Published in final edited form as:
Appl Spectrosc. 2011 March ; 65(3): 250–259. doi:10.1366/10-06109.

A Comparison of Analytical and Data Preprocessing Methods for Spectral Fingerprinting

DEVANAND L. LUTHRIA[*], **SUDARSAN MUKHOPADHYAY**, **LONG-ZE LIN**, and **JAMES M. HARNLY**
Food Composition and Methods Development Laboratory, Beltsville Human Nutrition Research Center, Agricultural Research Service, U.S. Department of Agriculture, Beltsville, Maryland 20705-3000

Abstract

Spectral fingerprinting, as a method of discriminating between plant cultivars and growing treatments for a common set of broccoli samples, was compared for six analytical instruments. Spectra were acquired for finely powdered solid samples using Fourier transform infrared (FT-IR) and Fourier transform near-infrared (NIR) spectrometry. Spectra were also acquired for unfractionated aqueous methanol extracts of the powders using molecular absorption in the ultraviolet (UV) and visible (VIS) regions and mass spectrometry with negative (MS−) and positive (MS+) ionization. The spectra were analyzed using nested one-way analysis of variance (ANOVA) and principal component analysis (PCA) to statistically evaluate the quality of discrimination. All six methods showed statistically significant differences between the cultivars and treatments. The significance of the statistical tests was improved by the judicious selection of spectral regions (IR and NIR), masses (MS+ and MS−), and derivatives (IR, NIR, UV, and VIS).

Index Headings

Spectral fingerprinting; Near-infrared spectroscopy; NIR spectroscopy; Ultraviolet–visible spectroscopy; UV-Vis spectroscopy; Direct mass spectrometry; Classification; Discrimination; Broccoli; Growing conditions; Analysis of variance; ANOVA; Principal component analysis; PCA; ANOVA-PCA

INTRODUCTION

Spectral fingerprinting is a rapid and analytically simple method for characterizing and comparing plant tissues.[1–4] Spectra of plant materials can be acquired from finely powdered solid samples using infrared (IR) or near-infrared (NIR) spectrometry or from unfractionated extracts of the solids using mass spectrometry (MS) or molecular absorption at ultraviolet (UV) or visible (VIS) wavelengths. Analysis of the solids or the extracts directly provides complex overlapping spectra that are integrated representations of the chemical composition of the plant materials. Meaningful results are derived from these complex spectra using multivariate analysis methods and/or classical statistical methods such as analysis of variance (ANOVA).

Principal component analysis (PCA) is probably the most commonly used of the many multivariate analysis methods.[5] It provides unsupervised pattern recognition, reduced

[*]Author to whom correspondence should be sent: D.Luthria@ars.usda.gov.

dimensionality, and easy visualization of the data. Harrington et al.[6,7] recently described a method called analysis of variance PCA (ANOVA-PCA), which uses a classical ANOVA approach to separate the original data matrix into additive matrices that characterize single factors of the experimental design and the residuals, as illustrated in Fig. 1. PCA of the sum of individual factor matrices added to appropriate residuals (either the limiting biological or analytical variance) provides clusters that, if compositionally different, will be distinctly separated based on the first principal component (horizontal axis). The difference between the clusters is easily observed visually and easily evaluated statistically using the Students' t-test.

Harnly's group[8–10] showed that the ANOVA matrices generated by Harrington's method[6,7] can be used to compute the relative variance of each experimental factor. This is analogous to classic ANOVA except that a full spectrum is substituted for each individual data point. The relative variance is computed by squaring and summing the appropriate ANOVA matrices (Fig. 1 and Table I). Depending on the experimental design, this approach can provide a nested[8,9] or a crossed multi-dimensional ANOVA[10] that permits the significance of each experimental parameter to be evaluated using an F-test.[10] In addition, the determination of the relative variance attributable to the analytical uncertainty is a valuable tool for the optimization of analytical parameters.

In two recent studies,[8,9] ANOVA-PCA was used to analyze spectra for methanol–water extracts of broccoli acquired using UV and positive and negative ionization MS (MS+ and MS−, respectively). The broccoli samples consisted of two cultivars grown under seven different conditions (four levels of Se fertilization, organically, and conventionally with full and 80% irrigation).[11,12] All three methods (UV, MS−, and MS+) were able to discriminate between the two cultivars and the seven growing treatments. No attempt was made to compare the quality of discrimination for the three methods. To the best of the authors' knowledge, there has been no direct comparison of spectral fingerprinting methods using common samples. In addition, no attempt has been made to statistically analyze the quality of the discrimination between clusters.

In the current study, spectra from six different sources (IR, NIR, UV, VIS, MS−, and MS+) are compared for their ability to discriminate between cultivars and growing treatments of a common set of broccoli samples.[8,9] Newly acquired IR, NIR, and VIS data are compared with the previously reported data for UV, MS−, and MS+.[8,9] The quality of discrimination obtained using ANOVA-PCA was statistically evaluated by applying a t-test to the separation of the PCA clusters and applying an F-test to the separated ANOVA matrices. In addition, the effect of preprocessing parameters for normalization, derivatization, and signal reproducibility was also evaluated.

EXPERIMENTAL

Plant Material

Samples were freeze-dried and powdered composites of two varieties of broccoli (*Brassica oleracea*): Majestic provided by Dr. John W. Finley (ARS, USDA) and Legacy provided by Dr. Gary Banuelos (ARS, USDA). The cv Majestic broccoli was grown in a greenhouse with four different concentrations of sodium selenate as previously described in detail.[12] These treatments resulted in 0.4, 5.7, 98.6, and 879.2 μg/g of selenium (dry weight) in the broccoli florets. In the rest of the text, the samples from the four selenium (Se) treatments are referred to as 0ppm, 5ppm, 100ppm, and 1000ppm, respectively. The cv Legacy broccoli was obtained from field studies from two different 4 ha field sites in central California (Harris Farms, Five Points, CA) as previously described in detail.[11] Conventionally and organically grown broccoli was collected at both sites. Conventionally grown broccoli was irrigated at two levels: 100% and 80% of the evapotranspiration (Eto) rate reported by the Westlands California Irrigation Management Information System weather station. Organically grown broccoli was produced using a single level of irrigation at 100% Eto rate. In the rest of the text, these three treatments are referred to as C100, C80, and Org, respectively.

Broccoli plants were harvested at each site and samples for each growing condition were processed separately. Samples from field crops were collected for at least four growing seasons.[11,12] Whole plants were separated into leaf, stems, and florets. Broccoli florets were then freeze-dried, later coarsely ground in the food processors, and composited. Ground samples were kept below −20 °C. Prior to analysis or extraction, samples were sieved through standard 20 mesh sieves (particle size < 0.850 mm) to obtain uniform homogenized particle size samples.

Chemicals

High-performance liquid chromatography (HPLC)-grade MeOH was purchased from Fisher Chemicals (Fair Lawn, NJ). HPLC-grade acetone was purchased from Burdick & Jackson (Muskegon, MI). Deionized water (18.2 MΩ·cm) was obtained in-house using a Nanopure Diamond analytical ultrapure water purification system (Model D11901, Branstead International, Dubuque, IA). Polyvinylidene difluoride (PVDF) syringe filters with pore size 0.45 μm were procured from National Scientific Company (Duluth, GA).

