An Approach to Comparative Analysis of Chromatographic Fingerprints for Assuring the Quality of Botanical Drugs

Yiyu Cheng* and Minjun Chen

College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310027, China

Weida Tong

Center for Toxicoinformatics, National Center for Toxicological Research (NCTR), The U.S. Food and Drug Administration (FDA), Jefferson, Arkansas 72079

Received February 24, 2003

The present study was focused on developing the chemometric methods for analysis of the chromatographic fingerprint to control the quality of botanical drugs, which has gained attention in Asia and other countries. We developed a novel approach to generate a set of fingerprint features, called Fisher components (FCs) that were extracted from the chromatographic fingerprint. The method greatly reduces the dimensionality of the fingerprint vector, and the resulting FCs still retain most discriminatory information of the original fingerprint. Choosing an example of relevance to contemporary botanical drugs, we applied the FCs to a set of Shenmai injection samples. We successfully identified the manufacturers of the samples using two classifiers, linear discriminant analysis (LDA) and k-Nearest Neighbor (k-NN) based on the FCs. We also applied a similarity assessment together with the visual analysis using the FCs to exam the products from different manufacturers. We found that the lot-to-lot consistency of products can be accurately determined using the FCs. Finally, we demonstrated that the application of chemometric methods for chromatographic fingerprinting offers reliability to detect suspected fraud samples. In summary, we demonstrated that the presented approaches could be useful to determine the identity, consistency, and authenticity of Shenmai injection through chromatographic fingerprinting. The methods are equally applicable to other botanical drugs.

INTRODUCTION

Chromatographic fingerprinting characterizes the chemical patterns that present unique chemical compositions of samples (e.g. botanical drugs) embedded in chromatograms. It is well established that the samples with similar chemical patterns likely have similar properties. In early practice, the similarity of samples in the chemical patterns is usually assessed by visual comparison of the spectra, which is often subjective and not always persuasive. The application of chemometric methods for chromatographic fingerprinting offers several benefits; it is not subjective, analysis is consistent, and the results are reproducible. Thus, such practice has been demonstrated in pharmaceutical fingerprinting,[1−3] fuel spill identification,[4] classification of wine,[5] common batches searching of illicit heroin samples,[6] etc.

Botanical drugs have been widely applied for clinical use in China, Japan, Korea, and other Asian countries as well as some Northern American and European countries. Botanical drugs are usually extracted from plants and used as mixtures. It is difficult to ensure the quality of a botanical drug because of a large variability associated with the products produced by different manufacturers as well as the products from different batches of the same manufacturer. Chromatographic fingerprinting has been recommended as a potential and reliable approach for the quality control of botanical drugs.[7] There are several reports on developing chromatographic fingerprints for Gingko biloba,[8,9] Ma Huang,[10] and other botanical drugs. However, little attention has been paid to develop thorough and robustic procedures to compare the chromatographic fingerprints, particularly, for quality assurance of botanical drugs.

Shenmai injection is one of the most commonly used traditional Chinese injections. It has been used to treat coronary atherosclerotic cardiopathy and viral myocarditis, and it is also capable of raising tumor patient's immunity. Shenmai injection is produced from Radix Ginseng and Radix Ophiopogonis. Its main effective components are ginsenosides, and most of the ginsenosides have been identified.[11,12] Currently, the quality of Shenmai injection is normally assured based on the total content of ginsenosides, which is set to the least 0.8 mg of ginsenosides in a 1 mL injection. It is known that this standard is neither sufficient to determine the identity of an extracted plant material as a Shenmai injection nor to ensure the quality of an alleged Shenmai injection owing to the ginsenosides being present in other plants.

In this paper, we report our primary effort to develop the chemometric methods for analysis of the chromatographic fingerprint of botanical drugs with respect to quality control. Choosing an example of relevance to contemporary botanical drugs, the analysis was conducted on the Shenmai injections from three different commercial manufacturers. We first

* Corresponding author phone: 0086-571-87951138; fax: 0086-571-87980668; e-mail: chengyy@zju.edu.cn.

10.1021/ci034034c CCC: $25.00
Published on Web 04/24/2003

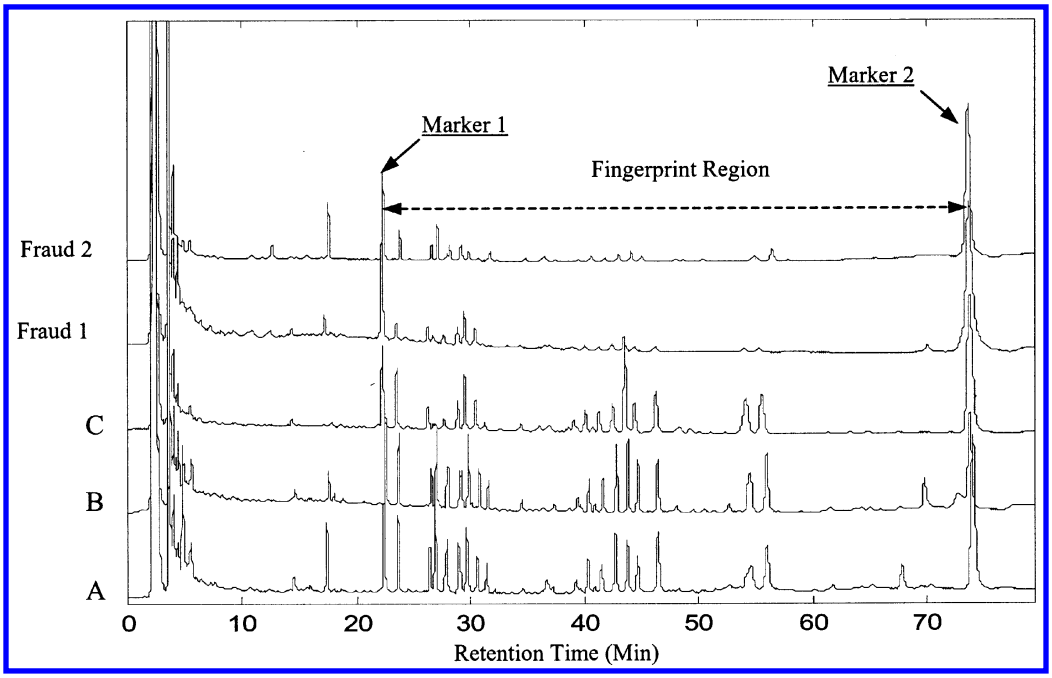

Figure 1. Representative chromatograms from each of the three manufacturers (A−C) and two suspected frauds.

