This is a pre-referred version of the paper published in Cellulose (2016) 23:901-913 DOI 10.1007/s10570-015-0848-z

Application of chemometric analysis to infrared spectroscopy for the identification of wood origin

Ara Carballo-Meilán†, Adrian M. Goodman ‡, Mark G. Baron*, Jose Gonzalez-Rodriguez*

† Department of Chemical Engineering of the University of Loughborough, Loughborough, LE11 3TU, UK

‡ School of Life Sciences of the University of Lincoln, Brayford Pool, Lincoln, LN6 7TS, UK

**School of Chemistry of the University of Lincoln, Brayford Pool, Lincoln, LN6 7TS, UK*

Telephone: +441522886878

E-mail: jgonzalezrodriguez@lincoln.ac.uk

Chemical characteristics of wood are used in this study for plant taxonomy classification based on the current Angiosperm Phylogeny Group classification (APG III System) for the division, class and subclass of woody plants. Infrared spectra contain information about the molecular structure and intermolecular interactions among the components in wood but the understanding of this information requires multivariate techniques for the analysis of highly dense datasets. This article is written with the purposes of specifying the chemical differences among taxonomic groups, and predicting the taxa of unknown samples with a mathematical model. Principal component analysis, t-test, stepwise discriminant analysis and linear discriminant analysis, were some of the chosen multivariate techniques. A procedure to determine the division, class, subclass, order and family of unknown samples was built with promising implications for future applications of Fourier Transform Infrared spectroscopy in wood taxonomy classification

Plant taxonomy classification, Infrared spectroscopy,Multivariate analysis, Wood, Angiosperm, Gimnosperm

Introduction

Trees belong to seed-bearing plants which are subdivided into two major botanical groupings: Gymnosperms (*Gymnospermae*) and Angiosperms (*Angiospermae* or flowering plants). Coniferous woods or softwoods belong to the first-mentioned category and hardwoods to the second group (Sjostrom, 1981). These groups are subdivided into class, subclass, orders, families, genera and species based on the current Angiosperm Phylogenetic System Classification (APG III System). Traditional methods of botanical classification include a taxonomic system based on structural and physiological connections between organisms and a phylogenetic system, based on genetic connections.

The method of "chemical taxonomy" consists of the investigation of the distribution of chemical compounds in series of related or supposedly related plants (Erdtman, 1963). Taxonomically, the species are difficult to classify because there is great inter-species variability as well as narrow gaps between the morphological characteristics of different species (Gidman et al., 2003). The chemical composition of softwoods (gymnosperms) differs from that of hardwoods (angiosperms) in the structure and content of lignin and hemicelluloses. Generally speaking Gymnosperms have less hemicelluloses and more lignin (Martin, 2007). In hardwood the predominant hemicellulose is a partially acetylated xylan with a small proportion of glucomannan. In softwoods, the main hemicellulose is partially acetylated galactoglucomannan and arabinoglucuronoxylan (Barnett and Jeronimidis, 2003; Ek et al., 2009). The composition of xylans from various plants appears as well to be related to their belonging to evolutionary families (Ek et al., 2009). With regards to lignin, softwoods mainly contains only guaiacyl *lignin*, while hardwood contains both guaiacyl (G) and syringyl (S) lignin and the syringyl/guaiacyl (S/G) ratio varies among species (Barnett and Jeronimidis, 2003; Obst, 1982; Stewart et al., 1995; Takayama, 1997) (e.g. species of the same genus can show a large variation in the S/G ratio (Barnett and Jeronimidis, 2003)).

Fourier transform infrared spectroscopy (FTIR) is a non-destructive technique suitable for representations of phylogenetic relationships between plant taxa, even those that are closely related (Shen et al., 2008). An advantage is that it can be applied in the analysis of wood without pre-treatment, thus avoiding the tedious methods of isolation which are normally required (Åkerholm et al., 2001; Obst, 1982). Infrared spectroscopy is quite extensively applied in plant cell wall analysis (Kacuráková et al., 2000). Furthermore, in combination with multivariate analysis, FTIR has been used for the chemotaxonomic classification of flowering plants, for example: the identification and classification of the *Camellia* genus using cluster analysis and Principal Component Analysis (PCA) (Shen et al., 2008); the taxonomic discrimination of seven different plants that belong to two orders and three families using a dendogram based on PCA (Kim et al., 2004); and the differentiation of plants from different genera using cluster analysis (Gorgulu et al., 2007). In woody tissues, FTIR has been used to characterize lignin (Obst, 1982; Takayama, 1997), characterise soft and hardwood pulps using Partial Least-Squares analysis (PLS) and PCA (Bjarnestad and Dahlman, 2002). In addition, the interaction of wood polymers using Partial Least-Squares regression (Åkerholm et al., 2001) and differentiation of wood species using Partial Least-Squares regression (Hobro et al., 2010) has also been investigated. This paper reports on the chemical differences between wood samples using spectral data and multivariate analysis. To the best of our knowledge, this is the first time that unknown samples from trees have been successfully classified into division, class, subclass, as well as, order and family through a linear model based on the chemical features of wood using FTIR spectroscopy.

Materials and Methods

Branch material was collected from 21 tree species in Lincoln (Lincolnshire, UK). Five Gymnosperm trees and 16 Angiosperm trees (12 from Rosids class and 4 from Asterids class) were analysed. Table 1 provides a detailed description of the samples. The samples were stored in a dry environment at ambient temperature conditions

Sample preparation

Sample preparation was reproduced in the same manner as described in detail in another publication (Carballo-Meilan et al., 2014). The dataset obtained from a PerkinElmer Spectrum 100 FTIR Spectrometer was integrated by 3500 variables and 252 observations recorded in pith, bark, rings and sapwood positions. Results from the ring dataset (101 observations) are shown in the present article.

Multivariate techniques

The data set was processed with Tanagra 1.4.39 software. A range of multivariable statistical methods were chosen to analyse spectra of the wood samples including: Principal Component Analysis (PCA), t-test, Stepwise Discriminant Analysis (STEPDISC) method, Partial-Least squares for Classification (C-PLS), Linear Discriminant Analysis (LDA) and PLS-LDA linear models. The statistical methodology from the previous research (Carballo-Meilan et al., 2014) was used in this work.

Results and discussion

Wood spectra dataset

The raw spectra of 16 wood samples that belong to the Angiosperm division and 5 wood samples from the Gymnosperm division were statistically analysed. The sample size available for chemometric analysis in the division dataset was 29 and 72 observations from Conifer and Angiosperm, respectively. From the total number of cases (101), 83 were assigned as training set and 18 as test set. Equivalent procedure was executed with class (74) and subclass (18) datasets; the former with 54 Rosids and 18 Asterids, and the later with 11 Euasterids I and 7 Euasterids II. In the case of the class dataset, the sample was divided to give 60 observations as training set and 14 as test set, and in the case of the subclass dataset 11 cases were assigned as training set and 7 as test set. Vibrational spectra from the growth rings of the wood samples are shown in Fig. 1-A, Fig. 2-A and Fig. 3-A for division, class and subclass dataset, respectively; the arrows indicate important bands in the discrimination of samples based on the STEPDISC results (See section below).

