease (CLD), and 78 healthy individuals. Glycans were enzymatically released from proteins in serum and permethylated prior to mass spectrometric quantification. A subset of the participants (35 HCC and 35 CLD cases) was used as a training set to select global and subgroup-specific peaks. The peak selection step is preceded by peak screening, where we eliminate peaks that seem to have association with covariates such as age, gender, and viral infection based on the 78 spectra from healthy individuals. To ensure that the global peaks have good generalization capability, we subjected the entire spectral preprocessing and peak selection step to a cross-validation; a randomly selected subset of the training set was used for spectral preprocessing and peak selection in multiple runs with resubstitution. In addition to global peak identification method, we describe a new approach that allows the selection of subgroup-specific glycans by searching for glycans that display differential abundance in a subgroup of patients only. The performance of the global and subgroup-specific peaks is evaluated via a blinded independent set that comprises of 38 HCC and 17 CLD cases. Further evaluation of the potential clinical utility of the selected global and subgroup-specific candidate markers is needed.

1. Introduction

Current diagnosis of hepatocellular carcinoma (HCC) relies on clinical information, liver imaging, and measurement of serum alpha-fetoprotein (AFP). The reported sensitivity (41-65%) and specificity (80-94%) of AFP is not sufficient for early diagnosis and additional markers are needed [1, 2].

Mass spectrometry (MS) provides a promising strategy for biomarker discovery. The feasibility of MS-based proteomic analysis to distinguish HCC

† Corresponding author

from cirrhosis, particularly in patients with hepatitis C virus (HCV) infection, has been studied [3-6]. Recent proteomic studies have identified potential markers of HCC including complement C3a [7], kappa and lambda immunoglobulin light chains [8], and heat-shock proteins (Hsp27, Hsp70, and GRP78) [9].

Many currently used cancer biomarkers including AFP are glycoproteins [10]. Fucosylated AFP was introduced as a marker of HCC with improved specificity [11, 12] and other glycoproteins including GP73 are currently under evaluation as markers of HCC [13, 14]. The analysis of protein glycosylation is particularly relevant to liver pathology because of the major influence of this organ on the homeostasis of blood glycoproteins [15, 16]. An alternative strategy to the analysis of glycoproteins is the analysis of protein associated glycans [17, 18]. The characterization of glycans in serum of patients with liver disease is a promising strategy for biomarker discovery [19].

Current methods allow quantitative comparison of permethylated glycan structures by matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) MS [20], which provide a rich source of information for molecular characterization of the disease process. Although MALDI-TOF MS continuously improves in sensitivity and accuracy, it is characterized by its high dimensionality and complex patterns with substantial amount of noise. Biological variability and disease heterogeneity in human populations further complicate the MALDI-TOF MS-based biomarker discovery. While various signal processing methods have been used to reduce technical variability caused by sampling or instrument error, reducing non-disease-related biological variability remains a challenging task. For example, peaks associated to known covariates such as age, gender, smoking status, and viral infection should be eliminated; we call this preprocessing step *peak screening* [5]. In addition, robust computational methods are needed to minimize the impact of biological variability caused by unknown intrinsic biological differences.

In this paper, we present computational methods for analysis of MALDI-TOF MS to discover glycan biomarkers for the detection of HCC in patients with chronic liver disease (CLD), consisting of fibrosis and cirrhosis patients [21, 22]. The objective is to improve the diagnostic capability of a panel of "whole population" level (global) biomarkers and to investigate the extraction of subgroup-specific biomarkers that are more patient specific than the global markers. Our proposed approach involves the following two steps.

The first step searches for a panel of global peaks that distinguishes HCC from CLD at the whole population level by treating all HCC patients as one group [4, 5]. We utilize a computational method that combines ant colony optimization and support vector machine (ACO-SVM), previously described in

[5], to identify the most useful global peaks. Although these peaks may include peaks that may be attributed to subgroups of patients, neither the subgroup-specific peaks nor the subgroups are likely to be isolated due to the unknown (mostly nonlinear) interaction of the global peaks.

The second step uses a genetic algorithm (GA) to search for subgroup-specific peaks and to discover subgroups of subjects from the training set. The disease state of an unknown individual is determined by the SVM classifier built in the first step. Then, the subgroup to which the individual belongs will be determined by comparing its intensity with each of the subgroup-specific peaks defined in the second step.

