Food Bioprocess Technol (2015) 8:359–370
DOI 10.1007/s11947-014-1407-6

ORIGINAL PAPER

Received: 6 February 2014 / Accepted: 8 September 2014 / Published online: 18 September 2014

Characterization of the botanical origin and quality of honeys is of great importance and interest in agriculture. In this study, an electronic nose (e-nose) was applied for identifying the botanical origin of honey as well as determining their main quality components such as glucose, fructose, hydroxymethylfurfural (HMF), amylase activity (AA), and acidity. Principal component analysis (PCA) and discriminant factor analysis (DFA) were employed to generate scatter plots of honey samples from 14 botanical origins. Origin discrimination models with 100 % overall accuracy were established by least squares support vector machines (LS-SVM). LS-SVM outperformed the linear regression method of partial least squares regression (PLSR) for quality prediction, showing that the non-linear correlations between e-nose responses were important for the analysis of honey. Moreover, three sensor selection algorithms, namely, uninformation variable elimination (UVE), successive projections algorithm (SPA), and competitive adaptive reweighted sampling (CARS) were applied for the first time to analyze e-nose fingerprints of honey. After the calculation of the above three algorithms and the comparison of their results, from a total of 18 sensors, the important ones were selected for glucose (three), fructose (five), HMF (three), AA (five), and acidity (four) prediction, respectively. The results of sensor selection show the advantages of reducing redundancy of e-nose data, optimizing the sensor array of an e-nose, and improving the performance of models in terms of robustness. The overall results show that the laborious, time-consuming, and destructive analytical methods like high-performance liquid chromatography (HPLC), acid-base titration, and spectrophotometry could be replaced by e-nose to provide a rapid and non-invasive determination of the botanical origin and quality of honey.

Honey · Botanical origin · Quality · Electronic nose · Multivariate analysis · Sensor selection

Honey is the natural sweet substance produced by *A.* bees. The bees collect the nectar of plants either from secretions of living plants or excretions of plant-sucking insects. After collection, the collected raw material is transformed by combining with specific substances from the bees, deposited, dehydrated, and stored in honeycombs to ripen and mature (European Commission 2002). Humans have been consuming honey for thousands of years as a product in its own right, and it is also used as an ingredient in baking and confectionary products (Hennessy et al. 2010).

The identification of the botanical origin is a main part of the quality analysis of honey (Etzold and Lichtenberg-Kraag 2008). Honey composition, flavor, aroma, color, and texture depend predominantly on the botanical source that it originates from (Ampuero et al. 2004; Oddo et al. 2004). The price and popularity differs greatly among honey of different floral origins (Chen et al. 2012). Besides the quality specification, there is also, in the minds of consumers, a perceived link between the quality of honey and its provenance (Hennessy et al. 2010). Honey so labeled with botanical origin can command a price premium; consequently, there is a potential

Lingxia Huang and Hongru Liu contributed equally to this work.

L. Huang
College of Animal Sciences, Zhejiang University, Hangzhou 310058, People's Republic of China

H. Liu · B. Zhang (✉) · D. Wu (✉)
Laboratory of Fruit Quality Biology, The State Agriculture Ministry Laboratory of Horticultural Plant Growth, Development and Quality Improvement, Zhejiang University, Hangzhou 310058, People's Republic of China
e-mail: bozhang@zju.edu.cn
e-mail: china.di.wu@gmail.com

for economic fraud (Hennessy et al. 2008). For these reasons, the identification of botanical origin for honey products is important not only because of specific legislation but also because of market demands including those of food processors, retailers, enforcement agencies, and consumers (Ulloa et al. 2013). Besides the identification of botanical origin, several quality features of honey have to be determined, which include water content, enzyme activities of invertase and α-amylase, hydroxymethylfurfural (HMF), electrical conductivity, and sugar composition (mainly glucose, fructose, maltose, and sucrose). These quality features vary significantly among different honey products (Oddo et al. 2004). As with the identification of botanical origin, the quality inspection of honey is also of interest to regulatory authorities, food processors, retailers, and consumer groups (Wang et al. 2010).

Melissopalynological analysis is the reference method for the identification of the botanical origin of honey. It is mainly based on the identification and quantification of pollen grains in the honey sediment. However, as this involves a laborious counting procedure requiring specialized knowledge and expertise in the interpretation of results, this method is rather difficult and very time-consuming (Ulloa et al. 2013). Analytical and quantitative methods such as high-performance liquid chromatography (HPLC) and high-performance anion-exchange chromatography are also routinely performed in quality determination of honey (Wang et al. 2010; Cozzolino et al. 2011). These methods are laborious and time-consuming, require considerable analytical skills, involve a lot of tedious and complex pretreatment of samples, and use many hazardous organic reagents that require high costs for storage and disposal. Moreover, each quality feature of interest needs a specific analytical method, and only a limited number of samples can be analyzed. Therefore, there is a trend to develop rapid, simple, efficient, non-invasive, and accurate analysis methods for the quality inspection of honey.