Exploratory data analysis

A PCA mathematical technique was applied to over 101 samples of individual spectra of trees to find the most relevant wavelengths, between the range 4000-500 cm^{-1}, which contribute to sample discrimination between Gymnosperm versus Angiosperm divisions, Rosids versus Asterids classes and Euasterids I versus Euasterids II subclasses. The data set was standardized so each variable received equal weight in the analysis. PCA of the spectra of wood from division, class and subclass dataset gave five main factor loading. Differences between groups, using the two first factors, led to poor structure of the data.

T-test was computed to determine which factors were more significant for differentiating groups. The factor rotated loading (FR) extracted from PCA were used for interpreting the principal components and to determine which variables are influential in the formation of PCs. Normality and homogeneity of variance was checked. Mann-Whitney test (i.e., non-parametric alternative to the t-test) was also performed, confirming the significance of the factors. The wavenumbers loading on those highlighted factors were chemically identified. In later computations, STEPDISC method confirmed the importance of those chemicals in the discrimination. The results of that probe showed that there are chemical differences between Gymnosperms and Angiosperms that were condensed only inside the fourth and fifth rotated factor (FR4 and FR5). The t-test was 2.902 with an associated probability of 0.00456 for FR4, and 4.6767 ($p= 0.000009$) for FR5. Then the null hypothesis may be rejected at the 99.54% and 99.99% levels for FR4 and FR5, respectively and, therefore, it is concluded that there is a significant difference in means due to the factor selected. A detailed band assignment of the factors highlighted in the t-test is presented in Table 2. Those factors seem to contain relevant meaning. The most highly correlated wavenumbers with those factors are 1762-1719, 1245-1220 and 1132-950 from FR4 and 2978-2832, 1713-1676 and 1279-1274 from FR5. As the STEPDISC method highlighted, it is highly likely that the C=O stretching in hemicelluloses and lignin, wavenumbers 1730, 1712 and 1684 cm^{-1} from feature selection (range 1762-1719 cm^{-1} in FR4 and 1713-1676 cm^{-1} in FR5) play a key role in the classification.

In the case of Rosids vs. Asterids, the t-test emphasized FR3 and FR5 as main descriptors of the chemical differences between class. The result seems not significantly different with 95% probability for FR5 ($t =1.7379$, $p=0.0865$). The difference was only significant for FR3 at the 5% significant level since the p value was 0.00148 (t was 3.3062). Major contributors to the FR3 formation are wavenumber between 1171 and 884 cm^{-1}, and 2860-2847 cm^{-1}. The most highly correlated wavenumbers with FR5 are 1687-1385. Then the C-H ring in glucomannan, 874 and 872 (associated with FR3), and the C=O stretching and C-H deformation in lignin and carbohydrates, wavenumbers 1678, 1619, 1617, 1613 and 1438 cm^{-1} associated with FR5 are all important chemical signals for differentiating Rosids from Asterids classes, based on PCA and STEPDISC analysis. With regards to the differences between Euasterids I and Euasterids II, FR4 was selected from the t-test analysis with a value of the probability greater than 0.05 ($t=1.9179$, $p=0.0731$). This factor is highly correlated with the wavenumbers 1763-1709 and 1245-1212 cm^{-1}. Based on the feature selection procedure, it could be that 1769, 1701 and 1697 cm^{-1} were significant for distinguishing among the subclass groups but the results were limited by the small sample size. The identity of the mentioned wavenumbers was associated with C=O stretching in hemicelluloses and lignin. The wavenumbers responsible for the classification between division, class, subclass, order and family are described in the next section (STEPDISC analysis).

A subset of wavenumbers from the STEPDISC method was used as input in PCA to emerge the underlined structure in division, class and subclass datasets. The scores extracted from PCA were used for interpreting the samples and the loading to determine which variables are in relation with the samples. The higher the loading of a variable, the more influence it has in the formation of the factor and vice versa. The score plot from division dataset (Fig. 1-B) showed that conifers were highly correlated with FR3, and the loading plot (Fig. 1-D) showed that the wavenumber 1684 could be related with conifers since it correlates more with its factor. A 3D plot (Fig. 1-C) with the individual observations is shown to highlight the underline structure of the dataset using the first three rotated factors. In the score plot from class dataset (Fig. 2-B), the Asterids sample correlated highly with FR2 and the Rosids sample better with FR1. The correlation plot (Fig. 2-D) suggested that the wavenumber 2031 is more highly correlated with FR2, and therefore would be more connected with the Asterids group. With respect to the subclass dataset, loading plot is shown in Fig. 3-B. In this case Euasterids I observations were positively correlated with FR2, and Euasterids II with FR1. The wavenumbers 1701, 1697 and 1769 were correlated with FR1, suggesting some closeness with Euasterids II.

STEPDISC analysis

Supervised approach, based on the Wilks' partial lambda, known as STEPDISC method was computed over the normalized wavenumbers to determine the most significant variables for the classification process. Groups based on the current

Angiosperm Phylogeny Group classification (APG III System) were used to find the discriminator wavenumbers. Forward strategy and computed statistic F to 3.84 as statistical criterion for determining the addition of variables was chosen. The cut-off value selected as minimum conditions for selection of the variables was 0.01 significant level to find the most relevant variables. Seven biomarkers (1730, 1712, 1420, 3068, 1684, 1610, and 1512 cm^{-1}) were successfully found to discriminate between Angiosperms and Gymnosperms. The wavenumbers, arranged in a descendent order based on their F-values (i.e., the variable's total discriminating power, the greater contributor to the overall discrimination in the STEPDISC method will show a better F-value (Klecka, 1980)), have the following band assignment: 1730 (C=O stretching in acetyl groups of hemicelluloses (xylan/glucomannan) (Åkerholm et al., 2001; Bjarnestad and Dahlman, 2002; Gorgulu et al., 2007; Marchessault, 1962; McCann et al., 2001; Mohebby, 2008, 2005; Rana et al., 2009; Stewart et al., 1995)), 1712(C=O stretch (unconjugated) in lignin (Hobro et al., 2010)), 1420 (aromatic ring vibration combined with C-H in-plane deformation lignin (Kubo and Kadla, 2005; Rhoads et al., 1987; Wang et al., 2009)), 3068 (C-H stretch aromatic (Larkin, 2011; Silverstein et al., 2005)), 1684 (C=O stretch in lignin (Coates, 2000; Silverstein et al., 2005; Sudiyani et al., 1999)), 1610 (aromatic skeletal vibration plus C=O stretching lignin (Kubo and Kadla, 2005; Wang et al., 2009)), and 1512 (aromatic skeletal vibration lignin(Hobro et al., 2010; Huang et al., 2008; Kubo and Kadla, 2005; Wang et al., 2009)). It seems that differences between groups can be attributed to the lignin region. These spectral differences between hard and softwood lignin were observed in the fingerprint region between 1800 and 900 cm-1 by other authors (Pandey, 1999).