identified a set of HPLC signals for constructing a chromatographic fingerprint. Then, we developed a method based on Fisher criterion to generate a set of fingerprint features, called Fisher components (FCs), from the chromatographic fingerprint. The method greatly reduces the dimensionality of the fingerprint vector, and the resulting FCs still retain most discriminatory information of the original fingerprint. We applied the FCs to identify the manufacturers of the samples using two classifiers, linear discriminant analysis (LDA) and k-Nearest Neighbor (k-NN). We also proposed a similarity assessment to evaluate the lot-to-lot consistency of products by comparing their chromatographic fingerprints. In this approach, the equivalence between the products was determined based on their relative position on the score plot of the FCs. We found that the visual analysis of the plot provides reasonable accuracy for assessing the consistency of products. At last, some suspected frauds were used to test the ability of fraud detection using the proposed fingerprinting in conjunction with the chemometric methods. In summary, we found that the presented approaches could be useful to determine the identity, consistency, and authenticity of Shenmai injection through chromatographic fingerprinting. The methods are also equally applicable to other botanical drugs.

EXPERIMENTAL SECTION

Samples. Seventy-three eligible samples of Shenmai injection were obtained from three different manufacturers in China, respectively. There were 29 lots of samples from the manufacturer A, 32 lots of samples from the manufacturer B, and 12 lots of samples from the manufacturer C. In addition, two lots of frauds produced from inappropriate materials were also used in the study to determine specificity and sensitivity of chromatographic fingerprinting for detecting the fraud samples.

The representative chromatograms of the Shenmai injections from each of three manufacturers and the two suspected frauds are shown in Figure 1. It is clear from visual inspection and comparison that the samples from three manufacturers had similar patterns in the marked region, the so-called fingerprint region.[1] Except the fingerprint region for the two frauds are much smaller compared to those of the samples from the three manufacturers. This also suggested that the two samples were different from other products, so we considered them as the frauds here.

Table 1. Solvent Composition of Gradient of the HPLC Analysis

time (min)	A[a] (%)	B[b] (%)	curve
0	75.0	25.0	
5	75.0	25.0	linear
45	31.5	68.5	
50	31.5	68.5	linear
65	15.0	85.0	
80	15.0	85.0	

[a] A = 50 mmol/L of KH_2PQ_4. [b] B = acetonitrile−water (80:20).

Experimentation. For establishing the fingerprint region, a Marker solution containing 0.4 mg/mL of prednisone and 0.16 mg/mL of biphenyl was prepared. A 5-mL Shenmai injection was mixed with a 300-μL Marker solution. The injection was filtered through a 0.45 μm membrane filter. A 100-μL volume of the injection was injected into the HPLC system. Each sample was measured three times. There is no exclusion of the results.

The chromatographic apparatus and reagents were obtained from standard commercial sources. The chromatograms were obtained using a liquid chromatographic system (a Shimadzu ***VP***Series modulus) that consisted of a dual plunger pump, an autosampler, a degasser, a column oven, and SPD-10A*VP* photodiode-array detector. The column was a Lichrospher C_{18} column (250 mm × 4 mm i.d.) (Huaiyin Hanbang Tech. Ltd., Jiangsu, China) with 5 μm spherical particles. A step gradient of (A) 50 mmol/L of KH_2PO_4 and (B) acetonitrile−water (80:20) was used. The gradient is presented in Table 1. The flow-rate was 1 mL/minute. Chromatographic runs were conducted at a temperature of 40 °C. The SPD-10A*VP* photodiode array detector was set at 202 nm.

Chromatographic Fingerprinting. A chromatographic fingerprint is a set of peaks of a chromatogram that represents the characteristics of a sample. To accurately capture the information encoded in a chromatogram, a chromatographic fingerprint was mathematically represented by a vector of peak areas in this study. The absolute area of each peak was calculated using the Shimadzu Class-VP 6.1 software. If a specific peak was absent in a particular chromatogram, the peak value was set to zero. To accurately and reliably quantify the peaks, only the peaks satisfying a user-defined threshold value (which is set equal to 0.5% of the total areas of all the peaks here) were selected to construct the fingerprint. Consequently, 18 peaks in the fingerprint region were used because they met the above criteria and appeared in most chromatograms. Thus, the chromatographic fingerprint for the sample i was an 18-dimensional vector $\mathbf{x}_i = [x_{i1},x_{i2},\cdots,x_{i18}]^t$, where x_{ij} was the absolute area of the peak j in the chromatogram and the superscript t indicates the transpose of matrix. To minimize the variability in different experimental conditions, the chromatographic fingerprints were normalized with respect to detector response (i.e. peak area). Accordingly, the ith chromatographic fingerprint ($\mathbf{x}_i$) was normalized on a linear scale based on the maximum area ($x_{i\max}$) and the minimum area ($x_{i\min}$) of the peak in the vector $\mathbf{x}_i$using $\mathbf{x}_i = (\mathbf{x}_i - x_{i\min})/(x_{i\max} - x_{i\min})$.

METHODOLOGY

A chromatographic fingerprint is usually high dimensional (>10 variables, here a variable such as a peak is considered as one-dimensional). Comparative analysis of the high dimensional vectors poses a challenge in implementation and interpretation. Transforming the high dimensional vector into a low dimensional space could greatly facilitate comparative analysis of chromatographic fingerprint and also be beneficial for ensuring the quality of botanical drugs. However, it is important that the low dimensional features resulted from the transformation should retain the critical information in the original fingerprint. These so-called fingerprint features extracted from the original fingerprint are often low dimensional and have high discriminatory power.

