JOURNAL OF AGRICULTURAL AND FOOD CHEMISTRY

J. Agric. Food Chem. **2008,** *56,* 11589–11595 **11589**

Application of a ^{1}H Nuclear Magnetic Resonance (NMR) Metabolomics Approach Combined with Orthogonal Projections to Latent Structure-Discriminant Analysis as an Efficient Tool for Discriminating between Korean and Chinese Herbal Medicines

JINHO KANG,[†,‡] MOON-YOUNG CHOI,[†,§] SUNMI KANG,[‡] HYUK NAM KWON,[‡] HE WEN,[‡] CHANG HOON LEE,[||] MINSEOK PARK,[§] SUSANNE WIKLUND,[⊥] HYO JIN KIM,[#] SUNG WON KWON,*[,§] AND SUNGHYOUK PARK*[,‡]

Department of Biochemistry and Center for Advanced Medical Education by BK21 project, College of Medicine, Inha University, Chungsuk Building, Room 505, Shinheung-dong, Chung-gu, Incheon 400-712, Korea, College of Pharmacy and Research Institute of Pharmaceutical Sciences, Seoul National University, Seoul 151-742, Korea, Molecular Oncology Branch, National Cancer Center, Madu 1-dong, Goyang-si, Gyeonggi-do 416-769, Korea, Department of Chemistry, Umeå University, SE-901 87 Umeå, Sweden, and Department of Pharmacy, Dongduk Women's University, 23-1, Sungbuk-gu, Hawolgok-dong Seoul 136-714, Korea

Correct identification of the origins of herbal medical products is becoming increasingly important in tandem with the growing interest in alternative medicine. However, visual inspection of raw material is still the most widely used method, and newer scientific approaches are needed. To develop a more objective and efficient tool for discriminating herbal origins, particularly Korean and Chinese, we employed a nuclear magnetic resonance (NMR)-based metabolomics approach combined with an orthogonal projections to latent structure-discriminant analysis (OPLS-DA) multivariate analysis. We first analyzed the constituent metabolites of *Scutellaria baicalensis* through NMR studies. Subsequent holistic data analysis with OPLS-DA yielded a statistical model that could cleanly discriminate between the sample groups even in the presence of large structured noise. An analysis of the statistical total correlation spectroscopy (STOCSY) spectrum identified citric acid and arginine as the key discriminating metabolites for Korean and Chinese samples. As a validation of the discrimination model, we performed blind prediction tests of sample origins using an external test set. Our model correctly predicted the origins of all of the 11 test samples, demonstrating its robustness. We tested the wider applicability of the developed method with three additional herbal medicines from Korea and China and obtained very high prediction accuracy. The solid discriminatory power and statistical validity of our method suggest its general applicability for determining the origins of herbal medicines.

KEYWORDS: Metabolomics; OPLS-DA; "oriental medicine"; prediction; NMR; *Scutellaria baicalensis*

INTRODUCTION

Metabolomics is an emerging -omics technology for examining the signature of low-molecular-weight compounds in a system (*1*−*3*) and has been applied to a variety of fields, such as plant science, toxicology, and clinical diagnosis (*4*−*7*). Just as with other -omics approaches, the aims are to categorize or classify samples and to understand the basic underlying principles that contribute to the differences among them. From a larger perspective, the metabolomic findings can be combined with other -omics data to gain a more system-wide understanding of the inter-relationships among genomes, proteomes, and metabolomes (*8, 9*). Metabolomics employs analytical small-molecule detection techniques, such as mass spectrometry and nuclear magnetic resonance (NMR), as well as statistical methods to analyze large amounts of data (*10*). Technical advances in analytical instrumentation as well as more theoretical statistical analysis have contributed to the recent rapid expansion of the metabolomics literature.

Plant metabolite profiling is a major metabolomics application field, particularly because plants produce a wide variety of

* To whom correspondence should be addressed. Telephone: +82-2-880-7844. Fax: +82-2-886-7844. E-mail: swkwon@snu.ac.kr (S.W.K.); Telephone: +82-32-890-0935. Fax: +82-32-884-6726. E-mail: spark@inha.ac.kr (S.P.).

[†] These authors contributed equally to this work.

[‡] Inha University.

[§] Seoul National University.

[||] National Cancer Center.

[⊥] Umeå University.

[#] Dongduk Women's University.