Aroma is an important parameter among the sensory properties of foods (Falasconi et al. 2012). In the present study, the acquisition and analysis of volatile compounds of honey were conducted using an electronic nose (e-nose) for the identification of botanical origin and determination of quality components of honey. E-nose is an instrument with an array of sensors to mimic the sense of smell, typically used to detect and distinguish odors precisely in complex samples and at low cost (Peris and Escuder-Gilabert 2009). As an objective, automated, and non-destructive technique to characterize food flavors, the e-nose has the advantage of high sensitivity and correlation with data from human sensory panels, ease of operation, and cost-effectiveness, requiring only a short time for analysis (Peris and Escuder-Gilabert 2009). However, the e-nose has not been considered for the quality measurement of honey using quantitative models. Much previous research established only qualitative discrimination models for honey classification using an e-nose (Ampuero et al. 2004; Benedetti et al. 2004; Kenjerić et al. 2009; Hong et al. 2011). There are several reports considering establishment of quantitative regression models for the prediction of quality features of honey. However, no reports have selected important sensors of the e-nose that are important to predict specific quality features of honey. The selection of important sensors is the key step for optimizing the sensor array of an e-nose, so that simple, fast, and low-cost e-nose systems with only the selected sensors can be designed.

Given the limited information on the usefulness of the e-nose for quality determination of honey, the main aim of this study was to investigate the feasibility of an e-nose for identifying the botanical origin and determining the main quality components of honey such as glucose and fructose and also other important components such as hydroxymethylfurfural (HMF), amylase activity (AA), and acidity. The specific objectives of the current study were to (1) acquire e-nose profiles of honey products from 14 botanical origins, (2) build origin discrimination models using pattern recognition and qualitative discrimination, (3) measure the reference values of investigated components of honey using traditional standard methods, (4) use the reference values of samples and their e-nose fingerprints to establish quantitative prediction models, and (5) identify the important sensors that were mostly correlated to the quality determination.

Materials and Methods

Sample Preparation and E-Nose System

Honey samples were purchased from local supermarkets in Hangzhou, China. The details of the geographic and botanical origins of these samples are shown in Table 1. There were two geographic origins and 14 botanical origins for these samples. Each botanical origin had six samples, resulting in 84 samples (six samples per origin × 14 origins). From these, 56 samples (four samples from each botanical origin) were selected for the model calibration, and the remaining 28 samples (two samples from each botanical origin) were used for validation. Two grams of sample was added to a 10-ml crimp-top vial with a diameter of 20 mm, sealed with an aluminum gasket containing a PTFE/silica gel septum, and then stored in a 0 °C icebox for further e-nose measurement.

The e-nose system used for this study was a Fox 4000 (ALPHA MOS, Toulouse, France) with three metal oxide sensor chambers equipped with 18 sensors. There are two types of sensors currently used: P & T sensors implemented in chambers A and B and LY2 sensors used in chamber CL. Their specific names are LY2/LG, LY2/G, LY2/AA, LY2/GH, LY2/gCTl, LY2/gCT, T30/1, P10/1, P10/2, P40/1, T70/2, PA/2, P30/1, P40/2, P30/2, T40/2, T40/1, and TA/2. P & T sensors are metal oxide sensors based on tin dioxide (SnO_2) (n-type

semiconductor). The difference resides in the geometry of the sensors. Type T has the sensitive layer placed on a tube of aluminum, while the sensitive layer of type P is placed on a plain substrate. The LY2 sensors are metal oxide sensors based on chromium titanium oxide ($Cr_2-xTixO_3+y$) and on tungsten oxide (WO_3). In the process of e-nose signal measurement, vials were heated at 40 °C for 18 min in a dry bath heater. Two-milliliter headspace gas in the vial was extracted by a syringe and injected into the Fox system. The headspace gas was pumped into the sensor chamber with a constant rate of 150 ml min^{-1}. The measurement phase lasted 120 s for each sample, and the clean phase was 240 s. The maximum or minimum response values of sensors in the e-nose were used

2 Statistics of five main quality components of honey samples measured by reference methods

Statistics	Calibration set (%)					Validation set (%)				
	Glucose	Fructose	HMF	AA	Acidity	Glucose	Fructose	HMF	AA	Acidity
Maximum	38.93	45.23	135.49	8.20	20.49	38.63	45.76	139.80	8.28	20.52
Minimum	24.89	33.61	2.88	2.69	9.47	25.83	33.72	2.91	2.65	9.58
Mean	32.35	39.10	25.95	5.38	14.22	32.26	38.80	26.70	5.35	14.18
SD	4.06	2.97	35.56	1.47	3.32	3.92	3.13	37.60	1.47	3.32
Range	14.04	11.62	132.61	5.51	11.02	12.80	12.04	136.89	5.63	10.93

hydroxymethylfurfural, *AA* amylase activity, *D* standard deviation

Chromatographic conditions are follows: The binary solvent system used was methanol/water (77/23, /), the elution of binary solvent was conducted in isocratic fashion. The flow rate was kept at 1.0 ml min^{-1}. The temperature of the column was 30 °C. The injection volume was 10 μl. The content of HMF was calculated by the following formula:

$$= \times - \times \frac{1000}{1000} \quad (3)$$

where is the content of HMF (mg/g), is the concentration of HMF (mg/ml) obtained from the established standard curve, is sample volume (ml), and is sample weight (g).

AA was determined by spectrophotometric method according to GB/T-18932.16-2003 2003. A total of 5 g sample was added to a mixture of 15 ml water and 2.5-ml acetate buffer. NaCl (1.5 ml) aqueous solution was added to the mixture, made to 25 ml with water in a volumetric flask, and used as the sample solution. Ten milliliters of sample solution, 5 ml of starch solution, and 10-ml iodine solution were kept in a water bath at 40 °C for 15 min, respectively. Then, the sample solution was incubated with the starch solution for 5 min, 1 ml of the mixture was added to 10-ml iodine solution, taking water as the control, and the absorbance was measured at 660 nm. The results were expressed as ml(g h)$^{-1}$. The diastase value was calculated with the following formula:

$$= \frac{300}{} \quad (4)$$

where is the diastase value and is the corresponding time.

