JOURNAL OF AGRICULTURAL AND FOOD CHEMISTRY

Article

pubs.acs.org/JAFC

Discrimination of the Production Season of Chinese Green Tea by Chemical Analysis in Combination with Supervised Pattern Recognition

Wenping Xu,[†,§,∥] Qiushuang Song,[†,∥] Daxiang Li,[†,∥] and Xiaochun Wan*[,†]

[†]Key Laboratory of Tea Biochemistry and Biotechnology, Ministry of Agriculture and Ministry of Education, Anhui Agricultural University, Hefei 230036, Anhui, People's Republic of China

[§]National Institute of Measurement and Testing Technology, Chengdu 610021, Sichuan, People's Republic of China

ABSTRACT: High-performance liquid chromatography (HPLC) has been used to quantify levels of free amino acids, catechins, and caffeine in Chinese green tea. Levels of free amino acids and catechins in green tea leaves show obvious variation from spring to summer, which is useful information to identify the production season of commercial green tea. Supervised pattern recognition methods such as the *K*-nearest neighbor (KNN) method and Bayesian discriminant method (a type of linear discriminant analysis (LDA)) were used to discriminate between the production seasons of Chinese green tea. The optimal accuracy of the KNN method was ≤97.61 and ≤94.80% as validated by resubstitution and cross-validation tests, respectively, and that of LDA was ≤95.22 and ≤93.54%, respectively. Compared with LDA, the KNN method did not require a Gaussian distribution and was more accurate than LDA. The KNN method in combination with chemical analysis is recommended for discrimination of the production seasons of Chinese green tea.

KEYWORDS: *green tea, chemical composition, tea production season, pattern recognition, K-nearest neighbor method*

INTRODUCTION

Green tea is a popular beverage that is consumed worldwide. To produce green tea, freshly harvested leaves are immediately steamed or roasted to prevent fermentation, thereby yielding a dry, stable product.[1] In the late 1980s to the early 1990s in China, tea-processing technology was simple, and the types and grades of green tea were easy to group. In this period, green tea was usually divided into nine types. Each type was usually separated into six grades according to the drying process and the shape and quality of the tea leaves (see withdrawn Chinese Standard GB/T 14456-1993). The basic principles of the processing of green tea have not changed appreciably; each company involved in tea production has its own unique process, which has resulted in the production of numerous types of green tea. Grading green-tea products according to the previously used rules is difficult. Hence, the grading rules have been rewritten in the new Chinese Standard for green tea (Chinese Standard GB/T 14456-2008). At present, the season of production, flavor, and maturity of the leaves (as judged by their shape and quality) as well as safety are the main concerns when consumers buy green tea. According to the growing season, green tea can be divided into "spring tea", "summer tea", and "autumn tea" in China, which refers to tea harvested and processed before late May, between early June and early July, and after mid-July, respectively. Autumn tea accounts for only a small part due to the slow growth of tea plants compared with spring tea and summer tea. In a particular tea plantation, the chemical composition of green tea changes dramatically in different growing seasons. Studies have shown that spring tea commonly has higher levels of amino acids and moderate levels of catechins (thereby yielding a heavy, mellow, and brisk flavor), whereas summer tea usually contains higher levels of catechins and lower levels of amino acids (leading to a more bitter and astringent flavor).[2,3] In many cases, consumers judge the quality of green tea mainly by production season (production date). The production season of green tea has been an important issue that determines its price. The identification of production season of undated green tea has recently become one of the most important challenges for tea researchers. Although sensory test methods are widely used in tea quality studies, it is difficult to distinguish spring tea from summer tea by sensory tests because the taste is affected by tea cultivar, processing methods, and agroclimatic conditions.[4] In addition, the result assessed by tasters is often less coherent and less impartial, because it is influenced easily by the physical or physiological factors of tasters.[5]

In international trading, besides sensory assessments, chemical components such as water-soluble extracts, total ashes, and total polyphenols (TP) are important supplementary markers for the quality of green tea. However, macroindicators are useful only for poor-quality products. The color, aroma, and taste are important aspects of tea quality and are determined by chemicals such as theaflavins, volatile organic compounds (VOCs), catechins, and caffeine.[6] Chemical analyses in combination with pattern recognition can overcome the limitations of sensory tests to provide an interesting approach to quality control in the food industry.[7] Pigments such as theaflavins have a significant influence on the color of black tea. Theaflavins have been used for the study of tea quality