10.1021/jf802088a CCC: $40.75
Published on Web 11/26/2008

Table 1

name	^{1}H	^{13}C	assignment	name	^{1}H	^{13}C	assignment
baicalin	6.55 (s)	106.4	H-3	xylose	5.11 (d, J = 3.8)		H-1α
	6.77 (s)	96.9	H-8		4.57 (d, J = 9.5)		H-1β
	7.75 (d, J = 7.9)	128.5	H-2′		3.18 (m)		H-2
	7.51 (t, J = 7.9)	134.7	H-3′	arginine	3.72 (t, J = 7.0)	56.8	Hα
	5.13 (d, J = 8.0)	102.9	H-1″		3.22 (t, J = 6.2)	42.9	Hδ
	3.92 (d, J = 7.7)	78.8	H-5″		1.90 (m)	30.0	Hβ
		132.5	C-6		1.69 (m)	26.4	Hγ
		153.8	C-7			159.4	Cε
		152.5	C-9			176.2	COOH
		167.3	C-2	alanine	1.48 (d, J = 7.2)		Hβ
		185.4	C-4		3.73 (q, J = 7.2)	53.2	Hα
		108.6	C-10			177.6	COOH
		132.5	C-1′	valine	1.01 (d, J = 7.0)		Hγ
		177.4	C-6″		2.28 (m)		Hβ
wogonoside	5.10 (d, J = 7.8)	102.8	H-1″		3.50 (d, J = 4.5)		Hα
	3.91 (s)	63.3	$-OCH_3$	isoleucine	0.87 (t, J = 7.4)		Hδ
	7.95 (d, J = 7.9)		H-2′		1.256 (m)		Hγ
	7.43 (t, J = 7.9)		H-3′		1.542 (m)		Hγ
	6.50 (s)	101.4	H-6		1.992 (m)		Hβ
	6.59 (s)	106.9	H-3	aspartate	2.65 (dd, J = 17.0, 8.6)		Hβ
	3.93 (d, J = 7.7)		H-5″		2.86 (dd, J = 17.0, 3.7)		Hβ
		185.2	C-4		3.58 (dd, J = 8.6, 3.7)		Hα
		167.0	C-2	glutamine	2.12 (m)		Hβ
		108.1	C-10		2.45 (m)		Hγ
		132.3	C-1′		3.72 (t, J = 6.2)		Hα
		158.4	C-7	glutamate	2.09 (m)		Hβ
		177.3	C-6″		2.35 (m)		Hγ
sucrose	5.40 (d, J = 3.8)	94.8	H-1α (Glc)		3.72 (dd, J = 7.2, 4.7)		Hα
	4.18 (d, J = 8.9)	79.3	H-3 (Frt)	proline	2.05 (m)		Hγ
	4.04 (t, J = 8.7)	76.6	H-4 (Frt)		2.35 (m)		Hβ
	3.86 (m)	84.3	H-5 (Glc)		3.32 (m)		Hδ
	3.75 (m)	75.3	H-3 (Glc)		4.09 (dd, J = 8.4, 6.6)		Hα
	3.53 (dd, J = 10.0, 3.7)	73.8	H-2 (Glc)	lactate	1.33 (d, J = 7.2)		Hβ
	3.45 (t, J = 9.1)	72.1	H-10 (Frt)		4.15 (d, J = 7.2)		Hα
glucose	5.25 (d, J = 3.7)		H-1α	malate	2.52 (dd, J = 16.4, 7.1)	43.7	Hβ
	4.60 (d, J = 7.9)		H-1β		2.75 (dd, J = 16.4, 4.5)		Hβ
raffinose	4.97 (d, J = 3.6)	101.0	H-1α (Gal)		4.30 (dd, J = 7.1, 4.5)		Hα
	5.44 (d, J = 3.8)	94.6	H-1α (Glc)	citrate	2.52 (d, J = 17.5)	46.4	H-4
	3.96	73.6	−CHOH		2.71 (d, J = 17.5)	46.4	H-4
	3.52	72.0	−CHOH			78.02	C-3
						180.3	C-2
						183.2	C-1

metabolites that are directly related to the economic and medical values of plant-derived materials (*11, 12*). Among the various factors that determine the quality of plant materials, their origins have become increasingly important. For example, it was recently reported that American and Asian ginseng roots have contradictory effects on the vascular system (*13*) and acute glycemia (*14*). Additionally, correct identification of plant or agricultural products is a significant socioeconomic issue, because their prices vary greatly depending upon the origins, and there are many cases of malpractice and fraud. Conventionally, postmarket determination of the origins of plant materials has been performed by visual or microscopic inspection of raw material. However, this method can be subjective and cannot be applied to powder samples. Therefore, there is a great need for new approaches for the determination of the origins of plant material.

Metabolomics has been applied to the classification of plant materials, with principal component analysis (PCA) as the main statistical approach (*15*−*18*). For example, the metabolite changes occurring after viral infection, origin-classification of chamomile flower, and grade differentiation of pine mushrooms were studied with a combination of NMR and PCA methods. However, these PCA-based approaches have limited practical use, because PCA cannot assign the class membership of unknown test samples, which is critical to the validation of statistical models and, thus, to the practical application of metabolomics. PCA is the basis of multivariate modeling and is very useful for outlier detection and for finding patterns and trends. However, as an unsupervised method, it rotates a data matrix to find the maximum variations in the observations. Therefore, the resulting principal components do not necessarily align with the best predictive components for class separation or, here, the origins. Moreover, the prediction of class membership is better quantified using a supervised prediction and regression method, because the misclassification error can be approximated. Therefore, PCA is not the method of choice for class differentiation, which is required for determination of plant-sample origins.

