SPANDIDOS PUBLICATIONS

MOLECULAR MEDICINE REPORTS 3: 821-824, 2010

Mass spectrometry-based proteomic analysis of Kashin-Beck disease

JIANQIANG DU[1], XIAOMIN WU[1], HUQIN ZHANG[1], SHUANG WANG[2], WUHONG TAN[2] and XIONG GUO[3]

[1]Key Laboratory of Biomedical Information Engineering of Ministry of Education, School of Life Science and Technology; [2]Department of Orthodontics, Stomatological Hospital, College of Medicine; [3]Department of Public Health, School of Medicine, Xi'an Jiaotong University, 710061 Xi'an, P.R. China

Received April 9, 2010; Accepted June 11, 2010

DOI: 10.3892/mmr.2010.327

Abstract. Kashin-Beck disease (KBD) is a degenerative osteoarticular disease of unknown etiology. The management of KBD would benefit from the identification of the biomarkers related to this disease. In this study, mass spectrometry (MS)-based proteomic profiling was used to identify potential biomarkers of the disease. One hundred and sixteen serum samples of KBD cases and healthy controls were collected and analyzed. A framework for data analysis was implemented, which included normalization, denoising using undecimated discrete wavelet transforms, baseline subtraction, peak detection and alignment, non-parametric testing and classification by support vector machine. The method identified correlative mass points and obtained a discriminative pattern with 90.91% sensitivity and 82.61% specificity. The results of this study, although preliminary, suggest that further proteomics study may be useful with a larger number of appropriate specimens, careful experiment manipulation and improved MS techniques.

Introduction

Kashin-Beck disease (KBD) is a degenerative osteoarticular disease involving growth and joint cartilage that constitutes a significant human and socio-economic burden in affected areas. The disease is mainly distributed in a diagonal belt with low selenium from the northeast to the southwest of China, in which over 2.50 million patients are affected with KBD and 30 million people are at risk (1,2). The management of the disease would benefit from the identification of biomarkers related to this disease.

In this study, a mass spectrometry (MS)-based proteomic profiling technique was used to identify potential biomarkers for KBD. In recent years, MS-based proteomic profiling has increasingly been used to locate disease-related proteomic patterns in complex mixtures of proteins derived from biological fluids, such as serum or urine. The methods used to study several types of diseases, including lung, ovarian and prostate cancer, have resulted in promising findings (3,4). In all these studies, high sensitivity and specificity were obtained from the discriminative pattern. Generally, a typical MS dataset contains tens or hundreds of spectra, with each spectrum containing tens of thousands of mass to charge (m/z) intensity. From a modeling viewpoint, these spectra are considered complex functional data, in which the key features of scientific interest are peaks. On the other hand, MS data are inherently noisy, representing a complex signal consisting of electronic and chemical noise with a varying baseline caused by matrix-associated chemical noise or ion overload. From a biological perspective, peaks constitute the most important features of the spectrum. In MS-based proteomic profiling studies, the aim should be to identify peaks related to specific outcomes of different diseases or specific clinical responses.

Typically, the data analytic approach includes two steps focused on the peaks. The first step involves feature extraction and quantification, in which the peak locations and quantification of each peak in each spectrum is identified, including baseline correction, normalization and denoising. After properly pre-processing a set of spectra, assuming that p peaks from n spectra are found, a p x n matrix of 'protein expression levels' is yielded. The second step consists of using this matrix to search for proteins that may be differentially expressed between different conditions or correlated with clinical outcomes by performing discrimination and classification.

The purpose of the present study was to investigate changes in serum proteome and to identify potential biomarkers for KBD. A spectra process algorithm was developed to process the MS spectra data of KBD cases and controls. A workflow that combined denoising with undecimated discrete wavelet transforms, baseline subtraction, peak detection and alignment, and finally support vector machine classification, was used to locate the discriminative pattern.

