


## Integration of SIMCA and near-infrared spectroscopy for rapid and precise identification of herbal medicines

Follow this and additional works at: <https://www.jfda-online.com/journal>

 Part of the [Food Science Commons](#), [Medicinal Chemistry and Pharmaceutics Commons](#), [Pharmacology Commons](#), and the [Toxicology Commons](#)



This work is licensed under a [Creative Commons Attribution-Noncommercial-No Derivative Works 4.0 License](#).

### Recommended Citation

Yang, I.-C.; Tsai, C.-Y.; Hsieh, K.-W.; Yang, C.-W.; Ouyang, F.; Lo, Y.M.; and Chen, S. (2013) "Integration of SIMCA and near-infrared spectroscopy for rapid and precise identification of herbal medicines," *Journal of Food and Drug Analysis*: Vol. 21 : Iss. 3 , Article 15.

Available at: <https://doi.org/10.1016/j.jfda.2013.07.008>

This Original Article is brought to you for free and open access by Journal of Food and Drug Analysis. It has been accepted for inclusion in Journal of Food and Drug Analysis by an authorized editor of Journal of Food and Drug Analysis.

Available online at [www.sciencedirect.com](http://www.sciencedirect.com)

SciVerse ScienceDirect

journal homepage: [www.jfda-online.com](http://www.jfda-online.com)

## Original Article

# Integration of SIMCA and near-infrared spectroscopy for rapid and precise identification of herbal medicines



I-Chang Yang<sup>a,b</sup>, Chao-Yin Tsai<sup>c,d</sup>, Kuang-Wen Hsieh<sup>e</sup>, Ci-Wen Yang<sup>c</sup>,  
Fu Ouyang<sup>c</sup>, Yangming Martin Lo<sup>a</sup>, Suming Chen<sup>b,c,d,\*</sup>

<sup>a</sup>Department of Nutrition and Food Science, University of Maryland, College Park, MD, USA

<sup>b</sup>Taiwan Agricultural Mechanization Research and Development Center, Taipei, Taiwan, ROC

<sup>c</sup>Department of Bio-Industrial Mechatronics Engineering, National Taiwan University, Taipei, Taiwan, ROC

<sup>d</sup>Bioenergy Research Center, National Taiwan University, Taipei, Taiwan, ROC

<sup>e</sup>Department of Bio-Industrial Mechatronics Engineering, National Chung-Hsing University, Taichung, Taiwan, ROC

## ARTICLE INFO

## Article history:

Received 9 November 2012

Received in revised form

6 January 2013

Accepted 11 March 2013

Available online 8 August 2013

## Keywords:

Classification

Herbal medicine

Near infrared

SIMCA

Spectroscopy

## ABSTRACT

The recognition, control, and monitoring of herbal medicinal materials is a crucial work and challenge in the pharmaceutical industry. Consequently, the development of a rapid and accurate inspection method and model is an important goal and job. The raw materials of a variety of herbal medicines were measured using nondestructive near-infrared spectroscopy with soft independent modeling of class analogy to build up the classification model. The adulterated samples could be eliminated by the analysis of the model, and identification rates were demonstrated in the range of 98–100%. The method could be applied not only to the pharmaceutical industry but also to the food industry. Food materials can be measured with the inspection model for effective identification and determination of adulteration.

Copyright © 2013, Food and Drug Administration, Taiwan. Published by Elsevier Taiwan LLC. Open access under [CC BY-NC-ND license](https://creativecommons.org/licenses/by-nc-nd/4.0/).

## 1. Introduction

The integration of Chinese herbal medicines into modern medical treatments has assumed a more prominent profile in recent years [1]. Without purification to single out specific ingredients, quality control of these raw materials of pharmaceutical value is crucial and yet remains a challenge,

especially when accurate determination of functional ingredients in raw materials often requires extensive analytical tasks. In fact, many Chinese herbal medicines are now preserved in dry powder form for ease of storage and distribution. The herbs are not easily differentiated visually for materials still in dried whole form without expert training; identification of the specific varieties in powder form is even more

\* Corresponding author. Department of Bio-Industrial Mechatronics Engineering, National Taiwan University, No. 1, Sec. 4, Roosevelt Road, Taipei, 10617, Taiwan, ROC.

E-mail address: [schen@ntu.edu.tw](mailto:schen@ntu.edu.tw) (S. Chen).

1021-9498 Copyright © 2013, Food and Drug Administration, Taiwan. Published by Elsevier Taiwan LLC. Open access under [CC BY-NC-ND license](https://creativecommons.org/licenses/by-nc-nd/4.0/).  
<http://dx.doi.org/10.1016/j.jfda.2013.07.008>

challenging. There is a dire need to develop a rapid and accurate detection method for these herbal materials in order to foster their application in modern medicine.

It is well recognized that the quality of herbal pharmaceuticals can vary considerably because of the inherent variability caused by sourcing from regions with different climatic and soil conditions. Adulteration problems are also likely to be encountered as economic incentives continue to grow for herbal materials. To date, sensory and chemical analyses are often required for the inspection and/or identification of herbal medicines. With morphological or histological techniques to differentiate herbal materials becoming impossible in the powder form, analytical approaches such as high-performance liquid chromatography [2], gas chromatography–mass spectrometry [3], thin layer chromatography [4], and capillary electrophoresis [5] are often used. However, it is impossible to apply these methods for online inspection because of the complicated sample preparation procedures and long analysis times required. Moreover, these sample-destructive methods are poorly suited for inspection because they inevitably damage and consume the sample materials.

Nondestructive inspection methods such as optical [6,7], ultrasonic [8], and electromagnetic [9] techniques that make online inspection and automation possible are becoming popular with practical applications in a variety of areas, especially in the pharmaceutical [10,11], food [12], agricultural [13], chemical [14], and biobased industries [15]. Because traditional Chinese medicines come from a variety of different plant parts, near-infrared spectroscopic analysis is well suited to the analysis of the highly varied chemical compositions of the herbal materials [16,17]. The above inspections with high-performance classification methods would be needed and appreciated. The soft independent modeling of class analogy (SIMCA) method provides a useful classification of high-dimensional variations and incorporates principal component analysis (PCA) to reduce the dimensions of the spectra [18,19]. A SIMCA model consists of a collection of PCA models, and the data sets are independent. In comparison with the nonlinear modeling method, the calculation speed of SIMCA with PCA can be increased by calculating the covariance matrices and the indices. Therefore, the SIMCA model with PCA was selected for the development of the offline calibration model.

In the previous study [16], a robust inspection model using near-infrared spectroscopy with artificial neural network (ANN) analysis was developed for the classification of herbal materials. Although the results were excellent, developing the calibration database was time-consuming, consequently limiting the applicability of the model. In the present study, a SIMCA [20,21] based on near-infrared spectroscopy was developed to improve the calculation speed and retain the capacity for accurate classification of the herbal medicine materials.