A recently developed approach, orthogonal projections to latent structure-discriminant analysis (OPLS-DA), is a type of supervised classification and regression method that correlates spectroscopic data to a certain property, such as class membership, in this case, Korean or Chinese origin (*19*). The correlated variation between the observations and the different class types is found by rotating the components, so that the variation of main scientific interest will be observed in the first component, referred to as the predictive component, t_p (scores) and p_p (loadings) (*20*). The additional component in the OPLS-DA model is referred to the orthogonal (uncorrelated) component, t_o and p_o. This separation of predictive and orthogonal compo-

nents facilitates the interpretation of class differences (e.g., for biomarker identification), class-orthogonal divergence, and also the prediction of the class membership of unknown samples, which together increase the overall model usability (*21, 22*). Therefore, OPLS-DA is more suited than PCA to differentiate origins in cases where many factors can affect metabolite profiles.

In the present study, a metabolomics approach combining NMR spectroscopy with OPLS-DA was developed for discriminating the origins of Korean and Chinese herbal plant materials. The method was first developed on *Scutellaria baicalensis* samples and yielded superb results. The discrimination model was statistically sound and allowed for identification of signals underlying the differences between the two groups. Importantly, validation by predictions on blind test samples gave a statistical measurement of the reliability of the approach, which is required for practical application. As a test of the wider usage, the same method was applied to predict the origins of three other herbal medical products. The excellent results suggest that this approach should be useful for the development of generally applicable metabolomics tools for discriminating origins of herbal medicines.

MATERIALS AND METHODS

Sample Collection. All of the plants were collected in person by visiting the actual culture locations to guarantee the genuineness of the origins of the samples. Chinese plant samples were collected during two periods around September 20th and October 4th in 2006, and Korean samples were collected during two periods around October 20th and November 7th in 2006. A total of 38 *S. baicalensis* samples (roots) were collected from six different locations in Korea (Yeosoo, Sooncheon, Kwangyang, Koheung, Iksan, and Youngjoo) and six different locations in China (Shenyang, Chengdu, Yanbian, Heilongjiang, Neimenggu, and Huangshan). For *Atractyloides japonica* (roots), a total of 20 samples were collected from two different locations in Korea (Kwangjoo and Najoo) and two different locations in China (Hangzhou and Neimenggu). A total of 33 *Pueraria lobata* samples (roots) were collected from six different Korean locations (Iksan, Sooncheon, Kwangyang, Geochang, Kimcheon, and Andong) and five different Chinese locations (YanJi, YanBian, HeBei, ShenYang, and Guangxi). For *Alisma orientale* (stem), a total of 25 samples were collected from four different locations in Korea (Sooncheon, Iksan, Koheung, and Youngjoo) and four different locations in China (Chengdu, Fujian, Shenyang, and Duping). The collected samples were dried in the shade. All of the collected samples were independently inspected and authenticated by an expert plant systematician (Professor Chang Soo Yook, Kyunghee Universisty, Korea). For *S. baicalensis*, used as the initial system, separate quality control was performed according to the Korean Pharmacopoeia, and all of the samples passed the test [baicalin contents > 10% measured by high-performance liquid chromatography (HPLC); data not shown].

Sample Preparation for NMR Spectroscopy. Dried samples were ground with an electric blender, and 100 mg of each sample was extracted with a CD_3OD and D_2O mixture (1:1) with 10 mM potassium phosphate as buffer (pH 6.0). The extraction was carried out by sonication at room temperature in a bath-type sonicator for 20 min. Tetradeuterated trimethylsilanepropionic acid (TSP, final 0.025%) was added as an internal standard. Any particulate material was removed by centrifugation at 13000*g* for 1 min, and the supernatant was transferred to a standard 5 mm NMR tube.

NMR Spectroscopy. One-dimensional NMR spectra were measured on a 500 MHz Bruker Avance spectrometer equipped with a cryogenic triple-resonance probe at the Korea Basic Science Institute (Ochang, Korea). For efficient water suppression at the high Q probe, one-dimensional nuclear Overhauser effect spectrometry (NOESY) pulse program with water presaturation was used. All of the spectra were recorded with 16 000 complex points at 25°. The time domain data were Fourier-transformed, phase-corrected, and baseline-corrected manually. The signal intensities were normalized against the intensity of the 0.025% TSP signal at 0.00 ppm. The intensity values of all of the spectra were saved in one text file for data binning.