With regards to class dataset, 10 biomarkers (2031, 1678, 1619, 1617, 1613, 784, 771, 874, 872, and 1438 cm-1) were found to successfully discriminate between the Rosids and Asterids classes within the Angiosperm division. Differences between groups can be attributed to C=O stretching in lignin and C-H deformation in carbohydrates and lignin, based on their literature assignments (in order of greater contribution to the overall discrimination): 2031 (-N=C=S (Donald L. Pavia & Gary M. Lampman & George S. Kriz & James A. Vyvyan, 2009; Larkin, 2011)), 1678 (C=O stretching aryl ketone of guaiacyl (G) (Rhoads et al., 1987)), 1619, 1617, 1613 (C-O stretching of conjugated or aromatic ketones, C=O stretching in flavones (Hobro et al., 2010; Huang et al., 2008)), 784 (Out of plane CH bend (Silverstein et al., 2005)), 771 (out of plane N-H wagging primary and secondary amides in carbohydrates or OH out of plane bending (Marchessault, 1962; Muruganantham et al., 2009; Peter Zugenmaier, 2007)), 874, 872 (C-H ring glucomannan (Åkerholm et al., 2001; Bjarnestad and Dahlman, 2002; Kacuráková et al., 2000; Marchessault, 1962)), and 1438 (C-H deformation in Lignin and carbohydrates (Mohebby, 2005)). Thiocyanate was also seen by other authors to discriminate among Angiosperms (Rana et al., 2009).

The last probe was run over subclass dataset; 5 biomarkers (1769, 1697, 3613, 3610, and 1701 cm-1) were found to successfully discriminate between Euasterids I and Euasterids II subclass from Asterids class. As mentioned before, C=O stretching in lignin and carbohydrates seems relevant for the classification. The greater contributor to the discrimination between subclass groups was the wavenumber 1769, attributed in the literature to C=O stretching in acetyl groups of hemicelluloses (xylan/glucomannan) (Åkerholm et al., 2001; Bjarnestad and Dahlman, 2002; Gorgulu et al., 2007; Marchessault, 1962; McCann et al., 2001; Mohebby, 2008, 2005; Rana et al., 2009; Stewart et al., 1995), this contributor was followed in order of importance (the second greatest F-value) by 1697 assigned to C=O stretching (Coates, 2000; Silverstein et al., 2005), 3613 and 3610 (O-H stretching (Coates, 2000)), and lastly 1701 related to Conj-CO-Conj lignin (Hobro et al., 2010; Larkin, 2011).

STEPDISC method was run over different split datasets from ring dataset, the imbalance effect on the results was also checked; in such a way, the discriminator wavenumbers from the output of STEPDISC method were selected and used to construct linear regression models.

Linear model and validation

The next step after selecting the discriminator wavenumbers was to compute and compare several linear models: C-PLS, LDA and PLS-LDA. The discrete class attribute are the taxons based on the current taxonomic classification of trees and the continuous attributes are the discriminator wavenumbers filtered through the STEPDISC previous method. Wilks's lambda is a multivariate measure of group differences over the predictors (Klecka, 1980) and it was used to measure the ability of the variables in the computed classification function from LDA to discriminate among the groups. Classification was done by using the classification functions computed for each group. Observations were assigned to the group with the largest classification score (Rakotomalala, 2005). LDA gave the lowest error in the classification and was for that reason the only one shown in this work. Bias-variance error rate decomposition was used to adjust the correct number of predictors in the model to the current sample size, as describe in our previous work (Carballo-Meilan et al., 2014). As shown in Fig. 4, the optimum model in division would have 4 wavenumbers instead of 7. In the case of the class model, the overfitting region showed up above 8 and underfitting below 7. Similar approach was taken for the subclass model where 4 wavenumber were selected as the optimum model. Table 3 shows the classification functions with their statistical evaluation for division, class and subclass datasets. The coefficients of the classification functions are not interpreted. Smallest lambda values (not shown) or largest partial F means high discrimination (Klecka, 1980). The significance of the difference was checked using Multivariate Analysis of Variance (MANOVA) and two transformations of its lambda, Bartlett transformation and Rao transformation (Rakotomalala, 2005). According to Rao's transformation (for small sample sizes, $p < 0.01$), it can be concluded that there is a significant difference between groups in the three cases: division (Rao-F $(7,75)=46.417$, $p=0.000$), class (Rao-F $(7,75)=21.975$, $p=0.000$) and subclass (Rao-F $(7,75)=35.028$, $p=0.000$). The discriminant functions scores were plotted in Fig. 5 to show the discrimination among division, class and subclass groups. The separation looks greater in the case of class and subclass.

Validation of the model was done to evaluate the statistical and the practical significance of the overall classification rate and the classification rate for each group. Cross-validation (CV), bootstrap method, leave-one-out (LOO), Wolper and Kohavi bias-variance decomposition, and an independent test set which was not used in the construction of the model (test size appears in brackets in Table 3) were used in the validation procedure. The bootstrap value shown in Table 3 is the higher error obtained by the .632 estimator and its variant .632+. This error was seen to be preferred for Gaussian population and small training samples size ($n \leq 50$) (Chernick, 2011). Error rate estimation is presented to evaluate the variance explained by the model; in division, 52% bias, 47% variance, 0.0671 error rate; in class, 64% bias, 36% variance, 0.1552 error rate; and in subclass, 57% bias, 43% variance, 0.0950 error rate. The model seems stable with a low classification error. Further validation of the method was performed with an unknown piece of wood. The division, class, subclass and order were determined correctly. The samples were taken from a willow tree and belonged to Angiosperm > Rosids > Eurosids I > Malpighiales.

Conclusion

A procedure was developed for the taxonomic classification of wood species using samples from different division, class and subclass. First, a STEPDISC method was used to select the predictor wavenumbers for classification. The chemical differences between taxonomic groups were attributed mainly to the differences in their lignin and hemicelluloses content, as well as some amide contribution. The results were also confirmed by a t-test applied on the output from PCA procedure. LDA, PLS-LDA and C-PLS linear models were computed to calculate the classification functions with the predictor variables as dependent variables and groups based on the APG III System as independent variables. LDA provided the lowest classification error based on different validation techniques such as bootstrap or LOO. For an unknown sample its division, class, subclass and order were successfully determined. This study demonstrates that spectra data obtained from wood samples have the potential to be used to discriminate trees taxonomically.

A scaffold for the taxonomic classification of woody plants has been produced. A procedure to statistically define differences among species and use them in a model that classifies unknown samples is possible. With additional work this may prove to be a useful tool to aid in the taxonomic classification of plants. Naturally the current models should only be applied to the species included in the model and, because of the differences in chemical composition among species, it is important that new models are developed to broaden its application.

