MALDI-TOF MS combined with magnetic beads for detecting serum protein biomarkers and establishment of boosting decision tree model for diagnosis of systemic lupus erythematosus

Zhuochun Huang1, Yunying Shi2, Bei Cai1, Lanlan Wang1, Yongkang Wu1, Binwu Ying1, Li Qin1, Chaojun Hu3 and Yongzhe Li3

**Objectives.** To discover novel potential biomarkers and establish a diagnostic pattern for SLE by using proteomic technology.

**Methods.** Serum proteomic spectra were generated by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) combined with weak cationic exchange magnetic beads. A training set of spectra, derived from analysis of sera from 32 patients with SLE, 43 patients with other autoimmune diseases and 43 age- and sex-matched healthy volunteers, was used to train and develop a decision tree model with a machine learning algorithm called decision boosting. A blinded testing set, including 32 patients with SLE, 42 patients with other autoimmune diseases and 40 healthy people, was used to determine the accuracy of the model.

**Results.** The diagnostic pattern with a panel of four potential protein biomarkers of mass-to-charge (m/z) ratio 4070.09, 7770.45, 28045.1 and 3376.02 could accurately recognize 25 of 32 patients with SLE, 36 of 42 patients with other autoimmune diseases and 36 of 40 healthy people.

**Conclusions.** The preliminary data suggested a potential application of MALDI-TOF MS combined with magnetic beads as an effective technology to profile serum proteome, and with pattern analysis, a diagnostic model comprising four potential biomarkers was indicated to differentiate individuals with SLE from RA, SS, SSc and healthy controls rapidly and precisely.

**Key words:** Boosting decision tree model, Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry, Proteomics, Systemic lupus erythematosus, Weak cationic exchange magnetic beads.

**Introduction**

SLE is a multisystem autoimmune connective tissue disorder with a broad range of clinical manifestations. Although physicians and researchers are trying to elucidate the pathogenesis of SLE, no definite conclusions have been reached so far. Genomic researches indicate that abnormal expression of susceptible genes plays an important role in the pathogenesis of SLE [1–3], but the detailed mechanisms remain enigmatic. The diagnosis of SLE is established according to the American College of Rheumatology (ACR) criteria, which comprise symptoms, signs and autoantibody detection. Although ANA detection is taken as a screening test because of its high sensitivity, the specificity is not satisfactory. Anti-double stranded DNA antibody and anti-Smith antibody are acceptable as laboratory confirmatory tests for SLE because of their high specificities (>85%), but their positive rates among patients with SLE are comparatively low (25–50%) [4]. Combination detection of these autoantibodies could improve the diagnosis of SLE, but problems still remain, such as added cost, complicated operation and prolonged detection time. Up till now, no better laboratory test has been introduced to facilitate the early diagnosis of SLE.

Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) is an important proteomic technology. Many protein biomarkers of certain diseases have been indicated by using MALDI-TOF MS to analyse serum proteome. Magnetic beads have large surface area and can capture more small molecular peptides and proteins [5, 6]. The combination of MALDI-TOF MS and magnetic beads can take advantages of both and therefore detect more low molecular weight proteins in serum.

Our research used MALDI-TOF MS combined with weak cationic exchange (WCX) magnetic beads to detect serum proteome of 64 patients with SLE, 85 patients with other autoimmune diseases (30 with RA, 30 with SS and 25 with SSc) and 83 healthy volunteers. All data were randomly divided into a training set and a blinded testing set for analysis. Training set was used to screen novel biomarkers specific for SLE and to establish an optimal decision tree model for SLE diagnosis. Blinded testing set was used to validate the sensitivity and specificity of the decision tree model.

**Materials and methods**

**Patients and controls**

This study was approved by the Ethics Committee of the Chinese Human Genome and the Ethics Committee of West China Hospital, and written informed consent was obtained from all participants. A total of 232 serum samples were included in this study, among which 64 were from patients with SLE, 85 from patients with other autoimmune diseases (30 with RA, 30 with SS and 25 with SSc), and 83 from healthy volunteers. All the samples were collected between January 2006 and December 2007 at West China Hospital, Sichuan University and Peking Union Medical College Hospital. All the patients were diagnosed according to the latest diagnosis criteria of ACR. Demographic features of all the patients and healthy volunteers are provided in Table 1 and sample numbers used for profiling (training set) and validation (blinded testing set) are listed in Table 2.