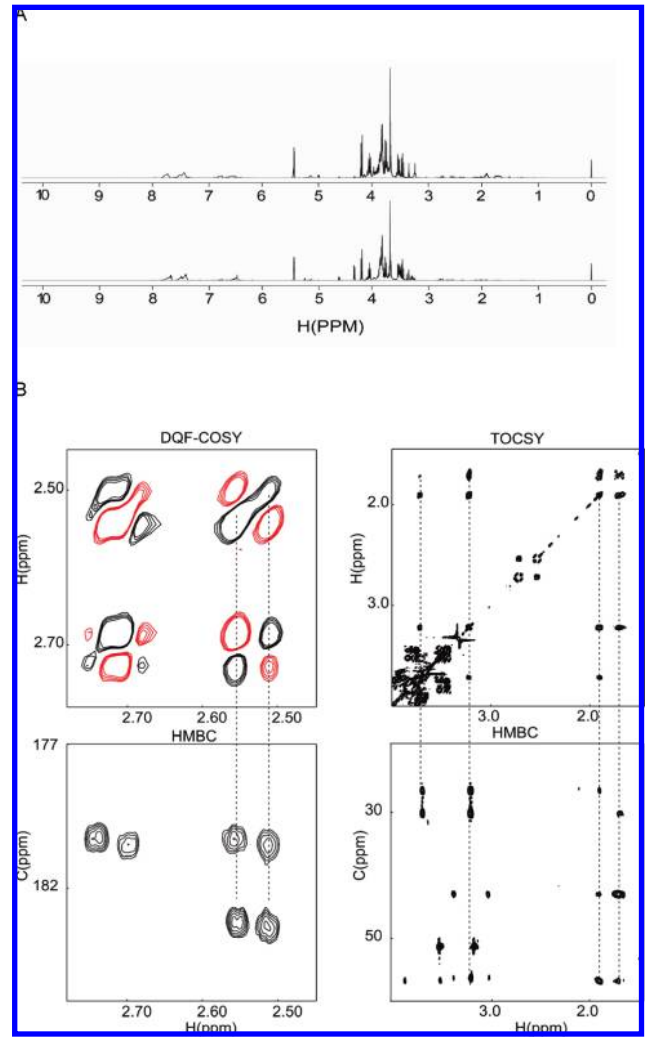

Figure 1. NMR spectra and assignments of key metabolites from *S. baicalensis*. (A) Representative ^{1}H NMR spectra of *S. baicalensis* sample extracts: (top) Korean *S. baicalensis* and (bottom) Chinese *S. baicalensis*. (B) Unambiguous signal assignments for citric acid (left, DQF-DOSY and HMBC spectra) and arginine (right, TOCSY and HMBC spectra) with 2D NMR spectra. For citric acid, the H-4 doublelet (2.52 ppm) has a large coupling constant (17.5 Hz) and, therefore, exhibits positive and negative cross-peaks in the phase-sensitive DQF-COSY. These two peaks correlate with two carboxyl peaks, C-1 and C-2, at 183.2 and 180.3, respectively, giving rise to the four peaks on the HMBC spectrum.

Two-dimensional homo- and heteronuclear correlation spectra [heteronuclear multiple-bond correlation (HMBC), heteronuclear multiple-quantum coherence (HMQC), total correlation spectroscopy (TOCSY), and double-quantum-filtered correlation spectroscopy (DQF-COSY)] were measured on a 400 MHz Varian Unity Inova spectrometer equipped with a triple-resonance inverse detection probe. HMBC was recorded with a one-bond *J*-filter and gradient selection in a nonphase-sensitive manner and then processed in magnitude mode to obtain purely absorptive lineshapes. All other spectra were recorded in a phase-sensitive mode and processed with complex Fourier-transform in the indirect dimension. The phases of the indirect dimension were corrected manually for these spectra. All of the 2D NMR data were processed with nmrPipe and analyzed by nmrview software.

Data Analysis. The spectra were integrated at every 0.04 ppm step using an in-house-developed Perl program. The integral values were

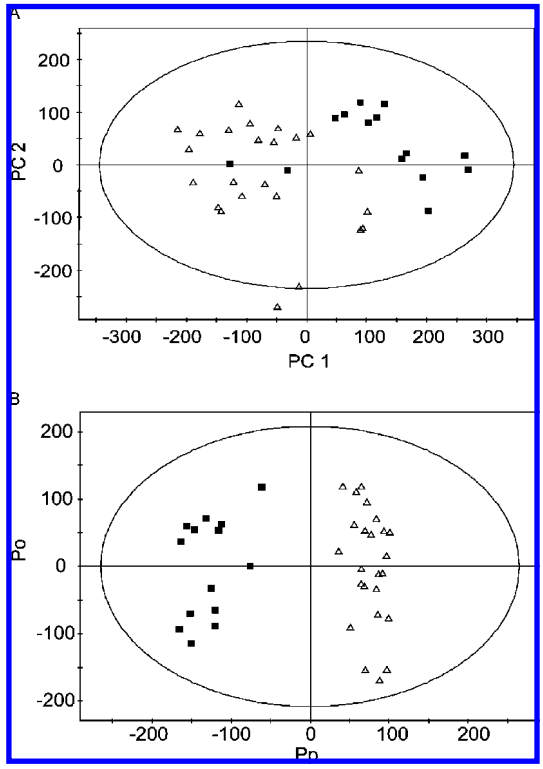

Figure 2. PCA and orthogonal projections to latent structure discriminant analysis (OPLS-DA) of Korean and Chinese *S. baicalensis* groups. (A) Score plot for PC1 and PC2 of PCA. (B) Score plot for OPLS-DA. △, Chinese *S. baicalensis*; ■, Korean *S. baicalensis*.

normalized against the TSP signal as an internal standard. The water region (4.6−5.8 ppm) was excluded from the raw data for the analysis. The data were formatted as an ascii-text file to be imported in statistical software. For both PCA and OPLS-DA multivariate analysis, data fitting was iterated right before the cross-validation coefficient starts to decrease. SPSS (general statistical analysis, SPSS, Chicago, IL), Matlab (STOCSY and OPLS-DA analysis, MathWorks, Natick, MA), SIMCA-P, version 11.0 (OPLS-DA analysis, Umetrics, Sweden), Chenomx (spectral database, Edmonton, Alberta, Canada), and Excel (data conversion, Microsoft, Seattle, WA) programs were used for data analysis. Statistical total correlation spectroscopy (STOCSY) was implemented in Matlab, as described previously (*21*).