Extractions

The extraction process was described in detail previously.[8,9] Briefly, 1 g of weighed freeze-dried and powdered broccoli samples were extracted three times with 5, 2.5, and 2.5 mL of MeOH:H_2O (60 : 40, % v/v). The supernates were combined and brought to a final volume of 10 mL with MeOH :H_2O (60 : 40, % v/v). All extracts were split and stored in 2 mL HPLC vials under nitrogen at −70 °C until analyzed. Aliquots of the extracts were filtered prior to UV and MS analysis.

Sample Analysis

Powdered samples for each of the seven treatments (0ppm, 5ppm, 100ppm, 1000ppm, C100, C80, and Org) were extracted four times for UV and VIS measurements and five times for MS measurements. For NIR measurements, five separate powdered samples were prepared and analyzed for each growing condition and, for IR measurements, three separate samples were analyzed.

Infrared Instrumentation

Fourier transformed infrared (FT-IR) spectra were acquired on solid freeze-dried broccoli samples (particle size < 0.85 mm). The FT-IR spectral scans were obtained with a Perkin Elmer Spectrum One spectrometer using a Universal attenuated total reflection (ATR) accessory (Boston, MA). Spectral scans were recorded between 4000 and 650 cm^{-1} at 4 cm^{-1} resolution. Spectra of the clean and dry ATR crystal against air were used for background. Between determinations, the crystal was carefully cleaned with Kim-wipes. The spectrometer's optics was sealed from the atmosphere but its compartment was not purged during measurement. The spectra were converted to ASCII format and transferred to an Excel file for statistical analysis. Sample analyses were repeated three times to produce a data matrix with 21 rows (7 samples × 3 repeats) and 3351 columns (wavenumber).

Near-Infrared Instrumentation

Freeze-dried broccoli samples were scanned with a near-infrared instrument (Thermo-Electron, Waltham, MA) fitted with an InGaAs detector and a CaF_2 beam splitter. Spectral scans were collected between the regions 4000 and 10000 cm^{-1} with 16 cm^{-1} resolution. Between determinations, the tubes were shaken well and packed by tapping on a soft pad 3 to 5 times. Sample analyses were repeated five times to produce a data matrix with 35 rows (7 samples × 5 repeats) and 3112 columns (wavenumber). The spectra were converted to ASCII format and transferred to a Excel file for statistical analysis.

Ultraviolet–Visible Instrumentation

UV and VIS spectra of the broccoli extracts were acquired as previously described. [8,11] All data were recorded on a Lambda 25 spectrophotometer (Perkin Elmer, Boston, MA). For VIS fingerprints (400–700 nm, acquired at 1 nm intervals), the extracts were analyzed directly. For UV spectra (220–400 nm, acquired at 1 nm intervals), the extracts were diluted by a factor of 50 due to strong absorbance in this region. For the UV spectra, sample analyses were repeated ten times to produce a data matrix with 70 rows (7 samples × 10 repeats) and 181 columns (wavelengths). For the VIS spectra, sample analyses were repeated ten times to produce a data matrix with 70 rows (7 samples × 10 repeats (5 separate extractions and each analyzed twice)) and 301 columns (wavelengths).

Mass Spectrometry Instrumentation

The instrumentation was described in detail previously.[9,11] Briefly, data were acquired with an Agilent 1100 HPLC (Agilent, Palo Alto, CA) coupled with a diode array detector (DAD) and mass spectrometer detector (MSD, SL mode). The MSD (SL) used electrospray ionization (ESI) and was programmed to acquire data in the positive (MS+) and negative (MS−) ionization modes, in rapid sequence, at a low (100 V) fragmentation voltage. The samples were injected directly into the ionizer with no column at 1 mL/min using an infusion pump. The MSD was programmed to scan masses from *m/z* 50 to 2000 with 1 amu resolution over a period of 10 min. For both ionization modes, analyses were repeated five times. The MS− data matrix consisted of 35 rows (7 samples × 5 repeats) and 99 columns (ions). The MS+ data matrix consisted of 35 rows (7 samples × 5 repeats) and 167 columns (ions).

Preprocessing and Analysis of Variance

The data were processed as previously described[8–10] based on the method of Harrington et al.,[6] which is outlined in Fig. 1. Spectra from each analytical instrument were exported to Excel (Microsoft, Inc., Belleview, WA) for organizing, preprocessing, and ANOVA processing. PCA was performed using either Pirouette 3.1 (Infometrix, Inc., Bothell, WA) or Solo (Eigenvector Research, Inc., Wenatchee, WA). In Fig. 1, each outlined block (M1–M9 and M5*–M9*) represents a data matrix. In matrices M5*–M9*, the data in each cell of the corresponding matrix (M5–M9) have been squared. The sum of values for every cell in the squared matrices is denoted by Σ and is a single numerical value.

In Excel, the imported spectra for each analytical instrument were organized into a matrix (M1 in Fig. 1) consisting of rows of spectra and columns of variables (wavelengths, wavenumber, or masses). In Fig. 1, the preprocessed data matrix (M2), represents a series of matrices that were different for the analytical instruments. For IR, NIR, UV, and VIS, first and second derivatives were calculated using the coefficients of Savitzky and Golay.[13] The order of the polynomial and the number of points used for the derivatives were systematically varied and evaluated. For MS− and MS+, the data were filtered based on signal reproducibility. The first filter (1/5 criteria) incorporated all the masses; i.e., if counts for a specific mass exceeded the detection threshold (1% of the highest mass count) for only one of the five repeat analyses, the mass was included in the calculations. The second filter (5/5 criteria) incorporated masses only if counts for all five repeats of a sample exceeded the threshold. If only four repeats had a measured intensity (the fifth sample was less than the threshold), then the intensities for all five repeats were listed as 0. The MS data were not derivatized.

The preprocessed spectra were then normalized to a unit vector (M3); i.e., the sums of the squares of the points in a spectrum (row) were set equal to 1. The grand means (M4) were computed as the average of all the data for each variable (column). The grand means residuals (M5) were obtained by subtracting M4 from M3. The transition from M3 to M5 is usually called mean centering. The squares of the grand means residuals (M5*) were summed to provide the total variance of the data set. The grand means residuals (M5) were used to compute the cultivars means matrix (M6). Each column was populated by two values, the means for each of the cultivars. The cultivar residuals matrix was used to compute the cultivar means residuals matrix (M7 = M5 − M6). The squares of the data (M7* and M8*) were summed to provide the variance between and within the cultivars, respectively. In a similar manner, the cultivar residuals (M7) were used to compute the treatment means (M8) and the treatment residuals (M9 = M7− M8). The squares of the data (M8* and M9*) were summed to provide the variance between and within the treatments, respectively.

Principal Component Analysis

PCA was used to discriminate between cultivar and treatment. For cultivar, the between cultivar matrix (M6) was added to the analytical uncertainty matrix (M9) and then analyzed by PCA. As shown in Fig. 2 for NIR, the cv Legacy and cv Majestic are well separated based on the first principal component (PC1 the horizontal axis). Similarly, for treatment, the between-treatment matrix (M8) was added to the analytical uncertainty matrix (M9) prior to analysis by PCA. The treatment data were analyzed in two different ways. PCA was applied to all the data and provided seven clusters best observed in three-dimensional space (PC1 versus PC2 versus PC3) as shown in Fig. 3. PCA was also applied to pairs of treatments and provided two clusters separated on the horizontal axis as shown in Fig. 2 for conventionally and organically grown cv Legacy.

Statistical Calculations

The ability of the six analytical instruments to discriminate between cultivars and treatments was evaluated using a t-test of the PCA data (Fig. 2) and an F-test of the ANOVA matrices (Fig. 1).