Materials and methods

Serum proteomic mass spectral data. A total of 116 sera samples (38 from KBD patients and 78 from controls) were

Correspondence to: Professor Xiong Guo, Department of Public Health, School of Medicine, Xi'an Jiaotong University, 710061 Xi'an, P.R. China
E-mail: guox@mail.xjtu.edu.cn

Key words: Kashin-Beck disease, mass spectrometry, proteomic

used for further data analysis after outlier screening. The KBD cases are from the disease areas of Yong-Shou County, Shaanxi, China. There were no significant differences in terms of age or gender among the groups. Peripheral blood samples (5 ml) were collected by a trained phlebotomist from the veins of the upper arm in the morning from each KBD case or control according to a standard protocol. Each sample was allowed to clot and centrifuged at 3,000 rpm for 20 min within 30 min of collection to remove cellular components. Aliquots of sera for the mass spectrometric analysis were frozen at -80˚C until use; measurements were performed on samples of second-time thawed serum. All subjects provided their informed consent prior to participation in the study, and the study protocols were approved by the ethics committee of Xi'an Jiaotong University. Samples were analyzed using a SELDI-TOF mass spectrometer interfaced with a CM10 protein array (Ciphergen Biosystems) (5).

Outlier screening. Several successive analyses were applied to pre-process the raw MS data. First was outlier screening, where spectra whose data distribution substantially deviated from others were manually removed by examining the average Pearson's correlation coefficient of each spectrum against all other spectra within the dataset. This step was performed using SpecAlign software version 2.4 (6).

Intensity normalization. When comparing mass specta, raw data cannot be used directly, since the amounts of the unknown may differ due for various physical and chemical reasons, leading to different total ion currents. Therefore, comparisons can usually only be performed after the normalization of the ion currents. Normalization reduces variations in signal intensity between the spectra. A commonly used normalization method for mass spectrometric data is rescaling each spectrum by its total ion current (7); we found that this method worked well with our data. In addition, the spectra were scaled to have an overall maximum intensity of 100. Normalization was performed using the msnorm function of Matlab.

Denoising using undecimated discrete wavelet transform. Wavelet transform has been successfully used in various applications to remove noise and recover the true signal. Discrete wavelet transform (DWT) is the most commonly used wavelet algorithm for scientific application, but the classical DWT is not a time-invariant transform. In order to obtain more complete characteristics of an analyzed signal and to restore the translation invariance – a desirable property lost with classical DWT – we used an undecimated discrete wavelet transform (UDWT). In UDWT, decimators are removed so that signals are no longer decimated after filtering and approximation and detail signals are of a size that is the same as that of the analyzed signal. Compared to DWT, UDWT provides much more precise information. In our study, each individual spectrum was denoised using UDWT, as implemented in version 2.4 of the rice wavelet toolbox (http://dsp.rice.edu/software/rice-wavelet-toolbox), also adopted by Morris *et al* (8,9). Denoising works by computing the wavelet coefficients for the MS signal, then performing hard thresholding. Coefficients less than the threshold value are set to zero, while coefficients greater than the threshold remain unchanged. The threshold is the product of a thresholding parameter η and a robust estimation of the noise, the median absolute deviation divided by 0.67. Our previous study found that the choice of wavelet basis does not strongly impact denoising performance. Mean square error (MSE) was used to evaluate the denoising performance. A lower value for MSE indicates less error and the best denoising performance.

Baseline subtraction. A baseline is a systematic artifact commonly observed in MS data. A drifting baseline results in serious distortion of ion intensities without adequate correction. In our study, the low-frequency baseline of each spectrum was estimated using multiple shifted windows of 200 bins. Spline approximation was used to regress the varying baseline. The regressed baseline was subtracted from the spectrum, yielding a baseline corrected spectrum. In the present study, this process was implemented utilizing the 'msbackadj' function of Matlab software.

Peak detection. Peak detection deals with the selection of m/z values that display a reasonable intensity compared to those that display noise. After the spectrums were denoised and the baseline corrected and normalized, Coombers's (7,8) methods were adopted to identify peaks. First, all local maxima in the average spectrum and the associated peak endpoints of the spectrum were identified. Then, the signal-to-noise ratio (S/N) at each local maximum was computed by comparing the ratio of the intensity at the maximum to the estimation of local noise. Local maxima with a S/N greater than the threshold were considered peaks. Of only those peaks, the individual peaks were labeled by the m/z value of the local maximum in the mean spectrum. The identified peaks were quantified in the individual spectra; a total of 961 peaks were identified using this method.

Statistical analysis. The Kolmogorov-Smirnov test was used to select peaks discriminating the two classes significantly. As these data are non-Gaussian distributed, a non-parametric test best suits the purpose. Fifty-nine peaks were retained after Kolmogorov-Smirnov testing with the significance level set to 0.05. The data from these peaks were used for classification in the classification step.