Received: April 5, 2012
Revised: June 18, 2012
Accepted: June 21, 2012
Published: June 21, 2012

dx.doi.org/10.1021/jf301340z | J. Agric. Food Chem. 2012, 60, 7064−7070

combined with principal component analysis (PCA) and correlation analysis.[8,9] VOCs identified by a metal oxide sensor (MOS)-based electronic nose (EN) such as (*E*)-2-hexenal have been used successfully to separate tea samples using partial least-squares (PLS) analysis, linear discriminant analysis (LDA), PCA, and the fuzzy c means clustering algorithm (FCM).[10,11] Chemicals that have a significant influence on flavor, such as catechins and caffeine, have been used for the identification of tea grades using the *K*-nearest neighbor (KNN) algorithm, artificial neural network (ANN), LDA, and support vector classification (SVC) pattern recognition with electronic tongue technology and high-performance liquid chromatography (HPLC).[4,5,12] Metal elements that relate to the safety of tea and can reflect differences in geographical origin have been used to classify teas and beverages by cluster analysis (CA), PCA, LDA, probabilistic neural networks (PNN), artificial neural networks trained by back-propagation (BP-ANN), and soft independent modeling of class analogy (SIMCA).[13−16] Near-infrared (NIR) spectroscopy with SIMCA and the support vector machine (SVM) has been used successfully to identify tea categories.[17,18] The studies mentioned also demonstrated that NIR spectroscopy with the SVM could be applied to help discriminate between types of green tea according to geographical origin.[19] Reports also showed that multispectral imaging techniques combined with the least-squares support vector machine (LS-SVM) method could be used to identify tea categories.[20]

As far as we know, studies using chemicals to discriminate the production season of green tea are lacking. In our earlier study, we had difficulty separating spring green tea from summer green tea by a single method of analysis because the levels of chemicals and qualities of collected tea samples varied considerably on the basis of the locations of tea gardens, agroclimatic conditions, and type of plantation, as well as cultivars and methods of manufacturing.

In the present study, a method of chemical analysis in combination with pattern recognition was introduced to discriminate between two types of green tea: spring green tea and summer green tea. The flavor was the most significant difference between spring green tea and summer green tea. Earlier studies showed that the bitterness and astringency of tea are attributed mainly to caffeine, catechins, and flavon-3-ol glycosides,[6,21] whereas several free amino acids (especially theanine) impart a umami taste.[22] Thus, in the present study, HPLC was used for the determination of the levels of catechins and free amino acids of 160 green tea samples produced in the spring or summer. Further quantitative data were analyzed by pattern recognition: the KNN method and the Bayesian discriminant method. The present study introduced a valuable chemometric pattern recognition technique to distinguish between spring green tea and summer green tea.

MATERIALS AND METHODS

Materials. One hundred and sixty green tea samples manufactured in the spring or summer between 2007 and 2011 were collected from tea factories across China (i.e., the provinces of Anhui, Jiansu, Henan, Zhejiang, Sichuan, Yunnan, and Hubei). In addition, the tea samples collected were representative samples of each tea-producing area. The processing of all samples was in accordance with the basic principles of green tea production: freshly harvested leaves were immediately steamed or roasted to prevent fermentation. All samples were stored at −20 °C to prevent the oxidation of chemicals such as catechins.

Chemical standards of epigallocatechin (EGC), catechin (C), epicatechin (EC), epigallocatechin gallate (EGCG), gallocatechin gallate (GCG), epicatechin gallate (ECG), caffeine, and theanine were purchased from Sigma-Aldrich (St. Louis, MO, USA). HPLC grade acetonitrile, methanol, and acetic acid were obtained from Tedia (Fairfield, OH, USA). Dissolved amino acids and an AccQ·Tag Kit were purchased from Waters (Milford, MA, USA). The reagent kit comprised Waters AccQ·Fluor borate buffer, Waters AccQ·Fluor powder (6-aminoquinolyl-*N*-hydroxysuccinimidyl carbamate (AQC)), Waters AccQ·Fluor reagent diluent, and Waters amino acid hydrolysate standard (each ampule contained a 2.5 mM mixture of 16 amino acids: aspartic acid (Asp), serine (Ser), glutamic acid (Glu), glycine (Gly), histidine (His), arginine (Arg), threonine (Thr), alanine (Ala), proline (Pro), tyrosine (Tyr), valine (Val), methionine (Met), lysine (Lys), isoleucine (Ile), leucine (Leu), and phenylalanine (Phe) plus cysteine (Cys), which was present at 1.25 mM). The rest of the reagents and solvents were of analytical grade. Water (18.2 mΩ) was purified by a Millipore Mill-Q Ultrapure Water System (Billerica, MA, USA).

Methods. *Preparation of Tea Infusions.* Tea infusions for the analysis of catechins and caffeine concentrations were prepared according to the procedures described in ISO standard 14502-1.[23] Tea infusions for the analysis of concentrations of free amino acids were prepared according to the procedures described in the Chinese Standard GB/T 8312-2002.[24] The infusions for the testing of amino acids were further diluted 1-fold with purified water and filtered using a 0.45 μm filter before derivatization.

Analyses of Chemical Composition. The chemicals in green tea were determined using a Waters 600E series HPLC equipped with a quaternary pump, a 2475 fluorescence detector, and a 2489 ultraviolet (UV)−visible detector. Catechins (EGC, C, EC, EGCG, GCG, and ECG) and caffeine in tea infusions were analyzed on a reverse-phase C_{18} column (Phenomenex Luna 5 μm, 250 mm × 4.6 mm). Detection was carried out at 278 nm, and the injection volume was 5 μL. The flow rate was 1.0 mL min^{-1}, and the column temperature was maintained at 20 °C. The mobile phase consisted of 2% aqueous acetic acid (v/v) as solvent A, acetonitrile as solvent B, and water as solvent C. The gradient conditions were as follows: 0−4 min, 92% A and 8% B; 32 min, 79% A and 21% B; 37 min, 71% A and 29% B; 38 min, 0% A and 29% B; 45−50 min, 0% A and 75% B; and 51−60 min, 92% A and 8% B. Results were recorded using a Waters Empower 2 ChemStation and quantified with external standards (ESTD).