Individual peaks with high discriminatory power can be considered as one type of the fingerprint features. Another type of fingerprint features could be a linear combination of the peaks, such as principal components (PCs) derived from principal component analysis (PCA). Normally, PCs retain the maximum amount of variance in a data matrix but do not warrant the maximum discriminatory information. In this study, we used a method based on Fisher criterion to extract the fingerprint features, i.e., FCs, to better reflect the discriminatory information of chromatographic fingerprints. The fingerprint features extracting method is derived from the Fisher's linear discriminant method, which has been widely used as a classifier for interpreting the data from all kinds of instruments.[13,14]

Fisher Components. Assume that we have n chromatograms, each corresponds to a botanical drug sample, where n_1 samples belong to class 1 and n_2 samples belong to class 2. Then we have

$$n = n_1 + n_2 \tag{1}$$

A chromatographic fingerprint with m peaks can be represented by a vector $\mathbf{x} = [x_1,x_2,\cdots,x_m]^t$, where x_i denotes the area of the ith peak. Thus, a data matrix $\mathbf{X}$ of $m \times n$ can be obtained as

$$\mathbf{X} = [\mathbf{X}^{(1)},\mathbf{X}^{(2)}] \tag{2}$$

where $\mathbf{X}^{(i)}(i = 1,2)$ is a matrix of $m \times n_i$ for the sample class i.

Assume that there exists a set of fingerprint features that can be generally expressed as

$$y = \mathbf{u}^t\mathbf{x} \tag{3}$$

where $\mathbf{u} = [u_1,u_2,\cdots,u_m]^t$. With (3), the fingerprint feature could be either a discriminatory peak or a linear combination of some/all peaks. For instance, the kth peak will be selected as a fingerprint feature when the vector $\mathbf{u}$ is expressed as the following:

$$\mathbf{u} = \left[\underbrace{0,0,\cdots 0}_{k-1},1,0,\cdots 0\right]^t$$

Theoretically, there exists a transformation vector $\mathbf{u}$ that satisfies the maximum of the following Fisher criterion

$$F(\mathbf{u}) = \frac{(\bar{y}^{(1)} - \bar{y}^{(2)})^2}{\bar{S}_w^{(1)} + \bar{S}_w^{(2)}} \tag{4}$$

where $\bar{y}^{(i)}$ and $\bar{S}^{(i)}$ are the mean value and the variance of the fingerprint feature (y) of the class i, respectively:

$$\bar{y}^{(i)} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_j^{(i)} \quad (i = 1,2) \tag{5}$$

$$\bar{S}^{(i)} = \frac{1}{n_i}\sum_{j=1}^{n_i} (y_j^{(i)} - \bar{y}^{(i)})^2 \quad (i = 1,2) \tag{6}$$

Clearly, when the distance between the centers of two classes is increased and/or the variance within the classes is decreased, the fingerprint feature (y) has a better discriminatory power to separate two classes.

Replacing the variable y in (4) by (3), we can rewrite eq 4 as follows:

$$F(\mathbf{u}) = \frac{\mathbf{u}^t\cdot\mathbf{S}_b\cdot\mathbf{u}}{\mathbf{u}^t\cdot\mathbf{S}_w\cdot\mathbf{u}} \tag{7}$$

$\mathbf{S}_b$ is the between-class scatter matrix while $\mathbf{S}_w$ is the sum of the within-class scatter matrix; both are defined in (8) and (9), respectively

$$\mathbf{S}_b = (\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})^t \tag{8}$$

$$\mathbf{S}_w = \sum_{i=1}^{2}\sum_{j=1}^{n_i} (\mathbf{x}_j^{(i)} - \bar{\mathbf{x}}^{(i)})(\mathbf{x}_j^{(i)} - \bar{\mathbf{x}}^{(i)})^t \tag{9}$$

where $\mathbf{x}_j^{(i)}$ is the vector of the jth sample for the ith class,

while $\bar{\mathbf{x}}^{(i)}$ is the mean vector for the class i that is defined in (10):

$$\bar{\mathbf{x}}^{(i)} = \frac{1}{n_i}\sum_{j=1}^{n_i} \mathbf{x}_j^{(i)} \quad (i = 1,2) \tag{10}$$

The maximum value of $F(\mathbf{u})$ was derived using the method of Lagrange multipliers in this study. The mathematical presentation is described as follows:

Let us assume

$$\mathbf{u}^t \cdot \mathbf{S}_w \cdot \mathbf{u} = c \neq 0$$

the Lagrange function could be defined as

$$L(\mathbf{u}, \lambda) = \mathbf{u}^t \cdot \mathbf{S}_b \cdot \mathbf{u} - \lambda \cdot (\mathbf{u} \cdot {}^t\mathbf{S}_w \cdot \mathbf{u} - c) \tag{11}$$

where λ is the Lagrange multiplier. We take the vector derivative of L (11) with respect to $\mathbf{u}$ and set the resultant equation to zero. This procedure generates the generalized equation given in (12)

$$\mathbf{S}_b \cdot \mathbf{u} = \lambda \cdot \mathbf{S}_w \cdot \mathbf{u} \tag{12}$$

If $\mathbf{S}_w$ is a nonsingular matrix, the vector $\mathbf{u}$ will be the eigenvector corresponding to the largest eigenvalue of $\mathbf{S}_w^{-1} \cdot \mathbf{S}_b$. Thus, the $\mathbf{u}$ is the vector that identifies the fingerprint feature with the largest value of $F(\mathbf{u})$. Denoting $\mathbf{u}_1 = \mathbf{u}$, the first FC, which has the best discriminatory power is obtained by

$$y_1 = \mathbf{u}_1{}^t \mathbf{x} \tag{13}$$

To obtain the second FC, we denote $\mathbf{X}_1 = \mathbf{X}$ and $\mathbf{t}_1 = \mathbf{X}^t\mathbf{u}_1$. Thus, we have

$$\mathbf{q}_1 = \mathbf{X}_1 \cdot \mathbf{t}_1 / (\mathbf{t}_1{}^t \cdot \mathbf{t}_1) \tag{14}$$

where $\mathbf{q}_1$ is a loading vector of the data matrix $\mathbf{X}_1$. To obtain the orthogonal transformation vectors, the remaining data matrix $\mathbf{X}_2$ is calculated

$$\mathbf{X}_2 = \mathbf{X}_1 - \mathbf{t}_1 \cdot \mathbf{q}_1{}^t \tag{15}$$

By replacing $\mathbf{X}$ by $\mathbf{X}_2$ to calculate the parameters $\mathbf{S}_b$ and $\mathbf{S}_w$ in (7), the second transformation vector $\mathbf{u}_2$ is derived from the eigenvector corresponding to the largest eigenvalue of $\mathbf{S}_w^{-1} \cdot \mathbf{S}_b$. Thus, the second FC is

$$y_2 = \mathbf{u}_2{}^t \mathbf{x} \tag{16}$$

Similarly, the rest of FCs can be obtained by applying the above procedure iteratively. When the F_i for the ith FC obtained from (7) is small enough, it is considered that the discriminatory information of the fingerprint has been extracted out. In this study, the iteration was ended if the value of F_i reaches to a user-defined threshold.