**Methods**

**Reagents and instrument.** The MALDI-TOF MS (PBS IIc) was provided by Ciphergen Biosystems, USA and Au-chip was purchased from Ciphergen Biosystems. The WCX magnetic...
bound molecules were eluted by incubation with 10% (v/v) TFA. Then, the elute (5 mM ammonio-1-propanesulfonate (CHAPS) were purchased from Beijing SED Science and Technology, China). After fixation of the magnetic beads for 1 min in PBS, the supernatant was aspirated. This washing procedure was repeated twice.

The activated magnetic beads, incubated for 1 h at 4°C and washed with PBS, were the available supporting software provided by Ciphergen Biosystems. The spectra of 232 serum samples were separated by a stratified sampling procedure using the all-in-one peptide molecular mass standards (Ciphergen Biosystems).

### Data analysis

The data analysis involved three stages: (i) peak detection and alignment; (ii) selection of differentially expressed peaks among the three groups that may represent potential biomarkers of SLE; and (iii) data analysis using a decision tree algorithm.

Peak detection was performed using Ciphergen ProteinChip software 3.0.2 (Ciphergen Biosystems). The protein peaks with m/z ranging from 2000 to 50000 were selected for analysis, while those with m/z ranging between 0 and 2000 were eliminated from analysis to avoid interference from adducts, artefacts of the energy-absorbing molecules and other possible chemical contaminants. Peak detection involved (i) baseline subtraction, (ii) mass accuracy calibration and (iii) automatic peak detection. Using Biomarker Wizard Version 3.1.0 (Ciphergen Biosystems), biomarkers were generated that represent consistent protein peaks across multiple spectra. Baseline subtraction was performed on all spectra. The peak of m/z 4901 was selected to normalize dimension. The settings for auto-detect peaks to cluster were as follows: signal-to-noise ratio was 5 and minimum peak threshold was 10% for the first pass; for cluster completion, cluster mass window was 0.3%, and signal-to-noise ratio for the second pass was 2.

Differently expressed peaks were selected by the differences in protein peak intensities between groups using the Biomarker Wizard Version 3.1.0, which used the non-parametric Kruskal–Wallis test and Mann–Whitney test.

Construction of the decision tree classification algorithm was performed by Biomarker Patterns Software 5.0 (BPS; Ciphergen Biosystems). BPS is an implementation of the Classification and Regression Trees (CART) decision tree system developed by Breiman et al. [7]. BPS uses the peak information generated by the training set of known samples to build a binary decision tree algorithm. The algorithm functions by assigning each sample in the data set into one of the two nodes with a rule based on the intensity of a particular peak or splitter [8, 9].

### Statistical analysis

The spectra of 232 serum samples were separated by a stratified random sampling into ‘training set’ data (a total of 118, 32 SLE, 15 RA, 15 SS, 13 SSc and 43 healthy volunteer samples) and ‘blinded testing set’ data (a total of 114, 32 SLE, 15 RA, 15 SS, 12 SSc and 40 healthy volunteer samples). Our aim was to develop a boosting decision classification using the training set data, for which both the clinical diagnosis and the proteomic spectrum profile data pairs of each sample were available. The resulting boosting decision tree classification was then applied to the blinded testing set data for estimating the accuracy of the decision tree model. Ciphergen ProteinChip Software 3.0.2, Biomarker Wizard software 3.1.0, and BPS 5.0 (Ciphergen Biosystems) were the available supporting software provided by Ciphergen Company for data analysis.

All the results were expressed as mean ± s.d., and P-values < 0.05 were considered statistically significant.
Results

Identification of serum protein profile

Protein spectra of 232 serum samples were generated by MALDI-TOF MS combined with WCX magnetic beads. The combination was particularly effective in resolving low molecular weight proteins and peptides (Fig. 1).

Comparison of serum protein spectra between SLE and control subjects [disease controls and healthy controls]

A total of 98 protein peaks were detected in SLE cases in the training set, among which 60 had a P-value <0.05 when compared with control subjects. Of the 60 differential protein peaks, 32 peaks were down-regulated and 28 peaks were overexpressed in SLE cases (data shown as supplementary data available at Rheumatology online).