Multivariate Analysis

One of the advantages in developing an e-nose is that the analytical process does not require separating samples into individual chemicals, but detects and analyzes the volatile fraction of the sample as a whole. The signal produced by the e-nose results in a matrix of semi-independent variables (the sensor array output) and a set of dependent variables (classes or quality features) (Scott et al. 2006). The matrix contains rich information of volatile fraction in the sample. However, it is difficult to directly tell which sensors are important for the analysis. As with the human olfactory system, sensors of the e-nose are not designed to detect a particular volatile, but learn new patterns and associate them with new odors via training and data storage functions as humans do (Ampuero and Bosset 2003). Therefore, the massive quantity of the e-nose matrix needs multivariate analysis to appropriately extract meaningful information in an efficient way to establish qualitative discrimination and quantitative prediction models. Multivariate analysis for the e-nose is similar to the process of pattern learning in humans. In order to evaluate if any single sensor could be used to determine any component of honey, the correlation coefficients were calculated between the response of each sensor and the reference values of five quality components.

D

Two classic pattern recognition techniques, namely, principal component analysis (PCA) and discriminant factor analysis (DFA), were employed to generate scatter plots in two dimensions to understand the cluster of samples. PCA is the most frequently used unsupervised technique that decomposes the data matrix into several principal components (PCs) to characterize the most important directions of variability in the high-dimensional data space (Wu and Sun 2013a). DFA is another classic pattern recognition tool in which the decision boundary between different groups is calculated (Papadopoulou et al. 2012). In DFA calculation, the contribution of data is maximized by the linear combinations, resulting in generating the largest difference between predetermined groups and small variance within the individual group (Ampuero and Bosset 2003; Papadopoulou et al. 2012). The difference between PCA and DFA is that the PCA calculation does not consider the relationship of the data to the group numbers, while DFA calculation includes the group

information (predetermined groups). Therefore, PCA is a non-supervised method with no information on the groups of samples but only the variance of the dataset, and DFA is a supervised method that is based on a priori data classification.

The least squares support vector machines (LS-SVMs) are employed to establish discrimination models for the identification of honey samples from 14 botanical origins and from two geographical origins. As an optimized version of the SVM, LS-SVMs employ non-linear map function and maps the input features to a high-dimensional space, thus changing the optimal problem into an equality constraint condition (Wu and Sun 2013b). Instead of solving a convex quadratic programming (QP) problem as in classical SVM, LS-SVMs find the solution by solving a set of linear equations. The optimal parameters of γ and σ^2 were found using the grid-search algorithm. The number of support vectors in the LS-SVM model is equal to the number of training data (Iplikci 2006; Abe 2007). The details of LS-SVMs are described by Wu et al. (2008b). In addition, for the discriminant analysis of botanical origins, the samples belonging to the same botanical origin were assigned an arbitrary number as their reference botanical origin value. This assignment was carried out according to the first column of Table 1. In order to solve multiclass categorization problems, a set of binary classifiers was used to encode a multiclass task with varieties (Allwein et al. 2001). The minimum output coding was used to obtain the minimal (Wu et al. 2008a; Chen et al. 2013). Specifically, 14 origins ( ) were encoded in the codebook using the minimum output coding, resulting in 14 combinations of binary numbers (−1 and +1) in . Table 1 represents the encoded binary matrix, where the columns represent the results of the binary classifiers (−1 and +1) and the rows indicate the different botanical origins. The LS-SVM discrimination was carried out by establishing binary classifiers in four dimensions separately. The classified results of the four binary classifiers were then decoded by the codebook into the arbitrary numbers of botanical origins, which were evaluated whether they were classified correctly or not.

C

Quantitative prediction was implemented by building calibration models to predict five quality components of honey using their corresponding e-nose data. Partial least squares regression (PLSR) was carried out to perform linear calibration between calibration sample matrix (C) and the values of one of the quality indices ( ). As a bilinear modeling technique, PLSR extracts a set of orthogonal factors called latent variables (LVs) and explores the optimal function by minimizing the error of sum squares (Wu et al. 2012a). The optimal numbers of latent variables were determined at the lowest value of the prediction residual error sum of squares

coefficients of the variable. A threshold of stability is then used to eliminate uninformative variables. The variables with absolute stability values less than the threshold are considered as uninformative variables and should be removed. The determination of the threshold is based on an artificial random variable matrix as a reference. The details of UVE calculation can be found in the literature (Wu et al. 2009). SPA is a variable selection algorithm designed to select variables with minimal redundancy (Araujo et al. 2001). In SPA calculation, a sequence of projection operations is carried out in the columns of the variable matrix (rows represent samples), resulting in candidate subsets of variables. These are then evaluated according to the prediction performances of their calibrated models established based on multiple linear regression (MLR). Details of SPA description are described by Wu et al. (2012b). CARS is a novel variable selection algorithm proposed by Li et al. (2009). CARS uses the absolute values of regression coefficients of a PLSR model as an index for evaluating the importance of each wavelength. Variables with large absolute coefficients have more probability to be selected. In this study, the processes of UVE, SPA, and CARS were performed with the aid of Matlab 2011a software (The Mathworks, Inc., Natick, MA, USA).