## 2. Materials and methods

This study used 48 different raw herbal medicines that were provided by Sun Ten Pharmaceutical Co., Ltd. (New Taipei City, Taiwan). In the form of dry powder. For each of the 48

medicines, there were 30 individual samples, producing a total of 1440 spectral measurements for this study. Each sample was loaded into a 20-mL vial such that the volume of the vial was approximately 2/3 full with a powder depth of at least 2 cm, in order to prevent light transmission during the spectral reflectance measurement.

This study included only raw unprocessed dry herbal sample types originating primarily from plant leaves, stems, roots, flowers, fruits, seeds, and nuts; other possible herbal ingredients that occur naturally in nonsolid forms, such as liquids or gels, were not included. All samples were ground to the crude powder form for the near-infrared spectroscopic inspection.

### 2.1. Sample preparation and grouping

Table 1 lists the pool of 48 medicinal herbs used in this study. The stability of the SIMCA model was evaluated using Group 1 data, which consisted of three nonoverlapping calibration subgroups (A, C, E) paired with three prediction subgroups (B, D, F), respectively. Thirty different kinds of herbs were randomly selected (from among the 48 available) and separated into the three independent calibration subgroups. The selection process also assigned 15 herbs to each prediction subgroup, allowing for some overlap between the prediction subgroups and also between nonpaired calibration and prediction subgroups. There was no overlap between A and B, between C and D, or between E and F. The selection process was as follows: (1) starting with the pool of 48 available herbs, subgroup A (10 herbs) was selected, leaving 38 herbs available in the pool; (2) both subgroups B (15) and C (10) were each selected randomly from the pool of 38 and then the members of subgroup C were eliminated, leaving 28 in the pool; (3) both subgroups D (15) and E (10) were each selected randomly from the pool of 28 and then the members of subgroup E were eliminated, leaving 18 in the pool; (4) finally, subgroup F (15) was selected from the pool of 18.

The above data in Group 1 was used to test SIMCA models using different (independent) sets of spectral data. For Group 2, the pool of all 48 herbs was separated into two subgroups: calibration subgroup G containing 30 herbs and prediction subgroup H containing 18 herbs. These Group 2 data were used to test SIMCA models using a larger set of calibration data; by design, the herbs in subgroup G (calibration set of 30) include all the herbs in Group 1 calibration subgroups A (10) and C (10) but none of those in subgroup E (10).

#### Group 1

- Subgroup A contained 10 herbs: *Angelicae Sinensis Radix*, *Artemisia scoparia Waldst et Kit.*, *Citrus Sinensis Exocarpium*, *Clematidis Radix*, *Ligustici Rhizoma*, *Magnoliae Flos*, *Nelumbinis Folum*, *Platycodi Radix*, *Polyporus*, and *Rhei Rhizoma*.
- Subgroup B contained 15 herbs: *Atractylodis Rhizoma*, *Bupleuri Radix*, *Cinnamomi Ramulus*, *Ephedrae Herba*, *Evodiae Fructus*, *Hoelen*, *Paeoniae Lactiflorae Radix*, *Paeoniae Veitchii Radix*, *Perillae Folium*, *Pinelliae Tuber*, *Puerariae Radix*, *Salviae Miltiorrhizae Radix*, *Saposhinkoviae Radix*, *Scutellariae Radix*, and *Zingiberis Siccum Rhizoma*.

**Table 1 – The list of 48 herbal medicines and their affiliated subgroups (A ~H).**

Herbal medicines	Group 1						Group 2	
	Cal.	Prd.	Cal.	Prd.	Cal.	Prd.	Cal.	Prd.
	A (10)	B (15)	C (10)	D (15)	E (10)	F (15)	G (30)	H (18)
<i>Achyranthis Radix</i>				✓			✓	
<i>Amomi Semen</i>				✓			✓	
<i>Angelicae Sinensis Radix</i>	✓			✓			✓	
<i>Artemisia scoparia</i> Waldst et Kit.	✓			✓			✓	
<i>Astragalus membranaceus</i> Bqe.					✓			✓
<i>Atractylodis Rhizoma</i>		✓	✓				✓	
<i>Bupleuri Radix</i>		✓					✓	
<i>Cinnamomi Ramulus</i>		✓					✓	
<i>Cinnamomum japonicum</i> Sieb.					✓			✓
<i>Citrus Sinensis Exocarpium</i>	✓			✓			✓	
<i>Citrus Undeveloped Exocarpium</i>				✓			✓	
<i>Clematidis Radix</i>	✓			✓		✓	✓	
<i>Coix lacryma-jobi</i> L.					✓			✓
<i>Curcumae Radix</i>				✓			✓	
<i>Cyperus rotundus</i>				✓			✓	
<i>Dioscorea opposita</i> Thumb.					✓			✓
<i>Dolichos lablab</i> L.								✓
<i>Ephedrae Herba</i>		✓	✓				✓	
<i>Evodiae Fructus</i>		✓	✓				✓	
<i>Foeniculum vulgare</i>						✓		✓
<i>Forsythia suspensa</i> Vahl						✓		✓
<i>Glycyrrhiza uralensis</i> FISCH.					✓			✓
<i>Hoelen</i>		✓	✓			✓	✓	✓
<i>Houttuynia cordata</i> Thumb.					✓			✓
<i>L. var. orientale</i> SAMUELS.					✓			✓
<i>Ligustici Rhizoma</i>	✓			✓			✓	
<i>Lithospermum Officinale</i> Root					✓			✓
<i>Magnolia officinalis</i>					✓			✓
<i>Magnoliae Flos</i>	✓			✓		✓	✓	
<i>Morus alba</i>						✓		✓
<i>Nelumbinis Folium</i>	✓			✓		✓	✓	
<i>Paeonia suffruticosa</i> Andr					✓			✓
<i>Paeoniae Lactiflorae Radix</i>		✓	✓			✓	✓	
<i>Paeoniae Veitchii Radix</i>		✓	✓			✓	✓	
<i>Perillae Folium</i>		✓					✓	
<i>Pinelliae Tuber</i>		✓	✓				✓	
<i>Platycodi Radix</i>	✓			✓			✓	
<i>Polyporus</i>	✓			✓		✓	✓	
<i>Prunus armeniaca</i>						✓		✓
<i>Puerariae Radix</i>		✓				✓	✓	
<i>Rhei Rhizoma</i>	✓			✓		✓	✓	
<i>Salviae Miltiorrhizae Radix</i>		✓	✓			✓	✓	
<i>Saposhinkoviae Radix</i>		✓					✓	
<i>Saussureae Radix</i>								✓
<i>Scrophularia ningpoensis</i> HEMSLE.								✓
<i>Scutellariae Radix</i>		✓	✓			✓	✓	
<i>Sophora flavescens</i> AIT						✓		✓
<i>Zingiberis Siccatum Rhizoma</i>		✓	✓				✓	

Cal. = calibration set; Prd. = prediction set.