Construction of boosting decision tree

Construction of the decision tree classification algorithm was performed by BPS 5.0 (Ciphergen Biosystems). The data used in this proceeding were from the Biomarker Wizard Version 3.1.0 software. The BPS produced many decision trees and evaluated the error cost (represented as ‘relative cost’ in the software) for each tree. Of the classification trees analysed by BPS software, the most optimal classification tree with the lowest error cost (‘relative cost’ being 0.214 and BPS settings as follows: mode = 1.70, advanced = 10 and testing = 10) was eventually established, that is, peaks of m/z 4070.09, 7770.45, 28045.1 and 3376.02 were selected in the classification tree. The protein peaks of m/z 4070.09, 7770.45 were down-regulated and m/z of 28045.1, 3376.02 were overexpressed in patients with SLE compared with control subjects (Fig. 2). All 118 training set serum samples were differentiated into five terminal nodes (Fig. 3). Samples differentiated into terminal nodes 1 and 4 were assigned to disease controls (DC), terminal node 2 to SLE and terminal nodes 3 and 5 to healthy controls (HC). For example, if an unknown sample has protein peak m/z 4070.09 (intensity <1.72757) and m/z 28045.1 (intensity >2.47333), then the sample is placed in terminal node 2 and classified as SLE. If the sample is placed in terminal node 1, it will be assigned to DC. Based on the stochastic nature of reality, misclassification of a new sample cannot be ruled out even for a pure node that contains only one sample type. For example, terminal node 5 contains only HC sample. To make clear whether an unknown sample could be correctly classified or misclassified, the expected probability was calculated for each class in the five terminal nodes (Table 3). For example, the expected probability for SLE samples to be misclassified in terminal node 5 is 3.03%. Although the figure is not 0, the likelihood of SLE samples being assigned to this node is still low, whereas HC specimens have 93.94% chance of being correctly classified to terminal node 5. The corresponding receiver operating characteristics curve (ROC) of the optimal decision tree was supplied by the BPS 5.0. The ROC integral was 0.955 (Fig. 4).

Diagnostic characteristics of the decision tree model

The decision tree could differentiate samples of SLE from other autoimmune diseases and healthy people. The accuracy of the optimal decision tree model was presented by comparison between the decision tree judgement and clinical diagnosis of each sample. In the training set, the decision tree model could accurately recognize 29 of 32 SLE, 37 of 43 other autoimmune diseases and 42 of 43 healthy people. Validation on the blinded testing set indicated that the decision tree could differentiate SLE, other autoimmune diseases and healthy people with the accuracy of 78.1, 85.8 and 90%, respectively (Table 4).

Discussion

The current common approaches for diagnosis of SLE in clinic are largely based on clinical manifestations and autoantibody detections. In this study, we described a high-throughput, multi-dimensional strategy using MALDI-TOF MS in combination with WCX magnetic beads for comprehensive profiling of serum proteomes for SLE patients.

MALDI-TOF MS is a key tool for proteomic analysis. It could achieve high-throughput protein analysis of biological samples with high sensitivity, reproducibility and resolving capability. Comprehensive MALDI-TOF MS analyses require efficient and selective pre-fractionation to facilitate analysis of peptides and proteins in samples. Magnetic beads are effective tools to capture proteins of interest with its large surface. Protein profiling by MALDI-TOF MS after proteome fractionation with magnetic beads is a robust, precise and rapid technique for the investigation of complex blood samples [10] and is appropriate for preliminary biomarker discovery [11], especially for the detection of low concentration of proteins and peptides in serum [10]. Protein profiling has been applied in various disease researches [12–15]. Martorella et al. [12] presented a serum peptide profile composed of 98 discriminatory low molecular weight peptides to distinguish metastatic thyroid cancer with high statistical accuracy. Freed et al. [13] identified 83 protein peaks having differential expressions among normal, healthy smokers and patients with head and neck squamous cell cancer (HNSCC). They found that the best model composed of eight peaks resulted in the correct classification of samples as follows: 89% normal, 93% benign and 98% HNSCC. Kim et al. [14] even used this technique as a fundamental tool to screen a potential bio-indicator which they would identify subsequently. Researchers of Alzheimer’s disease also employed this technique [15].