The proposed hybrid method will provide the ability to capture glycans that are differentially abundant in only a subset of patients in addition to those that are differentially abundant glycans at the whole population level. This will allow us to not only identify a panel of useful global peaks that lead to good generalization, but also to offer a more patient-specific approach for the identification of glycan biomarkers.

2. Methods

2.1. *Sample collection*

HCC cases and controls were enrolled in collaboration with the National Cancer Institute of Cairo University, Egypt, from 2000 to 2002, as described previously [22]. Briefly, adults with newly diagnosed HCC aged 17 and older without a previous history of cancer were eligible for the study. Diagnosis of HCC was confirmed by pathology, cytology, imaging (CT, ultrasound), and serum AFP. Controls were recruited from the orthopedic department of Kasr El Aini Faculty of Medicine, Cairo University [22]. 17 HCC cases were classified as early (Stage I and II) and 33 HCC cases as advanced (Stage III and IV) according to the staging system [23]; for the remaining 23 HCC cases the available information was not sufficient to assign the stage. Patients with CLD were recruited from Ain Shams University Specialized Hospital and Tropical Medicine Research Institute, Cairo, Egypt during the same period. The CLD group has a biopsy confirmed 21 fibrosis and 25 cirrhosis patients; 6 individuals in the CLD group did not have sufficient clinical information. Patients negative for hepatitis B virus (HBV) infection, positive for HCV RNA, and with AFP less than 100 mg/ml were selected for the study. Blood samples were collected by a trained phlebotomist each day around 10 am and processed within a few hours according to a standard protocol. Aliquots of sera were frozen at -80 ^{o}C immediately after collection until analysis; all mass spectrometric measurements were performed on twice-thawed sera. Each patient's HBV and HCV viral infection status was

assessed by enzyme immunoassay for anti-HCV, anti-HBC, and HBsAg, and by PCR for HCV RNA [22, 24].

2.2. *Sample preparation and MS data generation*

The sample preparation involved release of N-glycans from glycoproteins, extraction of N-glycans, and solid-phase permethylation as described previously [20]. The resulting permethylated glycans were spotted on a MALDI plate with DHB-matrix, MALDI plate was dried under vacuum, and mass spectra were acquired using a 4800 MALDI TOF/TOF Analyzer (Applied Biosystems Inc., Framingham, MA) equipped with a Nd:YAG 355-nm laser as described previously [17]. MALDI-spectra were recorded in positive-ion mode, since permethylation eliminates the negative charge normally associated with sialylated glycans. [25]. 203 raw spectra were exported as text files for further analysis [a]. Each spectrum consisted of approximately 121,000 m/z values with the corresponding intensities in the mass range of 1,500-5,500 Da.

2.3. *Global peak selection*

Figure 1 illustrates our approach for global peak selection, which begins by splitting the spectra into a labeled set and a blinded set. The labeled set consists of a subset of HCC cases, a subset of CLD cases, and all healthy individuals (normal). The blinded set comprises of masked HCC and CLD cases; it is used to evaluate the generalization capability of the selected peaks. Peak detection, peak screening, and peak selection are performed on the labeled set by subjecting the entire process to cross-validation. As illustrated in Figure 1, a subset of the labeled HCC and CLD spectra (~70% from each group) is randomly selected at each iteration as a training set, while the remaining HCC and CLD spectra are used as validation set. A spectrum in the training set is considered as an outlier, if its record count is more than two standard deviations away from the median record count of the spectra within the training set. Outliers are removed from the subsequent analyses. Each spectrum in the training set is binned, baseline corrected, and normalized as described previously [5]. After scaling the peak intensities to an over all maximum intensity of 100, local maximum peaks above a specified threshold are identified and peaks that fall within a pre-specified mass are coalesced into a single m/z window to account for drift in m/z location. The maximum intensity in each window is used as the variable of interest. The threshold intensity for peak detection is selected so that isotopic clusters are represented by a single peak.