Principal Component Analysis t-Test—Figure 2 shows a typical result for PCA of NIR data for cultivars (filled symbols) and a comparison of one pair of treatments, organic versus conventional (open symbols). The data have been arbitrarily scaled so they can be shown in the same plot. In each case, a mean and standard deviation is computed for each cluster. The differences in the means, the combined standard deviations of the mean, and the t-value are computed in the conventional manner. For Fig. 2, the t-value for the cultivar and the treatment data are 125 and 44, respectively. The separation, as is readily visible, is statistically, highly significant.

Variance F-Test—Table I summarizes the F-value calculations using the variances computed in Fig. 1. As stated previously, the seven composite samples represented two cultivars, one with four levels of Se and one treated grown organically and conventionally at two levels of irrigation. Since each cultivar is treated differently, a nested ANOVA was employed, as shown in Fig. 1. The primary difference between data from the six analytical instruments was the number of repeat analyses (Table I). Because repeat preparations and composite samples were analyzed, it was anticipated that variance would arise from cultivar, treatment, analytical uncertainty (sample preparation and instrumental error), and random error.

In Fig. 1, the between-cultivar variance (SS_{bC}) describes the variance due to the two different cultivars and the within-cultivar variance (SS_{wC}) describes the variance arising from treatment, analytical uncertainty, and random error. The between-treatment variance (SS_{bT}) describes the variance between treatments and the within-treatment variance (SS_{wT}) describes the variance due to analytical uncertainty and random error. This order can be reversed (treatment and then cultivar) without affecting the computed variances. Because each cultivar is treated differently, there is no cross-variance component between cultivar and treatment.

Table I shows that the F-value for cultivar is computed as the ratio of the-between cultivar variance (SS_{bC}) and the analytical uncertainty (SS_{wT}). Similarly, the F-value for treatment is computed as the ratio of the between-treatment variance (SS_{bT}) and the analytical uncertainty (SS_{wT}). In these calculations, each spectrum is treated as an individual data point. Thus, for UV and VIS measurements, the degree of freedom (df) associated with the between-cultivar variance (SS_{bC}) is 1 (the number of cultivars minus 1) and the df for analytical uncertainty (SS_{wT}) is 54 (the number of spectra minus the number of cultivars, or $7 \times 8 - 2$). The df associated with between-treatment variance (SS_{bT}) is 6 (the number of treatments minus 1) and the df for analytical uncertainty (SS_{bT}) is 49 (the number of spectra minus the number of treatments, or $7 \times 8 - 7$).

RESULTS AND DISCUSSION

Six sources of spectra were compared in this study; NIR, IR, VIS, MS−, MS+, and UV. Results for NIR, IR, and VIS are reported for the first time and are treated in detail. Results were previously reported for MS−, MS+, and UV with an emphasis on their ability to discriminate between cultivar and treatment. New results are presented here with an emphasis on the quality of the discrimination and the effect of preprocessing parameters.

Near-Infrared Analysis

NIR spectra were collected between 10000 and 4000 cm^{-1} (1000–2500 nm) at 2 cm^{-1} increments. Five separate subsamples of each of the finely powdered composites were analyzed. A typical spectrum, after transformation to the first derivative using a quadratic polynomial fit to seven points (1-Q-7) and normalized to a unit vector (Fig. 1, M3), is shown in Fig. 4A. The traces in Figs. 4B through 4D show the relative variance of the means for cultivars and treatments and the relative variance associated with analytical uncertainty (from M6*, M8*, and M9* in Fig. 1), as a function of wavenumber, calculated with ANOVA. The summed results (for the whole spectrum) for relative variance of cultivar, treatment, and analytical uncertainty are presented in Table II, row 1.

Spectral Band Selection—The NIR spectra for the broccoli samples consisted of three regions (Region I, 7300–10000; Region II, 5400–6600; and Region III, 4190–4950) separated by two strong water bands (6600–7300 cm^{-1} and 4950–5400 cm^{-1}) as shown in Fig. 4A. The data for the full spectra in Table II, row 1, show that approximately 27% of the relative sample variance came from the difference between cultivars, 60% from the difference between treatments, and 13% from the analytical uncertainty. The computed F-values were very high, indicating that the probability that either the cultivar means or the treatment means were similar was very low ($p < 0.0001$). For the seven treatments, the high

F-value doesn't provide information on the comparison of specific pairs. Such a comparison would require a calculation using only the spectra from the two treatments of interest. The traces in Figs. 4B and 4C show that the water bands contributed strongly to both the cultivar and treatment variance.

Research has shown that moisture content can have a strong influence on the NIR spectra.[14,15] In the current study, samples were lyophilized and powdered, but the moisture content was not determined afterwards. Each of the five repeat analyses for each treatment came from a single, composite sample. The strong water bands in Fig. 4A and the strong variance associated with the water bands (Figs. 4B and 4C) indicate that a contribution from the moisture content of the composite samples towards the differentiation of the cultivars and treatments cannot be ruled out.

Table II, row 2, shows the results for ANOVA after removal of the water bands. These modified data now show that approximately 13% of the relative sample variance came from the two cultivars, 69% from the treatment, and 18% from the analytical uncertainty. Removal of the water bands decreased the relative variance contributed by the cultivars by a factor of two. The computed F-values were lower than those obtained with the water bands present but still show a low probability ($p < 0.0001$) that either the cultivar means or the treatment means were similar.

Analysis of variance of the three spectral regions (Table II, rows 3–5) separated by the water bands showed that the information content of each region was significantly different. All three regions still showed that the probability of the cultivar means and treatment means being the same was low (<0.0001 to 0.12), but the F-values for Regions I and II were significantly lower (and probabilities higher) than the value for Region III. The F-values for Region III were almost the same as that for the full spectrum with water bands. Perhaps the most revealing data are the relative variances for analytical uncertainty: 65% for Region I, 34% for Region II, and only 11% for Region III (Table II, rows 3, 4, and 5, respectively).

Preprocessing Optimization—Computation of the first or second derivative of NIR, IR, UV, and VIS spectra is commonly used to remove dc offset between spectra.[15] The increased noise associated with the derivative calculation is reduced by using an increased number of data points in the calculation. First and second derivatives with smoothing can be easily computed using higher-order polynomials for up to 25 data points using the published coefficients of Savitzky and Golay.[13] All the data discussed to this point were processed using a 7-point first-derivative smooth with a quadratic polynomial (1-Q-7).

Table II lists the F-values computed for Region III for a quadratic first derivative using from 7 (1-Q-7) to 25 (1-Q-25) points (rows 5–11); for a cubic/quartic first derivative using 25 points (1-C/Q-25) (row 12); and for a cubic/quartic and quartic/quintic second derivative using 25 points (2-C/Q-25 and 2-Q/Q-25, respectively) (rows 13 and 14). As the number of points in the calculation increased from 7 to 25 (rows 5–11), the percent variance due to analytical uncertainty systematically decreased while the F-values increased for both cultivar and treatment. In every case, the probability was <0.0001 that the mean spectra were equal. Use of higher polynomial (1-C/Q-25) resulted in a higher relative analytical variance and poorer F-values. This trend continued with 2-Q/C-25 and 2-Q/Q-25. The second derivative and higher-order polynomials resulted in increased analytical uncertainty, which was not offset by inclusion of more data points in the calculation.