Acknowledgements

This work was supported by Europracticum IV (Leonardo da Vinci Programme). We gratefully acknowledge to the Consello Social from Universidade de Santiago de Compostela (Spain)

References

Åkerholm, M., Salmén, L., Salme, L., 2001. Interactions between wood polymers studied by dynamic FT-IR spectroscopy. Polymer (Guildf). 42, 963–969. doi:10.1016/S0032-3861(00)00434-1

Anchukaitis, K.J., Evans, M.N., Lange, T., Smith, D.R., Leavitt, S.W., Schrag, D.P., 2008. Consequences of a rapid cellulose extraction technique for oxygen isotope and radiocarbon analyses. Anal. Chem. 80, 2035–2041. doi:10.1016/j.gca.2004.01.006.Analytical

APG, I., 2003. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG II. Bot. J. Linn. Soc. 141, 399–436. doi:10.1046/j.1095-8339.2003.t01-1-00158.x

Barnett, J.R., Jeronimidis, G., 2003. Wood quality and its biological basis. Blackwell, Oxford, p. 226.

Bjarnestad, S., Dahlman, O., 2002. Chemical Compositions of Hardwood and Softwood Pulps Employing Photoacoustic Fourier Transform Infrared Spectroscopy in Combination with Partial Least-Squares Analysis. Anal. Chem. 74, 5851–5858. doi:10.1021/ac025926z

Carballo-Meilan, A., Goodman, A.M., Baron, M.G., Gonzalez-Rodriguez, J., 2014. A specific case in the classification of woods by FTIR and chemometric: Discrimination of Fagales from Malpighiales. Cellulose 21, 261–273. doi:10.1007/s10570-013-0093-2

Chen, J., Liu, C., Chen, Y., Chen, Y., Chang, P.R., 2008. Structural characterization and properties of starch/konjac glucomannan blend films. Carbohydr. Polym. 74, 946–952. doi:10.1016/j.carbpol.2008.05.021

Chernick, M.R., 2011. Bootstrap Methods: A Guide for Practitioners and Researchers. Wiley, Hoboken, N.J., p. 400.

Coates, J., 2000. Interpretation of infrared spectra, a practical approach. Encycl. Anal. Chem. 10815–10837.

Donald L. Pavia & Gary M. Lampman & George S. Kriz & James A. Vyvyan, 2009. Introduction to spectroscopy. Brooks/Cole, Cengage Learning, Belmont, CA, p. 727.

Ek, M., Gellerstedt, G., Henriksson, G., 2009. Wood Chemistry and Wood Biotechnology. Walter de Gruyter, Berlin, p. 308.

Erdtman, H., 1963. Some aspects of chemotaxonomy. Chem. Plant Taxon. 89–125.

Gidman, E., Goodacre, R., Emmett, B., Smith, A.R., Gwynn-Jones, D., 2003. Investigating plant–plant interference by metabolic fingerprinting. Phytochemistry 63, 705–710. doi:10.1016/S0031-9422(03)00288-7

Gorgulu, S.T., Dogan, M., Severcan, F., 2007. The characterization and differentiation of higher plants by fourier transform infrared spectroscopy. Appl. Spectrosc. 61, 300–8. doi:10.1366/000370207780220903

Hobro, A., Kuligowski, J., Döll, M., Lendl, B., 2010. Differentiation of walnut wood species and steam treatment using ATR-FTIR and partial least squares discriminant analysis (PLS-DA). Anal. Bioanal. Chem. 398, 2713–22. doi:10.1007/s00216-010-4199-1

Huang, A., Zhou, Q., Liu, J., Fei, B., Sun, S., 2008. Distinction of three wood species by Fourier transform infrared spectroscopy and two-dimensional correlation IR spectroscopy. J. Mol. Struct. 883-884, 160–166. doi:10.1016/j.molstruc.2007.11.061

Kacuráková, M., Kauráková, M., Capek, P., Sasinkova, V., Wellner, N., Ebringerova, A., Kac, M., 2000. FT-IR study of plant cell wall model compounds: pectic polysaccharides and hemicelluloses. Carbohydr. Polym. 43, 195–203. doi:10.1016/S0144-8617(00)00151-X

Kim, S.W., Ban, S.H., Chung, H.J., Cho, S., Choi, P.S., Yoo, O.J., Liu, J.R., 2004. Taxonomic discrimination of flowering plants by multivariate analysis of Fourier transform infrared spectroscopy data. Plant Cell Rep. 23, 246–50. doi:10.1007/s00299-004-0811-1

Klecka, W.R., 1980. Discriminant analysis. Sage Publications, Beverly Hills, Calif., p. 71.

Kubo, S., Kadla, J.F., 2005. Hydrogen bonding in lignin: a Fourier transform infrared model compound study. Biomacromolecules 6, 2815–21. doi:10.1021/bm050288q

Larkin, P., 2011. Infrared and Raman Spectroscopy; Principles and Spectral Interpretation. Elsevier, Amsterdam ; Boston, p. 230.

Liang, C.Y., Marchessault, R.H., 1959. Infrared spectra of crystalline polysaccharides. II. Native celluloses in the region from 640 to 1700 cm.1. J. Polym. Sci. 39, 269–278. doi:10.1002/pol.1959.1203913521

Liang, C.Y., Marchessault, R.H., 1959. Infrared spectra of crystalline polysaccharides. II. Native celluloses in the region from 640 to 1700 cm.1. J. Polym. Sci. 39, 269–278. doi:10.1002/pol.1959.1203913521

Marchessault, R.H., 1962. Application of infra-red spectroscopy to cellulose and wood polysaccharides. Pure Appl. Chem. 5, 107–130. doi:10.1351/pac196205010107

Marchessault, R.H., Liang, C.Y., 1962. The infrared spectra of crystalline polysaccharides. VIII. Xylans. J. Polym. Sci. 59, 357–378. doi:10.1002/pol.1962.1205916813

Marchessault, R.H., Pearson, F.G., Liang, C.Y., 1960. Infrared spectra of crystalline polysaccharides. I. Hydrogen bonds in native celluloses. Biochim. Biophys. Acta 45, 499–507.

Martin, J.W., 2007. Concise encyclopedia of the structure of materials. Elsevier, Amsterdam ; Boston, p. 512.

McCann, M.C., Bush, M., Milioni, D., Sado, P., Stacey, N.J., Catchpole, G., Defernez, M., Carpita, N.C., Hofte, H., Ulvskov, P., Wilson, R.H., Roberts, K., 2001. Approaches to understanding the functional architecture of the plant cell wall. Phytochemistry 57, 811–821. doi:10.1016/S0031-9422(01)00144-3

Mohebby, B., 2005. Attenuated total reflection infrared spectroscopy of white-rot decayed beech wood. Int. Biodeterior. Biodegradation 55, 247–251. doi:10.1016/j.ibiod.2005.01.003

Mohebby, B., 2008. Application of ATR Infrared Spectroscopy in Wood Acetylation. J. Agric. Sci 10, 253–259.