- Subgroup C contained 10 herbs: *Atractylodis Rhizoma*, *Ephedrae Herba*, *Evodiae Fructus*, *Hoelen*, *Paeoniae Lactiflorae Radix*, *Paeoniae Veitchii Radix*, *Pinelliae Tuber*, *Salviae Miltiorrhizae Radix*, *Scutellariae Radix*, and *Zingiberis Siccatum Rhizoma*.
- Subgroup D contained 15 herbs: *Achyranthis Radix*, *Amomi Semen*, *Angelicae Sinensis Radix*, *Artemisia scoparia* Waldst et Kit., *Citrus Sinensis Exocarpium*, *Citrus*

*Undeveloped Exocarpium*, *Clematidis Radix*, *Curcumae Radix*, *Cyperus rotundus*, *Ligustici Rhizoma*, *Magnoliae Flos*, *Nelumbinis Folum*, *Platycodi Radix*, *Polyporus*, and *Rhei Rhizoma*.

- Subgroup E contained 10 herbs: *Astragalus membranaceus* Bqe., *Cinnamomum japonicum* Sieb., *Coix lachrymal-jobi* L., *Dioscorea opposita* Thumb., *Glycyrrhiza uralensis* FISCH., *Houttuynia cordata* Thumb., *L. var.*

orientale SAMUELS., Lithospermum Officinale Root, Magnolia officinalis, and Paeonia suffruticosa Andr.

- Subgroup F contained 15 herbs: Clematidis Radix, Foeniculum vulgare, Forsythia suspensa Vahl, Hoelen, Magnoliae Flos, Morus alba, Nelumbinis Folium, Paeoniae Lactiflorae Radix, Paeoniae Veitchii Radix, Polyporus, Prunus armeniaca, Salivae Miltiorrhizae Radix, Rhei Rhizoma, Scutellariae Radix, and Sophora flavescens AIT.

#### Group 2

- Subgroup G contained 30 herbs: Achyranthis Radix, Amomi Semen, Anglica Sinensis Radix, Artemisia scoparia Waldst et Kit., Atractylodis Rhizoma, Bupleuri Radix, Cinnamomi Ramulus, Citrus Sinensis Exocarpium, Citrus Undeveloped Exocarpium, Clematidis Radix, Curcumae Radix, Cyperus rotundus, Ephedrae Herba, Evodiae Fructus, Hoelen, Ligustici Rhizoma, Magnoliae Flos, Nelumbinis Folum, Paeoniae Lactiflorae Radix, Paeonia Veitchii Radix, Perillae Folium, Pinelliae Tuber, Platycodi Radix, Polyporus, Puerariae Radix, Rhei Rhizoma, Sailvae Miltiorrhizae Radix, Saposhinkoviae Radix, Scutellariae Radix, and Zingiberis Siccatur Rhizoma.
- Subgroup H contained 18 herbs: Astragalus membranaceus Bqe., Cinnamomum japonicum Sieb., Coix lachrymal-jobi L., Dioscorea oppositifolia Thumb., Dolichos lablab L., Foeniculum vulgare, Forsythia suspensa Vahl, Glycyrrhiza uralensis FISCH., Houttuynia cordata Thumb., L. var. orientale SAMUELS., Lithospermum Officinale Root, Magnolia officinalis, Morus alba, Paeonia suffruticosa Andr., Prunus armeniaca, Saussureae Radix, Scrophularia ningpoensis HEMS., and Sophora flavescens AIT.

## 2.2. Apparatus and experiments

The spectra of the herb samples were measured on a FOSS NIRSystems instrument Model 6500 NIR reflectance spectrometer (FOSS NIRSystems, Inc., Laurel, MD, USA) configured with a rapid content analyzer module and a tungsten halogen lamp as the light source, and using the VISION 3.0 software (FOSS NIRSystems, Inc.) for system control and data acquisition. The samples were scanned at 2-nm intervals in the range of 400–2498 nm, encompassing the visible and near-infrared wavelengths. Silicon detectors were used below 1100 nm, followed by lead sulfide detectors above 1100 nm. Spectral analysis and the development of SIMCA models were carried out using Matlab 7.2 (The Mathworks, Inc., Natick, MA, USA) with PLS Toolbox 5.0 (Eigenvector Research, Inc., Wenatchee, WA, USA).

## 2.3. Data pretreatment

For all 1440 sample spectra, the standard normalized variate transformation [22] was applied to reduce spectral variation. For the Group 1 calibration subgroups (A, C, E), the 300 sample spectra in each subgroup were separated into two sets—a calibration set of 200 spectra and a validation set of 100 spectra—using the Kennard–Stone algorithm [23]. For Group 2, calibration subgroup G contained 900 sample spectra that were separated into a set of 600 calibration set samples and

300 validation set samples. SIMCA models were developed using the calibration data sets and then used to predict the spectral samples in the prediction sets.

## 2.4. Principal component analysis

PCA [20] is a useful chemometric analysis tool for spectral data compression and information extraction that allows the most important information contained in the spectra to be described using a small number of principal components (PC). In this study,  $\mathbf{M}$  is a near-infrared spectral data matrix with  $m$  rows ( $m$  samples) and  $n$  columns ( $n$  wavelengths). PCA decomposes  $\mathbf{M}$  as the sum of series combinations of  $\mathbf{t}_i$  and  $\mathbf{p}_i$ . The  $\mathbf{t}_i$  and  $\mathbf{p}_i$  pairs are ordered (i) by the amount of variance captured. The scores ( $\mathbf{t}_i$  vectors) contain information on how the samples relate to each other. The loadings ( $\mathbf{p}_i$  vectors) contain information on how the variables relate to each other. The PCA model is truncated after the  $k$  components and remaining variance factors are consolidated into a residual matrix  $\mathbf{E}$ :

$$\mathbf{M} = \mathbf{t}_1\mathbf{p}_1^T + \mathbf{t}_2\mathbf{p}_2^T + \cdots + \mathbf{t}_k\mathbf{p}_k^T + \mathbf{E} \quad (1)$$

Mathematically, PCA relies on an eigenvector decomposition of the covariance matrix of the process variables. For a given data matrix  $\mathbf{M}$  with  $m$  rows and  $n$  columns, the covariance matrix of  $\mathbf{M}$  is defined as:

$$\text{cov}(\mathbf{M}) = \frac{\mathbf{M}^T\mathbf{M}}{m-1} \quad (2)$$

The columns of  $\mathbf{M}$  have been “mean-centered” by subtracting the original mean of each column. In the PCA decomposition, the  $\mathbf{p}_i$  vectors are eigenvectors of the covariance matrix. For each  $\mathbf{p}_i$ :

$$\text{cov}(\mathbf{M})\mathbf{p}_i = \lambda_i\mathbf{p}_i \quad (3)$$

$\lambda_i$  is the eigenvalue associated with the eigenvector  $\mathbf{p}_i$ . The  $\mathbf{t}_i$  forms an orthogonal set ( $\mathbf{t}_i^T\mathbf{t}_j = 0$ , for  $i \neq j$ ), whereas  $\mathbf{p}_i$  is orthonormal ( $\mathbf{p}_i^T\mathbf{p}_j = 0$  for  $i \neq j$ ;  $\mathbf{p}_i^T\mathbf{p}_i = 1$ , for  $i = j$ ). Note that for  $\mathbf{M}$  and any  $\mathbf{t}_i$ ,  $\mathbf{p}_i$  pair,