---

[a] These files are available at http://microarray.georgetown.edu/web/files/psb.zip

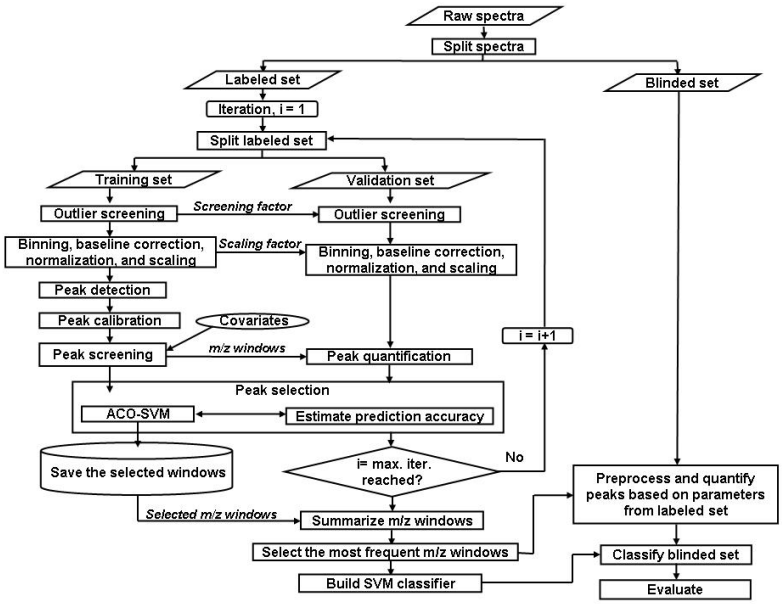

Figure 1. Methodology for global peak detection.

Logistic regression models are used to examine association of the glycans to known covariates including age, gender, smoking status, residency, HCV and HBV viral infections. This analysis is performed on the samples from healthy individuals to unambiguously isolate peaks associated to the covariates. The independent variables of a logistic regression model are the intensities of a given peak across all normal samples. The dependent variable is the status of a given covariate; all covariates in this study have binary values including age (young vs. old). The association of every peak to each covariate was determined on the basis of the corresponding statistical significance ($p<0.01$) in fitting a logistic regression model. Glycan intensities associated to the covariates are removed. From the remaining peaks, ACO-SVM selects the best peaks in terms of their ability to distinguish a subset of the HCC and CLD spectra in the validation set, which was not involved in the peak selection process. The spectra in the validation set are screened for outliers, binned, baseline corrected, normalized, and scaled on the basis of the parameters used to preprocess the spectra in the training set. The peaks in the validation set are quantified at the selected m/z windows and are presented to SVM classifier previously trained using the peaks from the training set. The performance of the SVM classifier in predicting the disease state of the subjects in the validation set is used by ACO-SVM to guide

its search for the optimal peak set. The above steps are repeated multiple times by randomly splitting the labeled spectra into training and validation sets.

The peaks selected in multiple runs are summarized to determine the most frequently selected m/z windows. Note that the number of peaks detected and their m/z windows could vary at each iteration due to the change in the population set in each iteration. After obtaining all peaks selected in multiple iterations, we summarize the peaks by merging overlapping m/z windows. The optimal peak set is determined based on the frequency of occurrence of the peaks in multiple runs.

To evaluate the peak selection process further, we quantify the glycan intensities at the m/z windows of the optimal peak set in the labeled and blinded sets. Note that the blinded set is not used during the peak detection and peak selection phases, thus it serves as an independent set to evaluate the generalization capability of the selected peaks. The spectra in the blind set are outlier screened, binned, baseline corrected, normalized, and scaled on the basis of parameters used to preprocess the spectra in the labeled set. We build an SVM using the labeled set and evaluate the capability of the SVM classifier in distinguishing HCC from CLD in the blinded set in terms of sensitivity, specificity, and area under the ROC (AuROC).

2.4. *Identification of subgroup-specific peaks*

Figure 2 illustrates our proposed method to identify subgroup-specific peaks by searching for peaks that are differentially abundant in a subset of patients. The method is described here in two phases: training and operation phase.

In the training phase, for each candidate peak we search a subgroups of HCC cases in which the peak is differentially abundant. The candidate peaks are the summarized peak set from the global peak selection process. Note that this peak list includes each summarized peak regardless of its frequency of occurrence. We apply GA to search the optimal subgroup of patient for each candidate peak. A chromosome in the GA assigns a binary bit for each HCC patient in the labeled set (“1” for a patient selected in the subgroup, “0” otherwise). The algorithm starts with randomly selected binary bits. GA evolves the chromosomes with the aim of maximizing a multi-objective fitness function, which involves two parameters (1) the AuROC obtained in using the peak to separate a selected subgroup of HCC patients from patients with CLD and (2) the number of HCC patients involved in the subgroup. The goal is to search for a peak and a subgroup that not only display good separation between the HCC subgroup and patients with CLD, but also assign a reasonable number of subjects to the subgroup. During the operation phase, the label of a spectrum from the

blinded set is predicted by an SVM classifier previously built using the global peaks in the labeled set. If the predicted label is HCC, its glycan intensities will be compared with the subgroup-specific peaks to determine which subgroup the individual belongs to. The subject is assigned to a previously identified subgroup, if its peaks intensity falls within the subgroup's intensity range.