RESULTS AND DISCUSSION

NMR Spectral Acquisition and Metabolite Identification. For the purpose of developing an efficient discrimination tool, we began with a model system using Korean and Chinese *S. baicalensis*. The plant has been used widely in both countries for a variety of pharmacological activities and, thus, is of significant socioeconomic value. Our initial trial to employ a well-established marker metabolite, baicalin, for identifying origins failed, because its content was not significantly different between the two groups of samples (data not shown). Therefore, we used a metabolomics approach to analyze a large number of metabolites in a systematic way. To compare the metabolite profiles between the Korean and Chinese samples, we obtained one-dimensional NMR spectra (**Figure 1A**). The overall features of the two representative spectra were quite similar, suggesting similar major-compound contents, consistent with our analysis of baicalin. As simple visual inspection of the spectra did not allow for discrimination of the origins, we attempted to identify the metabolites in the samples by analyzing the NMR spectra. A variety of metabolites were identified on the basis of their chemical-shift values (**Table 1**). Because the chemical shifts in these samples could differ from the literature values, as a result of either different solvent conditions or the presence of matrices, we also confirmed the identification using two-dimensional NMR spectra, including DQF-COSY, HMBC, and HMQC. Representative unambiguous signal assignments for important metabolites are shown in **Figure 1B**. We next investigated whether these metabolites could discriminate the origins of *S. baicalensis* samples.

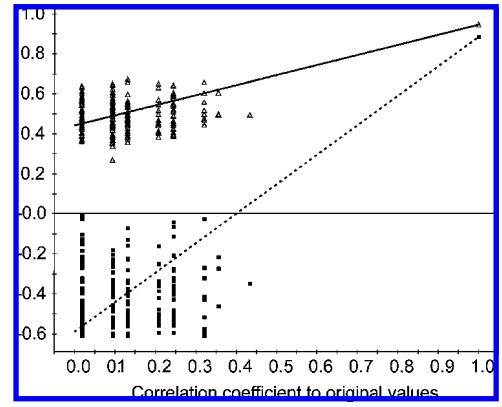

Figure 3. Statistical validation of the OPLS-DA analysis result by "*y*-scrambling". A total of 200 permutations were performed, and the resulting R^2 and Q^2 values were plotted: △, R^2; ■, Q^2. The solid line represents the regression line for R^2, and the dashed line represents the regression line for Q^2.

Multivariate Statistical Analysis for Discrimination of Origins. To make the most of the information contained in the NMR spectra, we performed a multivariate metabolomic statistical analysis on the entire spectra. First, we performed PCA, a widely used metabolomic profiling technique for plant metabolites. PCA analysis showed some class differentiation, but there were noticeable overlaps, possibly because of the structured variation within each group (**Figure 2A**). To address this problem, we used an OPLS-DA statistical approach. The method, a supervised approach, would also enable us to verify the statistical model by predicting the origins of test samples. This step is integral to stringent validation of the statistical model but has not been performed in most metabolomics studies employing PCA analysis. The OPLS-DA model for distinguishing sample origins was established using one predictive and four orthogonal components (**Figure 2B**). The score plot for t_p and t_o shows a cleaner separation between the groups by the first predictive component than the PCA-based approach. The model had an overall goodness of fit, $R^2(y)$, of 95% and an overall cross-validation coefficient, $Q^2(y)$, of 91%. Of the overall $R^2(x)$ value of 87%, 60% was structured but uncorrelated to the response and 27% was predictive, that is, responsible for the class separation. These results show that the OPLS-DA model can reliably differentiate classes even in the presence of large structured noise and that OPLS-DA is more appropriate than PCA in discriminating the origins of *S. baicalensis* samples. In most metabolomics data analyses, compounding factors orthogonal to the variables of interest may obscure the intended class separation. Therefore, OPLS-DA should give better separation along the particular dimension that one is interested in. Still, the presence of large unstructured noise would hamper the separation for OPLS-DA.