The levels of amino acids were measured using the Waters AccQ·Tag method on a Waters AccQ·Tag column (Nova-Pak C_{18}, 4 μm, 150 mm × 3.9 mm). The AccQ·Tag method is a highly sensitive, stable, and reproducible method for amino acid analyses.[25,26] The procedures were conducted as directed in the AccQ·Fluor Reagent Kit Care and Use Manual. The elution conditions employed were as follows: column temperature, 37 °C; fluorescence detector, λ_{ex} = 250 nm, λ_{em} = 395 nm; flow rate, 1.0 mL min^{-1}; injection volume, 5 μL; mobile phase A, AccQ·Tag eluent A; mobile phase B, 60% (v/v) acetonitrile. The gradient conditions were as follows: 0 min, 100% A; 0.5 min, 98% A; 14.5 min, 93% A; 18.5 min, 90% A; 31.5 min, 67% A; 32.5 min, 90% A; 33.5−36.5 min, 0% A; and 37−47 min, 100% A. The mixed solution of Waters amino acid hydrolysate standards supplemented with theanine was used as a stock solution to set up the calibration curves. In the series of standard solutions, the concentrations of theanine ranged from 10 to 400 pmol/μL, the concentrations of Cys were between 1.25 and 50 pmol/μL, and the concentrations of the other 16 amino acids ranged from 2.5 to 100 pmol/μL.

Data Processing. Data are the mean values of the duplicated analysis of each sample. Data for the amino acids, catechins, and caffeine in 160 green tea samples were separated into two groups according to season. Results were analyzed by SAS 9.1 software (SAS Institute, Cary, NC, USA) and Minitab 15 software (Minitab, University Park, PA, USA). The normal distribution (Gaussian distribution) test of data was conducted using the Shapiro−Wilk test using SAS software (univariate procedure). If the *p* value given by the Shapiro−Wilk test was <0.05, then the test data set did not follow a Gaussian distribution. The mean levels of chemicals in spring tea and summer tea were compared by SAS software using the Wilcoxon two-

Table 1. Average Levels of Amino Acids in Green Tea with Normal Distribution and *t* Test[a]

	spring		summer		
parameter	mean ± SD (mg g^{-1})	*p* value (Shapiro–Wilk)	mean ± SD	*p* value (Shapiro–Wilk)	*p* value (Wilcoxon)
Asp	3.950 ± 1.361	0.0218	3.514 ± 1.468	0.0598	0.2181
Ser	1.486 ± 0.84	<0.0001	1.986 ± 1.336	<0.0001	0.1715
Glu	6.094 ± 2.582	0.001	3.772 ± 1.811	0.0089	<0.0001
Gly	2.138 ± 1.506	<0.0001	1.133 ± 1.032	<0.0001	0.0003
His	2.673 ± 1.049	0.4253	1.865 ± 1.010	0.0296	0.0004
Arg	3.046 ± 2.360	<0.0001	1.221 ± 0.878	0.0002	<0.0001
Thr	0.588 ± 0.230	0.0144	0.498 ± 0.222	0.0202	0.0606
Ala	0.617 ± 0.230	<0.0001	0.47 ± 0.215	0.3103	0.0066
Pro	1.615 ± 0.701	0.0001	1.314 ± 0.560	0.317	0.0117
theanine	16.258 ± 6.990	0.0774	10.675 ± 3.824	0.2359	0.001
Tyr	0.379 ± 0.185	0.0018	0.522 ± 0.309	0.0018	0.006
Val	0.317 ± 0.228	<0.0001	0.589 ± 0.379	0.0006	0.0003
Met	3.101 ± 1.463	0.0004	1.608 ± 0.755	0.0355	<0.0001
Lys	0.466 ± 0.345	<0.0001	0.677 ± 0.483	0.0009	0.0349
Ile	0.202 ± 0.162	<0.0001	0.388 ± 0.267	0.0002	0.0002
Leu	0.369 ± 0.259	<0.0001	0.611 ± 0.402	<0.0001	0.0027
Phe	0.352 ± 0.273	<0.0001	0.727 ± 0.529	<0.0001	0.0003
AA	43.653 ± 14.93	<0.0001	31.569 ± 12.165	0.0442	<0.0001

[a]Results of amino acid levels are expressed as the mean ± SD. If the *p* value given by Shapiro–Wilk is <0.05, then the test data set is not Gaussian distribution. If the *p* value given by Wilcoxon two-sample test is >0.05, then there is no significant difference between the mean levels of chemical between the spring and summer teas. Abbreviations: aspartic acid (Asp), serine (Ser), glutamic acid (Glu), glycine (Gly), histidine (His), arginine (Arg), threonine (Thr), alanine (Ala), proline (Pro), tyrosine (Tyr), valine (Val), methionine (Met), lysine (Lys), isoleucine (Ile), leucine (Leu), phenylalanine (Phe), and cysteine (Cys).