Proteomic researches on autoimmune diseases (SLE, RA, SS) were also published [16–21]. For SLE, Bauer et al. [18] focused on IFN-regulated chemokines in SLE. They identified dysregulated levels of 30 cytokines, chemokines, growth factors, soluble receptors and coordinated up-regulation of 12 inflammatory and/or...
homoeostatic chemokines, molecules that direct the movement of leucocytes in the body by performing a comprehensive survey of the serological proteome. Researches reported by Suzuki et al. [19], Mosley et al. [20] and Rovin et al. [21] mainly concern the urine proteome of SLE nephritis. They profile urine proteomic signature to distinguish SLE nephritis from SLE without nephritis or active SLE nephritis from inactive nephritis. Unlike the researches published so far, our study profiled serum proteome

Fig. 2. Detection of four protein peaks of proteomic profile in serum. Protein spectra of serum samples from two different SLE patients (SLE), two other autoimmune disease patients (DC) and two HCs (Nor). The x-axis represents the molecular mass calculation (m/z), and the y-axis represents the relative intensity. The spectrographic profiles reveal down-regulation of the m/z 4070.09, 7770.45 peaks and overexpression of m/z 28045.1, 3376.02 peaks in SLE samples.
In our study, WCX magnetic beads were used as a proteome fractionation tool. Strong anionic exchange, hydrophobic/reversed phase, hydrophilic/normal phase and immobilized metal affinity capture magnetic beads are other magnetic beads available. However, hydrophobic/reversed phase, hydrophilic/normal phase and immobilized metal affinity capture magnetic beads are directed towards hydrophobic proteins, hydrophilic proteins and signalling proteins, respectively. And WCX magnetic beads could capture more proteins in serum than strong anionic exchange magnetic beads, especially in the low molecular weight range (data of our earlier experiment not shown).

The initial serum proteome profiles of SLE are generated by using MALDI-TOF MS combined with WCX magnetic beads and pattern recognition software in our study. Existence of 60 differential protein peaks between SLE and other subjects indicates that a broad pathological change of SLE in serum proteome, autoantibodies, complements, immunoglobulins and unidentified proteins may be involved. Instead of a single protein type, a group of proteins contribute to the pathogenesis of SLE. Based on the differential protein peaks, potential biomarkers specific to SLE are available. Generation and analysis of serum protein profiles are necessary and significant.

The initial decision tree classification for distinguishing SLE from other autoimmune diseases (RA, SS and SSc) and HCs was generated by analysing ‘training set’ data by a machine learning algorithm derived from CART [8, 9]. Four potential biomarkers were involved in the decision tree because their exact identities would be essential for understanding what biological role these proteins/peptides may take in the pathogenesis of SLE, potentially leading to novel therapeutic targets. Efforts are under way to purify, identity and characterize these four protein biomarkers.

FIG. 3. The boosting decision tree classification of SLE, DC and HC samples in the training set. The root node (top) and descendant nodes were shown as ellipses, and the terminal nodes (Nodes 1–5) were shown as rectangles. The mass value in the nodes was followed by lower or equal to intensity value. If the answer to the question in a node of the tree is yes, you proceed down to the left node, otherwise (i.e. no), you proceed down to the right node. When proceeding to the terminal nodes, the decision tree assigned samples to three groups. Samples in terminal nodes 1 and 4 were assigned to DC, terminal node 2 was to SLE and terminal nodes 3 and 5 were to HC. The numbers in rectangles represent the actual clinical diagnosis of samples assigned to this terminal node by decision tree (i.e. in terminal node 1, decision tree assigned 26 samples to DC, but actually, 24 of them were SLE according to the clinical diagnosis).

FIG. 4. The ROC of the boosting decision tree.