Muruganantham, S., Anbalagan, G., Ramamurthy, N., 2009. FT-IR and SEM-EDS comparative analysis of medicinal plants, Eclipta Alba Hassk and Eclipta Prostrata Linn. Rom. J. Biophys 19, 285–294.

Obst, J.R., 1982. Guaiacyl and Syringyl Lignin Composition in Hardwood Cell Components. Holzforschung 36, 143–152. doi:10.1515/hfsg.1982.36.3.143

Pandey, K.K., 1999. A study of chemical structure of soft and hardwood and wood polymers by FTIR spectroscopy. J. Appl. Polym. Sci. 71, 1969–1975. doi:10.1002/(SICI)1097-4628(19990321)71:12<1969::AID-APP6>3.3.CO;2-4

Pandey, K.K., Vuorinen, T., 2008. Comparative study of photodegradation of wood by a UV laser and a xenon light source. Polym. Degrad. Stab. 93, 2138–2146. doi:10.1016/j.polymdegradstab.2008.08.013

Peter Zugenmaier, 2007. Crystalline cellulose and derivatives: characterization and structures. Springer, Berlin ; New York, p. 285.

Rakotomalala, R., 2005. “TANAGRA : un logiciel gratuit pour l’enseignement et la recherche.”

Rana, R., Langenfeld-Heyser, R., Finkeldey, R., Polle, A., 2009. FTIR spectroscopy, chemical and histochemical characterisation of wood and lignin of five tropical timber wood species of the family of Dipterocarpaceae. Wood Sci. Technol. 44, 225–242. doi:10.1007/s00226-009-0281-2

Rana, R., Sciences, F., 2008. Correlation between anatomical/chemical wood properties and genetic markers as a means of wood certification. Nieders\"achsische Staats-und Universit\"atsbibliothek Göttingen. doi:978-3-9811503-2-2

Revanappa, S.B., Nandini, C.D., Salimath, P.V., 2010. Structural characterisation of pentosans from hemicellulose B of wheat varieties with varying chapati-making quality. Food Chem. 119, 27–33. doi:10.1016/j.foodchem.2009.04.064

Rhoads, C.A., Painter, P., Given, P., 1987. FTIR studies of the contributions of plant polymers to coal formation. Int. J. Coal Geol. 8, 69–83. doi:10.1016/0166-5162(87)90023-1

Sekkal, M., Dincq, V., Legrand, P., Huvenne, J., 1995. Investigation of the glycosidic linkages in several oligosaccharides using FT-IR and FT Raman spectroscopies. J. Mol. Struct. 349, 349–352.

Shen, J.B., Lu, H.F., Peng, Q.F., Zheng, J.F., Tian, Y.M., 2008. FTIR spectra of Camellia sect. Oleifera, sect. Paracamellia, and sect. Camellia (Theaceae) with reference to their taxonomic significance. Plantsystematics.com 46, 194–204. doi:10.3724/SP.J.1002.2008.07125

Silverstein, R.M., Webster, F.X., Kiemle, D., 2005. Spectrometric identification of organic compounds. Wiley, Hoboken, NJ, p. 502.

Sjostrom, 1981. Wood chemistry: fundamentals and applications. Academic Press, New York, p. 293.

Stewart, D., Wilson, H.M., Hendra, P.J., Morrison, I.M., 1995. Fourier-Transform Infrared and Raman Spectroscopic Study of Biochemical and Chemical Treatments of Oak Wood (Quercus rubra) and Barley (Hordeum vulgare) Straw. J. Agric. Food Chem. 43, 2219–2225. doi:10.1021/jf00056a047

Sudiyani, Y., Tsujiyama, S., Imamura, Y., Takahashi, M., Minato, K., Kajita, H., Sci, W., 1999. Chemical characteristics of surfaces of hardwood and softwood deteriorated by weathering. J. Wood Sci. 45, 348–353.

Takayama, M., 1997. Fourier transform Raman assignment of guaiacyl and syringyl marker bands for lignin determination. Spectrochim. Acta Part A Mol. Biomol. Spectrosc. 53, 1621–1628. doi:10.1016/S1386-1425(97)00100-5

Wang, S., Wang, K., Liu, Q., Gu, Y., Luo, Z., Cen, K., Fransson, T., 2009. Comparison of the pyrolysis behavior of lignins from different tree species. Biotechnol. Adv. 27, 562–7. doi:10.1016/j.biotechadv.2009.04.010

List of Figures

List of Tables

Figures

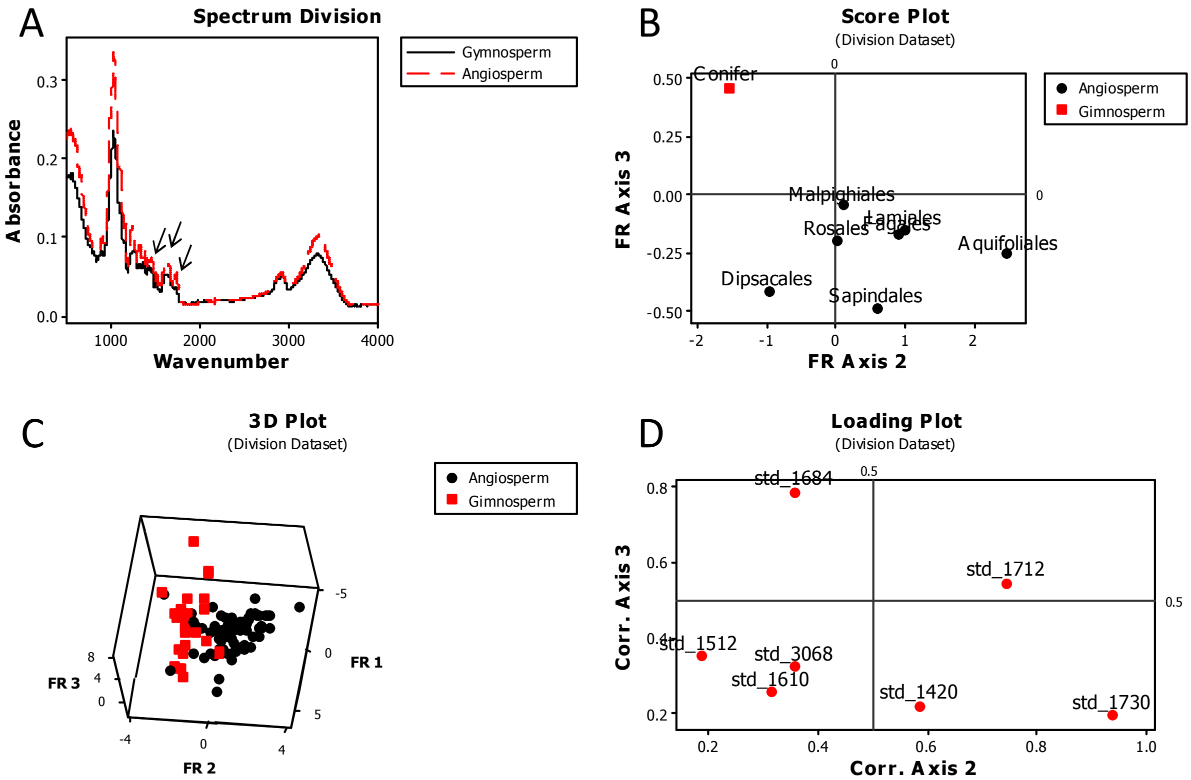

Fig. 1 Average FTIR spectrum of division: Gymnosperm versus Angiosperm (A), score plot (B), 3D plot (C) and loading plot (D) from Gymnosperm and Angiosperm dataset

