Pacific Symposium on Biocomputing 12:216-227(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

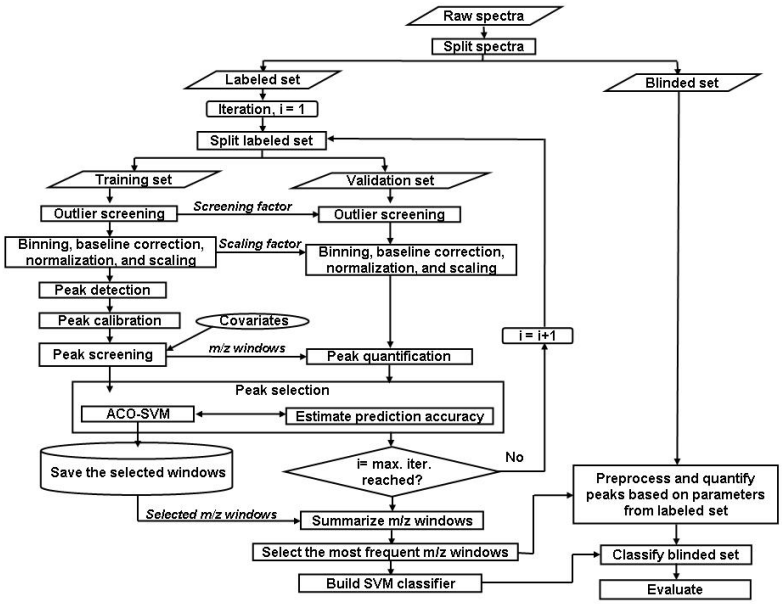

Figure 1. Methodology for global peak detection.

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