Table 2. Average Levels of Catechins and Caffeine in Green Tea with Normal Distribution and *t* Test[a]

	spring		summer		
parameter	mean ± SD (mg g^{-1})	*p* value (Shapiro–Wilk)	mean ± SD (mg g^{-1})	*p* value (Shapiro–Wilk)	*p* value (Wilcoxon)
EGC	4.493 ± 7.862	<0.0001	6.633 ± 7.655	<0.0001	0.1303
C	0.869 ± 0.737	<0.0001	0.989 ± 1.255	<0.0001	0.0178
EC	3.205 ± 2.473	<0.0001	4.469 ± 2.643	<0.0001	0.0008
EGCG	42.172 ± 19.383	<0.0001	62.243 ± 35.966	<0.0001	0.0014
GCG	7.107 ± 7.125	<0.0001	3.817 ± 5.665	<0.0001	0.0037
ECG	12.708 ± 4.654	0.2733	15.760 ± 6.714	<0.0001	0.0194
catechins	70.554 ± 34.020	<0.0001	93.911 ± 54.588	<0.0001	0.2626
caffeine	36.558 ± 5.024	0.0111	38.721 ± 6.059	0.8111	0.0626

[a]Results of catechin and caffeine levels are expressed as the mean ± SD. If the *p* value given by Shapiro–Wilk is <0.05, then the test data set is not Gaussian distribution. If the *p* value given by the Wilcoxon two-sample test is >0.05, then there is no significant difference between the mean levels of chemical between the spring and summer teas. Abbreviations: epigallocatechin (EGC), catechin (C), epicatechin (EC), epigallocatechin gallate (EGCG), gallocatechin gallate (GCG) and epicatechin gallate (ECG).

sample test method (NPAR1WAY procedure). If the *p* value given by the Wilcoxon two-sample test method was >0.05, then there was no significant difference between the mean levels of chemicals between spring tea and summer teas. The KNN method was employed to discriminate between the production season of green tea using original data without normal conversion, and the results of resubstitution and cross-validation tests were compared when the *K* value varied from 3 to 11. Furthermore, the original data of chemical compositions were analyzed by Minitab software for the conversion of normal distribution and translated as

$$\text{if } \lambda \neq 0, \text{ then } y' = (y^{\lambda} - 1)/\lambda \quad (1)$$

$$\text{if } \lambda = 0, \text{ then } y' = \log y \quad (2)$$

where y' represents translated data and y the original data. The values of λ were given automatically by Minitab software. The original and translated data were analyzed by SAS software (STEPDISC procedure, forward selection method, sle = 0.1, sls = 0.1) separately, and the significant variables were chosen for the Bayesian discriminant method (DISCRIM procedure). The accuracy of quadratic and linear discriminant functions, which were built by the original data and translated data, was compared.

RESULTS AND DISCUSSION

Analyses of Amino Acids. The free amino acids in tea were detected by the AccQ·Tag method. Using individual calibration curves, each of the free amino acids was quantified. Green tea was rich in Asp, Ser, Glu, Gly, His, Arg, Pro, theanine, and Met, which were >1.0 mg g^{-1} and together accounted for >92.4% of the amino acids (Table 1). Theanine was the most abundant free amino acid in tea, accounting for >37.2% of the amino acids. Cys was absent in all tea samples.

There was a significant difference between spring tea and summer tea with regard to the mean levels of 14 amino acids (Glu, Gly, His, Arg, Ala, Pro, Theanine, Tyr, Val, Met, Lys, Ile, Leu, and Phe). Six amino acids (Tyr, Val, Lys, Ile, Leu, and Phe) showed higher levels in summer tea, whereas the others showed higher levels in spring tea. The mean level of amino

dx.doi.org/10.1021/jf301340z | J. Agric. Food Chem. 2012, 60, 7064–7070

acids in spring tea was 43.653 mg g^{-1}, which was significantly different from that seen in summer tea (31.569 mg g^{-1}) and was 1.38 times higher than that in summer tea. The mean level of theanine in spring tea was 16.258 mg g^{-1}, which was 1.51 times higher than that in summer tea (10.675 mg g^{-1}). The amino acid content in tea (especially theanine) is considered to have a positive correlation with tea quality.[22] The seasonal variation of amino acid and theanine levels in the present study was in accordance with that given in previous papers,[27−29] but the earlier studies did not clearly show the seasonal variation of individual amino acids. The present study showed that only eight amino acids (Glu, Gly, His, Arg, Ala, Pro, theanine, and Met) had the same trend in seasonal variation. The Shapiro−Wilk test showed that most of the variables did not follow a normal distribution ($P < 0.05$). Nonparametric methods should be used in the discriminant analysis if the data are not translated into a normal distribution.