The relative variances attributable to cultivar and treatment were fairly constant for 1-Q-7 through 1-Q-25 derivatives (Table II, rows 5–11). Consequently, combined with the steady decrease in the analytical uncertainty, the F-values for cultivars and treatments for 25 points (row 11) were greater by factors of 2 and 1.5, respectively, compared to 7 points (row 5). The large increase in analytical uncertainty for the second derivatives (2-Q/C-25 and 2-Q/Q-25) resulted in decreases in the variance attributable to both cultivar and treatment. Use of more than 25 points for the derivative calculations might provide a decrease in the analytical variance to achieve equivalence with the first derivatives. However, in this study 25 points were the maximum used.

Principal Component Analysis—The data were analyzed and processed as described in the Experimental section and Fig. 1. The computed matrices for cultivar (Fig. 1, M6 + M9) and treatment (Fig. 1, M8 + M9) were subjected to PCA. Figure 2 presents the overlaid PCA scores for comparison of cultivars (cv Legacy and cv Majestic) (filled symbols) and one pair of treatments, conventionally (C100) and organically grown (Org) cv Legacy broccoli (open symbols). The scores for C100 versus Org are typical of the scores seen for all the treatment comparisons. It can be seen that in each case, the clusters are well separated, indicating that there is a statistically distinguishable difference in the spectra and hence the chemical composition of the samples.

The separation of the cultivars or pairs of treatments shown in Fig. 2 can be quantified using the Students' t-test based on the horizontal differences of the means and standard deviations of the clusters. Results for the comparison of the cultivars and one pair of treatments (C100 versus Org) are presented in Table II. Similar results were observed for every pair of treatments (data not shown).

The results obtained from t-tests of the PCA score plots were consistent with those obtained for the nested one-way ANOVA F-tests. In Table II, higher t-values were obtained for Region III as compared to the whole spectrum or the spectrum minus the water bands. The t-values increased with more points in the derivative calculation (rows 5–11) and decreased with use of higher-order polynomials and a second derivative (rows 12–14). This agreement was not unexpected since either the t-test or the F-test can be used to compare the means of two sample populations with the t-test restricted to two populations. In this study, however, the sample data subjected to the F-test were transformed by the PCA algorithm prior to performing the t-test.

Principal component analysis of all seven treatments together, rather than in pairs, produces the scores shown in Fig. 3A. The first three PCs are shown (at an optimal rotation) to clearly show the separation between the clusters for all seven treatments. It can be seen how treatment of the individual pairs leads to the scores shown in Fig. 2. The scores in Fig. 3A were based on preprocessing with a 1-Q-25 derivative. The score plots for all seven treatments make it easier to compare the separation obtained for the different analytical methods.

Infrared Analysis

Infrared data were collected from 4000 to 660 cm^{-1} at 1 cm^{-1} increments. Separately prepared solid samples were analyzed in triplicate. A typical sample spectrum, based on a 1-Q-11 derivative and normalized to a unit vector, is shown in Fig. 5A. The variance between cultivar and treatment means and for analytical uncertainty (Fig. 1, M6*, M8*, and M9*) as a function of wavenumber are shown in Figs. 5B through 5D.

Spectral Band Selection—The spectrum consists of three regions: a region of relatively broad absorption bands (Region I, 2800–4000 cm^{-1}), a region of little information (Region II, 1800–2800 cm^{-1}), and a region of narrow absorption bands (Region III, 800 to 1800 cm^{-1}, commonly known as the fingerprint region). Results for nested one-way ANOVA of the whole spectrum and for each region using a 1-Q-11 derivative are shown in Table II (rows 21–24). Many of the trends observed for the NIR data are also seen with the IR data. For each spectral region, the method provided discrimination between cultivar and treatment means at a statistically significant level and one spectral region (Region III) gave improved statistical results, lower analytical uncertainty, and larger F-values, compared to the full spectrum.

Preprocessing Optimization—As for NIR, the use of more data points for a quadratic first derivative (from 1-Q-7 to 1-Q-25) reduced the analytical uncertainty and improved the F-values (Table II, rows 24 and 25). Use of a higher-order polynomial with a first or second derivative (rows 26–28) led to increased analytical uncertainty and reduced F-values. The increased analytical uncertainty led to reduced relative variance attributable to both cultivar and treatment. These results are similar to those observed for NIR. However, the relative variance attributable to analytical uncertainty was consistently greater for IR.

Principal Component Analysis—PCA of the cultivar matrix and paired comparisons of treatments (not shown) provided excellent separation of the clusters on the horizontal axis. Figure 3B shows the PCA score plot for the first three components for all seven treatments in the treatment matrix (Fig. 1, M8 + M9) using a 1-Q-25 derivative. Seven clusters are clearly distinguished. The t-values from the PCA plots for Region III (Table II, row 24) were better than those for the whole spectrum (row 21). However, the t-values for the 1-Q-25 derivative were the same or slightly less than those for 1-Q-7. The cultivar t-values decreased with a second derivative. The t-values for the treatment pairs were less predictable. There was no observable pattern with the use of more points, a second derivative, or higher-order polynomials.

Visible Analysis

Molecular absorption was measured in the visible range, 400–700 nm, at 1 nm increments. The sample extracts were repeated ten times without dilution, as opposed to the 25× to 50× dilution factors used for UV. The VIS spectra (not shown) were broad and lacked any distinguishing structure. Consequently, the entire spectra were used. Because of the excellent reproducibility of the spectra (analytical uncertainty ≤2%), subtle differences in the broad spectral features provided F-values that were significantly greater than those computed for IR or NIR (Table II).

Preprocessing Optimization—The analytical uncertainty and the resultant F-values were relatively unaffected by the number of points used in the quadratic first derivative (Table II, rows 31 and 32). Use of 1-C/Q-25, 2-Q/C-25, and 2-Q/Q-25 derivatives (rows 33–35) produced little change in the analytical uncertainty, decreases in the F-values for cultivars, and increases in the F-values for treatments. However, the higher-order polynomials and derivatives resulted in a decreased relative variance for cultivar and an increased relative variance for treatment. These patterns were inconsistent with the results previously observed for IR and NIR measurements.

Principal Component Analysis—As for NIR and IR, excellent separation was achieved for the cultivars and individual pairs of treatments using ANOVA-PCA. The PCA score plot for all treatments (M8 + M9) is shown in Fig. 3D. The t-values for the PCA score plots for cultivars and one treatment pair (C100 versus Org) agreed well with the pattern of changes observed for the F-values.

Ultraviolet Analysis

We previously reported the use of ANOVA-PCA for the analysis of UV spectra (220–400 nm) for the same broccoli samples[8] analyzed in the current study. Sample analyses were repeated ten times. In the previous study, however, results for nested one-way ANOVA and optimization of derivatives were not presented. The results of that study were based on preprocessing of the data using the first derivative of a quadratic polynomial fit to seven data points (1-Q-7) and are shown in Table II, row 41. The spectra were relatively unstructured and the wavelength range was used. In our current study, we examined the impact of the use of more data points, higher-order polynomials, and a second derivative on the UV results.

Preprocessing Optimization—In general, the changes in analytical uncertainty and the computed F-values as the derivative parameters were changed were similar to the results obtained for VIS. The increase from 7 to 25 data points for the 1-Q derivative produced a slight decrease in the analytical uncertainty and a slight increase in the F-value for cultivar and treatment (Table II, rows 41 and 42). Use of a higher-order polynomial (1-C/Q-25) produced little change in any of the values (row 43). The use of the 2-C/Q-25 second derivative (row 44) also produced a decrease in the analytical uncertainty and a significant increase in the F-value for cultivars and treatments. Use of the 2-C/Q-25 second derivative (row 45) resulted in a further increase in the F-values for both cultivar and treatment and a shift of 15% in the relative variance from treatment to cultivar.

