Food Anal. Methods (2014) 7:360–365
DOI 10.1007/s12161-013-9634-4

Discrimination of Cocoa Beans According to Geographical Origin by Electronic Tongue and Multivariate Algorithms

Ernest Teye · Xingyi Huang · Fangkai Han · Francis Botchway

Received: 14 February 2013 / Accepted: 24 April 2013 / Published online: 12 May 2013

Abstract Electronic tongue as an advanced and novel emerging technology has been successfully utilized for the rapid identification of cocoa beans according to their geographical locations. Seven categories of cocoa beans from Ghana were used in this experiment. Electronic tongue system was used for data acquisition while three pattern recognition methods were applied comparatively to build discrimination model. The performances of the models were cross-validated to ensure its stability. Experimental results revealed that Fisher's discriminant analysis (FDA) is better than principal component analysis (PCA) for visualizing the cluster trends. *K*-nearest neighbor (KNN) model was slightly better than FDA model at an optimal performance of 100 % in the training set and 98.8 % in prediction set. Overall, support vector machine (SVM) was superior to both FDA and KNN with 100 % discrimination rate in both the training and prediction set at five PCs. This finding proves that electronic tongue technology coupled with SVM technique can rapidly, accurately, and reliably discriminate cocoa beans for quality assurance management.

Keywords Cocoa beans · Electronic tongue · Discrimination · Support vector machine

E. Teye · X. Huang (✉) · F. Han
School of Food and Biological Engineering,
Jiangsu University, Xuefu Road 301,
Zhenjiang 212013 Jiangsu, People's Republic of China
e-mail: ujshuang@gmail.com

X. Huang
e-mail: h_xingyi@163.com

E. Teye
School of Agriculture, Department of Agricultural Engineering,
University of Cape Coast, Cape Coast, Ghana

F. Botchway
Quality Control Company Limited, Cocobod Kade, Ghana

Introduction

Theobroma cacao is among the topmost commercially cultivated cash crops for many countries worldwide. Recently, cocoa bean is increasingly becoming extraordinarily famous consumer product because of its numerous health benefits (Jinap et al. 2005; Lee et al. 2003). Ghana is the second world leading producer of cocoa beans, and it is popularly called the "golden tree or pod" because of the pivotal role it plays in the nation's economy (Afoakwa et al. 2011). In this regard, stakeholders in the cocoa industry such as cocoa farmers, research institutes, and the government are working tirelessly to keep pace with the current international requirements for cocoa trade.

In addition, Ghana's cocoa beans has globally emerged amongst the best; it has become a standard by which other cocoa beans from other countries are judged and continues to enjoy better premium (Jinap et al. 1995; Othman et al. 2007). Factors such as favorable climatic conditions, well-defined pre- and postharvest activities coupled with vigilance quality control management at all production chains have synergistically contributed to the production of high-quality cocoa beans.

Traceability of cocoa bean quality is an essential activity for controlling fraud and accidental or deliberate mislabeling and adulteration for financial gains. Furthermore, the origin of cocoa bean has become an extremely relevant issue because a wide range of geographical areas has different chemical and organoleptic properties. To support this fact, studies have shown that there are various quality variations among agricultural commodities from different geographical locations (Rubayiza and Meurens 2005; Othman et al. 2007; Caligiani et al. 2007; Alonso-Salces et al. 2009; Cambrai et al. 2010; Pino et al. 1992). However, the conventional analytical methods used by other researchers (Hernández and Rutledge 1994; Luykx and van Ruth 2008; Anderson and Smith 2002;

Othman et al. 2012) are expensive, time consuming, cumbersome, require chemicals, and practically impossible when working with large samples among others. Therefore, simple and rapid analytical method for classification, identification, and detection of fraud regarding geographical origin of cocoa bean is needed to facilitate quality control management and enhance socioeconomic benefits.

Electronic tongue technology has emerged and proved to be a very remarkable scientific tool. It has successfully been used for qualitative and quantitative discrimination of food commodities. Chen et al. (2008) used electronic tongue to identify green tea grade, Chen et al. (2010) used taste sensor technique to determine caffeine and main catechins contents in green tea. Other studies on electronic tongue includes: identification of goat milk adulterated with bovine milk (Dias et al. 2009), for detection of sugars and acids in tomatoes (Beullens et al. 2006), for discrimination of honey according geographical origin and identification of honey (Escriche et al. 2012; Wei et al. 2009), recognition of six microbial species (Söderström et al. 2003), for classification and prediction of rice wine (Wei et al. 2011), and classification of red wine (Pigani et al. 2009), etc. Furthermore, electronic tongue has been primarily used in the food industry for process monitoring, freshness evaluation, authentic assessment, foodstuffs recognition, and quality analysis (Escuder-Gilabert and Peris 2010).