TABLE 3. Expected probability of classes assigned to five terminal nodes

<table>
<thead>
<tr>
<th>Terminal node</th>
<th>Class</th>
<th>Observation</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (n=26)</td>
<td>SLE</td>
<td>2</td>
<td>0.1034</td>
</tr>
<tr>
<td></td>
<td>DC</td>
<td>24</td>
<td>0.8621</td>
</tr>
<tr>
<td></td>
<td>HC</td>
<td>0</td>
<td>0.0354</td>
</tr>
<tr>
<td>2 (n=33)</td>
<td>SLE</td>
<td>29</td>
<td>0.8333</td>
</tr>
<tr>
<td></td>
<td>DC</td>
<td>4</td>
<td>0.1389</td>
</tr>
<tr>
<td></td>
<td>HC</td>
<td>0</td>
<td>0.0278</td>
</tr>
<tr>
<td>3 (n=14)</td>
<td>SLE</td>
<td>0</td>
<td>0.0588</td>
</tr>
<tr>
<td></td>
<td>DC</td>
<td>2</td>
<td>0.1765</td>
</tr>
<tr>
<td></td>
<td>HC</td>
<td>12</td>
<td>0.7647</td>
</tr>
<tr>
<td>4 (n=15)</td>
<td>SLE</td>
<td>0</td>
<td>0.1111</td>
</tr>
<tr>
<td></td>
<td>DC</td>
<td>13</td>
<td>0.7778</td>
</tr>
<tr>
<td></td>
<td>HC</td>
<td>1</td>
<td>0.1111</td>
</tr>
<tr>
<td>5 (n=30)</td>
<td>SLE</td>
<td>0</td>
<td>0.0303</td>
</tr>
<tr>
<td></td>
<td>DC</td>
<td>0</td>
<td>0.0303</td>
</tr>
<tr>
<td></td>
<td>HC</td>
<td>30</td>
<td>0.9397</td>
</tr>
</tbody>
</table>

TABLE 4. Sensitivity and specificity of decision tree model

<table>
<thead>
<tr>
<th>Actual class (according to clinical diagnosis)</th>
<th>Decision tree classification class</th>
</tr>
</thead>
<tbody>
<tr>
<td>SLE, n (%)</td>
<td>DC, n (%)</td>
</tr>
<tr>
<td>DC, n (%)</td>
<td>HC, n (%)</td>
</tr>
<tr>
<td>In the training set</td>
<td></td>
</tr>
<tr>
<td>SLE (n=32) 29 (90.6)</td>
<td>3 (9.4) 0 (0)</td>
</tr>
<tr>
<td>DC (n=43) 4 (9.3)</td>
<td>37 (86.0) 2 (4.7)</td>
</tr>
<tr>
<td>HC (n=43) 0 (0)</td>
<td>1 (2.3) 42 (97.7)</td>
</tr>
<tr>
<td>In the blinded testing set</td>
<td></td>
</tr>
<tr>
<td>SLE (n=32) 25 (78.1)</td>
<td>7 (21.9) 0 (0)</td>
</tr>
<tr>
<td>DC (n=42) 3 (7.1)</td>
<td>36 (85.8) 3 (7.1)</td>
</tr>
<tr>
<td>HC (n=40) 0 (0)</td>
<td>4 (10%) 36 (90)</td>
</tr>
</tbody>
</table>
The decision tree can classify unknown samples into three groups: SLE, other autoimmune diseases and healthy population. However, misclassification cannot be avoided. Expected probabilities are used to predict the chances of samples being correctly classified. SLE samples are most probably assigned to terminal node 2, while there is still a 3.03% chance of being misclassified to terminal node 5. The validation with 'blinded testing set' data indicates that the decision tree could correctly recognize 25 of 32 SLE, 36 of 42 other autoimmune diseases and 36 of 40 healthy people. The high sensitivity and specificity of the results are encouraging.

In conclusion, the preliminary data in our study suggested a potential application of MALDI-TOF MS combined with magnetic beads as an effective technology to profile serum proteome, and with pattern analysis, a diagnostic model comprising four potential biomarkers was indicated to differentiate individuals with SLE from RA, SS, SSc and HCs rapidly and precisely.

Rheumatology key messages

- Novel protein biomarkers can be found by MALDI-TOF MS combined with magnetic beads.
- Proteomic diagnostic model comprising four proteins with high sensitivity and specificity for diagnosis of SLE was indicated.

Acknowledgements

The authors thank Prof. Lin Dongtao for copyediting the manuscript.

Funding: This research was sponsored by National Natural Science Foundation of China (Nos 30670819 and 30772051).

Disclosure statement: The authors have declared no conflicts of interest.

Supplementary data

Supplementary data are available at Rheumatology Online.

References