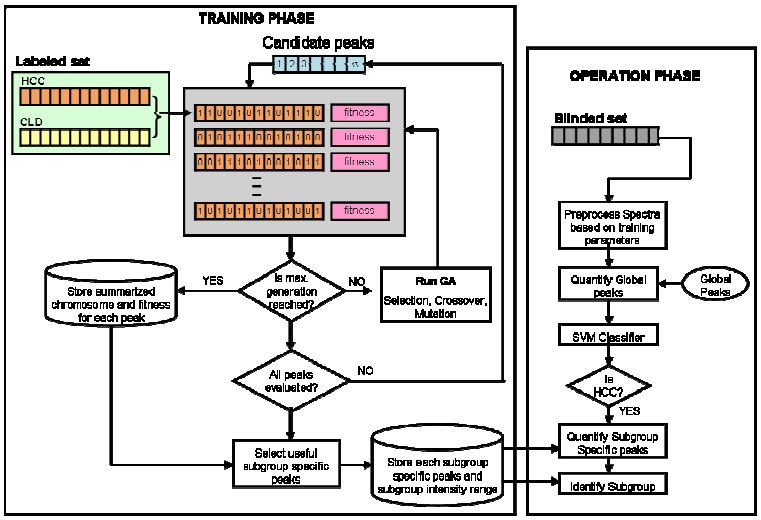

Figure 2: Methodology for subgroup-specific peak selection.

3. Results

MALDI-TOF mass spectrometric analysis of permethylated N-glycans enzymatically detached from serum proteins allowed relative quantification of about 100 oligosaccharides. We analyzed serum samples from 203 participants. Glycan analysis was performed as described previously [17, 20].

Spectral preprocessing and global peak detection were carried out following the methodology depicted in Figure 1. Briefly, we began the analysis by splitting the raw spectra into labeled set (35 HCC, 35 CLD, and 78 normal) and a blinded set (38 HCC and 17 CLD). From the labeled set, 25 HCC and 25 CLD spectra were randomly selected as a training set; the remaining 10 HCC and 10 CLD spectra were used as a validation set. Outlier screening was performed on the training set to determine if its record count of each spectrum is within two standard deviations from the median record count for the spectra within the training set. Outlier spectra were removed from the subsequent analyses. A binning algorithm reduced the dimension of each of these spectra from ~121,000 to 13,030 using a bin size of 100 ppm. The mean of the intensities within each

bin was used as the bin intensity. For each binned spectrum, we estimated the baseline by obtaining the minimum value within a shifting window size of 50 bins and a step size of 50 bins. Spline approximation was applied to regress the varying baseline. The regressed baseline was smoothed using the lowess smoothing method. The resulting baseline was subtracted from the spectrum. Then, each spectrum was normalized by dividing it by its total ion current. The spectra were scaled to have a maximum intensity of 100. Local maximum peaks above a specified threshold are identified and nearby peaks within 300 ppm mass separation are coalesced into a single m/z window and the maximum intensity in each window is used as the variable of interest. We adjusted the threshold intensity and the mass separation so that isotopic clusters resolved by the high resolution reflectron acquisition were represented by one glycan peak. The isotopic cluster at 1543-1547 Da was the only cluster resolved by the procedure to three individual peaks; we grouped this cluster to one variable prior to subsequent analyses. This procedure resulted in about 100 m/z windows. After performing peak screening on the basis of the 78 normal spectra, about 20 peaks were removed. From the remaining peaks, ACO-SVM algorithm selected the three most useful peaks. The capability of these peaks to predict the labels of the spectra in the validation set was used by ACO-SVM to search for the optimal peak set. The spectra in the validation set were preprocessed in the same way as the training set. For outlier screening and scaling, the parameters used by the training set were utilized. The intensity values within the detected windows were quantified and the maximum intensities within the windows were used as input to SVM classifier built previously using the training set. The above procedure was repeated 2000 times by randomly selecting (with resubstitution) 25 HCC and 25 CLD spectra from the labeled set as a training set and using the remaining 10 HCC and 10 CLD spectra as a validation set.