Despite the above development, upon a thorough literature search, no attempt has been made up till now on the use of electronic tongue technology for the discrimination of cocoa beans according to geographical origin. Moreover, there were no discussions on the identification analysis by linear and nonlinear algorithm methods. The purpose of this study was to identify cocoa bean samples from different geographical origins in Ghana. Principal components and Fisher's discriminant analysis were used to predict cluster trends, while Fisher's discriminant analysis (FDA), *K*-nearest neighbor (KNN), and support vector machine (SVM) were used comparatively to build classification models.

Materials and Methods

Sample Preparation

Cocoa bean samples from seven cocoa growing regions in Ghana were supplied by Quality Control Limited of Ghana. Each region had 25 samples and these were accurately labeled and transported to Jiangsu University, School of Food & Biological Engineering laboratory for further analysis. Twenty grams of the cocoa beans were ground separately for 15 s in a multipurpose grinder (QE-100, Zhejiang YiLi Tool Co., Ltd. China). The metal grinder container was then allowed to cool after grinding of each sample. This was done to reduce loss of volatile compounds. The powder of each sample was sieved with a 500-μm mesh before further analysis.

Data Acquisition

The experiments were performed with the α-Astree electronic tongue device (Alpha MOS Company, Toulouse, France). The sensor array of this device comprises of seven potentiometric chemical sensors namely; ZZ, BA, BB, CA, GA, HA, and JB, and a reference electrode. Each sensor is composed of silicon transistors with an organic coating. The sensitivity of the seven sensors are different from that of the five tastes; sourness, saltiness, sweetness, bitterness, and savory (Wei et al. 2009). Therefore, the sensors are sensitive to chemical and organoleptic properties in the samples. The responses produced by the sensors are transmitted through the transducer into signal data. These signals are the intensity values derived from the differences between the sensors and reference electrode (Ag/AgCl). The complete electronic tongue technology used in this study is shown in Fig. 1.

In this experiment, 1.0 g of each sample was accurately weighed and dissolved in 100 ml boiled distilled water. After that, the mixture was allowed to cool and then filtered with a filter paper (GB/T8314-2002). Eighty milliliter of the filtrate (liquid sample) was used for subsequent analysis. The reading time for each sample was set at 120 s (as seen in Fig. 2) and the sensors were rinsed in distilled water for 20 s after successive readings. Five samples were detected at a time. From Fig. 3, it could be verified that each sensor gave a different intensity based on the sensitivity of the sensors to the chemical properties in the cocoa bean samples used. After the measurements, the response values at last 10 (110–120)s of each sensor were extracted and further analyzed (Wei et al. 2009). At this time range, all the sensors were found to be more stable.