Model Evaluation

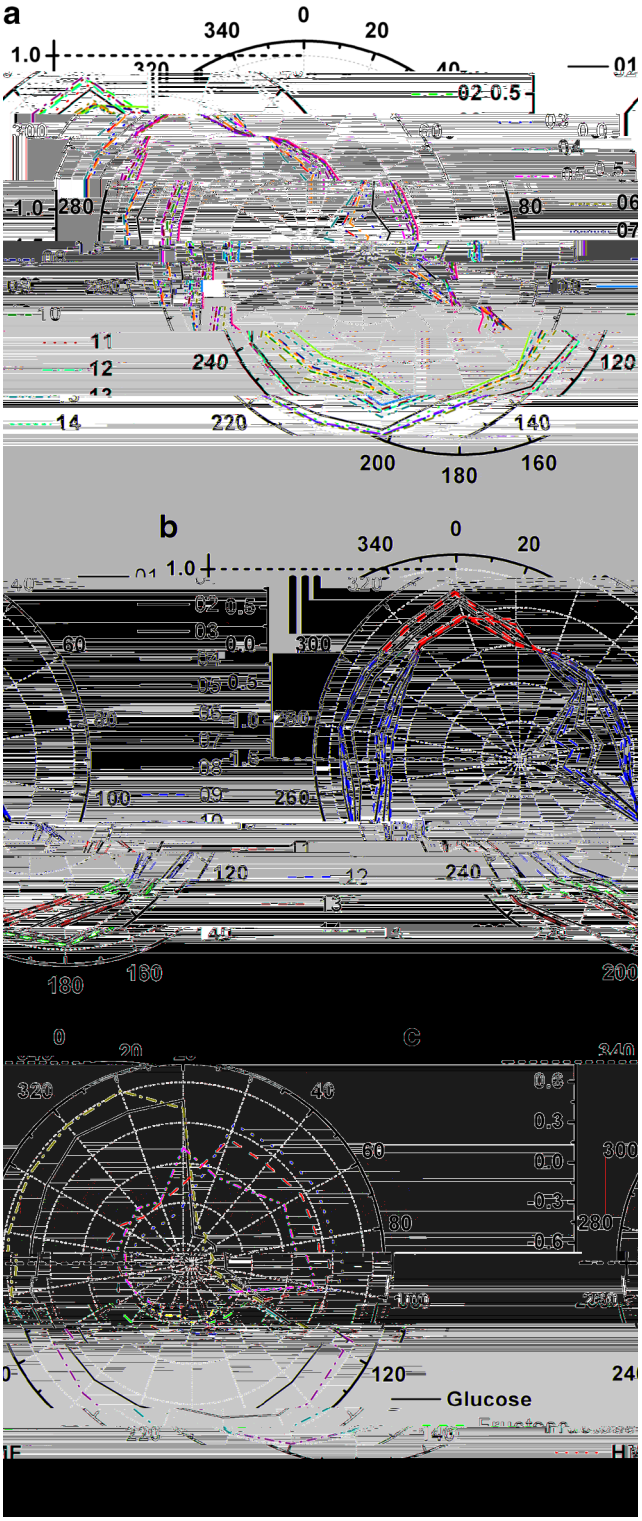
For the discrimination of geographical/botanical origins, the performance was evaluated by the overall accuracy and specific accuracy in both calibration and validation processes. The equations for overall accuracy and specific accuracy are shown as follows:

$$\mathrm{OA} = \frac{\mathrm{CC}}{\mathrm{TS}} \tag{5}$$

$$\mathrm{SA} = \frac{\mathrm{CC}}{\mathrm{TA}} \tag{6}$$

where OA is the overall accuracy, SA is the specific accuracy, CC is the number of correctly classified samples of all origins, TS is the total number of samples of all origins, CC is the number of correctly classified samples of botanical origin (=1 to 14), and TS is the total number of samples belonging to botanical origin (=1 to 14).

For the quantitative prediction of quality components, the predictive ability of the models was evaluated according to some statistic parameters, such as correlation coefficient of calibration ($_C$), coefficient of determination of calibration (^2_C

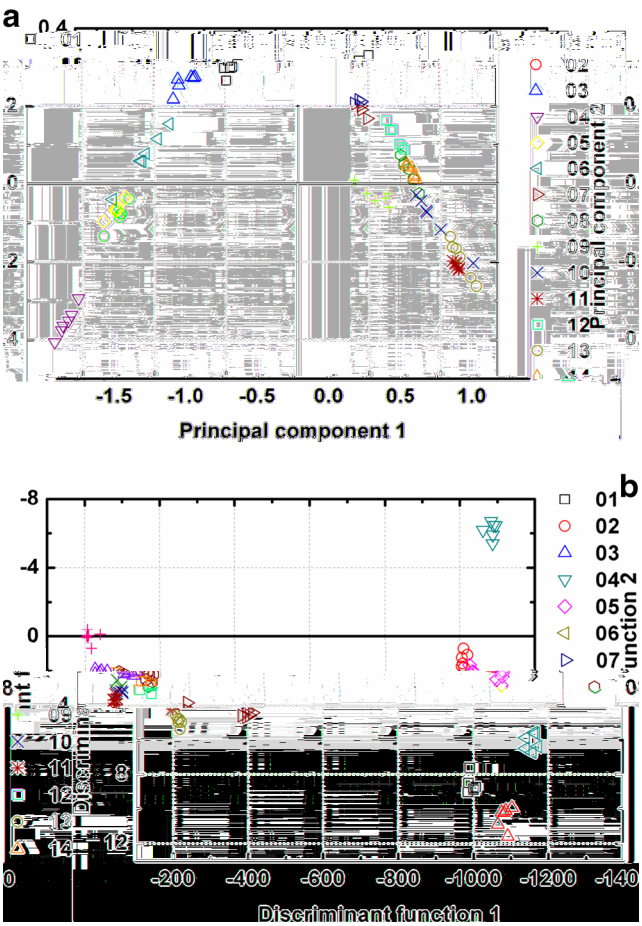


. . 1 Polar plots of the fingerprints (the maximum or minimum response values) of typical honey samples from 14 botanical origins ( ) or two geographical origins (.), and the correlation coefficients between the response of 18 sensors and the reference values of five quality components of honey ( ). No. of origins from *1* to *14* represent different botanical origins, whose specific corresponding relationships are shown in Table 1