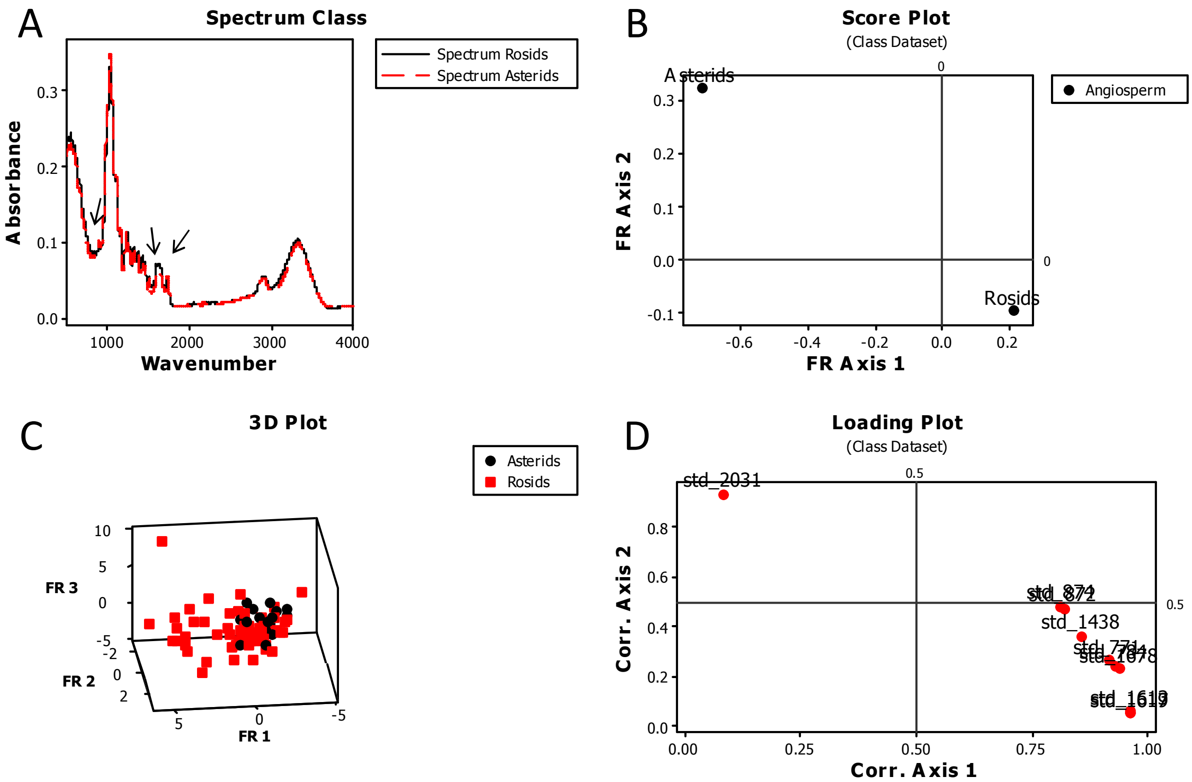

Fig. 2 Average FTIR spectrum of class: Rosids versus Asterids (A), score plot (B), 3D plot (C) and loading plot (D) from Rosids and Asterids dataset

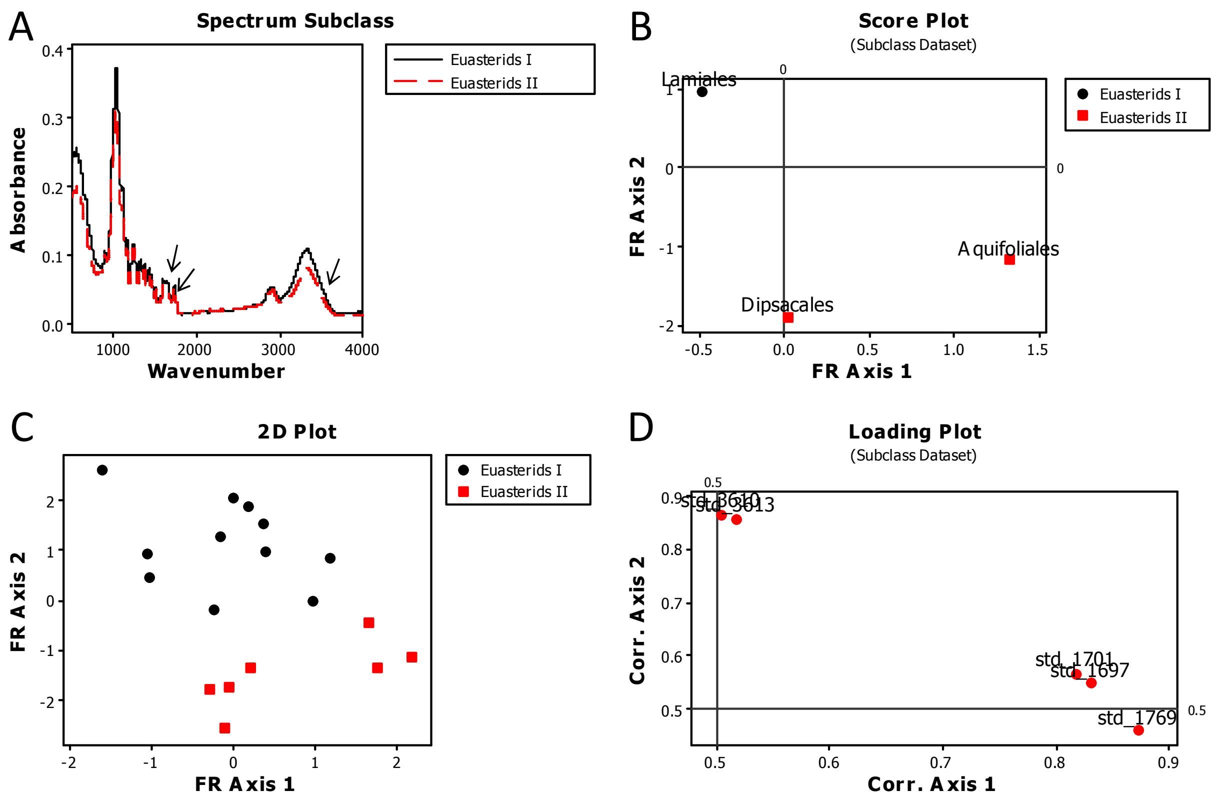

Fig. 3 Average FTIR spectrum of subclass: Euasterids I versus Euasterids II (A), score plot (B), 2D plot (C) and loading plot (D) from Euasterids I and Euasterids II dataset