Principal Component Analysis—Previously reported results for the analysis of UV data by ANOVA-PCA showed excellent separation of the clusters of cultivars and individual treatment pairs.[8] As stated earlier, those score plots were obtained with preprocessing using the 1-Q-7 derivative (Table II, row 41). The PCA score plots obtained with 1-Q-25 or 1-C/Q-25 derivatives were not significantly changed, as indicated by the t-values in rows 42 and 43. However, the use of second derivatives (2-Q/C-25 and 2-Q/Q-25, rows 44 and 45) produced significant increases in the t-values for both cultivar and treatment. These changes in the t-values were similar to the changes observed for the F-values.

The PCA score plots for all treatments are shown in Fig. 3C for data preprocessed with a 2-Q/C-25 derivative. Distinctive clusters are seen for all seven treatments. Complete separation for the clusters and pairs of treatments was achieved as indicated by the highly significant t-values in Table II.

Mass Spectrometric Analysis

As for the UV data, we previously reported results for the use of ANOVA-PCA for the analysis of MS spectra of the same broccoli samples.[9] At that time, however, results for nested one-way ANOVA and selection of masses based on the mass reproducibility were not reported. Table II presents the results of these mathematical treatments for negative (MS−) and positive (MS+) ionization, respectively. The MS data were not derivatized or smoothed and the full spectrum was used after preprocessing. All results are based on five repeat analyses of each sample.

Preprocessing Optimization—We have defined mass reproducibility as the frequency with which a signal exceeding 1% of the highest mass count is observed for repeat sample analyses. In this study, with five repeat analyses, the least stringent criteria for including a mass in a sample spectrum requires a signal for just one of the five repeat analyses (1/5 criteria). Inclusion of this mass in the sample spectrum means that it must be included for all samples to produce the well-defined matrix required by ANOVA or PCA. Thus, it is conceivable that a mass may be included in the calculations even though it is observed in only one of 35 spectra (five repeat analyses of seven samples).

The most stringent criteria is to include a mass only if a signal is observed for all five repeat analyses of a sample (5/5 criteria). In this case, any mass used in the calculations will appear in at least five of the spectra. However, we will have greater confidence that it is a legitimate analytical signal. Use of the 5/5 criteria reduced the number of ions used in the calculations from 99 to 88 for MS− and from 167 to 136 for MS+.

All the MS data in the previous study were used, i.e., the 1/5 reproducibility criteria was employed (Table I, rows 51 and 61). Using a 5/5 reproducibility criteria reduced the analytical uncertainty by almost a factor of two and increased the F-values for cultivar and treatment by factors of 1.5 to 2.5 (Table I, rows 52 and 62). For MS−, use of the 5/5 criteria increased the relative variance attributed to cultivar by 12% and reduced the relative variance from treatment by 8%. For MS+, there was little change.

The log transformation of MS data has been reported previously as a means of improving the analytical precision and the significance of the statistical analysis. We saw little change in the F-values when the matrices compiled using the 5/5 criteria were transformed as the natural log (data not shown).

Principal Component Analysis—The PCA score plots for all treatments (Fig. 1, M8+M9) are shown in Figs. 3E and 3F for MS+ and MS−, respectively, based on the 5/5 criteria. Again, excellent separation of cultivars and pairs of treatments was achieved using ANOVA-PCA. The t-values for all preprocessing modes were highly significant, indicating little possibility that the means were the same.

Comparison of Spectrometers

ANOVA and PCA allow statistical evaluation of the effect of the experimental parameters, in this case cultivar and treatment. Optimization of the spectral range and the preprocessing parameters provides for improved statistical results. For ease of comparison, Table III lists the optimum results for each method from Table II. The full spectral ranges were used for VIS, UV, and MS. Range III was selected for both NIR and IR. The NIR, IR, UV, and VIS data were processed using a 1-Q-25 derivative and the MS data were processed using a 5/5 criteria. Only the F-values are presented because the t-values for the PCA score plots mirrored the F-values.

Based on the F-values for the cultivars, three groups were observed. The first group, with the best F-values, consisted of VIS and UV. The second group was made up of MS+, MS−, and NIR. Finally, the worst F-value was obtained by IR. A similar grouping was observed for the F-values for treatments, although one might be tempted to move VIS from the first group into the second.

In this discussion, best and worst are relative terms since all six methods, even without optimization, were capable of distinguishing between the cultivars and treatments at a statistically significant level ($p \le 0.001$). In addition, the information content of the MS methods is undoubtedly greater than the other methods since it has the potential for identifying specific compounds. NIR and IR are both more informative than UV and VIS since specific functional groups can be identified. However, for the purpose of distinguishing between samples based on pattern recognition, UV and VIS gave the best values in this study.

The advantage of VIS and UV (molecular absorption spectrophotometry) would appear to derive from the excellent precision of the analytical signals. It is generally acknowledged that the short-term precision for UV and VIS can be as low as 0.3%. This compares to 5–10% for NIR and IR of solids, and 7–15% for MS with electrospray ionization of complex sample solutions (as expected with direct injection). This order of precisions is reflected in Table III, where the relative analytical uncertainties were approximately 2% for VIS and UV and 4–5% for NIR and MS. Only the relative uncertainty (9%) for IR appears out of order. If all six methods detected the same level of experimental variance, then the F-values would be inversely proportional to the analytical uncertainty, which is roughly the case.

The assumption that all six methods will detect the same experimental variance is not necessarily valid. Each analytical method measures different spectral qualities of the compounds making up the sample. We previously showed that MS− and MS+ had distinctly different spectra and pointed out that there was no correlation between the susceptibility of a compound to ionization and its absorbance spectrum.[9] Similarly, there is no correlation between the VIS and UV spectra or the IR and NIR spectra. It must be remembered that in this study, VIS, UV, and MS analyzed aqueous methanol extracts and NIR and IR analyzed solid samples. In addition, optimization of the spectral range for the best statistical differentiation between samples may further reduce the number of compounds that are measured by a method.

The data in Table III reflect the points made in the previous paragraph. For NIR, IR, MS−, MS+, and UV, the relative variance associated with cultivars ranged from 22% to 43%. Relative variance associated with treatment ranges from 52% to 74%. These limited ranges of variance suggest that, despite measuring different physical properties, the five methods are in reasonable agreement with respect to the variance associated with each experimental factor. This suggests that either the methods are monitoring the same compounds or the treatment and cultivar have profound chemical effects on all aspects of the chemistry of the plant and it doesn’t matter which compounds are measured. The VIS data, however, indicate that 73% of the variance arises from the cultivar and 25% comes from treatment. VIS analysis measures the fewest compounds of any of the methods. Many compounds with chromophores in the UV are transparent in the VIS.

The long-term precisions of the methods are a factor when future applications of the spectral fingerprinting method are considered. Historically, PCA has been performed on samples in a “batch” mode in a short period of time. Even with the shortest analysis intervals, it is good practice to analyze the samples randomly to eliminate any contribution from instrumental drift.[7] If we wish to establish a fingerprint database for materials such as food cultivars, botanical supplements, or genetically modified foods, it will be necessary to ensure that the spectra can be compared over extended periods of time. Such a project would favor methods with the greatest reproducibility. Ideally, the data should be as comprehensive as possible. This would suggest that the full, unprocessed spectra are stored as opposed to filtered data. In addition, this would suggest either the analysis of solid samples or the analysis of more than one extract to cover a wider polarity range.