The FCs have two important characteristics; the discriminatory power decreases in the order of the FCs, and all FCs are orthogonal.

The aforementioned approach can be equally applied to more than two classes, i.e., c classes. We can first generate a set of FCs for distinguishing the first class from the rest of the classes. Then, the similar procedure is employed to each of the rest of the class to determine their FCs. By the end of this stepwise approach through all the classes, we can obtain c sets of the FCs for distinguishing every class from the others.

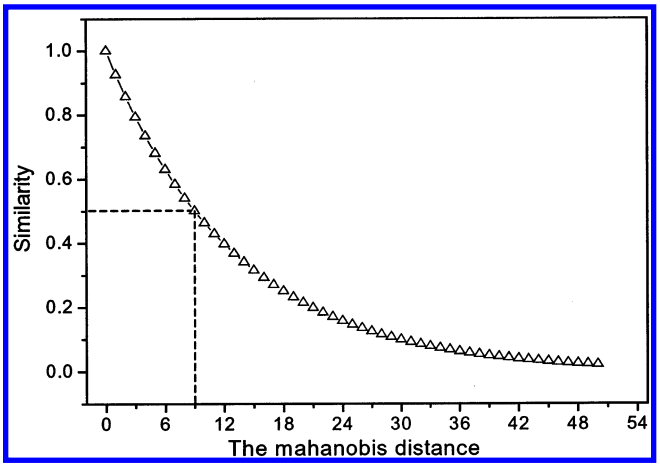

Figure 2. Illustration of relationship between the Mahanobis distance and similarity.

Discriminatory Power. According to (7), the discriminatory power of a fingerprint feature identified by the transformation vector $\mathbf{u}$ for either FC or others can be expressed using the following equation:

$$\xi = \frac{\mathbf{u}^t \cdot \mathbf{S}_b \cdot \mathbf{u}}{\mathbf{u}^t \cdot \mathbf{S}_w \cdot \mathbf{u}} \tag{17}$$

The value of ξ measuring the discriminatory power was used to compare the performance of the fingerprint features derived from different methods.

Similarity Measure of Chromatographic Fingerprints. In this study, the similarity between two chromatographic fingerprints was assessed based on their fingerprint features. We assume that $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_k]$ is the matrix of fingerprint feature transformation vectors. The original data matrix can be transformed as the following:

$$\mathbf{Y} = \mathbf{U}^t \cdot \mathbf{X} \tag{18}$$

Thus, each chromatographic fingerprint can be represented with a vector of fingerprint features $\mathbf{y} = [y_1, y_2, \cdots, y_k]^t$. Since the dimensionality of $\mathbf{y}$ is usually far less than that of $\mathbf{x}$, similarity between chromatographic fingerprints could be conveniently calculated using the exponential function of Mahalanobis distance

$$\psi(\mathbf{y}_1, \mathbf{y}_2) = \exp(-(\mathbf{y}_1 - \mathbf{y}_2)^t \cdot \mathbf{S}^{-1} \cdot (\mathbf{y}_1 - \mathbf{y}_2)/13) \tag{19}$$

where $\mathbf{y}_1$ and $\mathbf{y}_2$ is the vector of fingerprint features for samples 1 and 2, respectively, and $\mathbf{S}$ is a covariance matrix which is calculated from the samples belonging to the class of $\mathbf{y}_2$. It is worthwhile to point out that the Mahalanobis distance in (19) is scaled by dividing 13, which makes an even distribution of the similarity values of all the samples over the interval [0 1]. The relationship between Mahalanobis distance and similarity is plotted in Figure 2, where the similarity is 0.5 when the Mahalanobis distance equals to 9.

Classification Model. To identify a product manufacturer based on chromatographic fingerprinting, two classical

Table 2. Comparison of Three Types of Features on the Separation of Each Pair[a]

	peaks	ξ	FCs	ξ	PCs	ξ
manufacturer A and non-A	P17	1.6480	FC11	14.1575	PC1	0.1753
	P14	0.5819	FC12	9.6218	PC2	0.1693
manufacturer B and non-B	P14	1.0046	FC21	15.8826	PC1	0.0474
	P17	0.5639	FC22	9.096	PC2	0.0869
manufacturer C and non-C	P12	4.2847	FC31	15.1392	PC1	0.1804
	P6	0.6700	FC32	11.8414	PC2	0.0320

[a] P*i* is the *i*th peak in the vector **x**; FC*ij* is the the *i*th FC for the *j*th pair; PC*i* is the *i*th PC.

pattern recognition techniques, i.e., LDA and k-NN, were applied. LDA was performed using SPSS computer package (SPSS Inc.), while k-NN was performed with in-house programs. A brief description of both methods is given below, and the detailed mathematical basis can be found elsewhere.[15,16]

LDA hypothesizes that the distribution of the input variable is multivariate normal and that the covariance matrix of each class is not significantly different from each other. The Mahalanobis distances of a tested sample to the centroids of all classes are computed, and it is assigned to the class with the shortest distance. In this method, a linear function is used to delimit between classes, e.g., a straight line for two variables, a plane for three variables, etc.

The k-NN method is a nonparametric classification technique, which does not formulate a hypothesis on distribution of the input variables. The classification is performed by using a training data set that contains the input variable and the target variable (i.e. class information). After a model is developed on the training data set, it is then challenged by the tested samples. Once the *k* nearest samples are defined, the tested sample is classified to the class in which the majority of the *k* nearest samples belong to.

For evaluating the performance of the classifiers, a leave-one-out cross-validation method was used. In this approach, one of 73 samples is excluded, and then its classification is predicted by the model developed from the remaining 72 samples. The process has to be repeated until all samples in the training set are left out once, and the prediction results of the tested samples will be compared with its known results.