Analyses of Catechin and Caffeine. The levels of catechins and caffeine are shown in Table 2. There were significant differences between spring tea and summer tea with respect to the levels of all catechins except EGC. For the variables that showed a significant difference, only GCG showed lower content in spring tea. The total catechins observed were EGC, C, EC, EGCG, GCG, and ECG. On average, EGCG was the most abundant catechin and accounted for ≈60% of the total catechins in tea samples. Previous studies also showed that summer green tea usually has higher levels of tea polyphenols (TP) and galloylated catechins, whereas spring green tea usually has a higher level of nongalloylated catechins.[27,30] Studies indicate that catechins show bitterness and astringency and that caffeine is a bitter-tasting compound.[4] However, the present study showed that there was no significant difference in caffeine levels between spring green tea and summer green tea, a finding consistent with previous studies.[30] Previous studies also indicated that summer green tea showed appreciable bitterness and astringency, which was due to an increase in the levels of galloylated catechins and a decrease in the level of amino acids.[30] Therefore, amino acids and catechins are the optimal parameters to use to discriminate the production season of green tea. The Shapiro−Wilk test showed that most of the catechins and caffeine did not follow a normal distribution, so conversion to a normal distribution was also necessary for parametric discriminant analysis.

Discriminant Analyses Using the KNN Method. The KNN method is a nonparametric discriminant method. The unknown sample of the prediction set is classified according to the majority of its *K*-nearest neighbors in the training set.[31] The parameter *K* obviously influences the identification accuracy of the KNN model, and the optimum value of *K* is chosen on the basis of the cross-validation test with the lowest error rate.[32] Some original data in the present study did not follow a normal distribution, so the KNN method was employed to discriminate the production season of green tea without normal conversion. The arranging *K* number is often an odd number such as 3, 5, or 7 to avoid a failure in discrimination, so in the present study we arranged the *K* number as 3, 5, 7, 9, and 11. The best accuracy of resubstitution was achieved when the *K* number was 5 or 7, with only four samples misclassified (Table 3). The accuracy of cross-validation was enhanced if the *K* number increased from 3 to 9, whereas the error rate was highest if *K* was 9 (error rate = 6.74%). Hence, the highest accuracy of discriminant analysis was achieved if the *K* number was 7, and the accuracy of resubstitution and cross-validation was 97.61 and 94.80%, respectively.

Table 3. Discriminant Analysis Using *K*-Nearest Neighbor Method

	resubstitution results		cross-validation results	
K	misclassified observations[a]	error count estimates rate (%)	misclassified observations[a]	error count estimates rate (%)
3	14, 33, 56, 64, 94, 100, 111	4.36	12, 14, 31, 33, 52, 56, 62, 63, 64, 94, 100, 108, 111, 116	8.58
5	52, 56, 63, 64	2.25	12, 14, 33, 52, 56, 62, 63, 64, 108, 127	5.90
7	52, 56, 64, 127	2.39	6, 12, 13, 14, 52, 56, 63, 64, 127	5.20
9	6, 12, 13, 14, 16, 19, 23, 52, 56, 62, 63, 64	6.74	6, 12, 13, 14, 23, 52, 56, 64	4.49
11	6, 12, 13, 14, 19, 23, 33, 52, 56, 63, 64	6.18	4, 6,12, 13, 14, 16, 19, 23, 33, 46, 52, 56, 63, 64	7.87

[a]The misclassified observations are identifications of each sample.

Conversion to a Normal Distribution for the Parametric Discriminant Method. Discriminant methods include parametric methods and nonparametric methods. Parametric methods such as LDA require a normal distribution of data (Gaussian distribution). However, many studies neglect the requirement of a normal distribution. If the data do not follow a normal distribution, then there would be no clear linear boundaries for the separation of classes.[13] In the present study, conversion to a normal distribution for the parametric discriminant method was accomplished using Minitab software. It was very important for the conversion to a normal distribution to find a coefficient (λ) that could guarantee the chemical levels of spring tea and summer tea were distributed normally after translation. The suitable intervals of λ for such a conversion are shown in Tables 4 and 5. The median values of the intersections were the final estimated values of λ for the conversion to a normal distribution. Most of the data followed a Gaussian distribution (Shapiro−Wilk test, $p > 0.05$) after conversions (Tables 4 and 5), but translating levels of Gly, Ala, Tyr, Phe, EGC, C, EC, EGCG, GCG, and ECG into a normal distribution for spring tea and summer tea was not possible. This was because there was no intersection of λ for these variables, and then the medians of the intersections for spring tea were chosen as alternatives for the conversions to a normal distribution.