Statistical Validation of the Model. In comparison to PCA, OPLS-DA is a supervised method and the goodness of fit

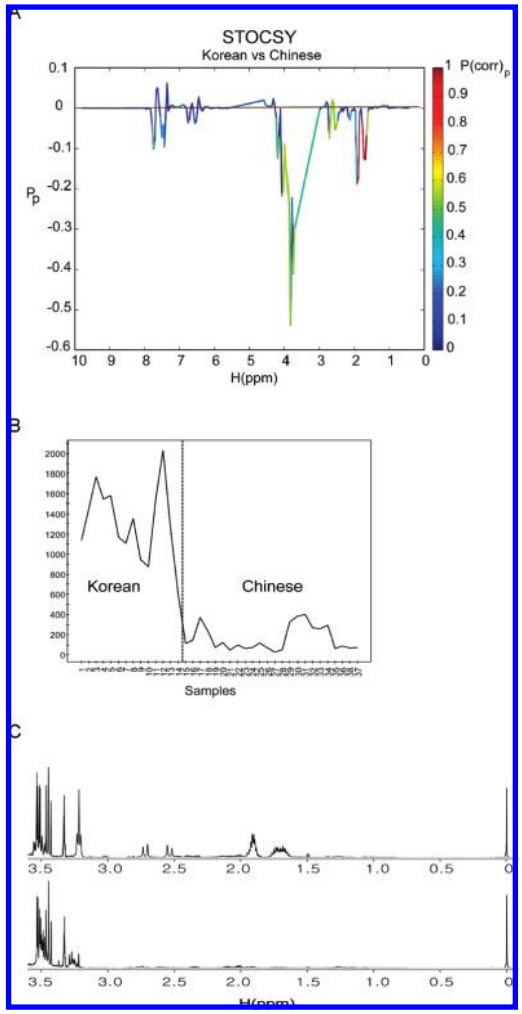

Figure 4. Identification of marker signals responsible for differences between Korean and Chinese *S. baicalensis* samples. (A) STOCSY spectrum from the OPLS-DA model for Korean and Chinese *S. baicalensis* samples. (B) Relative peak intensities, hence, contents for the 1.90 ppm signal in all of the samples. (C) Enlarged view of ^{1}H NMR spectra for *S. baicalensis* sample extracts showing the regions for the signals of citric acid and arginine: (top) Korean sample and (bottom) Chinese sample.

and the predictability of its result can be subjected to validation to test the possibility of correlation by chance. This statistical validation step is especially important for metabolomics data, because most of them have a larger number of observations than variables. For the OPLS-DA method, we can apply "*y*-scrambling" validation, where the *y* variable values are randomly shuffled and the models are rebuilt and analyzed. We performed this permutation procedure using the PLS-DA model with the same number of components. Nevertheless, the results are valid, because the solutions of PLS-DA and OPLS-DA are the same, with their main differences being the improved model interpretation of the latter. The procedure amounts to redistributing the class memberships of each sample randomly and observing the decrease in the predictive power and goodness of fit. We performed 200 rounds of random permutations of the *y* variable, which resulted in a substantial decrease in both parameters (**Figure 3**). Moreover, the extrapolated value of the Q^2 regression line was −0.589. Generally, an intercept value larger than 0.05 indicates overfitting in the original

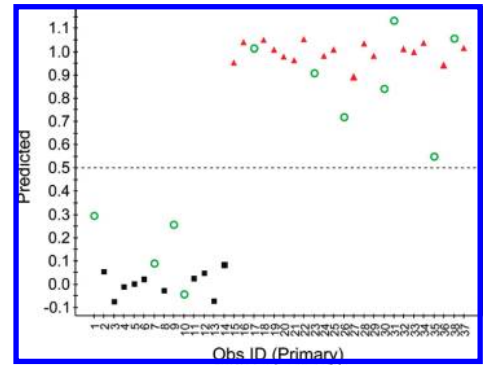

Figure 5. Prediction of origins of the Korean and Chinese *S. baicalensis* samples using the OPLS-DA model. Filled red triangle, Chinese samples (training set); filled black box, Korean samples (training set); blank green circle, results for the blind test samples. The dashed line represents the cutoff for the class membership prediction (*7*).

model. Therefore, these analyses show that our model is statistically valid and that the high value of predictability does not arise from overfitting.

Marker Compound Identification and Variable Importance Plot. Having achieved efficient separation and statistical validation, we explored the OPLS-DA model to identify the marker signals underlying the separation of origins. For this purpose, we constructed a STOCSY plot. While other correlation spectroscopic methods correlate peaks by physical coherence transfer during the mixing time, STOCSY does that in a statistical way by observing correlations in the peak intensity changes across the different samples used in metabolomics studies. In its one-dimensional format, it is represented by a line plot combining the modeled covariation (p_p) with the modeled correlation ($p(\mathrm{corr})_p$) in one graph (*21*). One of the advantages of this combination is that the modeled covariation will maintain the line shapes from the NMR data in the loading plot. This is an advantage, because it is easy to observe how close the differentiating metabolite is to the noise level. The correlation, meanwhile, will indicate the size of separation, which can also be compared to other metabolites. These analyses have been applied successfully to other types of data [e.g., gas chromatography/mass spectrometry (GC/MS) in a scatter plot manner] in a recent publication that one of us co-authored (*22*). The STOCSY plot shows that the signals at 1.70, 1.90, 2.54, 2.72, and 3.72 ppm make the largest contributions to the differentiation of the two sample groups (**Figure 4A**). On the basis of the signal assignments above, it was easy to find that the signals belong to citric acid or arginine. To demonstrate the actual biased distribution of the signals from these marker metabolites in one sample group, we built a plot with the intensities of one of those signals in the Chinese and Korean samples. **Figure 4B** shows that the signal at 1.90 ppm from arginine is consistently much higher in the Korean samples than in the Chinese. The differential contents of these compounds are also clear in the enlarged view of the representative spectra of the Korean and Chinese samples (**Figure 4C**), demonstrating the validity of the STOCSY analysis.