CONCLUSION

All six methods, NIR, IR, VIS, UV, MS−, and MS+, are capable of acquiring spectral profiles that allow discrimination between the two broccoli cultivars and the seven treatments at a statistically significant level. Both ANOVA and PCA can be used to analyze the data and provide F- and t-values that are highly correlated. Judicious selection of preprocessing parameters can enhance the significance of the statistical tests. In this study, the highest levels of discrimination were achieved by VIS, UV, and NIR, followed by MS and then IR.

Acknowledgments

This research was supported by the Agricultural Research Service of the U.S. Department of Agriculture and the Office of Dietary Supplements at the National Institutes of Health.

References

1. Goodacre R, York EV, Heald JK, Scott IM. Phytochemistry. 2003; 62:859. [PubMed: 12590113]
2. Dunn WB, Overy S, Quick WP. Metabolomics. 2005; 1:137.
3. Dunn WB, Ellis DI. Trends Anal Chem. 2005; 24:285.
4. Mattoli L, Cangi F, Maidecchi A, Ghiara C, Ragazzi E, Tubaro M, Stella L, Tisato F, Traldi P. J Mass Spectrom. 2006; 41:1534. [PubMed: 17051519]
5. Wold S. Chemom Intell Lab Syst. 1987; 2:37.
6. de Harrington PB, Vieira NE, Espinoza J, Kien JK, Romero R, Yergey AL. Anal Chim Acta. 2005; 544:118.
7. de Harrington PB, Vieira NE, Ping C, Espinoza J, Kien JK, Romero R, Yergey AL. Chemom Intell Lab Syst. 2006; 82:283.
8. Luthria DL, Mukhopadhyay S, Robbins RJ, Finley JW, Banuelos GS, Harnly JM. J Agric Food Chem. 2008; 56:5457. [PubMed: 18572954]
9. Luthria DL, Lin LZ, Robbins RJ, Finley JW, Banuelos GS, Harnly JM. J Agric Food Chem. 2008; 56:9819. [PubMed: 18841983]
10. Harnly JM, Pastor-Corrales MS, Luthria DL. Food Chem. 2008; 107:399.
11. Robbins RJ, Keck AS, Banuelos G, Finley JW. J Med Food. 2005; 8:204. [PubMed: 16117613]
12. Finley JW, Sigrid-Keck A, Robbins RJ, Hintze KJ. J Nutr. 2005; 135:1236. [PubMed: 15867310]
13. Savitzky A, Golay MJE. Anal Chem. 1964; 36:1627.
14. Candalfi A, Massart DL, Heuerding S. Anal Chim Acta. 1997; 345:185.
15. Chen Q, Zhao J, Zhang H, Wang X. Anal Chim Acta. 2006; 572:77. [PubMed: 17723463]

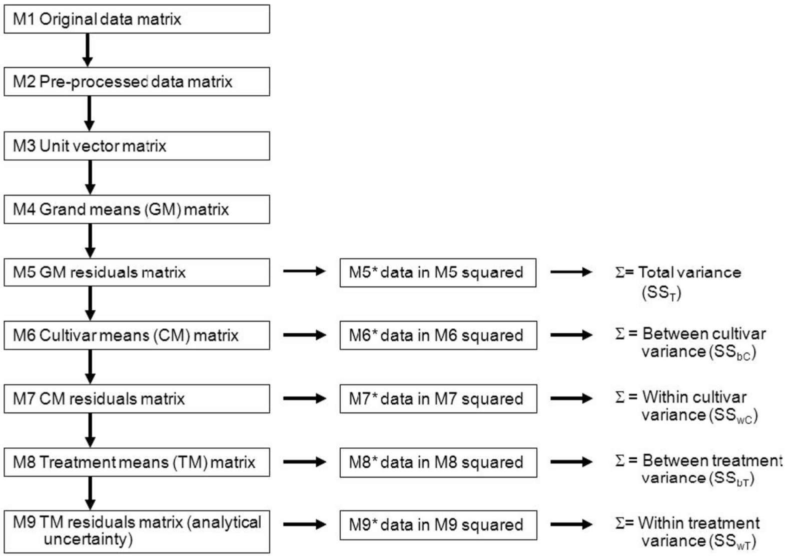
M1 Original data matrix
M2 Pre-processed data matrix
M3 Unit vector matrix
M4 Grand means (GM) matrix
M5 GM residuals matrix
M5* data in M5 squared
Σ= Total variance (SS$_T$)
M6 Cultivar means (CM) matrix
M6* data in M6 squared
Σ = Between cultivar variance (SS$_{bC}$)
M7 CM residuals matrix
M7* data in M7 squared
Σ = Within cultivar variance (SS$_{wC}$)
M8 Treatment means (TM) matrix
M8* data in M8 squared
Σ = Between treatment variance (SS$_{bT}$)
M9 TM residuals matrix (analytical uncertainty)
M9* data in M9 squared
Σ= Within treatment variance (SS$_{wT}$)

Fig. 1.

General diagram for ANOVA preprocessing scheme used for all methods. The data for IR, NIR, UV, and VIS were filtered using different derivatives as discussed in the text. The MS data was filtered based on the mass reproducibility.