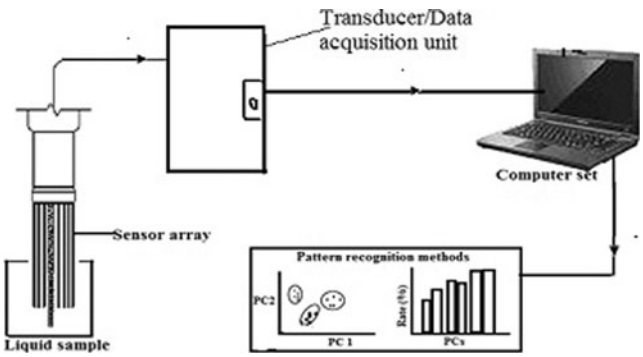

Fig. 1 Schematic diagram of the electronic tongue technology

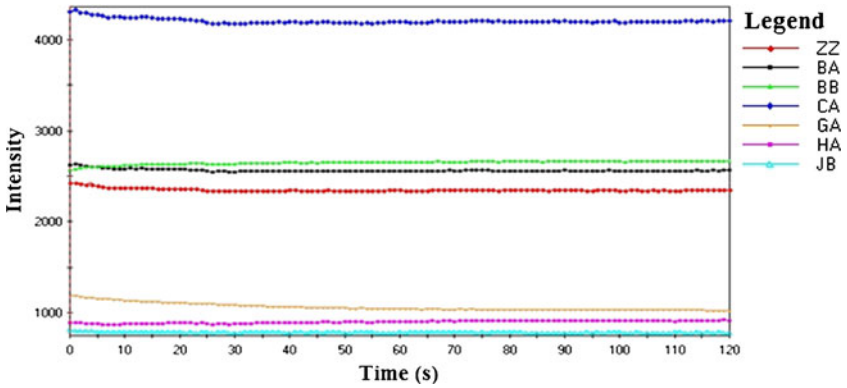

Fig. 2 Electronic tongue data acquisition from cocoa bean samples infusion

Software Device and Data Processing

All statistical calculations and pattern recognition methods were carried out with Microsoft Office Excel 2007 and Matlab Version 7.14 (Mathworks Inc., USA) in Windows 7 ultimate for data processing.

Pattern Recognition Methods

Supervised and unsupervised pattern recognition methods were investigated. Principal component analysis (PCA), FDA, KNN, and SVM were attempted to derive discrimination models. In each pattern recognition method, the 175 samples used were randomly grouped into two sets; 105 samples as training set and 70 samples as prediction set.

Results and Discussion

PCA Versus FDA

In this study, PCA and FDA were used comparatively to derive a visual cluster trend. PCA is an unsupervised pattern recognition method which is used for visualizing data trends in a dimensional space. It works by reducing the dimension of the data matrix and compressing the information into interpretable variables called principal components (PCs), which are orthogonal (Luna et al. 2013).

In this experiment, all the sensors data of the cocoa bean samples from the seven cocoa-growing regions were used for the PCA. PCA is not a classification tool, its properties provided a classification trends as a result of the seven cocoa-growing regions. To visualize the data trends, a score plot was obtained by using the topmost three principal components (PC1, PC2, and PC3). Figure 4 shows the outcome of the principal component analysis. The cluster trends were not very satisfactory though the PC1, PC2, and PC3 gave a total accumulative contribution of 92.95 % variance for the 175 samples used in the study. The first principal component accounted for the maximum variance while the second accounted for the maximum residual variance. As seen in Fig. 4, most of the data points overlapped resulting to unsatisfactory cluster trends. Hence, supervised pattern recognition methods were also attempted.