RESULTS AND DISCUSSION

Comparison of the Performance of the Fingerprint Features. We compared three types of the fingerprint features (FCs, PCs, and the features composed of peaks) against the samples that were divided into three pairs, i.e., manufacturer A and non-A, manufacturer B and non-B, and manufacturer C and non-C. For the sake of discussion and visualization, only the first two components for each type of the fingerprint features were used. For FCs and peaks, two features with highest ξ for each pair of the samples were used, while two PCs possessing the maximum amount of variance were selected for comparison. The results are summarized in Table 2.

Overall speaking, the performance of each type of the fingerprint features follows the order of FCs > Peaks > PCs. The FCs exhibit the best discriminatory performance (largest ξ) for distinguishing almost all the samples of each pair. This result indicates that the FCs are able to extract most, if not all, discriminatory information distributed among the peaks in the chromatographic fingerprint. In contrast, the PCs yield the smallest ξ for almost all three pairs. Despite the fact that the PCs are the commonly used features in chemometric analysis, they might not be good fingerprint features in application of chromatographic fingerprinting. This is largely due to the fact that the variance possessed by the PCs is not necessary to correlate with the discriminatory power.

Table 3. Summary of Results Using Different Classifiers and Different Input Features

		manufacturer				
input features	classifiers	A	B	C	total	correct %
peaks	k-NN	27/29	32/32	7/12	66/73	90.4
	LDA	26/29	32/32	10/12	68/73	93.2
FCs	k-NN	29/29	32/32	12/12	73/73	100.0
	LDA	29/29	32/32	12/12	73/73	100.0
PCs	k-NN	26/29	31/32	7/12	64/73	87.7
	LDA	26/29	31/32	8/12	65/73	89.0

Another way to compare the discriminatory performance of the fingerprint features could be achieved by visual analysis of the score plots derived from these features. The score plot illustrates the natural clustering among the samples of Shenmai injection based on their manufacturers. The discriminatory power of the features could be demonstrated by a good separation that is discerned in the score plot between the Shenmai injection manufacturers.

The score plots derived from the peaks, FCs and PCs, for manufacturer A, B, and C are shown in Figure 3a−g, respectively, where each sample is represented as a marker. It was noticeable that the samples from each of three manufacturers were clustered together much tighter in the score plot defined by the FCs (Figure 3d−f) than that by the peaks (Figure 3a−c) and the PCs (Figure 3g). In the PC-based score plot, all the points (samples) formed a single large cluster, where the samples from the different manufacturers were not well separated at all.

Among the three types of features compared, the results suggest that the FCs have the best discriminatory power to distinguish the samples from different manufacturers.

Identification of the Product Manufacturers. We further evaluated two classifiers, LDA and k-NN, as potential tools to determine the manufacturers of the samples based on the FCs. The first two FCs from each pair were used to construct the classification models. For comparison purposes, the two peaks with the largest ξ from each pair and the first six PCs were also used. The classification results are summarized in Table 3.

The results demonstrate that the choice of the type of fingerprint features as input variable is an important factor for the classification results. The two FC-based classification models correctly classified all the samples. In contrast, the peak-based models and PC-based models yielded only about 90% correct classification using both classifiers. The results further confirm our early observation that the information encoded in the FCs has more sample-specific characteristics compared to other fingerprint features, i.e., the peaks and PCs.

While the LDA and k-NN methods differ in a number of ways, they produced similar results. This suggests that the nature of fingerprint features (input variables) used, and more specifically the effectiveness in which the features encode