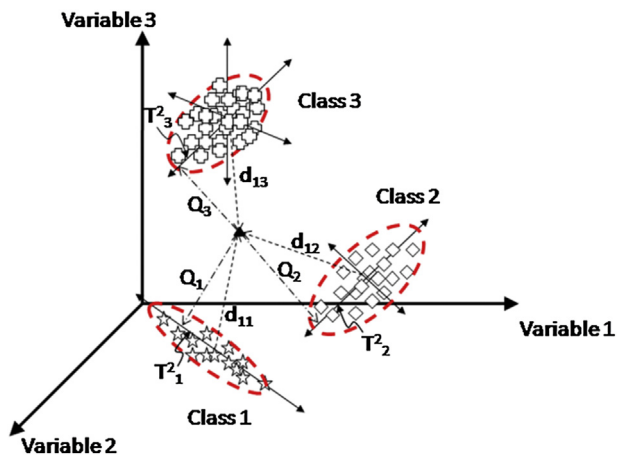
$$\mathbf{M}\mathbf{p}_i = \mathbf{t}_i \quad (4)$$

Here,  $\mathbf{t}_i$  is the projection of  $\mathbf{M}$  onto  $\mathbf{p}_i$ . The  $\mathbf{t}_i$  and  $\mathbf{p}_i$  pairs are arranged in descending order according to the associated, which is a measure of the amount of variance described by the  $\mathbf{t}_i$  and  $\mathbf{p}_i$  pair. The  $\mathbf{t}_1$ ,  $\mathbf{p}_1$  pair captures the greatest amount of variation in the data that can be captured with a linear factor; and then each subsequent pair captures the greatest possible amount of variance remaining after subtracting  $\mathbf{t}_i\mathbf{p}_i^T$  from  $\mathbf{M}$ .

## 2.5. Monitoring indices associated with PCA models

Hotelling's  $T^2$  and the  $Q$  residual [24] are two indices commonly used to evaluate new data using a previously developed PCA model. Hotelling's  $T^2$  can be viewed as the distance from a sample's projection into the  $k$ -dimensional subspace to the centroid of the subspace. The  $Q$  residual is the distance between a sample point in  $n$ -space and its projection in the  $k$ -dimensional subspace of the model (as shown in Fig. 1).





**Fig. 1 – The concept of Hotelling's  $T^2$  and Q residual under the three-variable condition with their own principal component distributions.**

### 2.5.1. Hotelling's $T^2$

Hotelling's  $T^2$  statistic gives a measure of significant variation of the process. It is the sum of normalized squared scores divided by their variance:

$$T^2 = \mathbf{t}^T \boldsymbol{\lambda}^{-1} \mathbf{t} = \sum_{i=1}^k \frac{t_i^2}{\lambda_i} \quad (5)$$

where  $\boldsymbol{\lambda}^{-1}$  is a diagonal matrix of the inverse of the  $k$  largest eigenvalues  $\lambda_i$  of covariance matrix  $\text{cov}(\mathbf{M})$  in descending order, and  $t_i$  is the  $i$ th score. The statistical thresholds for  $T^2$  can be calculated using the  $F$  distribution as follows:

$$T_{\alpha}^2 = \frac{k(m-1)}{(m-k)} F_{\alpha}(k, m-k) \quad (6)$$

where  $T_{\alpha}^2$  is the threshold value with an  $\alpha$  significance level of confidence (95% in this case),  $m$  is the number of samples used to build the PCA model, and  $k$  is the number of principal components retained in the model.  $F_{\alpha}(k, m-k)$  is the  $\alpha$  confidence interval of the  $F$  distribution with  $k$  and  $(m-k)$  degrees of freedom.

### 2.5.2. The Q residual

The Q residual is a measure of the variation of the data outside of the principal components included in the PCA model. The mismatch between measured and estimated sensor readings results in the residual  $e$ , which forms the basis of the Q statistic, which is formulated as follows:

$$\mathbf{e} = \mathbf{x} - \mathbf{t}\mathbf{p}^T = \mathbf{x}[\mathbf{I}_n - \mathbf{p}\mathbf{p}^T] \quad (7)$$

and

$$Q = \mathbf{e}^T \mathbf{e} = \sum_{j=1}^n e_j^2 \quad (8)$$

where  $e_j$  is the  $j$ th residual. The statistical thresholds for the Q residual can be calculated as follows:

$$Q_{\alpha} = \theta_1 \left[ \frac{h_0 c_{\alpha} \sqrt{2\theta_2}}{\theta_1} + \frac{\theta_2 h_0 (h_0 - 1)}{\theta_1^2} + 1 \right]^{\frac{1}{h_0}} \quad (9)$$

where

$$\theta_i = \sum_{j=k+1}^n \lambda_j^i, \text{ for } i = 1, 2, 3 \quad (10)$$

and

$$h_0 = 1 - \frac{2\theta_1\theta_3}{3\theta_2^2} \quad (11)$$

In Equation 10,  $c_{\alpha}$  is the standard normal deviate corresponding to the upper  $(1 - \alpha)$  percentile. In Equation 11,  $k$  is the number of principal components retained in the model and  $n$  is the total number of principal components.

### 2.5.3. SIMCA

A soft independent method of class analogy (SIMCA) model consists of a collection of PCA models, one for each class in the data set. PCA with Hotelling's  $T^2$  and Q residual is shown graphically in Fig. 1. Each class can have a different number of principal components; the number depends on the data in the class.

Some discussion of the geometric interpretation of Q residual and Hotelling's  $T^2$  is perhaps in order. As noted above, Q residual is a measure of the variation of the data outside of the principal components included in the PCA model. Hotelling's  $T^2$  is a measure of the distance from the multivariate mean (the intersection of the PCs in the figure) to the projection of the sample onto the two principal components. Hotelling's  $T^2$  limit defines an ellipse on the plane within which the data normally project. A sample with a large Hotelling's  $T^2$  value (but a small Q residual) is shown on the upper right-hand side of Fig. 1.