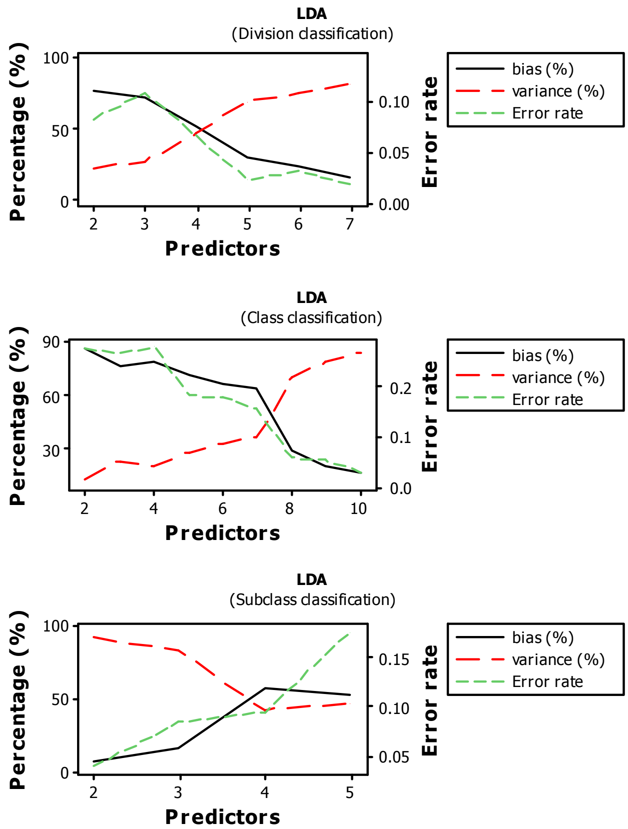

Fig. 4 Bias-variance decomposition from division, class and subclass models

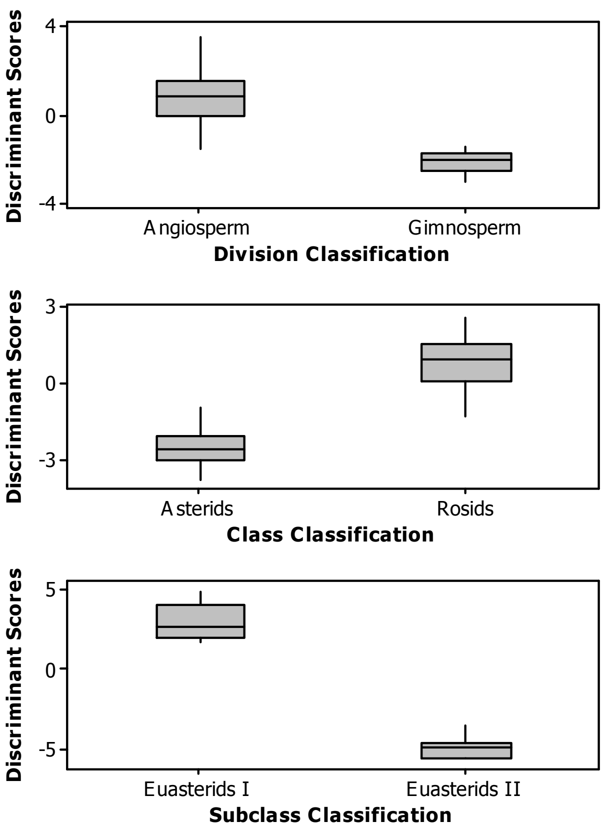

Fig. 5 Boxplot of the discrimination function scores in division, class and subclass linear models

Tables

Table 1 Tree species based on APG III System Classification **(APG, 2003)**

Division	Class	Subclass	Order	Family	Genus	Specie	Common name
Gymnosperm	Pinophyta	Pinopsida	Pinales	Taxaceae	Taxus L.	*Taxus baccata*	Yew
				Pinaceae	Pinus L.	*Pinus sylvestris*	Scot Pine (3 varieties)
					Larix	*Larix decidua*	Larch
Angiosperms	Rosids	Eurosids I	Rosales	Moraceae	Ficus	*Ficus carica*	Fig
				Ulmaceae	Ulmus L.	*Ulmus procera*	Elm
			Fagales	Betulaceae	Alnus M.	*Alnus glutinosa*	Black Alder
					Corylus L.	*Corylus avellana*	Hazel
					Betula L.	*Betula pubescens*	Birch
				Fagaceae	Castanea	*Castanea sativa*	Sweet Chestnut
					Fagus L.	*Fagus sylvatica*	Beech
					Quercus	*Quercus robur*	English Oak
			Malpighiales	Salicaceae	Populus	*Populus*	Poplar
					Populus	*Poplar nigra*	Black Poplar
					Salix	*Salix fragilis*	Willow
		Eurosids II	Sapindales	Sapindaceae	Acer	*Acer pseudoplatanus*	Sycamore
	Asterids	Euasterids I	Lamiales	Oleaceae	Fraxinus L.	*Fraxinus excelsior*	Ash (2 varieties)
		Euasterids II	Aquifoliales	Aquifoliaceae	Illex L.	*Illex aquifolium*	Holly
			Dipsacales	Adoxaceae	Sambucus	*Sambucus nigra*	Elder

Table 2 Band assignments of the third (FR3), fourth (FR4) and fifth (FR5) factor rotated loadings related to the variables obtained by PCA from ring dataset