Bayesian Discriminant Analysis. LDA uses linear combinations of data to form discriminant functions (DFs) for the separation of categories by minimization of the within-class and between-class ratios of the sum of squares.[16] In the present study, the Bayesian discriminant method (which is a type of LDA) was used to separate spring tea and summer tea using translated variables and original variables (which were not distributed normally). The forward stepwise analysis (STEPDISC procedure, achieved by SAS software) was used for selecting significant variables for discriminant analysis. There was an obvious difference in the chosen variables using translated variables or original variables (Table 6). The variables Met, Tyr, Thr, Val, Lys, Pro, Phe, Ala, EC, and EGCG were chosen for the discriminant analysis when using the original variables, whereas Met, Tyr, Arg, EGC, caffeine, Ser, Thr, EGCG, GCG, and theanine were used for the

Table 4. Normality Conversion and Normal Distribution Test for Translated Data of Amino Acid Levels[a]

	spring		summer			p value (Shapiro–Wilk)	
parameter	lower CL	upper CL	lower CL	upper CL	estimated λ	spring	summer
Asp	−0.30	0.74	0.14	0.96	0.44	0.5283	0.3769
Ser	−0.65	0.04	−0.28	0.35	−0.12	0.4611	0.2113
Glu	−0.21	0.59	0.01	0.76	0.30	0.9696	0.1196
Gly	0.46	0.73	0.22	0.51	0.49	<0.0001	0.0028
His	0.24	1.00	0.24	0.86	0.55	0.9557	0.5743
Arg	0.10	0.31	0.02	0.54	0.21	0.0732	0.1422
Thr	0.09	0.72	−0.08	0.74	0.41	0.8317	0.5564
Ala	0.04	0.64	0.36	1.02	0.50	0.0067	0.4441
Pro	−0.19	0.51	0.4	1.25	0.46	0.0741	0.0607
theanine	0.32	1.42	0.06	1.03	0.68	0.1514	0.7284
Tyr	0.03	0.66	0.00	0.68	0.35	0.9178	0.0256
Val	0.06	0.45	−0.20	0.61	0.26	0.4238	0.1095
Met	−0.19	0.55	0.14	0.86	0.35	0.7192	0.5171
Lys	−0.12	0.32	−0.14	0.41	0.10	0.9676	0.1666
Ile	−0.07	0.34	0.10	0.63	0.22	0.9183	0.2128
Leu	0.05	0.49	0.22	0.65	0.36	0.8731	0.3225
Phe	−0.18	0.28	0.00	0.52	0.14	0.8689	0.0244

[a]Abbreviations: aspartic acid (Asp), serine (Ser), glutamic acid (Glu), glycine (Gly), histidine (His), arginine (Arg), threonine (Thr), alanine (Ala), proline (Pro), tyrosine (Tyr), valine (Val), methionine (Met), lysine (Lys), isoleucine (Ile), leucine (Leu), phenylalanine (Phe), and cysteine (Cys). The lower CL and upper CL are the estimated bounds of λ for normal distribution conversion. If the p value given by Shapiro–Wilk is <0.05, then the translated data set is not Gaussian distribution.

Table 5. Normality Conversion and Normal Distribution Test for Translated Data of Catechin and Caffeine Levels[a]

	spring		summer			p value (Shapiro–Wilk)	
parameter	lower CL	upper CL	lower CL	upper CL	estimated λ	spring	summer
EGC	−0.62	−0.27	0.72	−0.27	−0.44	0.0156	<0.0001
C	−0.50	−0.03	−1.06	−0.50	−0.70	0.0008	0.0880
EC	−0.90	−0.26	−1.72	−0.68	−0.57	0.0505	<0.0001
EGCG	−0.17	0.38	−1.35	−0.49	0.11	0.0107	<0.0001
GCG	−0.05	0.28	−0.81	−0.32	0.11	0.0295	<0.0001
ECG	0.29	1.08	−1.01	0.04	−0.49	<0.0001	0.0990
caffeine	−2.64	0.13	−0.02	2.32	0.06	0.1957	0.2223

[a]Abbreviations: epigallocatechin (EGC), catechin (C), epicatechin (EC), epigallocatechin gallate (EGCG), gallocatechin gallate (GCG), epicatechin gallate (ECG). The lower CL and upper CL are the estimated bounds of λ for normal distribution conversion. If the p value given by Shapiro–Wilk is <0.05, then the translated data set is not Gaussian distribution.

Table 6. Discriminant Analysis Using Bayes Discriminant Method[a]

				resubstitution		cross-validation	
data type	chosen variables	pooled	function	misclassified observation	error rate (%)	misclassified observation	error rate (%)
without translation	Met, Tyr, Thr, Val, Lys, Pro, Phe, Ala, EC, EGCG	no	quadratic	15, 16, 19, 23, 28, 33, 34, 56, 84, 110, 127, 132	7.17	15, 16, 19, 23, 24, 28, 33, 34, 43, 51, 56, 64, 84, 102, 110, 111, 117, 127, 132, 143, 151	12.94
without translation	Met, Tyr, Thr, Val, Lys, Pro, Phe, Ala, EC, EGCG	yes	linear	11, 16, 23, 33, 110, 111, 120, 127, 137	5.77	11, 16, 23, 33, 110, 111, 120, 127, 128, 137	6.47
translated	Met, Tyr, Arg, EGC, caffeine, Ser, Thr, EGCG, GCG, theanine	no	quadratic	13, 19, 33, 36, 56, 64, 127, 151	4.78	13, 14, 19, 23, 33, 36, 56, 63, 64, 100, 110, 111, 116, 127, 132, 143, 151	10.69
translated	Met, Tyr, Arg, EGC, caffeine, Ser, Thr, EGCG, GCG, theanine	yes	linear	13, 14, 19, 23, 40, 52, 56, 64, 111, 127	5.90	13, 14, 19, 23, 40, 52, 56, 63, 64, 111, 127	6.46