The nearest class to a sample is defined as the class model that results in a minimum distance of the sample  $i$  to model  $j$ ,  $d_{ij}$ :

$$d_{ij} = \sqrt{(Q_r)^2 + (T_r^2)^2} \quad (12)$$

with a reduced Q residual:

$$Q_r = \frac{Q}{Q_{0.95}} \quad (13)$$

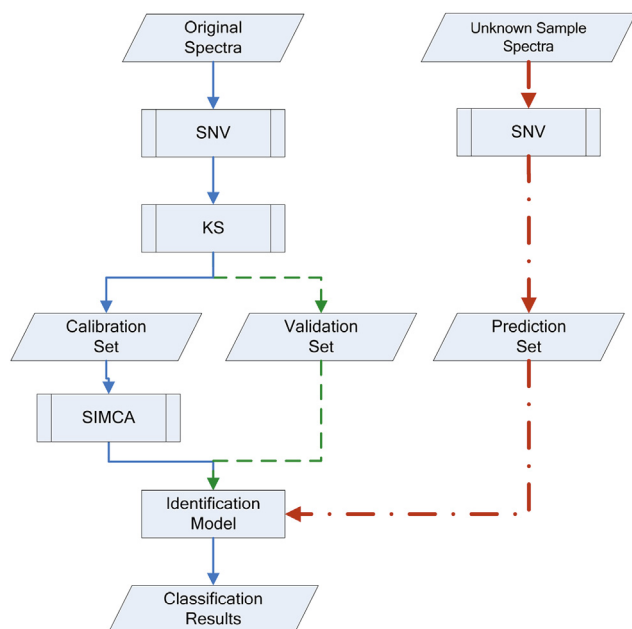
where  $Q_{0.95}$  is the 95% confidence interval for the model under consideration, and reduced Hotelling's  $T^2$ :

$$T_r^2 = \frac{T^2}{T_{0.95}^2} \quad (14)$$

with a similarly defined  $T_{0.95}^2$ . This distance measure gives equal weight to distance in the model space ( $T^2$ ) and in the residual space (Q). The use of reduced Q residual and  $T^2$  allows a direct comparison of the values of specific samples on different SIMCA submodels. Without the use of these reduced statistics, each model's  $T^2$  and Q residual values would be on very different numerical scales and not easily comparable.

## 3. Results and discussion

The purpose of the experimental design (Fig. 2) in this study was to allow for the comparison of the stability and performance of multiple SIMCA models built on different



**Fig. 2 – The pretreatments and the logical flow of different calibration, validation, and prediction sets.**

independent combinations of medicinal herbs, and to determine the effectiveness and stability of a SIMCA model when expanded to include a larger set of samples. Therefore, different calibration, validation, and prediction sets and pretreatments were designed to be operated.

### 3.1. Step 1. Results for three independent calibration subgroups

The average spectra of all the herbs, organized by subgroups, are shown in Fig. 3, and are not easily separated by the naked eye. The large variations in the visible range (400–700 nm) were attributable to the basic colors of the medicines. In Group 1, the calibration subgroups A, C, and E were paired to prediction subgroups B, D, and F, respectively, with no overlap in membership between the paired calibration and prediction subsets. The calibration and prediction subgroups were completely independent combinations. For each of the A, C, and E calibration subgroups, the calibration and validation sets were selected (using the Kennard–Stone algorithm) to perform the first test of the SIMCA models.

In calibration subgroups A, C, and E of Group 1, the number of selected principal components was based on the cumulative variance (%) of each herb, as shown in Table 2. Consequently, the number of principal components (PCs) in the calibration set differed for different PCA submodels. There were two to five PCs for subgroup A, two to four PCs for subgroup C, and three to four PCs for subgroup E. The larger range for subgroup A indicated a larger variance between the samples in set A.

For subgroups A, C, and E, three to four PCs described more than 95% of the cumulative variance presented by most of the samples. In subgroup A, there were three medicines—*Artemisia scoparia* Walodst et Kit., *Clemaridis Radix*, and *Ligustici Rhizoma*—that were well described by only two PCs,

and only one—*Plyporus*—that needed five PCs. In set C, only one sample—*Hoelen*—needed only two PCs to account for 95.17% of the cumulative variance, whereas all other samples in the subgroup required more PCs. All the herbs in subgroup E needed three to four PCs to explain above 95% cumulative variance.

The first step in the development of SIMCA models was to calculate the sub-PCA model of a single medicine, and then, for the remaining samples, the principal component distribution would be calculated using the sub-PCA model in the principal component space. The differentiation between herbs could then be determined by Hotelling's  $T^2$  and  $Q$  residual values to implement classification and identification. Fig. 4 shows the samples of the 10 herbs (A01 through A10) in subgroup A as distributed by their Hotelling's  $T^2$  and  $Q$  residual values calculated using the sub-PCA model for the 20 calibration set samples of herb A01 (*Angelicae Sinensis Radix*). The A01 samples appear nearest to the origin coordinates (0, 0) and can be clearly differentiated by Hotelling's  $T^2$  and  $Q$  residual values <1, as shown by the expanded view in the upper left inset box.

To complete the SIMCA model for subgroup A, a sub-PCA model was calculated for each set of the 20 calibration set samples for the remaining nine herbs (A02 through A10) in subgroup A. These models, based on the calibration set for subgroup A, were used to calculate Hotelling's  $T^2$  and  $Q$  residual values for the validation set samples for subgroup A, and then also for the samples in prediction subgroup B. The samples could then be assigned a binary identification of either 0 (belonging to subgroup A, based on Hotelling's  $T^2$  and  $Q$  residual values less than 1) or 1 (not belonging to subgroup A, based on Hotelling's  $T^2$  and  $Q$  residual values greater than 1). Fig. 5A and B show the identification of the subgroup A validation set samples (which belong to the subgroup A database) and of the prediction subgroup B samples (which do not belong to the subgroup A database), respectively.

The SIMCA model also calculates the nearest class of subgroup A for any given sample. Fig. 5C shows that the subgroup A validation set matches the classes defined by the subgroup A calibration set. For the samples in subgroup B that do not match the classes defined by the subgroup A calibration set, Fig. 5D shows the nearest class (defined within subgroup A) that was calculated for each of the “unknown” subgroup B samples, based on the mean values of Hotelling's  $T^2$  or  $Q$  residual or both.

Fig. 5 illustrates the information that can be provided by the two-stage SIMCA analysis for unknown samples. The first stage identifies whether an unknown sample belongs to the set of known samples as defined by the calibration set, which is useful for the authentication of an herbal sample. If an unknown sample does belong to the calibration set, then the second stage of analysis can then determine the specific match in the calibration set for the unknown sample.

Table 3 shows the results for the Group 1 SIMCA models when applied to identify whether an unknown sample belongs to the set of calibration samples on which the model was built (application of the first-stage SIMCA analysis). The subgroup-A SIMCA model correctly identified 100% of the 200 calibration set samples and 100% of the 100 validation set samples of subgroup A as positive matches (belonging to