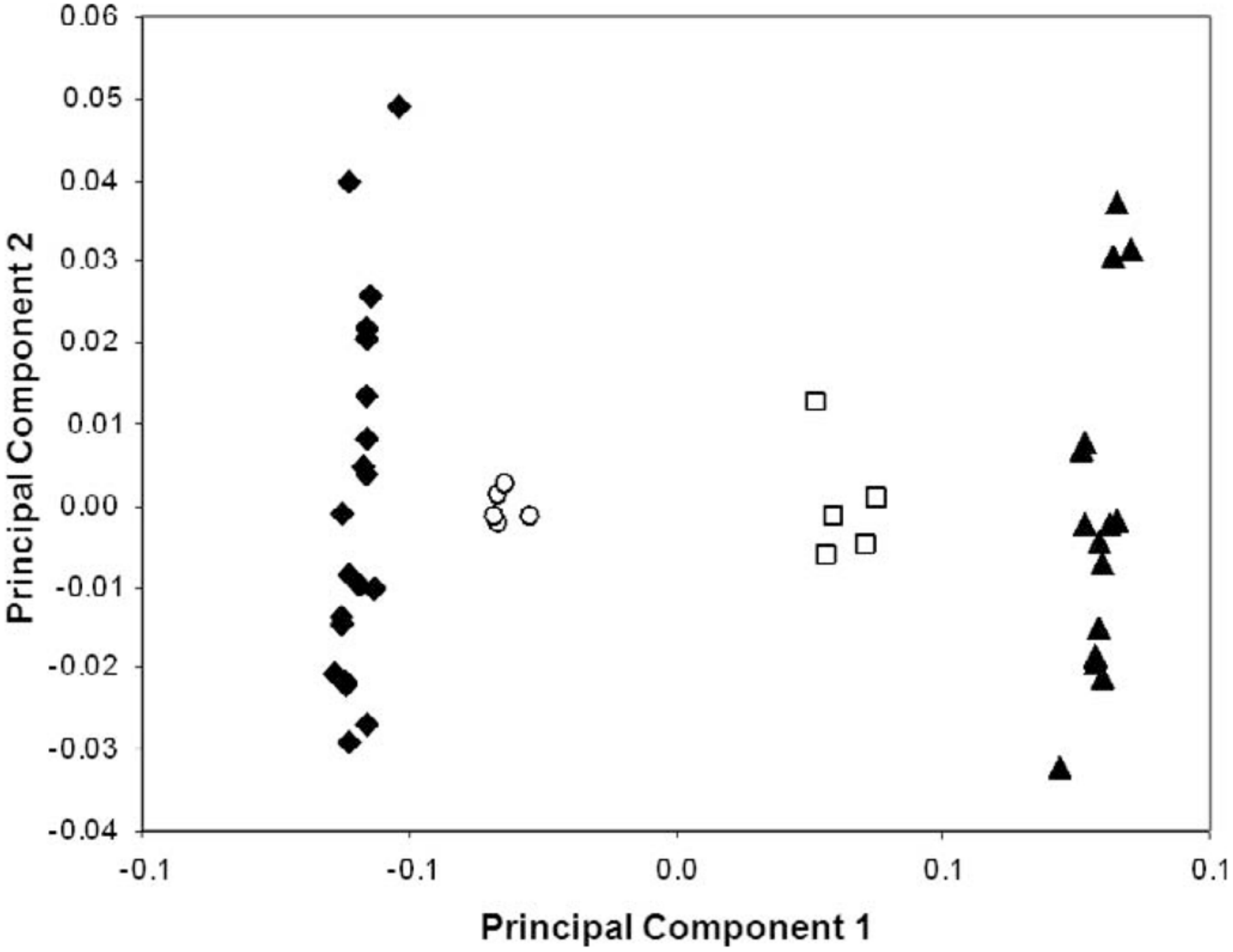

Fig. 2.

NIR: PCA scores for data acquired at 4190–4950 cm^{-1} with a 1-Q-25 derivative. Overlaid plots show comparison of cv Legacy (◆) versus cv Majestic (▲) and cv Legacy conventionally grown with 100% water (○) versus organically grown (□). Data have been scaled to permit both PCA plots to be seen together.

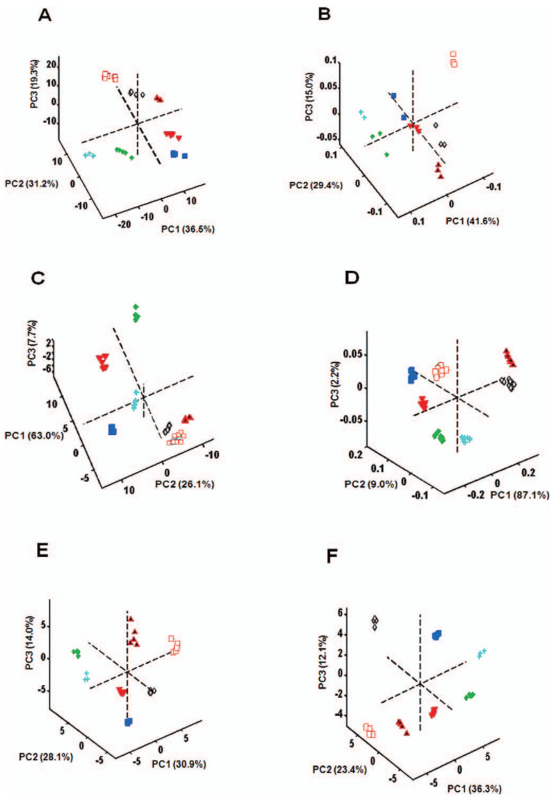

Fig. 3.

PCA score plots for the first three components for all the broccoli data for (**A**) NIR, (**B**) IR, (**C**) UV, (**D**) VIS, (**E**) MS−, and (**F**) MS+. The axes have been rotated to provide optimum cluster separation. Clusters are (solid red triangles) cv Majestic with 0 ppm Se, (teal crosses) cv Majestic with 5 ppm Se, (green asterisks) cv Majestic with 100 ppm Se, (solid blue squares) cv Majestic with 1000 ppm Se, (black open diamonds) cv Legacy conventionally grown with 100% water, (redoutlined black diamonds) cv Legacy conventionally grown with 80% water, and (red open squares) cv Legacy organically grown.

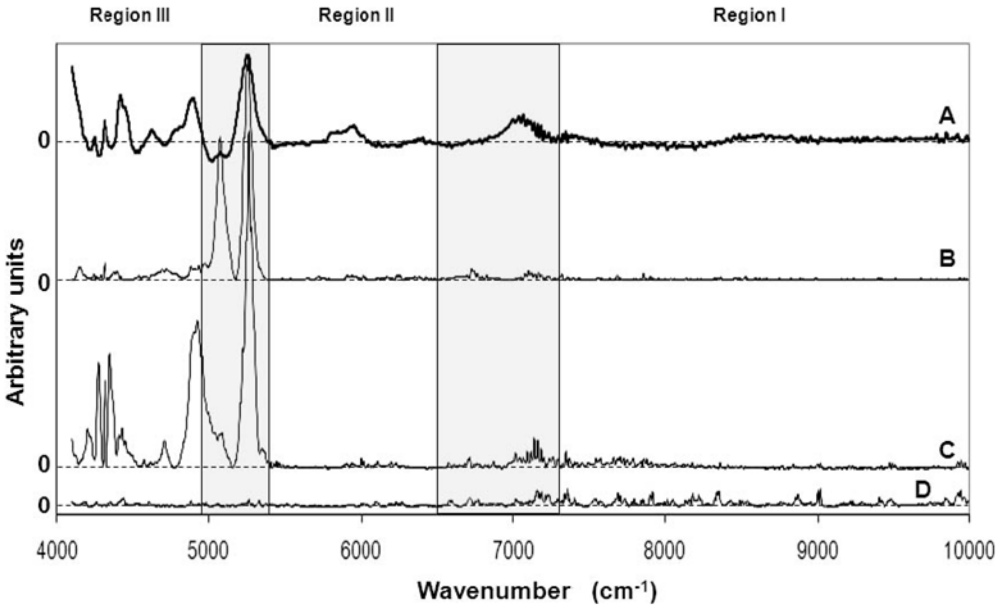

Fig. 4.

NIR: (*A*) Sample spectra after conversion to the first derivative and normalized to a unit vector, (*B*) relative variance between cultivars, (*C*) relative variance between treatments, and (*D*) relative analytical uncertainty. The shaded areas indicate the water bands.

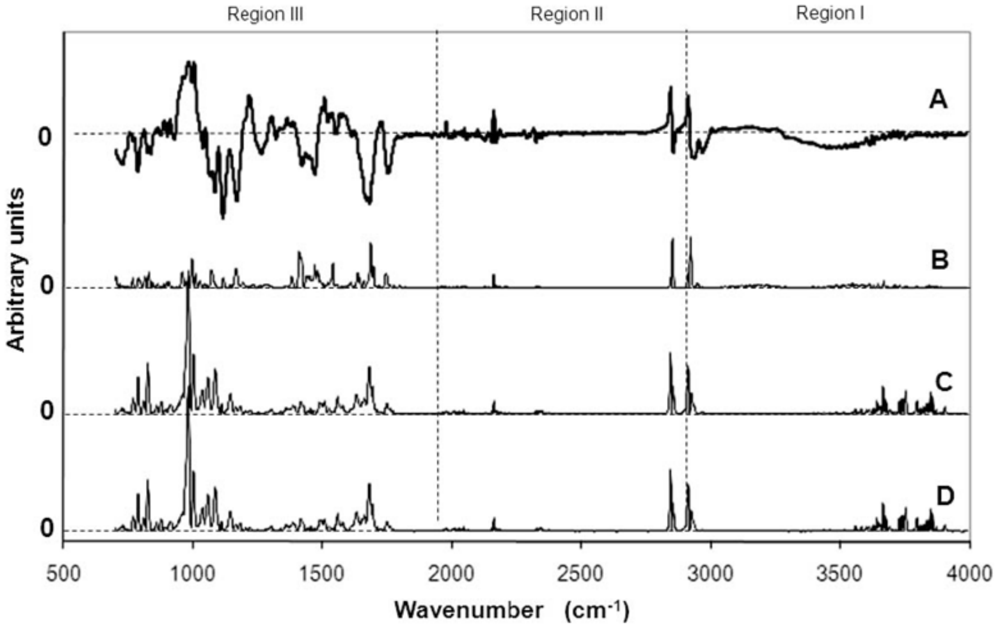

Fig. 5.