Validation of the Working Model by Prediction. Another critical step in a statistical multivariate analysis is to validate a model on samples not used in building the model itself. The process can be performed by leaving some of the data out (a test data set) and constructing new models with only the remaining data set. In this case, the test samples can be

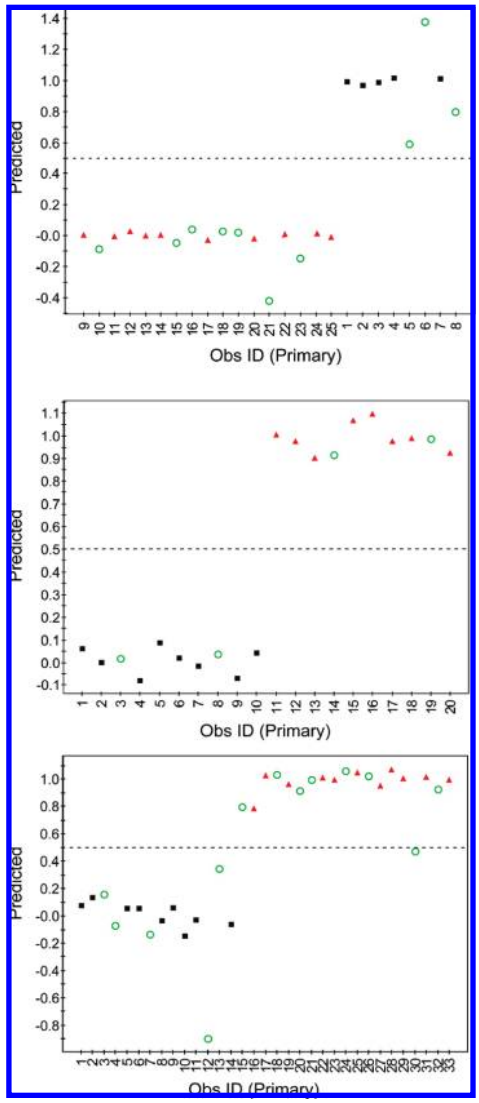

Figure 6. Prediction of origins of three additional herbal medicines: *A. orientale* (upper), *A. japonica* (middle), and *P. lobata* (lower). Filled red triangle, Chinese samples (training set); filled black box, Korean samples (training set); blank green circle, results for the blind test samples. The dashed line represents the cutoff for the class membership prediction (*7*).

considered unknown samples, whose class memberships will be predicted blindly by the model. One can compare the predicted memberships to the original values and thereby evaluate the accuracy/reliability of the metabolomics model. This method was used as a stringent judgment tool in recent clinical metabolomics studies for the diagnosis of heart disease and drug toxicity prediction (*7, 23*). For the prediction test, we randomly left out a total of 11 test data (4 Korean and 7 Chinese samples) and built the OPLS-DA prediction model without them. The approach yielded similar statistical characteristics to those obtained by cross-validation using the entire data set (data not shown) and was able to correctly predict the origins of the 11 test samples (**Figure 5**). The solid prediction results with a large number of test samples, as large as one-third of the total number of samples used in the model, show the reliability and robustness of the prediction model.

Application of the Approach to Other Herbal Medical Products. Having developed the statistical approach using the *S. baicalensis* system, we determined if it could also be applied to differentiate the origins of other herbal medical products from Korea and China. We performed the same sample-handling and data analysis on three other herbal medicines, including *Alisma orientale*, *Atractyloides japonica*, and *Pueraria lobata*. All of them showed good separation in the score plot for t_p and t_o without overlap (data not shown). As a stringent test of the practical applicability, we performed a blind prediction test for all three herbal medical products. We obtained 100, 100, and 92% accuracies in the prediction of origins of the *A. orientale*, *A. japonica*, and *P. lobata* samples, respectively (**Figure 6**). These results show that our method is robust and could thus be applicable to the discrimination of other herbal medical products. Although the analysis with *P. lobata* showed less than 100% accuracy, a larger sample size would allow for a more reliable statistical model.

CONCLUSIONS

In the present study, we applied NMR-based metabolomics combined with OPLS-DA multivariate analysis to develop an efficient tool for discriminating between Korean and Chinese herbal medicines. Although our data do not cover all of the possible diversities of the herbal medical samples, the robust statistical validation results, test sample prediction, and applicability to samples collected in various places at different times suggest that our approach could be effectively used for testing actual market samples. Additionally, the marker metabolites identified here could later be used for rapid differentiation of Korean and Chinese *S. baicalensis* samples without the need for statistical analysis. Because our approach was also successfully applied to three other herbal medicines, we believe that it might be used in the development of a generally applicable tool for determining the origins of herbal medical products.