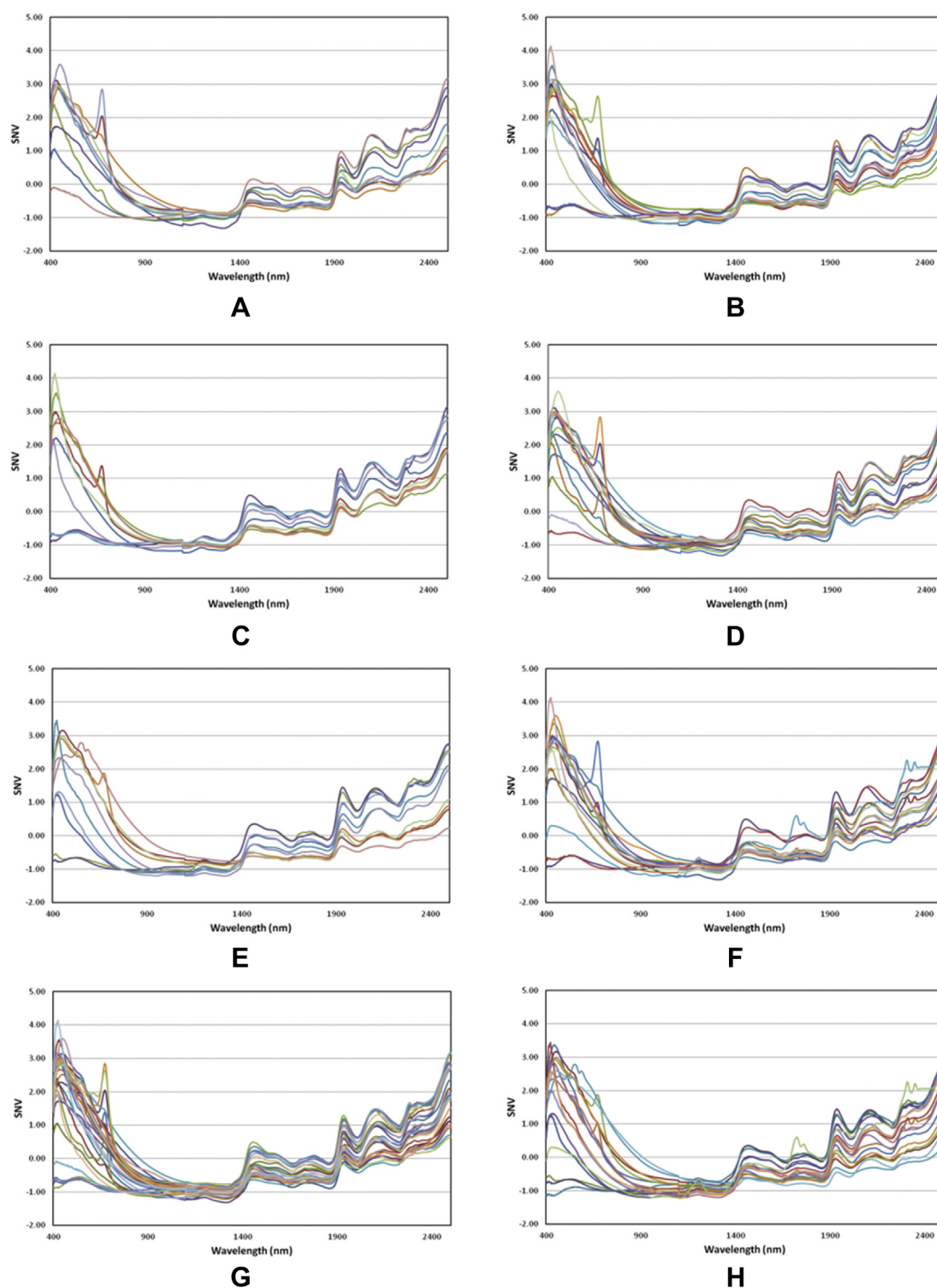


Fig. 3 – (A) The spectra of the calibration sets A, C, E, and G, and (B) the spectra of the prediction sets B, D, F, and H.

subgroup A), and correctly identified 100% of the 450 samples in prediction subgroup B as nonmatches. The subgroup-C SIMCA model correctly identified 99% of the 200 subgroup C calibration set samples and 98% of the 100 subgroup C validation set samples as positive matches, and 100% of the 450 samples in prediction subgroup D as nonmatches. The

subgroup-E SIMCA model correctly identified 100% of the 200 subgroup E calibration set samples and 99% of the 100 subgroup E validation set samples as positive matches, and 100% of the 450 samples in prediction subgroup F as nonmatches.

The calibration data sets in Group 1 consisted of three independent sets of 10 different Chinese medicinal herbs. The



**Table 2 – The herbal membership principal components and cumulative variance (%) of the calibration subgroups A, C, E, and G.**

Set	Herbal name	Principal component	Cumulative variance (%)
A	Angelicae Sinensis Radix	3	96.07
	Artemisia scoparia Waldst et Kit.	2	96.78
	Citrus Sinensis Exocarpium	4	96.44
	Clematidis Radix	2	96.96
	Ligustici Rhizoma	2	96.26
	Magnoliae Flos	3	96.45
	Nelumbinis Folium	3	98.51
	Platycodi Radix	4	97.39
	Polyporus	5	97.54
C	Rhei Rhizoma	4	95.01
	Atractylodis Rhizoma	3	95.79
	Ephedrae Herba	4	97.74
	Evodiae Fructus	4	96.11
	Hoelen	2	95.17
	Paeoniae Lactiflorae Radix	4	95.76
	Paeoniae Veitchii Radix	2	97.23
	Pinelliae Tuber	4	97.77
	Salviae Miltiorrhizae Radix	4	97.64
	Scutellariae Radix	3	96.87
	Zingiberis Siccum Rhizoma	3	96.79
E	Astragalus membranaceus Bge.	3	95.07
	Cinnamomum japonicum Sieb.	3	97.81
	Coix lacryma-jobi L.	3	97.58
	Dioscorea opposita Thumb.	4	95.74
	Glycyrrhiza uralensis FISCH.	3	96.12
	Houttuynia cordata Thumb.	3	96.91
	L. var. orientale SAMUELS.	4	97.92
	Lithospermum Officinale Root	4	96.75
	Magnolia officinalis	4	95.41
	Paeonia suffruticosa Andr	3	95.45
G	Achyranthis Radix	3	95.32
	Amomi Semen	3	98.74
	Anglicae Sinensis Radix	3	96.07
	Artemisia scoparia Waldst et Kit.	2	96.78
	Atractylodis Rhizoma	3	95.79
	Bupleuri Radix	4	97.05
	Cinnamomi Ramulus	4	95.87
	Citrus Sinensis Exocarpium	4	96.44
	Citrus Undeveloped Exocarpium	4	96.64
	Clematidis Radix	2	96.96
	Curcumae Radix	3	96.24
	Cyperus rotundus	3	98.25
	Ephedrae Herba	4	97.74
	Evodiae Fructus	4	96.11
	Hoelen	2	95.17
	Ligustici Rhizoma	2	96.26
	Magnoliae Flos	3	96.45
	Nelumbinis Folum	3	98.51
	Paeoniae Lactiflorae Radix	4	95.76
	Paeoniae Veitchii Radix	2	97.23
	Perillae Folium	6	95.31
	Pinelliae Tuber	4	97.77
	Platycodi Radix	4	97.39
	Polyporus	5	97.54
	Puerariae Radix	2	95.82
	Rhei Rhizoma	4	95.01
	Salviae Miltiorrhizae Radix	4	97.64
	Saposhinkoviae Radix	2	95.01
	Scutellariae Radix	3	96.87
	Zingiberis Siccum Rhizoma	3	96.79