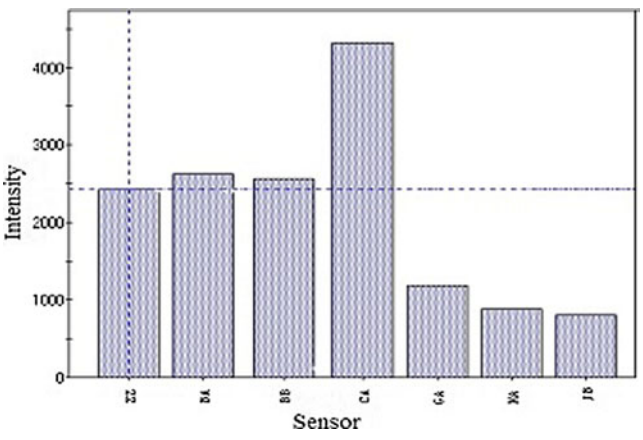

Fig. 3 The intensity of the sensors from the cocoa bean sample infusion

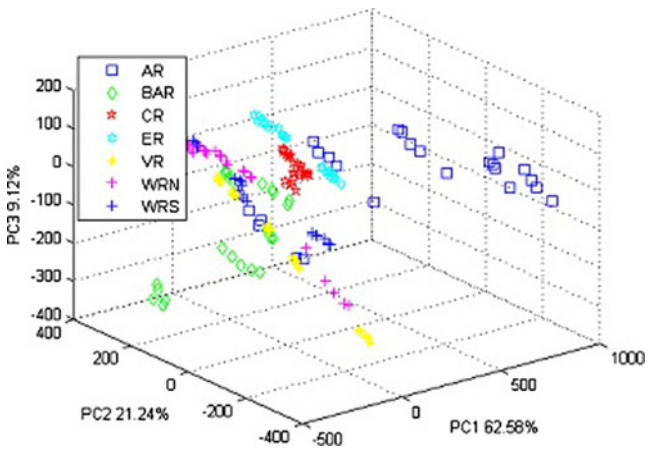

Fig. 4 Classification results by PCA for the seven geographical locations

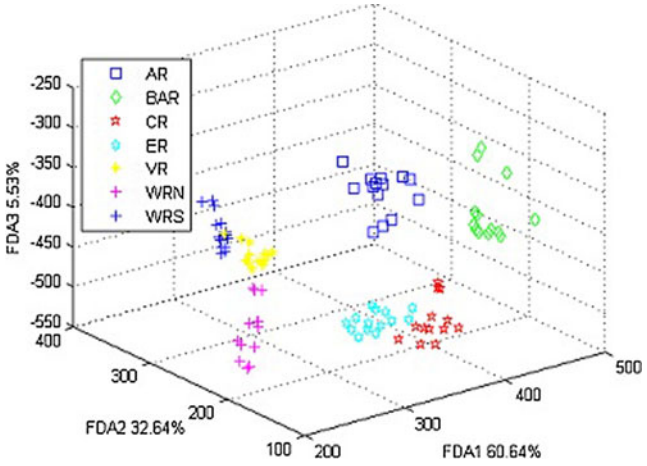

Fig. 5 Classification results by FDA for the seven geographical locations

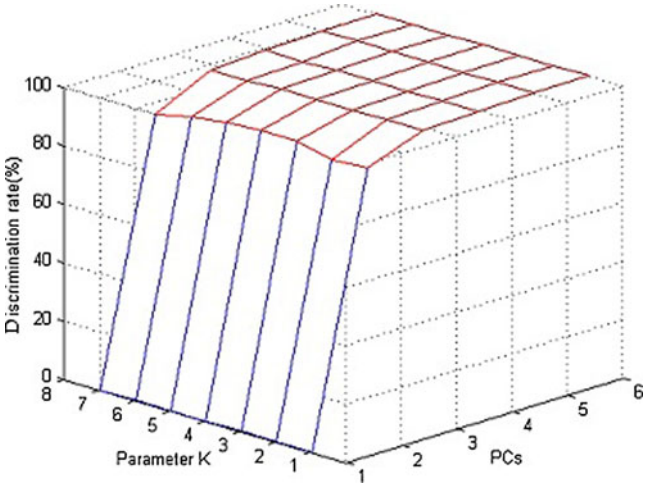

Fig. 6 Discrimination rate of *K*-NN model by cross validation under different PCs and *K*

FDA is a well-known supervised pattern recognition method which derived its name from its inventor Ronald Fisher. It functions by seeking the best projection subspace such that the ratio between class scatter to within class scatter is maximized (Yang et al. 2009); this brings about a smaller variance and a clearer discrimination. In this experiment, the top three Fisher's discriminant functions were selected and a 3-dimensional plot constructed with FDA1, FDA2, and FDA3 as seen in Fig. 6. The seven cocoa-growing regions clustered satisfactorily with a clear cluster trend. The top three FDA accounted for a total of 98.81 % (FDA1 60.64 %, FDA2 32.64 %, and FDA3 5.53 %). All the seven groups of samples were well discriminated with no overlap. FDA was further used to model data classification and the outcome was 95.74 % for the training set and 95.65 % for prediction set.

K-Nearest Neighbor

K-nearest neighbor is a linear discriminant technique that can be used to categorize an unknown sample in the prediction set according to the majority of its *K*-neighborhood members in the training set (Alcázar et al. 2007). KNN performs considerably well when working with multiclass simultaneous problem solving. The optimal performance of KNN model is significantly influenced by parameter *K*, and the choice is often determined by cross-validation method (Xu et al. 2012). This process is extremely important because it gives an estimate of the robustness of the model and normally, the best *K* gives the lowest error rate and the maximum number of neighbors (Japon-Lujan et al. 2006). The lower values of parameter *K* and PCs are preferred. Therefore in this study, seven *K* values and six PCs were selected and tested concurrently. The classification rate by KNN model was 98.8 and 100 % in the training set and prediction set, respectively. To check the robustness and stability of the model on a new set of data, the model was cross-validated. Figure 6 shows the optimal KNN model by cross-validation at $K=3$ and PCs=4 with classification rate of 99.4 %. This further proves that the model is stable on future dataset.

