JOURNAL OF NATURAL PRODUCTS

Article

pubs.acs.org/jnp

Metabolic Profiling and Classification of Propolis Samples from Southern Brazil: An NMR-Based Platform Coupled with Machine Learning

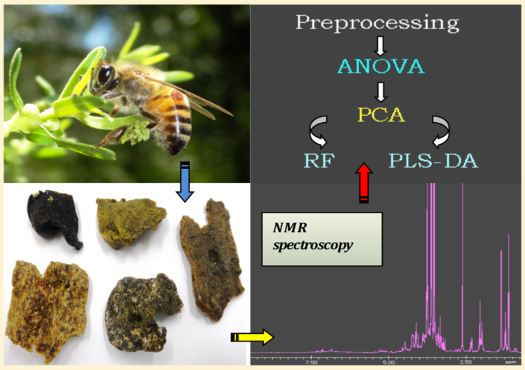
Marcelo Maraschin,*[†,⊥] Amélia Somensi-Zeggio,[†,⊥] Simone K. Oliveira,[†] Shirley Kuhnen,[†] Maíra M. Tomazzoli,[†] Josiane C. Raguzzoni,[†] Ana C. M. Zeri,[‡] Rafael Carreira,[§] Sara Correia,[§] Christopher Costa,[§] and Miguel Rocha[§]

[†]Plant Morphogenesis and Biochemistry Laboratory, Federal University of Santa Catarina, Florianópolis, SC, Brazil

[‡]Brazilian Biosciences National Laboratory (LNBio-CNPEM/MCTI), Campinas, São Paulo, Brazil

[§]CEB−Centre Biological Engineering, University of Minho, Campus of Gualtar, Braga, Portugal

S *Supporting Information*



ABSTRACT: The chemical composition of propolis is affected by environmental factors and harvest season, making it difficult to standardize its extracts for medicinal usage. By detecting a typical chemical profile associated with propolis from a specific production region or season, certain types of propolis may be used to obtain a specific pharmacological activity. In this study, propolis from three agroecological regions (plain, plateau, and highlands) from southern Brazil, collected over the four seasons of 2010, were investigated through a novel NMR-based metabolomics data analysis workflow. Chemometrics and machine learning algorithms (PLS-DA and RF), including methods to estimate variable importance in classification, were used in this study. The machine learning and feature selection methods permitted construction of models for propolis sample classification with high accuracy (>75%, reaching ∼90% in the best case), better discriminating samples regarding their collection seasons comparatively to the harvest regions. PLS-DA and RF allowed the identification of biomarkers for sample discrimination, expanding the set of discriminating features and adding relevant information for the identification of the class-determining metabolites. The NMR-based metabolomics analytical platform, coupled to bioinformatic tools, allowed characterization and classification of Brazilian propolis samples regarding the metabolite signature of important compounds, i.e., chemical fingerprint, harvest seasons, and production regions.

Propolis, or bee glue, is a sticky dark-colored substance produced from the collected buds or exudates of plants (resin) by bees (*Apis mellifera* L.). The resin is masticated, salivary enzymes are added, and the partially digested material is mixed with beeswax and used in the hive to seal the walls, strengthen the borders of combs, and embalm dead invaders.[1] Humans have used propolis as a remedy since ancient times.[2] In the last years, this product has been the subject of intensive studies highlighting its biological and pharmacological properties, such as antimicrobial,[3−6] antioxidative,[7] antiviral,[8] antitumoral,[9−11] anti-inflammatory,[3,4] and antineurodegenerative.[12] Propolis was also tested as a food preservative, due to its bactericidal and bacteriostatic properties. As some constituents of propolis are naturally found in foods, they are recognized as safe substances.[13]

The success of propolis as a dietary supplement led to an increased interest in its chemical composition. In general, resin comprising flavonoids and related phenolic acids represent approximately half of the propolis constituents, while beeswax, volatiles, and pollen represent approximately 30%, 10%, and 5%, respectively.[14] Still, the chemical composition of the bee glue is extremely dependent on the plants found around the hive, as well as on the geographic and climatic characteristics of the collection site. Buds from *Populus* species are the main source of resins in European and North American propolis ("poplar-type" propolis[2]). Alternatively, in regions where these plants are not native, other species from the genera *Clusia* (in Cuba) and *Baccharis* (in Brazil) are used as resin sources, increasing its diversity and complexity.[15] Less commonly, species from genera such as *Betula*, *Ulmus*, *Pinus*, *Quercus*, *Salix*, and *Acacia* are also used.[16]

More than 300 constituents have been identified in propolis,[14] with the phenolics being the most abundant compounds. In propolis from temperate zones, the most frequently reported phenolic components include the flavonoids pinocembrin, galangin, and chrysin and the phenolic acids caffeic acid, ferulic acid, and cinnamic acid.[2] Instead, the propolis samples from