LITERATURE CITED

(1) Griffin, J. L. *Philos. Trans. R. Soc. London, Ser. B* **2006**, *361*, 147–161.
(2) Griffin, J. L. *Philos. Trans. R. Soc. London, Ser. B* **2004**, *359*, 857–871.
(3) Fiehn, O. *Plant Mol. Biol.* **2002**, *48*, 155–171.
(4) Kikuchi, J.; Shinozaki, K.; Hirayama, T. *Plant Cell Physiol.* **2004**, *45*, 1099–1104.
(5) Hall, R.; Beale, M.; Fiehn, O.; Hardy, N.; Sumner, L.; Bino, R. *Plant Cell* **2002**, *14*, 1437–1440.
(6) Griffin, J. L.; Bollard, M. E. *Curr. Drug Metab.* **2004**, *5*, 389–398.
(7) Brindle, J. T.; Antti, H.; Holmes, E.; Tranter, G.; Nicholson, J. K.; Bethell, H. W.; Clarke, S.; Schofield, P. M.; McKilligin, E.; Mosedale, D. E.; Grainger, D. J. *Nat. Med.* **2002**, *8*, 1439–1444.
(8) Weeks, M. E.; Sinclair, J.; Butt, A.; Chung, Y. L.; Worthington, J. L.; Wilkinson, C. R.; Griffiths, J.; Jones, N.; Waterfield, M. D.; Timms, J. F. *Proteomics* **2006**, *6*, 2772–2796.
(9) Weckwerth, W. *Annu. Rev. Plant Biol.* **2003**, *54*, 669–689.
(10) Dunn, W. B.; Bailey, N. J.; Johnson, H. E. *Analyst* **2005**, *130*, 606–625.
(11) Mesnard, F.; Ratcliffe, R. G. *Photosynth. Res.* **2005**, *83*, 163–180.
(12) Sumner, L. W.; Mendes, P.; Dixon, R. A. *Phytochemistry* **2003**, *62*, 817–836.
(13) Sengupta, S.; Toh, S. A.; Sellers, L. A.; Skepper, J. N.; Koolwijk, P.; Leung, H. W.; Yeung, H. W.; Wong, R. N.; Sasisekharan, R.; Fan, T. P. *Circulation* **2004**, *110*, 1219–1225.

(14) Sievenpiper, J. L.; Arnason, J. T.; Leiter, L. A.; Vuksan, V. *J. Am. Coll. Nutr.* **2004**, *23*, 248–258.
(15) Choi, Y. H.; Kim, H. K.; Linthorst, H. J.; Hollander, J. G.; Lefeber, A. W.; Erkelens, C.; Nuzillard, J. M.; Verpoorte, R. *J. Nat. Prod.* **2006**, *69*, 742–748.
(16) Frederich, M.; Cristino, A.; Choi, Y. H.; Verpoorte, R.; Tits, M.; Angenot, L.; Prost, E.; Nuzillard, J. M.; Zeches-Hanrot, M. *Planta Med.* **2004**, *70*, 72–76.
(17) Wang, Y.; Tang, H.; Nicholson, J. K.; Hylands, P. J.; Sampson, J.; Whitcombe, I.; Stewart, C. G.; Caiger, S.; Oru, I.; Holmes, E. *Planta Med.* **2004**, *70*, 250–255.
(18) Cho, I. H.; Kim, Y. S.; Choi, H. K. *J. Pharm. Biomed. Anal.* **2007**, *43*, 900–904.
(19) Bylesjo, M.; Rantalainen, M.; Cloarec, O.; Nicholson, J.; Holmes, E.; Trygg, J. *J. Chemom.* **2006**, *20*, 341–351.
(20) Trygg, J.; Wold, S. *J. Chemom.* **2002**, *16*, 119–128.
(21) Cloarec, O.; Dumas, M. E.; Craig, A.; Barton, R. H.; Trygg, J.; Hudson, J.; Blancher, C.; Gauguier, D.; Lindon, J. C.; Holmes, E.; Nicholson, J. *Anal. Chem.* **2005**, *77*, 1282–1289.
(22) Wiklund, S.; Johansson, E.; Sjostrom, L.; Mellerowicz, E. J.; Edlund, U.; Shockcor, J. P.; Gottfries, J.; Moritz, T.; Trygg, J. *Anal. Chem.* **2008**, *80*, 115–122.
(23) Clayton, T. A.; Lindon, J. C.; Cloarec, O.; Antti, H.; Charuel, C.; Hanton, G.; Provost, J. P.; Le Net, J. L.; Baker, D.; Walley, R. J.; Everett, J. R.; Nicholson, J. K. *Nature* **2006**, *440*, 1073–1077.

Received for review July 9, 2008. Revised manuscript received September 26, 2008. Accepted September 29, 2008. This work was supported by the Korea Food and Drug Administration (Grant 07092-298), the INHA University Research Grant (INHA-32729-01), and the Korea Research Foundation Grants funded by the Korean Government (MOEHRD, Basic Research Promotion Fund) (KRF-2008-331-E00468, and KRF-2007-313-C00439 and KRF-2007-331-E00313). This study made use of the NMR facility at the Korea Basic Science Institute, which is supported by the Bio-MR Research Program of the Korean Ministry of Science and Technology (E28070).

JF802088A