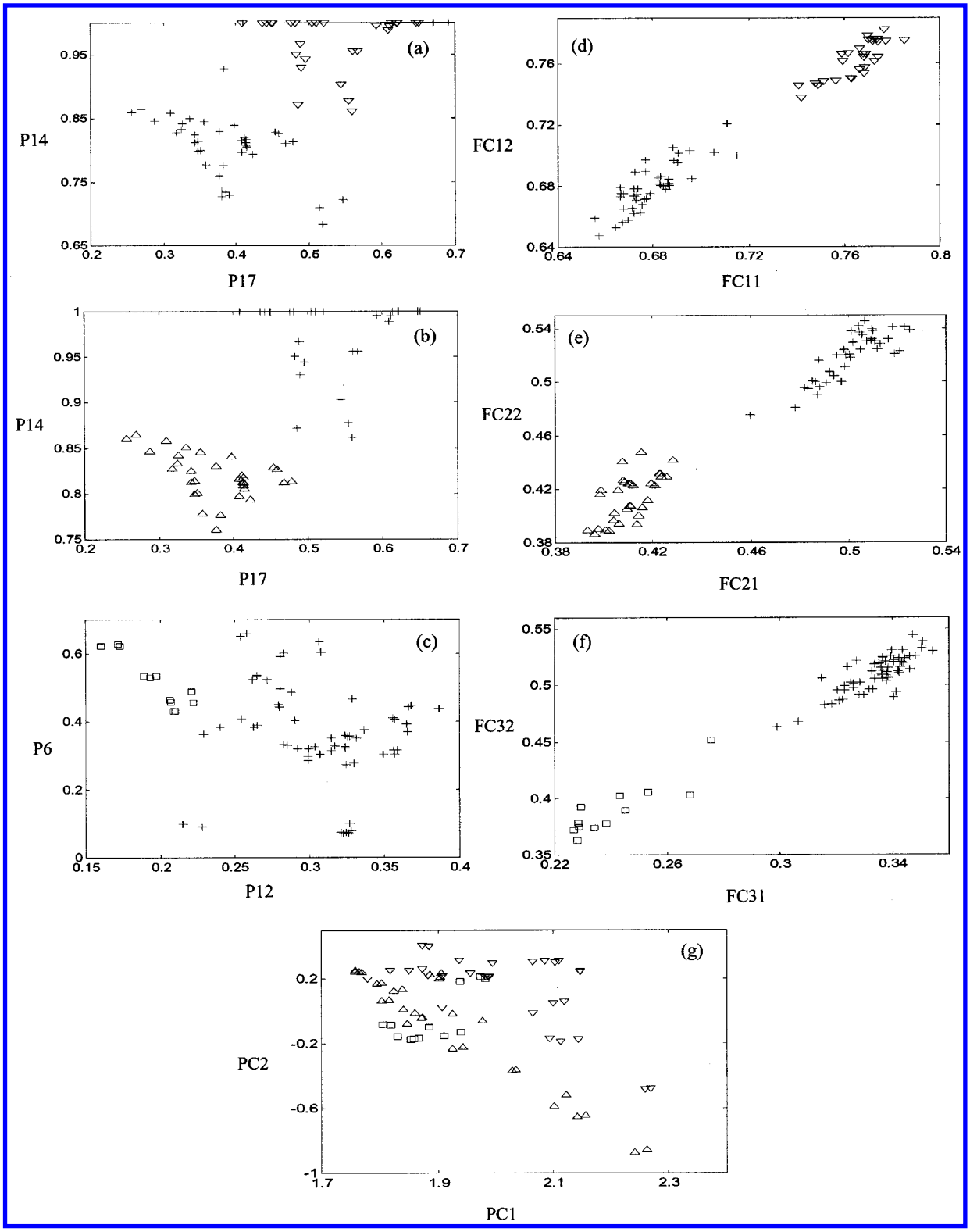

Figure 3. Score plots of the fingerprint features for different manufacturers. (a)−(c) Features of peaks for manufacturer A, B, and C; (d)−(f) Fisher components for manufacturer A, B, and C; (g) Principal components for three manufacturers. (□, samples of manufacturer A; Δ, samples of manufacturer B; ▽, samples of manufacturer C; +, samples of the other manufacturers of each pair.)

the information specified to the Shenmai injection, sometimes is far more important than the specific classifier employed. This finding emphasizes that a robust approach to extract meaningful fingerprint features is the critical step to develop a procedure for quality control of botanical drugs.

In summary, we found that the manufacturers of the Shenmai injection could be correctly identified using either LDA or k-NN based on the FCs. The FCs contains sufficient discriminatory power to distinguish the Shenmai injection manufacturers and should also be useful for monitoring the product consistency and detecting frauds.

Monitoring the Lot-to-lot Consistency of Products. Using the FCs, the vector of a chromatographic fingerprint can be reduced to two or three dimensions. Importantly, most

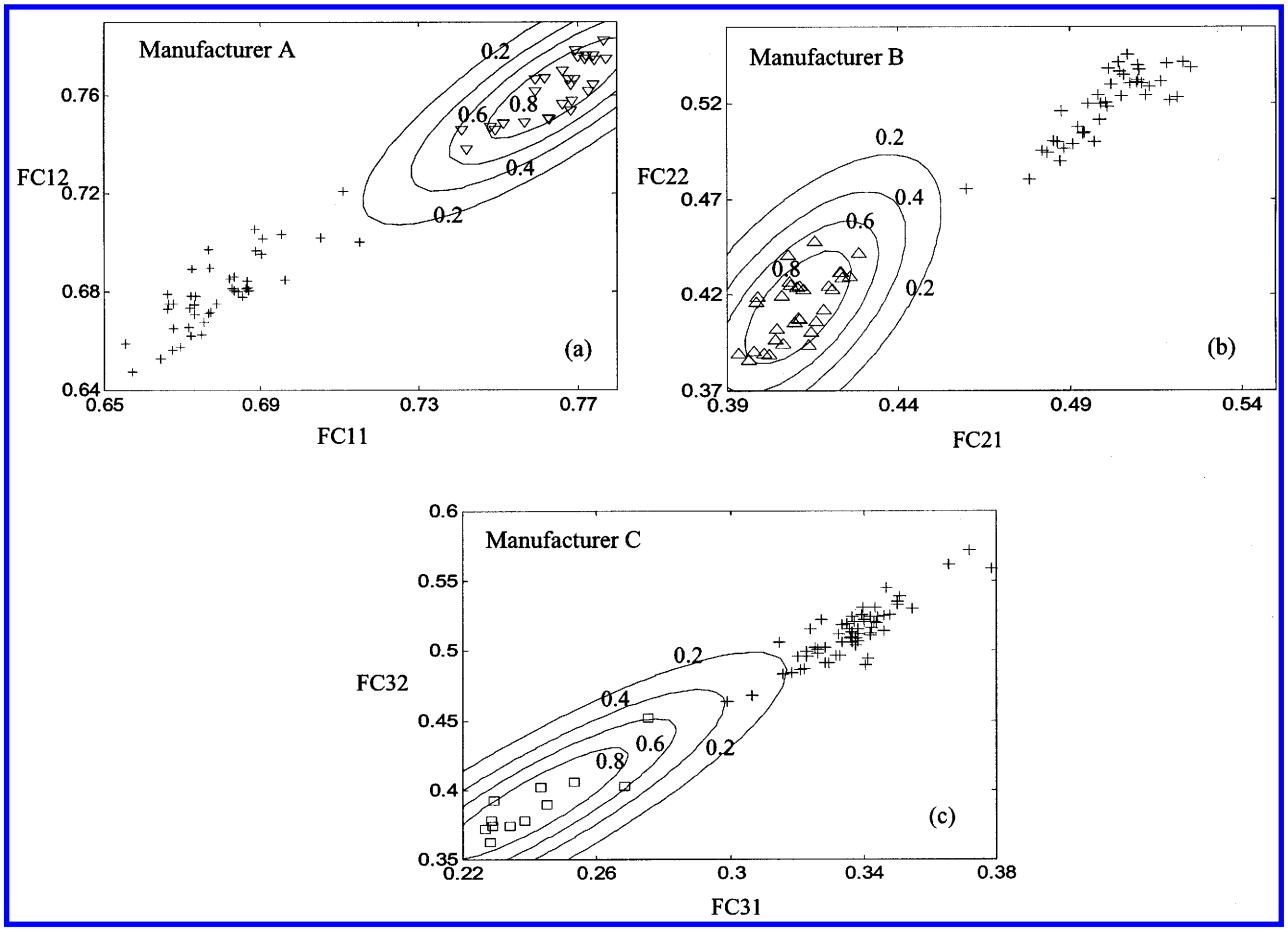

Figure 4. Similarity contour maps for each manufacturer based on the manufacturer-specific FCs. The markers are the same as those in Figure 3.