IR: (*A*) Sample spectra after conversion to the first derivative and normalized to a unit vector, (*B*) relative variance between cultivars, (*C*) relative variance between treatments, and (*D*) relative analytical uncertainty.

TABLE I

Analysis of variance calculations of ultraviolet (UV), visible (VIS), near-infrared (NIR), mass spectrometry (MS), and infrared (IR) spectral data. Subscripts are identified in Fig. 1 and in the text.

	UV and VIS	NIR and MS	IR
Cultivars	2	2	2
Treatments	7	7	7
Extract	4	5	3
Repeat analyses	2	1	1
Between C	SS_{bC}	SS_{bC}	SS_{bC}
df	1	1	1
MS	SS_{bC}	SS_{bC}	SS_{bC}
Within C	SS_{wC}	SS_{wC}	SS_{wC}
df	54	33	19
MS	$SS_{wT}/54$	$SS_{wT}/33$	$SS_{wT}/19$
F	$54SS_{bT}/SS_{wT}$	$33SS_{bT}/SS_{wT}$	$19SS_{bT}/SS_{wT}$
Between T	SS_{bT}	SS_{bT}	SS_{bT}
df	1	1	1
MS	SS_{bT}	SS_{bT}	SS_{bT}
Within T	SS_{wT}	SS_{wT}	SS_{wT}
df	49	28	14
MS	$SS_{wT}/49$	$SS_{wT}/28$	$SS_{wT}/14$
F	$49SS_{bT}/SS_{wT}$	$28SS_{bT}/SS_{wT}$	$14SS_{bT}/SS_{wT}$

TABLE II

Analysis of variance and principal component analysis statistics of ultraviolet (UV), visible (VIS), near-infrared (NIR), mass spectrometry (MS), and infrared (IR) spectral data.

					ANOVA			ANOVA		PCA	
		Experimental parameters			% Total variance			F-value[a]		t-value[a]	
		Range (cm^{-1})	Region	Derivative	Cultivar	Trtmnt	Analyte	Cultivar	Trtmnt	Cultivar	Trtmnt
NIR	1	10000–4000	All	1-Q-7	26.8	60.5	12.7	69.6	22.3	89.6	36.6
	2		$-H_2O$ Bands	1-Q-7	12.9	68.6	18.5	23.0	17.3	66.0	22.3
	3	10000–7300	Region I	1-Q-7	5.4	30.1	64.6	2.8	2.2		
	4	6600–5400	Region II	1-Q-7	23.5	42.6	33.9	22.8	5.9		
	5	4950–4190	Region III	1-Q-7	18.6	70.5	10.9	56.6	28.4	69.7	28.8
	6			1-Q-11	19.3	72.1	8.6	74.1	39.1		
	7			1-Q-17	20.7	73.3	5.9	115.2	57.6		
	8			1-Q-19	21.0	73.8	5.3	131.7	65.5		
	9			1-Q-21	21.2	74.1	4.7	148.1	73.2		
	10			1-Q-23	21.5	74.3	4.3	164.6	80.6		
	11			1-Q-25	21.6	74.4	4.0	179.4	87.5	125.3	43.6
	12			1-C/Q-25	20.0	71.8	8.3	79.9	40.5	78.3	33.6
	13			2-Q/C-25	9.2	66.0	24.7	12.3	12.5	12.9	2.9
	14			2-Q/Q-25	10.5	42.6	46.8	7.4	4.2	7.5	0.5
IR	21	4000–660	All	1-Q-11	26.6	50.8	22.6	21.0	5.3	11.9	12.2
	22	4000–3000	Region I	1-Q-11	24.6	44.6	31.0	15.0	3.2		
	23	3000–1800	Region II	1-Q-11	23.7	40.9	35.4	12.7	2.6		
	24	1800–800	Region III	1-Q-11	29.5	58.1	12.4	30.7	10.7	39.2	27.0
	25			1-Q-25	31.6	59.2	9.2	65.3	15.1	38.0	21.9
	26			1-C/Q-25	29.9	59.1	11.0	51.6	12.5	39.4	27.9
	27			2-Q/C-25	27.8	59.2	13.0	40.6	10.6	33.5	30.5
	28			2-Q/Q-25	23.8	48.5	27.7	16.4	4.1	24.2	11.2
VIS	31	423–768		1-Q-7	70.1	27.9	2.0	1899	114	64	8.5
	32			1-Q-25	73.4	24.6	2.0	1996	101	69	8.5
	33			1-C/Q-25	70.3	27.8	2.0	1925	115	67	8.5

		Experimental parameters			ANOVA			ANOVA		PCA	
					% Total variance			F-value[a]		t-value[a]	
		Range (cm^{-1})	Region	Derivative	Cultivar	Trtmnt	Analyte	Cultivar	Trtmnt	Cultivar	Trtmnt
	34			2-Q/C-25	60.1	38.0	1.9	1731	166	62	7.5
	35			2-Q/Q-25	58.6	39.3	2.1	1489	151	59	6.5
UV	41	232–388		1-Q-7	33.4	64.5	2.0	880	257	99	8.2
	42			1-Q-25	33.2	64.9	1.9	949	281	118	8.5
	43			1-C/Q-25	33.5	64.6	1.9	942	275	101	8.1
	44			2-Q/C-25	33.4	65.1	1.5	1192	352	181	9.7
	45			2-Q/Q-25	48.1	50.1	1.8	1435	226	237	12.5
Repeatability											
MS(−)	51	111–613		1/5	31.5	59.5	9.0	98	31	63.8	26.6
	52			5/5	43.0	51.9	5.1	238	48	63.5	29.6
MS(+)	61	101–518		1/5	33.9	59.4	6.7	168	42	102.9	32.5
	62			5/5	34.6	61.4	3.9	292	73	111.5	68.1

[a] All probabilities were less than 0.01 except for those specifically noted.

TABLE III

Ultraviolet (UV), visible (VIS), near-infrared (NIR), mass spectrometry (MS), and infrared (IR) spectral methods comparison summary.

	Experimental parameters				% Total variance			F-value[a]	
Method	**Range**	**Units**	**Variables**	**Derivative**	**Cultivar**	**Trtmnt**	**Analyte**	**Cultivar**	**Trtmnt**
NIR	4930–4210	cm^{-1}	361	1-Q-25	21.6	74.4	4.0	179	87
IR	1788–813	cm^{-1}	976	1-Q-25	31.6	59.2	9.2	65	15
VIS	423–768	nm	346	1-Q-25	73.4	24.6	2.0	1996	101
UV	232–388	nm	157	1-Q-25	33.2	64.9	1.9	949	281
MS(−)	111–612	*m*/*z*	88		43.0	51.9	5.1	238	48
MS(+)	101–518	*m*/*z*	103		34.6	61.4	3.9	292	73

[a] All probabilities were less than 0.01.