FR	ν (cm^{-1})	Literature assignments and band origin
Division		
4	1762-1719	1740-1730, 1725 C=O stretching in acetyl groups of hemicelluloses (Åkerholm et al., 2001; Bjarnestad and Dahlman, 2002; Gorgulu et al., 2007; Marchessault and Liang, 1962; Marchessault, 1962; McCann et al., 2001; Mohebby, 2008, 2005; Rana et al., 2009; Stewart et al., 1995)
	1245-1220	1245-1239 C-O of acetyl stretch of lignin and xylan
		1238-1231 common to lignin and cellulose, S ring breathing with C-O stretching C-C stretching and OH in-plane bending (C-O-H deformation) cellulose, C-O-C stretching in phenol-ether bands of lignin(Åkerholm et al., 2001; Anchukaitis et al., 2008; Bjarnestad and Dahlman, 2002; Hobro et al., 2010; CY Y Liang and Marchessault, 1959; Marchessault, 1962; Pandey and Vuorinen, 2008; Rhoads et al., 1987)
	1132-950	1125,1123,1113 aromatic C-H in-plane deformation syringyl in lignin(Kubo and Kadla, 2005; Rhoads et al., 1987; Wang et al., 2009)
		1110,1112 antisymmetrical in-phase ring stretch cellulose(CY Y Liang and Marchessault, 1959)
		1090, 1092 C-C glucomannan(Kacuráková et al., 2000; McCann et al., 2001)
		1090 antisymmetric β C-O-C hemicelluloses(Sekkal et al., 1995)
		1064 C=O stretching glucomannan(Gorgulu et al., 2007)
		1059,1033 C-O stretch (C-O-H deformation) cellulose(CY Y Liang and Marchessault, 1959; Rhoads et al., 1987)
		1030 aromatic C-H in-plane deformation guaiacyl plus C-O(Kubo and Kadla, 2005; Rhoads et al., 1987; Wang et al., 2009)
		1034,941,898 C-H, ring glucomannan(Åkerholm et al., 2001; Bjarnestad and Dahlman, 2002; Gorgulu et al., 2007; Kacuráková et al., 2000; McCann et al., 2001)
5	2978-2832	2957 2922, 2873, 2852 CH_3 asymmetric and symmetric stretching: mainly lipids and proteins with a little contribution from proteins, carbohydrates, and nucleic acids(Gorgulu et al., 2007)
		2945,2853 CH_2 antisymmetric stretching cellulose(Marchessault and Liang, 1962; Marchessault et al., 1960)
		2853 CH_2 symmetric stretching xylan(Marchessault and Liang, 1962; Marchessault et al., 1960)
		2940 (S), 2920(G), 2845-2835(S), 2820(G) C-H stretching (methyl and methylenes) lignin(Rhoads et al., 1987)
	1713-1676	1711 C=O stretch (unconjugated) in lignin(Hobro et al., 2010)
		Conj-CO-Conj(Larkin, 2011)
	1279-1274	1282,1280 C-H bending (CH_2-O-H deformation) cellulose(CY Y Liang and Marchessault, 1959; Rhoads et al., 1987)
Class		
3	2860-2847	2852 CH_2 symmetric stretching: mainly lipids with a little contribution from proteins, carbohydrates, and nucleic acids(Gorgulu et al., 2007)
		2853 CH_2 stretching xylan and cellulose(Marchessault and Liang, 1962; Marchessault et al., 1960)
	1171-884	1168-1146 C-O-C antisymmetric stretching in cellulose and xylan;
		and characteristic pectin band(Gorgulu et al., 2007; CY Y Liang and Marchessault, 1959; Marchessault and Liang, 1962; Marchessault, 1962; Mohebby, 2005; Pandey and Vuorinen, 2008; Rana and Sciences, 2008; Rhoads et al., 1987; Sekkal et al., 1995)
		1129-1088 out-of-plane ring stretch in cellulose and glucomannan, aromatic C-H in plane syringyl and C-O-C antisymmetric stretching hemicelluloses(Kubo and Kadla, 2005; C. Y. Liang and Marchessault, 1959; Sekkal et al., 1995; Wang et al., 2009)
		1076-883 C-O-C symmetric stretching in hemicelluloses and celluloses; C-O stretch glucomannan and celluloses; and aromatic C-H deformation guaiacyl, amorphous cellulose and glucomannan(Bjarnestad and Dahlman, 2002; Gorgulu et al., 2007; Kacuráková et al., 2000; Kubo and Kadla, 2005; CY Y Liang and Marchessault, 1959; Mohebby, 2005; Pandey and Vuorinen, 2008; Rana et al., 2009; Rhoads et al., 1987; Sekkal et al., 1995; Wang et al., 2009)

5	2929-2927	2922 CH_2 asymmetric stretching: mainly lipids with a little contribution from proteins, carbohydrates, and nucleic acids(Gorgulu et al., 2007)
	1687-1385	1683-1512 C-O ketones, flavones and glucuronic acid; amides in proteins; water; OH intramolecular H-bonding glucomannan; lignin skeletal(Chen et al., 2008; Gorgulu et al., 2007; Hobro et al., 2010; Huang et al., 2008; Kubo and Kadla, 2005; CY Y Liang and Marchessault, 1959; Marchessault and Liang, 1962; Rana and Sciences, 2008; Revanappa et al., 2010; Wang et al., 2009)
Subclass		
4	1763-1709	1740-1730, 1725 C=O stretching in acetyl groups of hemicelluloses (Åkerholm et al., 2001; Bjarnestad and Dahlman, 2002; Gorgulu et al., 2007; Marchessault and Liang, 1962; Marchessault, 1962; McCann et al., 2001; Mohebby, 2008, 2005; Rana et al., 2009; Stewart et al., 1995)
	1245-1212	1245-1239 C-O of acetyl stretch of lignin and xylan 1238-1231 common to lignin and cellulose, S ring breathing with C-O stretching C-C stretching and OH in-plane bending (C-O-H deformation) cellulose, C-O-C stretching in phenol-ether bands of lignin(Åkerholm et al., 2001; Anchukaitis et al., 2008; Bjarnestad and Dahlman, 2002; Hobro et al., 2010; CY Y Liang and Marchessault, 1959; Marchessault, 1962; Pandey and Vuorinen, 2008; Rhoads et al., 1987)

Table 3 Classification functions for Gymnosperm, Rosids and Euasterids I, and validation from division, class and subclass models

Classification functions		Statistical Evaluation	
Descriptors	LDA	F(1,5)	p-value
Division			
1730	3.3377	21.52445	0.000015
1712	-3.0887	9.14461	0.003414
1684	0.7958	1.6519	0.202655
1512	-2.9963	46.30463	0.000000
constant	-1.1877	-	
Class			
1678	-2.80427	23.71985	0.000011
1619	25.07698	14.33562	0.000398
1617	-22.13934	10.37686	0.002203
1438	0.917706	2.02774	0.160424
874	-1.413472	6.36166	0.014761
784	-6.00400	14.4103	0.000386
771	6.421311	21.53428	0.000024
constant	-0.52498		
Subclass			
3614	179.3411	4.59063	0.08504
3610	-224.9511	7.89394	0.037565
1768	58.8748	5.71739	0.062302
1701	-102.0568	6.67082	0.049265
constant	-22.1101	-	

Validation and test (ring samples)			
	Division	Class	Subclass
CV	0.0400	0.0900	0.0000
.632+ Bootstrap	0.0508	0.0899	0.0513
LOO	0.0396	0.1081	0.0000
Train test	0.0452	0.0435	0.0500
Independent test (size)	0.0556(18)	0.2143(14)	0.0000(7)
Error rate	0.0671	0.1552	0.0950