Table 1 Comparison of the identification results from FDA, KNN, and SVM models

Models	Optimum	Discrimination results of models	
	PCs	Training set (%)	Prediction set (%)
FDA	7	95.74	95.65
KNN	4	98.86	100
SVM	5	100	100

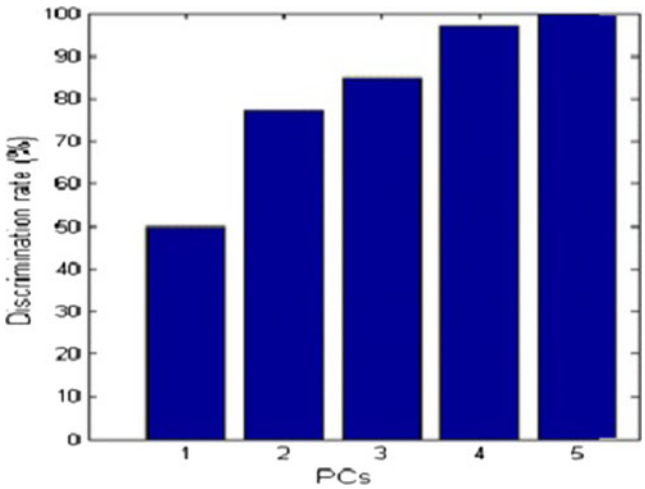

Fig. 7 Discrimination rate of SVM model by cross validation under different PC's

Support Vector Machine

Support vector machine technique was also applied in this experiment since the linear models could not provide a complete solution to the discrimination problem. SVM is a nonlinear supervised learning method developed by Vapnik and co-workers for two-group classification problems (Cortes and Vapnik 1995). It works by obtaining the optimal boundary of two groups in a vector space independent on the probabilistic arrangements of vectors in the training set. When the linear boundary in the low dimension input space is not enough to separate the two classes, SVM can create a hyper plane that allows linear separation in the higher dimension feature space. The optimal performance of SVM is dependent on the choice of the kernel function. There are three kernel functions: polynomial kernel, Gaussian kernel, or RBF and sigmoid kernel. However, Gaussian kernel function was chosen because it is comparatively simple and quick in its computation (Chen et al. 2009; Lin et al. 2009). The discrimination rate for SVM model was 100 % for both training and prediction set. To ensure the stability of the model, it was cross-validated as shown in Fig. 7. Cross-validation classification rate of the model was also100% when PCs=5. This revealed that the SVM model used for the discrimination of cocoa bean samples is robust and stable in this experiment. Therefore, this model will perform considerably well with new dataset.

General Discussion

As shown in Figs. 4 and 5, FDA gave a well-distinguished cluster trend; thus, it showed all the seven regional samples used in this experiment as compared to PCA where most of the data points overlapped. It could be explained that though PCA and FDA are all linear mathematical tools, PCA functions by reducing the dimensionality while preserving as much as possible the variance in a high dimensional space, whereas FDA also performs dimensionality reduction but preserves as much as possible the class or group discriminatory information. In other words, FDA works by reducing within-class distance and enhances between-class distance resulting to smaller variance and a better discrimination. Therefore, as seen from Figs. 4 and 5, FDA was comparatively better than PCA and this agrees with Yang et al. (2009). Furthermore, the neat classification could be explained by the chemical and organoleptic composition of the cocoa beans used. Cocoa beans exhibit significant differences in their own internal and external characteristics according to different geographical locations. This is because geographical origins have peculiar factors, which influences, pre- and postharvest activities. It further leads to differences in quality compositions of the cocoa beans. However, the conventional analytical methods are very cumbersome, expensive, and sometimes less sensitive. Interestingly, the electronic tongue technology has successfully discriminated cocoa beans from seven cocoa growing regions.

Table 1 compares the identification rates of FDA, KNN, and SVM. The identification rate (in percent) is an important factor for testing the performance of the models, and it was calculated by Eq. (1).

$$R_t = [(N_1)/(N_2)]x100 \quad (1)$$

Where R_t is the identification rate (in percent), N_1 is the number of samples correctly identified in either the training set or prediction set, and N_2 is the total number of samples used in either the training set or prediction set. The number of PCs was optimized while building the model as can be seen from Figs. 6 and 7. After the cross-validation, KNN and SVM gave an optimal performance. From Table 1, it could be seen that SVM was slightly superior to both FDA and KNN models. This could be explained that SVM as a nonlinear pattern recognition method is stronger in self-learning and adjustment than the linear counterparts. This technology could successfully be used for traceability management or quality assurance in the cocoa bean trade. However, it must be stated that further study is required to verify this claim.

Conclusion

Generally, the electronic tongue technology (electronic tongue device and multivariate algorithms) has a high potential for discriminating pure cocoa bean for quality control examinations. FDA gave a well-defined separation or cluster of the seven groups in the three-dimensional space. Among the supervised pattern recognition methods used for building discriminatory model, the performance of SVM was superior to all. Optimal SVM model was derived at five principal components and identification rate of 100 % for both training set and prediction set.