quality components of honey are shown in Fig. 1c. The highest absolute values of correlation coefficients were only 0.529, 0.303, 0.427, 0.637, and 0.357 for glucose, fructose, HMF, AA, and acidity, respectively, showing that no single sensor could be used to predict any of the five quality components accurately. Therefore, the combination of several sensors was considered for the quality prediction, which was achieved through multivariate analysis.

Identification of Geographical/Botanical Origin

PCA and DFA were used to check the capability of the e-nose in assigning honey samples to a specific botanical origin. The scores of the first two PCs or discriminant functions (DFs) were displayed in two two-dimensional views (Fig. 2), where similar samples were located close to each other and the differences between origins could be observed. The total explained variance rates (TEV, %) were 99.01 and 96.99 % for the first two PCs and the first two DFs, respectively, which shows that most of the information from e-nose data was included in the first two PCs/DFs. In Fig. 2a, sample points were generally clustered into two groups based on their first



. . 2 Scatter plots of samples from 14 botanical origins based on PCA ( ) and DFA (.). Names of origins from *1* to *14* see Fig. 1

two PCs that were relevant to their e-nose response. Values with positive scores on PC1 were found for all samples from China, while all samples from Australia and the samples from lychee (7) and longan (8) had values with negative scores on PC1. In general, samples from jujube (1), black locust (3), Chinese milkvetch (4), lychee (7), and red stringybark (9) were well separated from each other. Samples from other botanical origins overlapped with each other. In Fig. 2b, samples from jujube (1), black locust (3), Chinese milkvetch (4), miqueliana linden (6), lychee (7), and red stringybark (9) were well separated from each other; samples of polyfloral honey (2) and mandarin orange (5) overlapped; and samples from other botanical origins were clustered together. As with the PCA plot, all samples from Australia and the samples from lychee (7) and longan (8) were located at the left side of Fig. 2b, while the other samples from China were distributed at the right side of Fig. 2b.

The PCA and DFA results showed that the e-nose could discriminate honey from two geographical origins with reasonable accuracy. DFA discriminated better than PCA for botanical origins. However, samples from some botanical origins overlapped each other in PCA/DFA scatter plots. The successful discrimination of the samples from some botanical origins (origins 1, 3, 4, 7, and 9 in the PCA plot and 1, 3, 4, 6, 7, and 9 in the DFA plot) was because their PCs/DFs were different from each other and also from those of the samples from other botanical origins (origins 2, 5, 6, 8, 10, 11, 12, 13, and 14 in the PCA plot and origins 2, 5, 8, 10, 11, 12, 13, and 14 in the DFA plot). This is probably because the differences in e-nose signal values of the successful distinguished samples could be well extracted by the calculation of PCA/DFA and sufficient to be detected in PCA/DFA scatter plots. Therefore, the successfully distinguished samples were well separated from the other samples, but it was difficult to distinguish the other samples by PCA/DFA. This was probably because both PCA and DFA are linear approaches. Non-linear correlation between e-nose responses could not be retained after the PCA/DFA calculation, which might explain the difficulties in discrimination. Therefore, in order to improve the discrimination between samples of different botanical origins, LS-SVM, which is a non-linear modeling method, was investigated. When the reference arbitrary numbers of samples were assigned according to their geographical origins, binary numbers of −1 and +1 were used to represent China and Australia. A LS-SVM discrimination model was established based on the e-nose signals of samples and their reference arbitrary numbers, and 100 % OA for geographical discrimination was obtained based on the established LS-SVM discrimination model in both calibration and validation processes. The samples from lychee (7) and longan (8) were correctly classified into the geographical origin of China, although they were more close to the samples from Australia in both PCA and DFA plots (Fig. 2).

When samples were assigned to the reference arbitrary numbers according their botanical origins as shown in Table 1, the established LS-SVM discrimination model also had 100 % OA for all botanical origins and 100 % SA for each botanical origin in both calibration and validation processes. The samples of polyfloral (2) and mandarin orange (5) honey were correctly distinguished from each other, and all samples produced in Australia from different botanical origins were also identified correctly. This could not be achieved based on either PCA plot or DFA plot (Fig. 2). These results show that the non-linear correlations between e-nose responses were important for the discrimination of botanical and geographical origins of honey. Therefore, the discrimination based on LS-SVM algorithm, which is a non-linear modeling method, was better than PCA and DFA methods, which are both linear pattern recognition tools.

Furthermore, of the 18 sensors being used for the LS-SVM discrimination, those critical for origin discrimination were selected using the strategies UVE, SPA, and CARS. The CARS-LS-SVM model obtained 96.4 % OA in both calibration and validation processes. Although the other two methods had similar results, only four sensors were selected by CARS, while there were six and eight sensors selected by SPA and UVE, respectively. Therefore, the best discrimination model for botanical origins was the CARS-LS-SVM model. The four sensors selected by CARS were LY2/AA, LY2/gCTl, P40/2, and T40/2.