[a]There is an option in the DISCRIM procedure in SAS software that determines whether the pooled or within-group covariance matrix is the basis of the measure of the squared distance. If POOL=YES is specified, the procedure uses the pooled covariance matrix in calculating the (generalized) squared distances. Linear discriminant functions are computed. If POOL=NO is specified, the procedure uses the individual within-group covariance matrices in calculating the distances. Quadratic discriminant functions are computed.

translated variables. The lowest value for misclassification (4.78%) was achieved in resubstitution tests using translated variables and unpooled covariance matrices. The lowest value for misclassification (6.46%) was achieved in cross-validation tests using translated variables and pooled covariance matrices. The results showed that conversion to a normal distribution

could increase the accuracy of discriminant analysis when Bayesian discriminant analysis was used.

In the present study, a method of chemical analysis in combination with pattern recognition methods (KNN and LDA) was introduced successfully to discriminate between two types of green tea: spring green tea and summer green tea. The results showed that levels of amino acids, catechins, and caffeine were suitable parameters for discriminating between the production seasons of Chinese green tea. The accuracy of the KNN method was ≤97.61 and ≤94.80% as validated by resubstitution and cross-validation tests, respectively. LDA is a parametric method, so perhaps it would be unnecessary to apply it in situations when the variables do not follow a Gaussian distribution.[13] We found that there were significant differences in the chosen variables and accuracy of discriminant functions between the original data and translated data (Gaussian distribution). Compared with LDA, the KNN method did not require a Gaussian distribution and was more accurate, so the KNN method is recommended to discriminate between the production seasons of Chinese green tea.

AUTHOR INFORMATION

Corresponding Author

*Phone: 86-551-5786401. Fax: 86-551-5786765. E-mail: xcwan@ahau.edu.cn.

Author Contributions

[‖]These authors contributed equally to this work and should be considered co-first authors.

Funding

This project was supported by an earmarked fund for modern agro-industry (tea) research system of the Ministry of Agriculture, People's Republic of China.

Notes

The authors declare no competing financial interest.

REFERENCES

(1) Chacko, S. M.; Thambi, P. T.; Kuttan, R.; Nishigaki, I. Beneficial effects of green tea: a literature review. *Chin. Med.* **2010**, *5*, 13.

(2) Zhaopeng, S.; Zhonghua, L. Probe into mathematical model of chemical essence of bitterness and astringency in summer green tea. *J. Tea Sci.* **1987**, *27*−12.

(3) Chen, Y.; Jiang, Y.; Duan, J.; Shi, J.; Xue, S.; Kakuda, Y. Variation in catechin contents in relation to quality of 'Huang Zhi Xiang' Oolong tea (*Camellia sinensis*) at various growing altitudes and seasons. *Food Chem.* **2010**, *119*, 648−652.

(4) Yu, H.; Wang, J.; Zhang, H.; Yu, Y.; Yao, C. Identification of green tea grade using different feature of response signal from E-nose sensors. *Sens. Actuators, B* **2008**, *128*, 455−461.

(5) Chen, Q.; Zhao, J.; Vittayapadung, S. Identification of the green tea grade level using electronic tongue and pattern recognition. *Food Res. Int.* **2008**, *41*, 500−504.

(6) Scharbert, S.; Hofmann, T. Molecular definition of black tea taste by means of quantitative studies, taste reconstitution, and omission experiments. *J. Agric. Food Chem.* **2005**, *53*, 5377−5384.

(7) Valera, P.; Pablos, F.; Gustavo González, A. Classification of tea samples by their chemical composition using discriminant analysis. *Talanta* **1996**, *43*, 415−419.

(8) Liang, Y. R.; Lu, J. L.; Zhang, L. Y.; Wu, S.; Wu, Y. Estimation of black tea quality by analysis of chemical composition and colour difference of tea infusions. *Food Chem.* **2003**, *80*, 283−290.

(9) Wright, L. P.; Mphangwe, N. I. K.; Nyirenda, H. E.; Apostolides, Z. Analysis of caffeine and flavan-3-ol composition in the fresh leaf of *Camellia sinesis* for predicting the quality of the black tea produced in Central and Southern Africa. *J. Sci. Food Agric.* **2000**, *80*, 1823−1830.

(10) Togari, N.; Kobayashi, A.; Aishima, T. Pattern recognition applied to gas chromatographic profiles of volatile components in three tea categories. *Food Res. Int.* **1995**, *28*, 495−502.

(11) Dutta, R.; Hines, E. L.; Gardner, J. W.; Kashwan, K. R.; Bhuyan, A. Tea quality prediction using a tin oxide-based electronic nose: an artificial intelligence approach. *Sensors Actuators B−Chem.* **2003**, *94*, 228−237.

(12) Chen, Q.; Guo, Z.; Zhao, J. Identification of green tea's (*Camellia sinensis* (L.)) quality level according to measurement of main catechins and caffeine contents by HPLC and support vector classification pattern recognition. *J. Pharm. Biomed. Anal.* **2008**, *48*, 1321−1325.