Classification by support vector machine. Support vector machine (SVM) (10) is a statistical learning method that separates labeled data points into their respective classes. A separating hyperplane that yields the best expected separation on new data is fitted. By using an implicit transformation into high-dimensional feature space, datasets which are not linearly separable can be tackled. In this study, the SVM implementation libSVM by Chang and Lin was used (11), following the proposed procedure for performing SVM classification. First, data were converted to the format of the libSVM package, and then data scaling was conducting. Only the radial basis function kernel was considered; cross-validation was used to obtained the best parameter. This parameter was used to train a model from the training dataset. Finally, this model was used to predict the class label of each sample in the testing dataset.

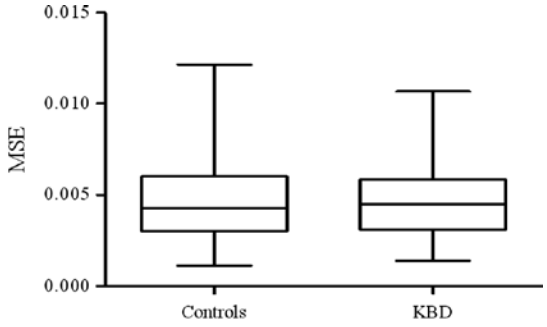

Figure 1. Mean square error of the data of KBD cases and controls.

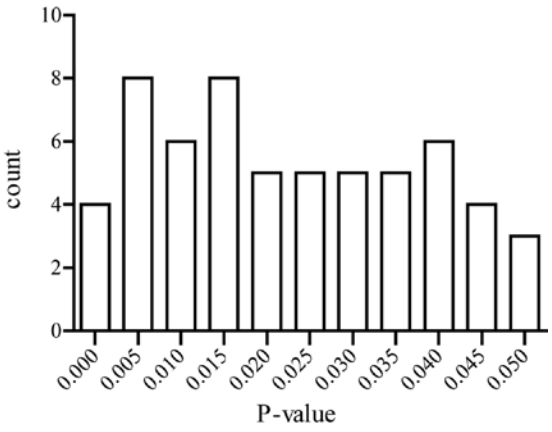

Figure 2. The distribution of the P-value of 59 peaks.

The 116 peaks data were randomly split into training and testing datasets. The training dataset consisted of 82 samples (27 KBD cases and 55 controls), while the testing dataset included 34 samples (11 KBD cases and 23 controls).

Results

Pre-process. In the pre-process step, Pearson's correlation coefficient was first used to pick out the outlier data. Spectra whose data distribution substantially deviated from others were manually removed (data not shown), then the data were denoised using UDWT. MSE was used to quantitively evaluate the denoising performance. The MSE of the MS data of KBD cases and controls are illustrated in Fig. 1. The mean value of the MSE was <0.005, indicating that most information in the MS data was retained after denoising. Denoising using UDWT was therefore efficient and appropriate in this circumstance. In the following steps, after normalization and baseline subtraction, the bias and artifact of MS data were eventually removed, allowing for the intensity value to be quantified and compared. Additionally, peaks were detected, aligned and quantified. Lastly, the Kolmogorov-Smirnov test was performed to select peaks significantly discriminating between the two classes, since the peak data were non-Gaussian distributed. A total of 56 peaks were obtained. The distribution of the P-value of these peaks is illustrated in Fig. 2. Sixteen of the peaks had a P-value <0.01. Table I lists the m/z, P-values, mean, standard deviation and coefficient of variation of the 16 peaks.

Classification. Another key issue examined in the study was the choice of classification procedure. Various classification methods have been proposed; we found SVM to be more robust and efficient. SVM can be trained very efficiently

Table I. Significant peaks between KBD cases and controls selected by the Kolmogorov-Smirnov test.

Peaks (m/z)	P-value	KBD			Control		
		Mean	SD	CV	Mean	SD	CV
1181.3993	0.0014	3.44	4.34	1.26	4.42	3.02	0.68
1602.1161	0.0076	1.64	2.30	1.41	2.97	2.81	0.95
2296.9581	0.0028	1.72	1.99	1.16	2.69	1.99	0.74
2797.2571	0.0080	2.79	2.44	0.88	3.88	2.63	0.68
3778.7086	0.0024	4.68	2.64	0.56	3.60	2.66	0.74
5045.1213	0.0057	2.47	2.94	1.19	4.11	3.07	0.75
5904.5922	0.0024	71.15	15.49	0.22	58.07	23.38	0.40
5944.9703	0.0087	12.33	6.34	0.51	10.48	9.11	0.87
6111.1388	0.0035	21.46	9.95	0.46	18.13	13.76	0.77
7154.4733	0.0016	3.44	1.86	0.54	2.49	1.65	0.66
8016.7201	0.0041	2.19	2.47	1.13	3.52	2.60	0.74
8375.7354	0.0043	4.08	1.93	0.47	3.27	2.17	0.66
9875.8331	0.0027	0.66	0.68	1.04	1.14	0.84	0.73
10864.8980	0.0062	0.87	0.72	0.83	1.24	0.94	0.76
16696.5780	0.0062	0.51	0.56	1.09	0.86	0.77	0.89
23611.5750	0.0087	3.16	1.69	0.54	2.71	1.92	0.71