Acknowledgments This work has been financially supported by the National Natural Science Foundation of China (no. 31071549) and Priority Academic Program Development of Jiangsu Higher Education Institutes. Winifred Akpene Teye is acknowledged for proof reading this manuscript.

Conflict of Interest Xingyi Huang has received research grant from National Natural Science Foundation of China (no. 31071549) and Priority Academic Program Development of Jiangsu Higher Education Institutes. Ernest Teye declares that he has no conflict of interest. Fangkai Han has received no research grant from any company. Francis Botchway declares that he has no conflict of interest. The authors declare that this article does not contain any studies with human or animal subjects.

References

Afoakwa EO, Quao J, Takrama J, Budu AS, Saalia FK (2011) J Food Sci Tech 1–9

Alcázar A, Ballesteros O, Jurado JM, Pablos F, Martín MJ, Vilches JL, Navalón A (2007) J Agric Food Chem 55:5960–5965

Alonso-Salces RM, Serra F, Reniero F, HÉberger KR (2009) J Agric Food Chem 57:4224–4235

Anderson KA, Smith BW (2002) J Agric Food Chem 50:2068–2075

Beullens K, Kirsanov D, Irudayaraj J, Rudnitskaya A, Legin A, Nicolaï BM, Lammertyn J (2006) Sensors and Actuators B: Chemical 116:107–115

Caligiani A, Cirlini M, Palla G, Ravaglia R, Arlorio M (2007) Chirality 19:329–334

Cambrai A, Marcic C, Morville S, Sae-Houer P, Bindler F, Marchioni E (2010) J Agric Food Chem 58:1478–1483

Chen Q, Zhao J, Vittayapadung S (2008) Food Res Int 41:500–504

Chen Q, Zhao J, Lin H (2009) Spectrochim Acta A Mol Biomol Spectrosc 72:845–850

Chen Q, Zhao J, Guo Z, Wang X (2010) J Food Compos Anal 23:353–358

Cortes C, Vapnik V (1995) Mach Learn 20:273–297

Dias LA, Peres AM, Veloso ACA, Reis FS, Vilas-Boas M, Machado AASC (2009) Sensors and Actuators B: Chemical 136:209–217

Escriche I, Kadar M, Domenech E, Gil-Sánchez L (2012) J Food Eng 109:449–456

Escuder-Gilabert L, Peris M (2010) Anal Chim Acta 665:15–25

Hernández C, Rutledge D (1994) Analyst 119:1171–1176

Japon-Lujan R, Ruiz-Jimenez J, De Castro ML (2006) J Agric Food Chem 54:9706–9712

Jinap S, Dimick P, Hollender R (1995) Food Control 6:105–110

Jinap S, Jamilah B, Nazamid S (2005) J Sci Food Agric 85:917–924

Lee KW, Kim YJ, Lee HJ, Lee CY (2003) J Agric Food Chem 51:7292–7295

Lin H, Chen Q, Zhao J, Zhou P (2009) J Pharm Biomed Anal 50:803–808

Luna AS, da Silva AP, Pinho JSA, Ferré J, Boqué R (2013) Spectrochim Acta A Mol Biomol Spectrosc 100:115–119

Luykx DMAM, van Ruth SM (2008) Food Chem 107:897–911

Othman A, Ismail A, Abdul Ghani N, Adenan I (2007) Food Chem 100:1523–1530

Othman A, Jalil AMM, Weng KK, Ismail A, Ghani NA, Adenan I (2012) African Journal of Biotechnology 9(7):1052–1059

Pigani L, Foca G, Ulrici A, Ionescu K, Martina V, Terzi F, Vignali M, Zanardi C, Seeber R (2009) Anal Chim Acta 643:67–73

Pino J, Nu de Villavicencio MÑ, Roncal E (1992) Food Nahrung 36(3):302–306

Rubayiza AB, Meurens M (2005) J Agric Food Chem 53:4654–4659

Söderström C, Winquist F, Krantz-Rülcker C (2003) Sensors and Actuators B: Chemical 89:248–255

Wei Z, Wang J, Liao W (2009) J Food Eng 94:260–266

Wei Z, Wang J, Ye L (2011) Biosens Bioelectron 26:4767–4773

Xu W, Song Q, Li D, Wan X (2012) J Agric Food Chem 60:7064–7070

Yang WY, Liang W, Xin L, Zhang SW (2009) Acta Automatica Sinica 35:1513–1519