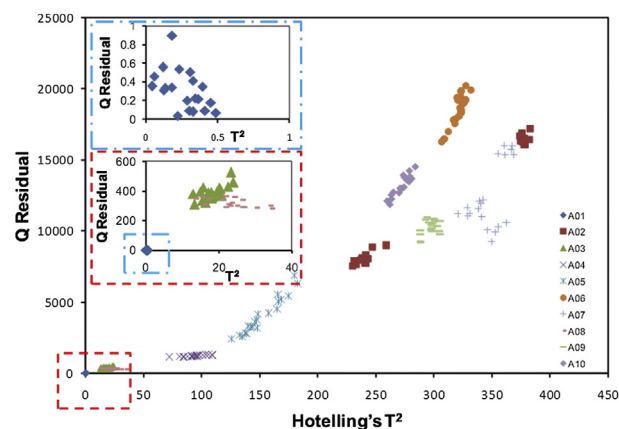
purpose of this experiment design was to test the stability of the SIMCA models—that is, whether different herbal medicines could be recognized by using different combinations of herbs on which the SIMCA models were built. The results showed that the herbs could be identified rapidly and accurately by the SIMCA models built on the different herb combinations.

### 3.2. Step 2. Results of one extended group for the practical application

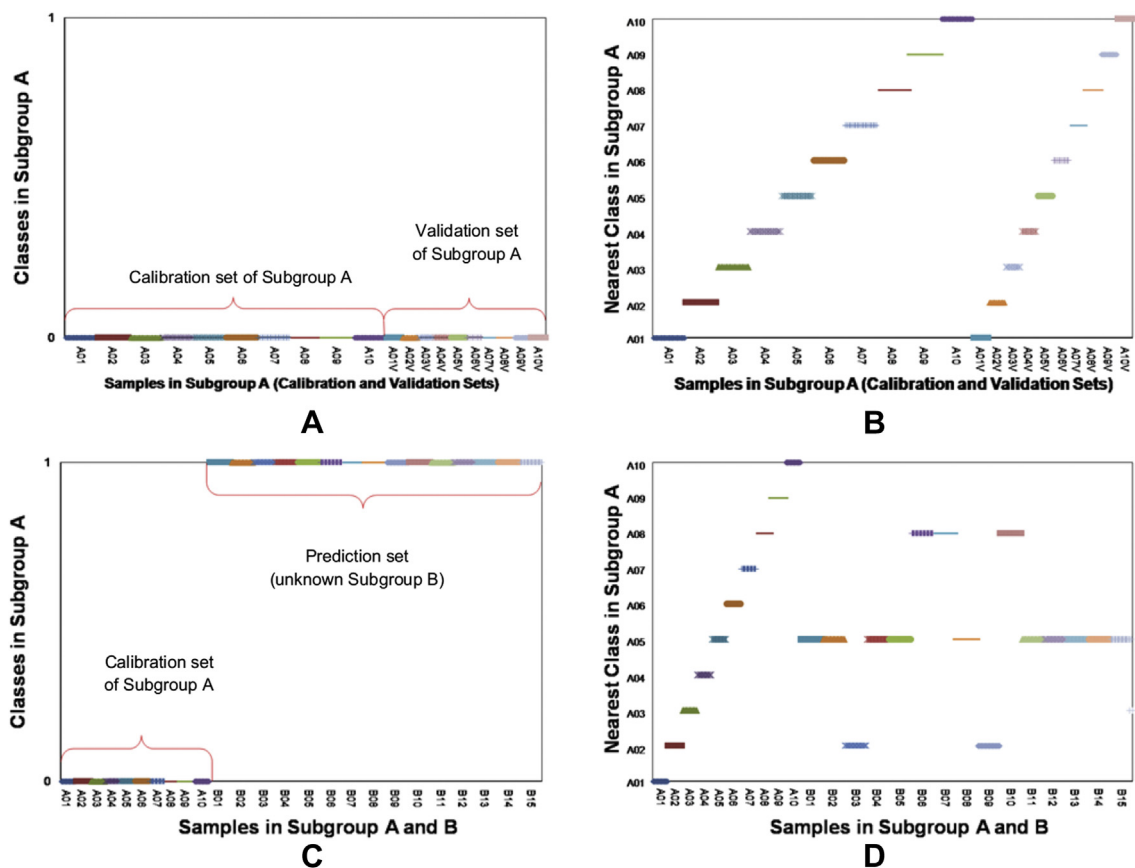
Table 3 also shows the results for the Group 2 SIMCA model, an extended model built on the larger sample calibration set of subgroup G based on 30 different herbs (600 samples in the subgroup G calibration set). For the experimental design, subgroup G included all of the samples from Group 1 subgroups A, B, C, and D, but none of the samples in subgroup E and only 10 of the 30 samples in subgroup F. There was no sample overlap between Group 2 subgroups G and H.

The subgroup-G SIMCA model correctly identified 99.5% of the 600 subgroup G calibration set samples and 99% of the 300 subgroup G validation set samples as positive matches, and correctly identified all subgroup H samples (540 samples spanning 18 herb varieties) as nonmatches, to subgroup G. Examining the samples by their Group 1 organization, the extended subgroup-G SIMCA model correctly identified the match or nonmatch status of 100%, 98%, and 100%, of the samples in subgroups A, C, and E, respectively, and 98.67%, 100%, and 99.33% of the samples in subgroups B, D, and F, respectively.

There were three misidentified samples in the calibration set of subgroup G: *Hoelen* and *Paeoniae Lactiflorae Radix* from subgroup C, and *Puerariae Radix* from subgroup B. The selection of the number of principal components to use in modeling each medicinal herb was a crucial step in SIMCA model development. Too many principal components can cause overfitting, whereas too few principal components will inadequately describe the spectral characteristics of any given group. In Group 2, 30 medicines were selected



**Fig. 4 – The distribution of Hotelling's  $T^2$  and Q residual values for the 10 sample herbs in subgroup A, according to the sub-PCA model calculated for sample A01. PCA = principal component analysis.**



**Fig. 5** – Results of using the SIMCA model built on subgroup A calibration set to (identify/classify) samples in subgroup A validation set and in prediction subgroup B. A and B show the identification of the subgroup A validation set samples and of the prediction subgroup B samples, respectively. C shows that the subgroup A validation set matches the classes defined by the subgroup A calibration set. D shows the nearest class (defined within subgroup A) that was calculated for each of the “unknown” subgroup B samples, based on the mean values of Hotelling’s T2 or Q residual or both. SIMCA = soft independent modeling of class analogy.

from the available pool of 48 to test the limitation of the classification and the calculation speed of the SIMCA model, including a variety of different plant parts—roots, stems, leaves, whole plant, flowers, fruits, etc. Because the powder

forms in which herbal medicines are most commonly stored prior to processing are even more difficult to visually differentiate than dried roots, stems, leaves, etc., the development of a rapid and accurate method for their

Table 3 – The SIMCA model identification rates for the calibration, validation, and prediction sample subsets.										
Set	Calibration			Validation			Set	Prediction		
	Total Sample	Misidentified Sample	Correct identified rate (%)	Total sample	Misidentified sample	Correct identified rate (%)		Total sample	Misidentified sample	Correct identified rate (%)
A	200	0	100.00	100	0	100.00	B	450	0	100.00
C	200	2	99.00	100	2	98.00	D	450	0	100.00
E	200	0	100.00	100	1	99.00	F	450	0	100.00
G	600	3	99.50	300	3	99.00	A	100	0	100.00
							B	450	6	98.67
							C	100	2	98.00
							D	450	0	100.00
							E	100	0	100.00
							F	450	3	99.33
							H	540	0	100.00
SIMCA = soft independent modeling of class analogy.										

classification and identification is becoming an increasingly important and urgent need.