The peaks selected in 2000 runs were summarized by merging overlapping windows. Figure 3 depicts a frequency plot of the summarized 66 peaks (m/z windows). As shown in the figure, two m/z windows dominated the selection, where the first and second m/z windows were selected in 76% and 35% of the runs, respectively. We quantified the peaks in the labeled set (35 HCC and 35 CLD spectra) within these two summarized windows and applied the maximum intensity values within the windows to build an SVM classifier.

To evaluate the performance of the SVM classifier, we preprocessed the spectra in the blinded set in same way as the training set and quantified the glycan intensities within the selected two summarized windows. These intensities were presented to the previously built SVM classifier, which predicted the samples with 95% sensitivity and 100% specificity; two HCC subjects in the blinded set were wrongly classified as CLD. For comparison, we repeated the

entire peak selection process (Figure 1) by replacing ACO-SVM with the SVM-recursive feature elimination (SVM-RFE) method [26]. Comparing the top 10 peaks in both methods, we observed five overlaps; the peak with the highest frequency was the same in both methods. The top 10 m/z windows in both methods gave 95% sensitivity and 100% specificity in classifying the samples in the blinded set (both methods wrongly classified the same HCC subjects as CLD). However, the top two m/z windows in SVM-RFE (selected in 87% and 30% of the runs, respectively, frequency plot not shown here) distinguished the HCC cases from CLD with 92% sensitivity and 94% specificity in the blind validation set; 1 CLD patient and 3 HCC cases were misclassified.

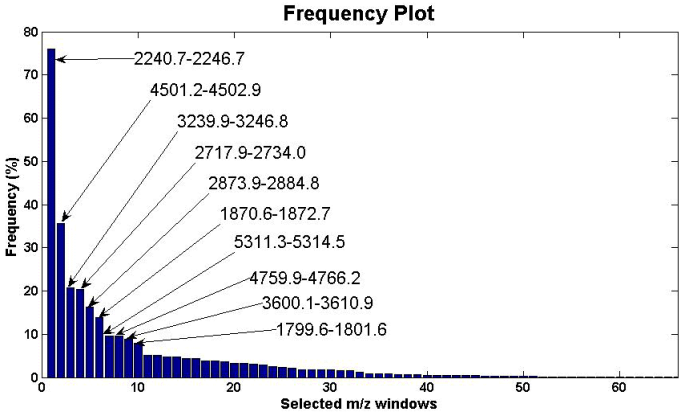

Figure 3. Frequency of occurrence of peaks selected by ACO-SVM in 2000 runs.

Glycan structures for nearly 50% of the peaks detected by the MALDI-TOF MS were determined. Out of 10 peaks selected by ACO-SVM, five have a known sugar composition. Similarly, five out of 10 peaks selected by SVM-RFE have known composition. Figure 4 depicts an overlay of the average HCC and CLD spectra. The five peaks, with known composition, identified by ACO-SVM are shown in the figure; four of these were also among the top 10 peaks selected by SVM-RFE. These five peaks yielded 87% sensitivity and 100% specificity in distinguishing HCC cases from CLD patients in the blinded set.

Finally, we used the methodology illustrated in Figure 2 to identify subgroup-specific peaks from the 66 peaks summarized from the previous 2000 ACO-SVM runs; all summarized peaks are considered as candidate peaks regardless of the frequency of occurrence in the 2000 runs. The subgroup-specific peak selection method identified four peaks that represent four subgroups (S1, S2, S3, and S4) consisting of 23, 21, 17, and 15 HCC patients, respectively. These four peaks were particularly selected, because they had

better AuROC, more number of HCC patients in the subgroup they represent, and less number of overlapping subjects than other candidate peaks.