(13) Herrador, M. Á.; González, A. G. Pattern recognition procedures for differentiation of green, black and oolong teas according to their metal content from inductively coupled plasma atomic emission spectrometry. *Talanta* **2001**, *53*, 1249−1257.

(14) Fernández, P. L.; Pablos, F.; Martín, M. J.; González, A. G. Multi-element analysis of tea beverages by inductively coupled plasma atomic emission spectrometry. *Food Chem.* **2002**, *76*, 483−489.

(15) Moreda-Piñeiro, A.; Fisher, A.; Hill, S. J. The classification of tea according to region of origin using pattern recognition techniques and trace metal data. *J. Food Compos. Anal.* **2003**, *16*, 195−211.

(16) McKenzie, J. S.; Jurado, J. M.; de Pablos, F. Characterisation of tea leaves according to their total mineral content by means of probabilistic neural networks. *Food Chem.* **2010**, *123*, 859−864.

(17) Chen, Q.; Zhao, J.; Zhang, H.; Wang, X. Feasibility study on qualitative and quantitative analysis in tea by near infrared spectroscopy with multivariate calibration. *Anal. Chim. Acta* **2006**, *572*, 77−84.

(18) Zhao, J.; Chen, Q.; Huang, X.; Fang, C. H. Qualitative identification of tea categories by near infrared spectroscopy and support vector machine. *J. Pharm. Biomed. Anal.* **2006**, *41*, 1198−1204.

(19) Chen, Q.; Zhao, J.; Lin, H. Study on discrimination of roast green tea (*Camellia sinensis* L.) according to geographical origin by FT-NIR spectroscopy and supervised pattern recognition. *Spectrochim. Acta Part A: Mol. Biomol. Spectrosc.* **2009**, *72*, 845−850.

(20) Wu, D.; Yang, H.; Chen, X.; He, Y.; Li, X. Application of image texture for the sorting of tea categories using multi-spectral imaging technique and support vector machine. *J. Food Eng.* **2008**, *88*, 474−483.

(21) Scharbert, S.; Holzmann, N.; Hofmann, T. Identification of the astringent taste compounds in black tea infusions by combining instrumental analysis and human bioresponse. *J. Agric. Food Chem.* **2004**, *52*, 3498−3508.

(22) Kaneko, S.; Kumazawa, K.; Masuda, H.; Henze, A.; Hofmann, T. Molecular and sensory studies on the umami taste of Japanese green tea. *J. Agric. Food Chem.* **2006**, *54*, 2688−2694.

(23) BSI. *Determination of Substances Characteristic of Green and Black Tea. Content of Total Polyphenols in Tea. Colorimetric Method Using Folin-Ciocalteu Reagent*; British Standards Institution: London, U.K., 2005; Vol. ISO 14502-1.

(24) SAC. *Tea − Determination of Caffeine Content*; China Standards Institution: Beijing, China, 2002; Vol. GB/T 8312.

(25) Callejon, R. M.; Tesfaye, W.; Torija, M. J.; Mas, A.; Troncoso, A. M.; Morales, M. L. HPLC determination of amino acids with AQC derivatization in vinegars along submerged and surface acetifications and its relation to the microbiota. *Eur. Food Res. Technol.* **2008**, *227*, 93−102.

(26) Paramas, A. M. G.; Barez, J. A. G.; Marcos, C. C.; Garcia-Villanova, R. J.; Sanchez, J. S. HPLC-fluorimetric method for analysis of amino acids in products of the hive (honey and bee-pollen). *Food Chem.* **2006**, *95*, 148−156.

(27) Ye, Y.; Chen, X.; Tang, D.; Su, L.; Yin, J. Comparative study of taste compounds levels of yuelu green tea produced in different season. *J. Zhejiang Agric. Sci.* **2008**, *6*, 705−706.

(28) Fang, S.; Zhang, X.; Xia, T.; Wan, X. Influence of tea cultivars, processing technology and seasons on quality of oolong tea. *J. Tea Sci.* **2002**, *22*, 135−139.

(29) Wang, X.; Chunlei, M.; Yao, M.; Jin, J.; Yang, Y. Biochemical components affected the seasonal differences of green tea quality. *Acta Bot. Boreal.—Occident. Sin.* **2011**, *31*, 1229—1237.
(30) Shi, Z.; Chen, B.; Zeng, Q.; Sun, H. Study of the chemicals yielding bitterness and astringency of summer green tea. *J. Tea Sci.* **1984**, 61—62.
(31) Derde, M. P.; Buydens, L.; Guns, C.; Massart, D. L.; Hopke, P. K. Comparison of rule-building expert systems with pattern-recognition for the classification of analytical data. *Anal. Chem.* **1987**, *59*, 1868—1871.
(32) Alsberg, B. K.; Goodacre, R.; Rowland, J. J.; Kell, D. B. Classification of pyrolysis mass spectra by fuzzy multivariate rule induction-comparison with regression, *K*-nearest neighbour, neural and decision-tree methods. *Anal. Chim. Acta* **1997**, *348*, 389—407.