Quality Determination

$A_{\_}$ A

The calibration of multivariate models was performed by PLSR and LS-SVM algorithms based on the matrix C with the fingerprints of honey from all e-nose sensors. The matrix was then analyzed as a new test set based on the established calibration models. Table 3 shows the predicted results of five quality components (glucose, fructose, HMF, AA, and acidity) of honey samples by analyzing the fingerprints from all e-nose sensors using the calibration algorithms PLSR and LS-SVM. It is obvious that LS-SVM models outperformed the corresponding PLSR models. Compared with the PLSR models, the RMSEV of LS-SVM models decreased by 29.83 to 62.07 % with an average of 47.08 %, while the RPD increased by 44.74 to 163.77 % with an average of 101.03 %. These results show that the non-linear correlations between e-nose responses were important for the quality determination of honey as found for the identification of botanical origin. Therefore, the established LS-SVM models, which retain the non-linear information, make better predictions than the corresponding PLSR models for the determination of glucose, fructose, HMF, AA, and acidity of honey. Therefore, LS-SVM was an effective method for both identification of botanical

Table 3 Prediction results of five quality components of honey samples considering all 18 sensors

Quality	Calibration model	Number[a]	Calibration			Validation			
			C	$^{2}_{C}$	RMSEC		^{2} ^{2}	RMSEV	RPD
Glucose	PLSR	4	0.639	0.409	3.090	0.664	0.440	2.879	1.336
	LS-SVM	56	0.988	0.975	0.631	0.959	0.919	1.092	3.524
Fructose	PLSR	4	0.674	0.455	2.174	0.624	0.374	2.434	1.274
	LS-SVM	56	0.974	0.937	0.738	0.843	0.692	1.708	1.844
HMF	PLSR	7	0.711	0.505	24.783	0.732	0.526	25.426	1.459
	LS-SVM	56	1.000	1.000	0.119	0.926	0.851	14.247	2.600
AA	PLSR	7	0.841	0.708	0.787	0.814	0.662	0.841	1.720
	LS-SVM	56	0.993	0.986	0.175	0.974	0.948	0.331	4.377
Acidity	PLSR	5	0.864	0.747	1.653	0.838	0.703	1.779	1.834
	LS-SVM	56	0.981	0.962	0.643	0.943	0.889	1.087	3.007

hydroxymethylfurfural, *AA* amylase activity

[a] Number of latent variables or support vectors

origin and quality determination of honey. With the exception of fructose, the values of the LS-SVM models for the other four components were higher than 0.9, showing that good determination of these components was obtained. The determination of fructose using the e-nose was also with reasonable accuracy with of 0.843. This confirmed that the e-nose with all 18 sensors could be used for determining these quality components of honey in a rapid and non-invasive way.

Establishing a simplified e-nose model involves the identification of a reduced number of appropriate sensors. The foregoing analysis of the fingerprints of honey from all e-nose sensors did not take into account the possibility that some sensors might contain useless information with regard to the quality prediction of honey samples. Therefore, the important sensors reflecting the characteristics of the e-nose for predicting quality components of honey were selected using the strategies UVE, SPA, and CARS. Table 4 shows the statistical results of LS-SVM models developed using the fingerprints from only the selected e-nose sensors for the determination of glucose, fructose, HMF, AA, and acidity of honey samples in the calibration and validation processes.

For glucose determination, similar results were obtained for the LS-SVM models based on the sensors selected by UVE, SPA, and CARS, respectively. However, it was noticeable that the sensors selected by CARS were fewer than that by the other two methods. This indicated that the CARS-LS-SVM model was more robust than the other two models for glucose determination. The AV_RMSE of the CARS-LS-SVM model was only 0.044, which was only about 10 % of those of the UVE-LS-SVM model (AV_RMSE=0.433) and SPA-LS-SVM model (AV_RMSE=0.508). Considering that the CARS-LS-SVM model had less AV_RMSE and fewer sensors, the important sensors for glucose determination were those selected by CARS. The sensors selected by CARS for glucose analysis were LY2/gCTl, LY2/gCT, and P30/2.

For fructose determination, the AV_RMSE values from the three models were similar. SPA selected only three sensors, which was the fewest. However, the result from the SPA-LS-SVM model was worse than the other two models. Considering that the CARS-LS-SVM model had fewer sensors than the UVE-LS-SVM model, the important sensors for fructose determination were determined as those selected by CARS, which were LY2/LG, LY2/G, P30/2, T40/2, and T40/1.

For the HMF determination, the CARS-LS-SVM model had a better prediction, fewer sensors, and less AV_RMSE than those of other two models. Therefore, the sensors (LY2/AA, P10/1, and T40/2) selected by CARS were determined as the important sensors for HMF determination.

For the AA determination, the SPA-LS-SVM model and CARS-LS-SVM model gave similar results, which were better than that from the UVE-LS-SVM model. Considering that the SPA-LS-SVM model (three sensors) had fewer sensors than the CARS-LS-SVM model (five sensors), the important sensors for AA determination were determined as those selected by SPA, which were LY2/AA, LY2/gCT, and T40/2.

For the acidity determination, the UVE-LS-SVM model and SPA-LS-SVM model gave similar results based on ten and six sensors, respectively. When CARS was used for the sensor selection, only four sensors were selected. Furthermore, the CARS-LS-SVM model outperformed the other two models. Therefore, the sensors (LY2/G, P40/1, T70/2, and P30/2) selected by CARS were determined as the important sensors for acidity determination.