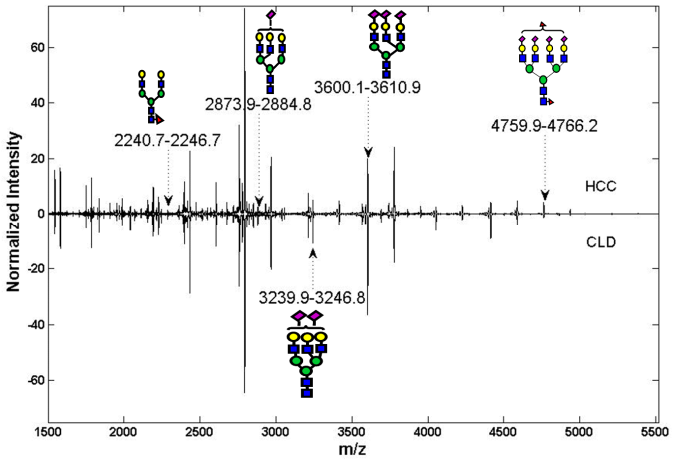

Figure 4. Mean HCC and CLD spectra and sugar composition of five peaks selected by ACO-SVM.

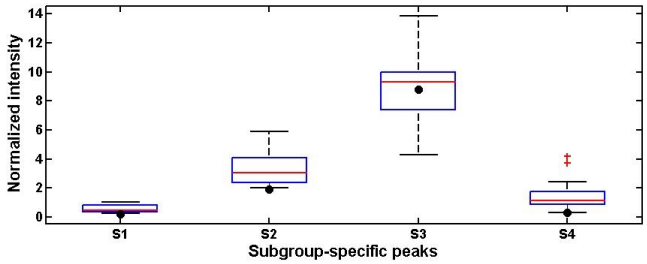

Figure 5. Box plots of peak intensities for four HCC subgroups. Dots represent glycan intensities of a blinded sample detected as HCC by a panel of global peaks. The intensity of the sample falls within the range of the peak for subgroup S3.

Figure 5 depicts box plots for the glycan intensity levels of the four subgroup-specific peaks in their respective subgroups of subjects. Note that only intensities of the subjects that belong to the subgroup the peak represents are shown by the box plots. We considered a subject from the blinded set that was correctly predicted as HCC case by the global peaks. Figure 5 shows the glycan intensities of this subject at the four subgroup-specific peaks (dots in the figure). These intensities are compared with the peak intensity distribution (box plot) of the four subgroups of HCC patients that the peaks represent. From the figure, we see that the HCC patient can be assigned to the subgroup labeled as S3.

4. Discussion

This paper introduces computational methodologies for quantitative comparison of glycans in serum and selection of biomarkers of hepatocellular carcinoma. Candidate glycan biomarkers were obtained by comparing MALDI-TOF spectra of permethylated glycan structures derived from HCC and CLD patient sera. Prior to peak selection, we removed peaks associated to covariates such as age, gender, residency, smoking, and viral infection. We showed that the algorithm has the ability to select a small set of glycan peaks that achieve high sensitivity and specificity in distinguishing HCC cases from patients with CLD in Cairo, Egypt. In addition, we proposed a method that can potentially discover subgroups of patients by searching for subgroup-specific peaks that are differentially abundant in a subset of patients only. Further analysis is needed to determine the implication of the subgroups of subjects and the subgroup-specific biomarkers. It will be interesting to see if the subgroups represent different disease stages or molecular pathways. In addition, the potential clinical utility of the selected candidate markers needs to be evaluated through independent laboratory experiments.

5. Acknowledgments

This work was supported in part by seed grants awarded to HWR and RG from NCI's Early Detection Research Network, NCI grant R03 CA119313 to HWR, and NCI grants R03 CA119288 and R01 CA115625 to RG.

References

1. J. A. Marrero: *Clin Liver Dis* 2005, **9**(2):235-251, vi.
2. S. Gupta, S. Bent, J. Kohlwes: *Ann Intern Med* 2003, **139**(1):46-50.
3. E. Orvisky, S. K. Drake, B. M. Martin, M. Abdel-Hamid, H. W. Ressom, R. S. Varghese, Y. An, D. Saha, G. L. Hortin, C. A. Loffredo *et al*: *Proteomics* 2006, **6**(9):2895-2902.
4. H. W. Ressom, R. S. Varghese, M. Abdel-Hamid, S. Abdel-Latif Eissa, D. Saha, L. Goldman, E. F. Petricoin, T. P. Conrads, T. D. Veenstra, C. A. Loffredo *et al*: *Bioinformatics* 2005, **21**(21):4039-4045.
5. H. W. Ressom, R. S. Varghese, S. K. Drake, G. L. Hortin, M. Abdel-Hamid, C. A. Loffredo, R. Goldman: *Bioinformatics* 2007, **23**(5):619-626.
6. E. E. Schwegler, L. Cazares, L. F. Steel, B. L. Adam, D. A. Johnson, O. J. Semmes, T. M. Block, J. A. Marrero, R. R. Drake: *Hepatology* 2005, **41**(3):634-642.
7. I. N. Lee, C. H. Chen, J. C. Sheu, H. S. Lee, G. T. Huang, D. S. Chen, C. Y. Yu, C. L. Wen, F. J. Lu, L. P. Chow: *Proteomics* 2006, **6**(9):2865-2873.