Table 4. Summary of Similarity Results of the Products and the Suspected Frauds

	products			fraud	
manufacturer-specific template	manufacturer A	manufacturer B	manufacturer C	sample 1	sample 2
manufacturer A	0.9933/0.6375	0.0884/3.063e-4	0.1269/3.237e-4	9.8634e-144	5.4234e-140
manufacturer B	0.0134/5.304e-6	0.9827/0.6695	0.1068/2.398e-5	5.1343e-59	2.0233e-109
manufacturer C	0.3906/0.0048	0.2079/0.0266	0.9814/0.5927	4.6342e-8	7.9324e-6

discriminatory information of the chromatographic fingerprint is retained in this low dimensionality. Thus, the similarity between chromatographic fingerprints (i.e. samples) could be inspected visually, which greatly facilitates the process to monitor the lot-to-lot consistency of the products.

To determine the lot-to-lot consistency, we first constructed a template FC that was the mean of the FCs of all the samples from a specific manufacturer. We then constructed similarity contour maps for each manufacturer based on the manufacturer-specific template FCs, which are depicted in Figure 4a−c for the manufacturer A, B, and C, respectively. The similarity between a tested sample and the samples of a specific manufacturer could be easily inspected based on the contour maps. If the products from different lots of a specific manufacturer are consistent, they should aggregate into a single cluster in the score plot.

As shown in Figure 4a, the products from the manufacturer A were tightly clustered together. This was also true for the products from the manufacturer B (Figure 4b). The results indicate the lot-to-lot consistencies for both manufacturer A and B. Relatively speaking, the samples from the manufacturer C are a little less consistent (Figure 4c).

The lot-to-lot consistency can also be quantitatively evaluated using the similarity value calculated from eq 19 that is also displayed in Figure 4. All the samples from the manufacturer A fell into a domain that has a similarity value greater than 0.6 (Figure 4a). In the same figure, other samples were far away from this domain and had much low values of the similarity (less than 0.2). This implies that the products between lots from the manufacturer A are consistent, and they are different from the products of other manufacturers. At the same time, although the similarity values of most samples from the manufacturer C were greater than 0.8 (Figure 4c), the similarity values of two samples were below 0.6. Furthermore, at least three samples from other manufacturers had the similarity between 0.2 and 0.4.

The maximum/minimum similarities between the manufacturer-specific templates and the samples of a manufacturer are summarized in Table 4. As shown in the table, the minimum similarity in the samples from the manufacturer A is 0.6375, while maximum similarities between the samples from manufacturers B,C and the manufacturer A-specific template are 0.088 and 0.1268, respectively. This indicates that samples from manufacturers B and C are

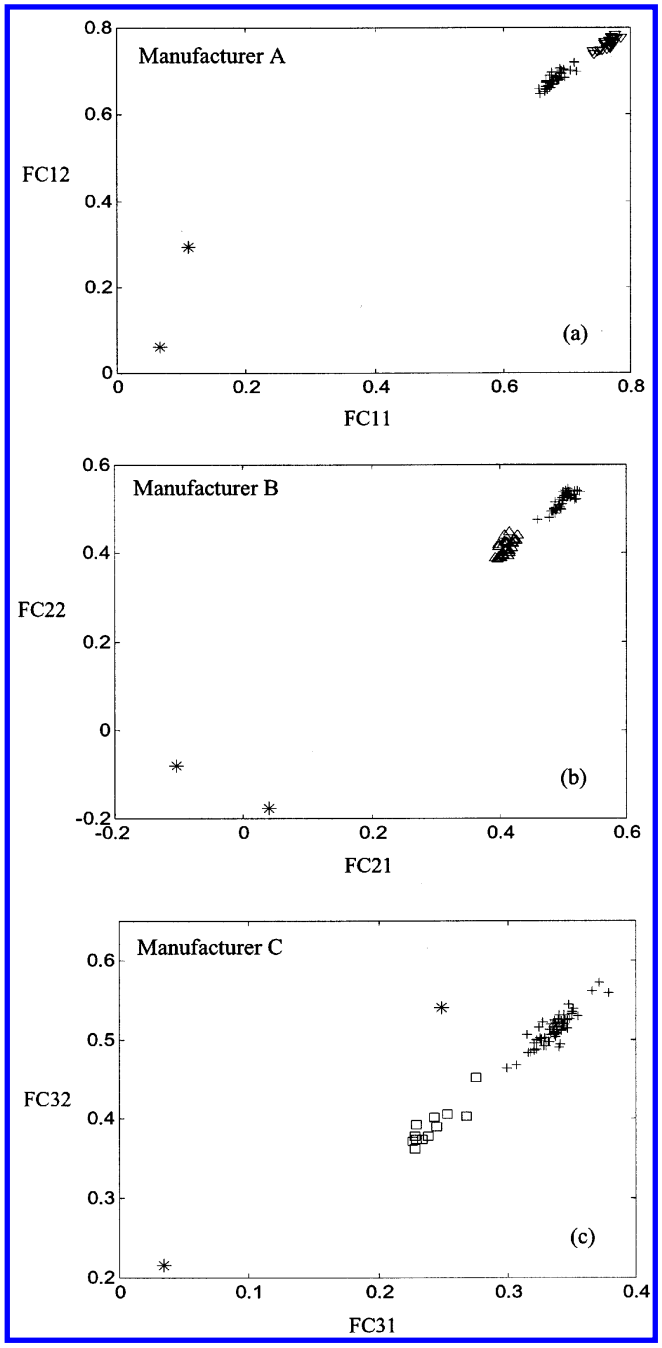

Figure 5. Illustration of the frauds on the score plots for each manufacturer based on the manufacturer-specific FC. The marker * represents frauds, and the others are the same as those in Figure 3.

different from those from manufacture A. The similar results could be observed for the samples from manufacturers B and C. The conclusions support those obtained from the visual inspection of FCs.

Therefore, similarity determination based on the FCs in conjunction with the visual inspection offers a powerful way to monitor the lot-to-lot consistency of botanical drugs.