SD, standard deviation; CV, coefficient of variation.

using supervised learning. The peaks retained after the test were used for SVM classification. The data were randomly divided into training and testing; the former were used to train a model, while the latter were used to test the performance of the model. Finally, in the test dataset, 10 KBD cases and 19 controls were correctly classified. The sensitivity was 90.91% (10/11), the specificity 82.61% (19/23) and the accuracy 85.29% (29/34).

Discussion

We believe that the careful pre-processing of mass spectra is crucial to developing a successful classification procedure of proteomics data. With this as our basis, we initially used raw data and applied methods that aimed to remove the various nuisance effects and additional errors present within the spectral data.

In this study, we presented a successive data analysis scheme for MS data that combines outlier screening, normalization, denoising using UDWT, baseline subtraction, peak detection and classification by SVM, and demonstrated that the proposed approaches may be used to select mass points from the MS dataset. For the KBD dataset presented in this study, 59 peaks were selected, yielding up to 90.91% sensitivity and 82.61% specificity in distinguishing cases from controls. The final selected peaks are more likely to represent identifiable proteins, protein fragments or peptides, which is crucial for the ultimate goal of identifying proteins or peptides that distinguish cases from controls. Once the proteins are identified, focus may turn to their validation through other sample-sets and analytical platforms. However, there are too many peaks in the final discriminative model; it is probably impractical to test each peak, so other methods should be used to reduce the number of peaks.

The use of computational methods alone may not provide a final solution to the analysis of proteomic MS data. The model in this paper was developed based on peaks with small P-values. However, sometimes a single P-value does not tell the whole story, and significant results with small P-values may be obtained by chance. Besides advanced computational methods capable of extracting useful information from this high dimensional and complex data, careful experiment manipulation and improved mass spectrometry are required.

Acknowledgements

This study was supported by the National Natural Science Foundation of China (grant nos. 30630058, 30972556, 60601017).

References

1. Moreno-Reyes R, Suetens C, Mathieu F, *et al*: Kashin-Beck osteoarthropathy in rural Tibet in relation to selenium and iodine status. N Engl J Med 339: 1112-1120, 1998.
2. National Kashin-Beck disease surveillance group: The monitoring report of Kashin-Beck disease prevalence rate of the whole country in 2003. Chin J Endemiol 23: 147-149, 2003.
3. Yan W and Chen SS: Mass spectrometry-based quantitative proteomic profiling. Briefings in Functional Genomics and Proteomics 4: 27-38, 2005.
4. Steel LF, Haab BB and Hanash SM: Methods of comparative proteomic profiling for disease diagnostics. J Chromatogr B Analyt Technol Biomed Life Sci 815: 275-284, 2005.
5. Wang S, Guo X, Tan WH, *et al*: Detection of serum proteomic changes and discovery of serum biomarkers for Kashin-Beck disease using surface-enhanced laser desporption ionization mass spectrometry (SELDI-TOF MS). J Bone Miner Metab 26: 385-393, 2008.
5. Wong JWH, Cagney G and Cartwright HM: Specalign – processing and alignment of mass spectra datasets. Bioinformatics 21: 2088-2090, 2005.
6. Alfassi ZB: On the normalization of a mass spectrum for comparison of two spectra. Journal of the American Society for Mass Spectrometry 15: 385-387, 2004.
7. Coombes KR, Tsavachidis S, Morris JS, Baggerly KA, Hung MC and Kuerer HM: Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform. Proteomics 5: 4107-4117, 2005.
8. Morris JS, Coombes KR, Koomen J, Baggerly KA and Kobayashi R: Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum. Bioinformatics 21: 1764-1775, 2005.
9. Burges CJC: A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery 2: 121-167, 1998.
10. Chang C-C and Lin C-J: LIBSVM: a library for support vector machines, 2001. Available at http://www.csie.ntu.edu.tw/~cjlin/libsvm