8. D. G. Ward, Y. Cheng, G. N'Kontchou, T. T. Thar, N. Barget, W. Wei, L. J. Billingham, A. Martin, M. Beaugrand, P. J. Johnson: *Br J Cancer* 2006, **94**(2):287-292.
9. J. M. Luk, C. T. Lam, A. F. Siu, B. Y. Lam, I. O. Ng, M. Y. Hu, C. M. Che, S. T. Fan: *Proteomics* 2006, **6**(3):1049-1057.
10. J. A. Ludwig, J. N. Weinstein: *Nat Rev Cancer* 2005, **5**(11):845-856.
11. K. Taketa, Y. Endo, C. Sekiya, K. Tanikawa, T. Koji, H. Taga, S. Satomura, S. Matsuura, T. Kawai, H. Hirai: *Cancer Res* 1993, **53**(22):5419-5423.
12. K. Shiraki, K. Takase, Y. Tameda, M. Hamada, Y. Kosaka, T. Nakano: *Hepatology* 1995, **22**(3):802-807.
13. J. A. Marrero, P. R. Romano, O. Nikolaeva, L. Steel, A. Mehta, C. J. Fimmel, M. A. Comunale, A. D'Amelio, A. S. Lok, T. M. Block: *J Hepatol* 2005, **43**(6):1007-1012.
14. M. A. Comunale, M. Lowman, R. E. Long, J. Krakover, R. Philip, S. Seeholzer, A. A. Evans, H. W. Hann, T. M. Block, A. S. Mehta: *J Proteome Res* 2006, **5**(2):308-315.
15. G. A. Turner: *Clin Chim Acta* 1992, **208**(3):149-171.
16. S. J. Lee, S. Evers, D. Roeder, A. F. Parlow, J. Risteli, L. Risteli, Y. C. Lee, T. Feizi, H. Langen, M. C. Nussenzweig: *Science* 2002, **295**(5561):1898-1901.
17. Z. Kyselova, Y. Mechref, M. M. Al Bataineh, L. E. Dobrolecki, R. J. Hickey, J. Vinson, C. J. Sweeney, M. V. Novotny: *J Proteome Res* 2007, **6**(5):1822-1832.
18. J. Zhao, W. Qiu, D. M. Simeone, D. M. Lubman: *J Proteome Res* 2007, **6**(3):1126-1138.
19. N. Callewaert, H. Van Vlierberghe, A. Van Hecke, W. Laroy, J. Delanghe, R. Contreras: *Nat Med* 2004, **10**(4):429-434.
20. P. Kang, Y. Mechref, I. Klouckova, M. V. Novotny: *Rapid Commun Mass Spectrom* 2005, **19**(23):3421-3428.
21. O. Nada, M. Abdel-Hamid, A. Ismail, L. El Shabrawy, K. F. Sidhom, N. M. El Badawy, F. A. Ghazal, M. El Daly, S. El Kafrawy, G. Esmat *et al*: *J Clin Virol* 2005, **34**(2):140-146.
22. S. Ezzat, M. Abdel-Hamid, S. A. Eissa, N. Mokhtar, N. A. Labib, L. El-Ghorory, N. N. Mikhail, A. Abdel-Hamid, T. Hifnawy, G. T. Strickland *et al*: *Int J Hyg Environ Health* 2005, **208**(5):329-339.
23. *AJCC Cancer Staging Manual, 6th Edition American College of Surgeons Philadelphia, Lippincott-Raven* 2002.
24. M. Abdel-Hamid, D. C. Edelman, W. E. Highsmith, N. T. Constantine: *J Hum Virol* 1997, **1**(1):58-65.
25. Y. Mechref, M. V. Novotny: *Anal Chem* 1998, **70**(3):455-463.
26. I. Guyon, J. Weston, S. Barnhill, V. Vapnik: *Machine learning* 2002, **46**:389-422.