In conclusion, the CARS-LS-SVM models were the best-selection-LS-SVM models for the determination of glucose, fructose, HMF, and acidity, while the SPA-LS-SVM model was the best-selection-LS-SVM model for AA determination. The optimal sets of sensors were determined according to the prediction accuracy, the number of selected sensors, and the

Prediction results of LS-SVM models for determining five quality components of honey samples considering only selected sensors

Quality	Sensor selection	No. of sensors	Calibration			Validation			
			$_C$	^2_C	RMSEC		^{2}	RMSEV	RPD
Glucose	UVE	6	0.988	0.975	0.640	0.961	0.922	1.073	3.606
	SPA	10	0.987	0.973	0.660	0.953	0.908	1.168	3.311
	CARS	3	0.966	0.928	1.078	0.959	0.915	1.122	3.447
Fructose	UVE	7	0.935	0.854	1.127	0.842	0.686	1.724	1.811
	SPA	3	0.893	0.790	1.350	0.805	0.625	1.885	1.643
	CARS	5	0.923	0.833	1.202	0.857	0.686	1.724	1.827
HMF	UVE	8	0.995	0.988	3.798	0.914	0.791	16.881	2.188
	SPA	6	0.996	0.988	3.913	0.787	0.583	23.855	1.568
	CARS	3	0.995	0.988	3.832	0.942	0.886	12.457	2.965
AA	UVE	8	0.991	0.981	0.201	0.964	0.929	0.384	3.766
	SPA	3	0.989	0.977	0.221	0.974	0.948	0.331	4.387
	CARS	5	0.987	0.974	0.233	0.977	0.954	0.310	4.681
Acidity	UVE	10	0.984	0.966	0.609	0.937	0.876	1.150	2.847
	SPA	6	0.969	0.934	0.842	0.934	0.869	1.180	2.771
	CARS	4	0.975	0.947	0.753	0.948	0.894	1.062	3.119

hydroxymethylfurfural, *AA* amylase activity

robustness of the established models. The reason for the selection of these optimal sensors was because the odor fingerprints detected by the selected optimal sensors might have some relationships with the odor of the predicted component of honey. The selected sensors were proven to be useful and important to establish the prediction models.

The performance of best-selection-LS-SVM models was compared with the LS-SVM models established using the fingerprints of honey from all 18 e-nose sensors (all-sensors-LS-SVM model). It was found that the best-selection-LS-SVM models had similar results compared with the corresponding all-sensors-LS-SVM models for the determination of glucose, fructose, AA, and acidity, whose RMSEV values of the best-selection-LS-SVM models increased by 2.75, 0.94, and 0.06 % and decreased by 2.27 %, respectively. On the other hand, the sensor selection improved the result of predicting HMF, where the RMSEV of its best-selection-LS-SVM model decreased by 12.57 % compared with the corresponding all-sensors-LS-SVM model. It should be noticed that instead of using all 18 sensors in the all-sensors-LS-SVM models, only three, five, three, five, and four sensors were selected for the determination of glucose, fructose, HMF, AA, and acidity, respectively. That means only 16.67, 27.78, 16.67, 16.67, and 22.22 % of the sensors were used for determining five quality components of honey. The above results show that the sensor selection in this study was efficient in terms of maintaining the model's accuracy and decreasing the sensor numbers. Moreover, the sensor selection was also able to improve the model's robustness, where the AV_RMSE values of the best-selection-LS-SVM models decreased by 90.46, 46.19, 38.95, 29.25, and 30.28 % for the determination of glucose, fructose, HMF, AA, and acidity, respectively, compared with those of the corresponding all-sensors-LS-SVM models.

As shown in Table 4, the e-nose is an efficient alternative for determining the quality of honey rapidly and non-invasively. For the analysis of five quality components, the best performance of the e-nose based on the best-selection-LS-SVM models was achieved for the determination of AA, which had an RPD value higher than 4.5, followed by the determination of glucose, HMF, and acidity that had RPD values around 3, and the RPD value of fructose determination was the lowest, but still over 1.5.

Mislabeling the botanical origin and quality components of honey is economically advantageous to unscrupulous suppliers, so labeling must be provided correctly with the aim of guaranteeing the authenticity of botanical origin and protecting the consumer from commercial exploitation. Traditional methods for identifying the botanical origin and determining the quality of honey are rather complex and time-consuming processes. In this study, the e-nose technique with multivariate analysis algorithms was investigated as an efficient analytical tool for identifying the botanical origin and determining quality components of honey. Compared with the linear pattern recognition methods like PCA and DFA, LS-SVM, which could retain the non-linear information of the e-nose, had better ability for discrimination of both geographical origins and botanical origins with 100 % OA. Similar to the analysis of origin identification, LS-SVM also proved to be

better than the linear regression method of PLSR for the quality prediction of honey. These results show that the non-linear correlations between e-nose responses were important for the origin and quality analysis of honey. Moreover, sensor selection was conducted for the first time to analyze e-nose fingerprints of honey, resulting in only three, five, three, five, and four sensors selected from 18 sensors in the e-nose for the determination of glucose, fructose, HMF, AA, and acidity, respectively. Sensor selection was shown to be efficient in terms of maintaining the model's accuracy, decreasing the sensor numbers, and improving the model's robustness. The best-selection-LS-SVM models had an ^{2} of 0.915, 0.686, 0.886, 0.948, and 0.894 for the determination of glucose, fructose, HMF, AA, and acidity, respectively, showing that simple, fast, and low-cost e-nose systems with only the selected sensors could be designed to refine this technique for the quality assessment of honey without additional laborious analysis. To the best of our knowledge, this is the first use of an e-nose for measuring glucose, fructose, HMF, and amylase activity of food products. The results of this study show that the use of e-nose fingerprints combined with chemometrics could identify the botanical origin and determine quality components of honey accurately and efficiently, so that the fraudulent labeling of honey could be prevented.