### 3.3. Comparisons and discussions

Using Hotelling's  $T^2$  and Q residual values, this SIMCA method effectively identified herbal matches and nonmatches relative to the herbs in the database. For the herbs identified as non-matches (outside of the database), Hotelling's  $T^2$  and Q residual values could still be useful for general classification in terms of resemblance to database members. Our previous study [16] described the effective classification of medicinal herbs implemented by using an ANN method; however, this ANN method needed longer calculation times that were not practical for online inspection applications. As the numbers of samples and herbal varieties increased, both the ANN calculation loading and required calculation times increased significantly. By contrast, the SIMCA method described in this study can be used to classify or identify medicinal herbs with greater flexibility in accommodating larger numbers of sample types and with shorter calculation times when compared to the ANN model.

The SIMCA method can be applied to identify medicinal herb samples that fall outside of those defined within the database, and find the nearest class in the database that shows any resemblance to nonmatching samples. This method can also be applied to other materials in the food and pharmaceutical industries. In theory, raw food materials prior to processing can be inspected using near-infrared spectroscopy with SIMCA to detect abnormal agents (e.g., plasticizers) that would be detected as being outside of the current database of regular materials that should be found in the processing operations. This inspection method could be developed as a rapid and powerful tool for the early detection of adulterants or contaminants in raw materials.

## 4. Conclusions

The method provided a flexible modeling database for the control of raw materials when processing Chinese medicinal herbs. Hotelling's  $T^2$  and Q residuals are useful indicators in the SIMCA models for evaluating differences between the samples. The use of near-infrared spectroscopy integrated with SIMCA can be developed into an accurate and high-performance industrial method for online inspection applications.

## Acknowledgments

The authors appreciate the support from Sun Ten Pharmaceutical Co., Ltd. (New Taipei City, Taiwan), which provided all the medicinal herbs used in this study, and from FOSS NIR-Systems, Inc. (Silver Spring, MD, USA), which supplied the Model 6500 NIR reflectance spectrometer configured with a rapid content analyzer (RCA) module. We also thank Ms Diane E. Chan in the Agricultural Research Service, BARC, United

States Department of Agriculture, for the proofreading and providing comments.

## REFERENCES

- [1] Liu C, Tseng A, Yang S. Chinese herbal medicine: modern applications of traditional formulas. Boca Raton, FL: CRC Press; 2005.
- [2] Chuang CC, Su CH, Huang WY, et al. Classification of *Fangchi Radix* samples by multivariate analysis. *J Food Drug Anal* 2008;16:48–56.
- [3] Gong F, Liang YZ, Xie PS, et al. Information theory applied to chromatographic fingerprint of herbal medicine for quality control. *J Chromatogr A* 2003;1002:25–40.
- [4] Simonvska B, Vovk I, Andersek S, et al. Investigation of phenolic acids in yacon (*Smallanthus sonchifolius*) leaves and tubers. *J Chromatogr A* 2003;1016:89–98.
- [5] Yang JJ, Long H, Liu HW, et al. Analysis of terandrine of fangchinoline in traditional Chinese medicines by capillary electrophoresis. *J Chromatogr A* 1998;811:274–9.
- [6] Kondo N. Machine vision based on optical properties of biomaterials for fruit grading system. *Environ Control Biol* 2006;44:151–9.
- [7] Woo YA, Kim HJ, Ze KR, et al. Near-infrared (NIR) spectroscopy for the non-destructive and fast determination of geographical origin of *Angelicae gigantis Radix*. *J Pharmaceut Biomed Anal* 2005;36:955–9.
- [8] Cai L, Wang Q, Gu C, et al. Vascular and micro-environmental influences on MSC-coral hydroxyapatite construct-based bone tissue engineering. *Biomaterials* 2011;32:8497–505.
- [9] Mukhopadhyay SC, Gupta GS, Woolley JD, et al. Saxophone reed inspection employing planar electromagnetic sensors. *IEEE Trans Instrum Meas* 2007;56:2492–503.
- [10] Reich G. Near-infrared spectroscopy and imaging: basic principles and pharmaceutical applications. *Adv Drug Deliv Rev* 2005;57:1109–43.
- [11] Zou HB, Yang GS, Qin ZR, et al. Progress in quality control of herbal medicine with IR fingerprint spectra. *Anal Lett* 2005;38:1457–75.
- [12] Williams P, Norris K. Near-infrared technology in the agricultural and food industries. St. Paul, MN, USA: American Association of Cereal Chemists; 1987.
- [13] Watson C. Near infrared reflectance spectrophotometric analysis of agricultural products. *Anal Chem* 1977;49:835–40.
- [14] Murray I, Williams P. Chemical principles of near-infrared technology; 1987.
- [15] Grift T, Zhang Q, Kondo N, et al. A review of automation and robotics for the bioindustry. *J Biomechatron Eng* 2008;1:37–54.
- [16] Yang CW, Chen S, Ouyang F, et al. A robust identification model for herbal medicine using near infrared spectroscopy and artificial neural network. *J Food Drug Anal* 2011;19:9–17.
- [17] Sun S-Q, Zhou Q, Chen J-B. Infrared spectroscopy for complex mixtures. Applications in food and traditional Chinese medicine. Beijing, China: Chemistry Industry Press; 2011.
- [18] Wold S. Pattern recognition by means of disjoint principal components models. *Pattern Recogn* 1976;8:127–39.
- [19] Branden KV, Hubert M. Robust classification in high dimensions based on the SIMCA method. *Chemometr Intell Lab* 2005;79:10–21.
- [20] Wold S, Esbensen K, Geladi P. Principal component analysis. *Chemometr Intell Lab* 1987;2:37–52.

- 
- [21] Dunn III WJ, Wold S. An assessment of the carcinogenicity of N-nitroso compounds by the SIMCA method of pattern recognition. *J Chem Inf Comp Sci* 1981;21:8–13.
- [22] Barnes R, Dhanoa M, Lister SJ. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Appl Spectrosc* 1989;43:772–7.
- [23] Kennard RW, Stone LA. Computer aided design of experiments. *Technometrics* 1969;11:137–48.
- [24] Chen Q, Kruger U, Meronk M, et al. Synthesis of  $T^2$  and  $Q$  statistics for process monitoring. *Control Eng Pract* 2004;12:745–55.