Detection of Fraud. It was demonstrated in the previous examples that the presented approach is capable of identifying the samples of the manufacturer X from others that might be hypothesized as the "frauds" of the manufacturer X. Thus, the instance of a fraud could be identified using chromatographic fingerprinting for determining the equivalence between samples. Two samples of Shenmai injection that were suspected to be produced from inappropriate materials were used to verify the detecting capability on the frauds. The association/disassociation of the fraud samples to a particular manufacturer was measured by the similarity value that was compared to the maximum and minimum similarity values for the manufacturer.

As shown in Table 4, the maximum and minimum similarities in the samples from the manufacturer A are 0.9933 and 0.6375, respectively. However, the similarities between the two suspected samples and the samples from the manufacturer A are almost equal to zero. The similar results were also observed in the relationship between the suspected samples and other two manufacturers B and C. Thus, it can be concluded that the two suspected samples did not belong to any of three manufacturers. This conclusion could be further confirmed by visual inspection of the score plots (Figure 5) of three manufacturers, where the fraud samples are marked with *.

The evaluation of the two suspected samples suggested that the combination of a proper chemometrics method with chromatographic fingerprint could be useful for fraud detection.

CONCLUSIONS

In this paper, we report our preliminary results to apply the chemometrics approaches for quality control of botanical drugs using the chromatographic fingerprint. We first introduced a method to derive the FCs that encode most sample-specific information. The method greatly facilitates comparative analysis of the chromatographic fingerprints of botanical drug samples. The classifiers based on the FCs provided a means to identify the manufacturer of a sample. Based on the FCs, the lot-to-lot consistency and frauds can be determined using either a similarity measure or visual inspection or both in conjunction. The successful application of the chemometric methods for quality control of botanical drugs has been demonstrated for 73 Shenmai injections. We concluded that the application of chemometrics methods for chromatographic fingerprinting offers potential for ensuring the identity, consistency, and authenticity of botanical drugs.

ACKNOWLEDGMENT

This project was financially supported by the Chinese Natural Science Found (No. 20245002), the Chinese National Basic Research Priorities Program (No. G1999054405), and the Chinese key technologies R&D Program (No. 2001-BA701A01).

REFERENCES AND NOTES

(1) Welsh, W. J.; Lin, W.; Tersigni, S. H.; Collantes, E.; Duta, R.; Carey, M. S.; Zielinski, W. L.; Brower, J.; Spencer, J. A.; Layloff, T. P. Pharmaceutical Fingerprinting: Evaluation of Neural Networks and Chemometric Techniques for Distinguishing among Same-Product Manufactures. *Anal. Chem.* **1996**, *68*, 3473−3482.

(2) Collantes, E. R.; Duta, R.; Welsh, W. J.; Zielinski, W. L.; Brower, J. Preprocessing of HPLC Trace Impurity Patterns by Wavelet Packets for Pharmaceutical Fingerprinting Using Artificial Neural Networks. *Anal. Chem.* **1997**, *69*, 1392−1397.

(3) Aksenova, T. I.; Tetko, I. V.; Ivakhnenko, A. G.; Villa, A. E. P.; Welsh, W. J.; Zielinski, W. L. Pharmaceutical Fingerprinting in Phase Space. 1. Construction of Phase Fingerprints. *Anal. Chem.* **1999**, *71*, 2423−2430.

(4) Lavine, B. K.; Brzoowski, D.; Moores, A. J.; Davidson, C. E.; Mayfield, H. T. Genetic algorithm for fuel spill identification. *Anal. Chim. Acta* **2001**, *437*, 233−246.
(5) Berente, B.; García, D. D. C.; Reichenäbcher, M.; Danzer, K. Method Development for the Determination of Anthocyanins in Red Wines by High-performance Liquid Chromatography and Classification of German Red Wines by Means of Multivariate Statistical Methods. *J. Chromatogr. A* **2000**, *871*, 95−103.
(6) Klemenc, S. Common Batch Searching of Illicit Heroin Samples—Evaluation of Data by Chemometrics Methods. *Forensic Sci. Int.* **2001**, *115*, 43−52.
(7) World Health Organization. *Guildlines for the Assessment of Herbal Medicine*; WHO: Geneva, 1996.
(8) Pietta, P.; Mauri, P.; Bruno, A.; Rava, A.; Manera, E.; Ceva, P. Identification of Flavonoids from *Ginkgo Biloba L.*, *Anthemis Nobilis L.* and *Equisentum Arvense L.* by High-performance Liquid Chromatography with Diode-array UV Detection. *J. Chromatogr.* **1991**, *553*, 223−231.
(9) Hasler, A.; Sitcher, O.; Meier, B. Identification and Determination of the Flavonoids from *Ginkgo Biloba* by High-performance Liquid Chromatography. *J. Chromatogr.* **1992**, *605*, 41−48.
(10) Hansen, L. B. A Stable Gas Chromatography-Mass Spectrometry Analysis System to Characterize Ma Huang Products Found in Health Foods and Supplements. *J. Pharm. Sci*. **2001**, *90(7)*, 943−948.
(11) Cheng, Y. Y.; Chen, M. J.; Wu, Y. J. The Measures for Determining the Similarity of Chemical Fingerprint and a Method of Evaluating the Measures. *Acta Chim. Sinica* **2002**, *60(11)*, 2017−2021.
(12) Zhang, H. J.; Wu, Y. J.; Cheng, Y. Y. Analysis of 'SHENMAI' Injection by HPLC/MS/MS. *J. Pharm. Biomed. Anal.* **2003**, *31(1)*, 175−183.
(13) Sahota, R. S.; Morgan, S. L. Recognition of Chemical Markers in Chromatographic Data by an Individual Feature Reliability Approach to Classification. *Anal. Chem.* **1992**, *64*, 2382−2392.
(14) Qi, Y. H.; Xu, L. Classification of Valence Changes of Trivalent Rare Earth Ions in Alkaline Earth Orates Using Artificial Neural Networks. *Chemometrics Intelligent Laboratory Systems* **1999**, *45(1−2)*, 287−293.
(15) Massart, D. L.; Vandeginste, B. G. M.; Deming, S. N.; Michotte, Y.; Kaufman, L. *Chemometrics: A Textbook*; Elsvier: Amsterdam, 1998.
(16) Varmuze, K. *Pattern Recognition in Chemistry*; Springer: Berlin, 1980.

CI034034C