We thank Donald Grierson from the University of Nottingham (UK) for his kind suggestions and efforts in language editing. The work was supported by the National High-tech Research and Development Program of China (863 Program, 2011AA100807), Zhejiang Provincial Natural Science Foundation of China (LY14C200009), Zhejiang Silkworm Industry Science and Technology Innovation Team (2011R50028), and the Fundamental Research Funds for the Central Universities (2014QNA6016).

Abe, S. (2007). Sparse least squares support vector training in the reduced empirical feature space. *A A , 10*(3), 203–214.

Allwein, E. L., Schapire, R. E., & Singer, Y. (2001). Reducing multiclass to binary: a unifying approach for margin classifiers. *, 1*, 113–141.

Ampuero, S., & Bosset, J. (2003). The electronic nose applied to dairy products: a review. *A , B: C , 94*(1), 1–12.

Ampuero, S., Bogdanov, S., & Bosset, J.-O. (2004). Classification of unifloral honeys with an MS-based electronic nose using different sampling modes: SHS, SPME and INDEX. *, 218*(2), 198–207.

Araujo, M. C. U., Saldanha, T. C. B., Galvao, R. K. H., Yoneyama, T., Chame, H. C., & Visani, V. (2001). The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. *C , 57*(2), 65–73.

Benedetti, S., Mannino, S., Sabatini, A. G., & Marcazzan, G. L. (2004). Electronic nose and neural network use for the classification of honey. *A , 35*(4), 397–402.

Centner, V., Massart, D. L., deNoord, O. E., deJong, S., Vandeginste, B. M., & Sterna, C. (1996). Elimination of uninformative variables for multivariate calibration. *A C , 68*(21), 3851–3858.

Chen, L., Wang, J., Ye, Z., Zhao, J., Xue, X., Heyden, Y. V., & Sun, Q. (2012). Classification of Chinese honeys according to their floral [illegible]

Sun, T., Xu, W., Lin, J., Liu, M., & He, X. (2012). Determination of soluble solids content in navel oranges by Vis/NIR diffuse transmission spectra combined with CARS method. , , , *32*(12), 3229–3233.

Ulloa, P. A., Guerra, R., Cavaco, A. M., Rosa da Costa, A. M., Figueira, A. C., & Brigas, A. F. (2013). Determination of the botanical origin of honey by sensor fusion of impedance e-tongue and optical spectroscopy. *C* , , *94*, 1–11.

Wang, J., Kliks, M. M., Jun, S., Jackson, M., & Li, Q. X. (2010). Rapid analysis of glucose, fructose, sucrose, and maltose in honeys from different geographic regions using Fourier transform infrared spectroscopy and multivariate analysis. , *75*(2), C208–C214.

Williams, P. C. (2001). Implementation of near-infrared technology. In Williams & Norris (Eds.), - (2nd ed., pp. 145–169). Saint Paul: American Association of Cereal Chemists.

Wu, D., & He, Y. (2014). Potential of spectroscopic techniques and chemometric analysis for rapid measurement of docosahexaenoic acid and eicosapentaenoic acid in algal oil. *C* , *158*, 93–100.

Wu, D., & Sun, D.-W. (2013a). Advanced applications of hyperspectral imaging technology for food quality and safety analysis and assessment: A review—part i: fundamentals. & , *19*, 1–14.

Wu, D., & Sun, D.-W. (2013b). Potential of time series-hyperspectral imaging (TS-HSI) for non-invasive determination of microbial spoilage of salmon flesh. , *111*, 39–46.

Wu, D., Feng, L., He, Y., & Bao, Y. (2008a). Variety identification of Chinese cabbage seeds using visible and near-infrared spectroscopy. *A AB* , *51*(6), 2193–2199.

Wu, D., He, Y., Feng, S. J., & Sun, D.-W. (2008b). Study on infrared spectroscopy technique for fast measurement of protein content in milk powder based on LS-SVM. , *84*(1), 124–131.

Wu, D., Chen, X., Shi, P., Wang, S., Feng, F., & He, Y. (2009). Determination of α-linolenic acid and linoleic acid in edible oils using near-infrared spectroscopy improved by wavelet transform and uninformative variable elimination. *A C A* , *634*(2), 166–171.

Wu, D., Chen, J., Lu, B., Xiong, L., He, Y., & Zhang, Y. (2012a). Application of near infrared spectroscopy for the rapid determination of antioxidant activity of bamboo leaf extract. *C* , *135*(4), 2147–2156.

Wu, D., Shi, H., Wang, S., He, Y., Bao, Y., & Liu, K. (2012b). Rapid prediction of moisture content of dehydrated prawns using online hyperspectral imaging system. *A C A* , *726*, 57–66.

Wu, D., Chen, X., Cao, F., Sun, D.-W., He, Y., & Jiang, Y. (2014). Comparison of infrared spectroscopy and nuclear magnetic resonance techniques in tandem with multivariable selection for rapid determination of ω-3 polyunsaturated fatty acids in fish oil. *B* , *7*(6), 1555–1569.