Received: April 9, 2015
Published: December 22, 2015

DOI: 10.1021/acs.jnatprod.5b00315
J. Nat. Prod. 2016, 79, 13−23

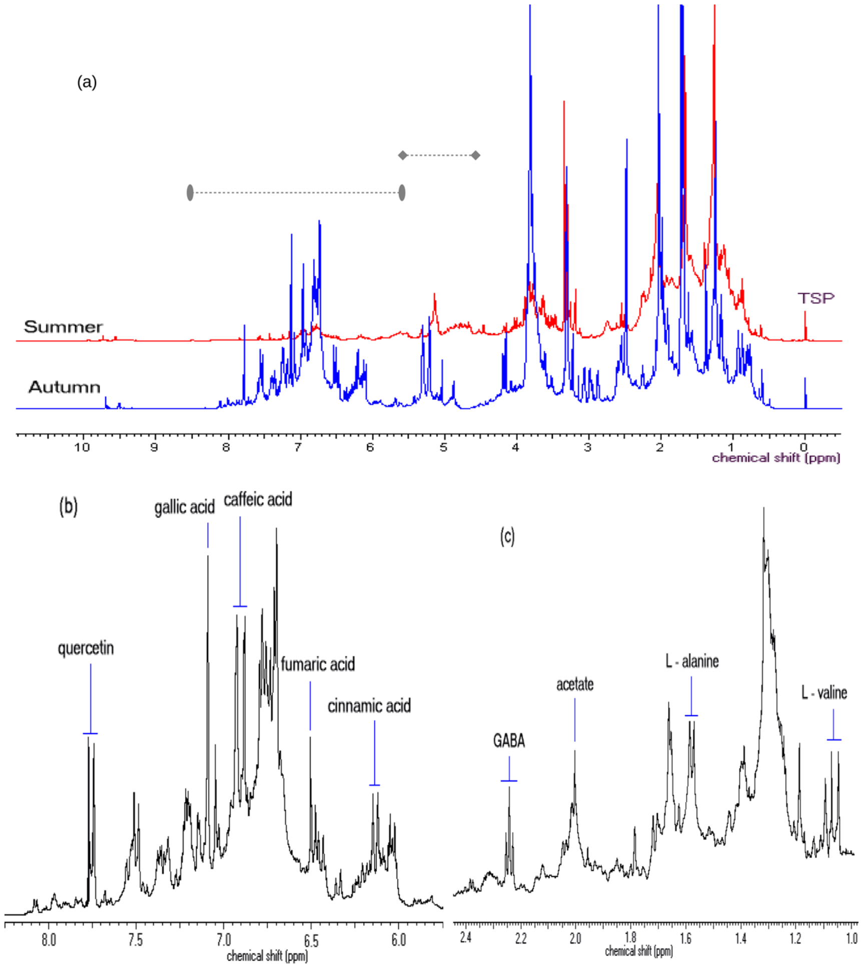


Figure 1. (a) ^{1}H NMR spectra of propolis samples collected in the highlands region of Santa Catarina state, southern Brazil (São Joaquim county, 28°17′38″ S, 49°55′54″ W, 1360 m altitude), in the autumn and summer of 2010. Discrepancies over the spectroscopic profiles might be found, being more pronounced at the anomeric (4.50−5.50 ppm, ◆—◆) and aromatic (5.50−8.50 ppm, ◀—▶) regions. A partial assignment of certain resonances to some of the identified compounds is shown in panels (b) and (c) for the aromatic and aliphatic regions, respectively.

tropical zones, in particular those from the southeastern region of Brazil, were shown to be rich in prenylated phenylpropanoids,[14] although nontypical flavonoids from "poplar-type" propolis, such as kaempferide and isosakuranetin, have also been found.[17,18] Moreover, Cuban propolis has recently caught the attention of scientists because of its peculiar enrichment in polyisoprenylated benzophenones, which makes it chemically distinct from both the European and the Brazilian bee glues.[19]

By using the set of features produced by NMR spectroscopy, several types of data analysis may be performed in connection with the aims of an investigation. For the purpose of sample classification, for instance, the selection of certain features is used to eliminate irrelevant and redundant information, possibly causing noise. If sample class labels are unavailable, or in cases where the presence of novel classes is suspected, unsupervised classification methods may be used to discover sample groups. Data transformation methods, such as principal component analysis (PCA), may be sufficient to reveal class structure within the samples.

On the other hand, when class labels are available, they may be used to support supervised classification. Predictive models or classifiers may then be built to classify new unlabeled data. For example, partial leastsquares-discriminant analysis (PLS-DA),[20−22] support vector machines (SVMs), and artificial neural networks (ANNs)[23−25] have been applied to build classification models from metabolomics data.

A topic that overlaps with supervised classification is that of feature selection, which may be employed to improve a classification model in terms of generalization performance and accuracy, by eliminating noninformative features. Feature selection may also be used to gain further insight into the rationale underlying class division within a particular domain. In the context of metabolomics, retrieving the set of class-discriminating features may aid in the identification of the class-determining metabolites. This may allow further elucidation of the system under investigation.

NMR metabolomics data are replete with feature correlations (multicollinearity), within both the signal (features relevant to class explanation) and noise (irrelevant features).[21] To overcome such constraints, we adopted a strategy where accuracy-based approaches are complemented with feature selection methods less prone to the bias effects of multicollinear data, including those based on *variable influence on the projection* (VIP) values, derived from PLS-DA, and *variable importance* produced by a Random Forest (RF) ensemble classifier.

As stated above, propolis is a source of valuable compounds for human health.[26] However, due to its considerable chemical heterogeneity, the production of standardized and homogeneous extracts is a difficult task. Indeed, chemical characterization and standardization of propolis extracts are technically tedious, time-consuming, and non-cost-effective when adopting traditional analytical selective techniques, such as HPLC.

Over the past years, NMR spectroscopy has been recognized as a powerful tool for characterizing chemically complex matrices.[27] However, the amount of information afforded by an NMR sample is large, as a typical high magnetic field ^{1}H NMR spectrum contains 32 000 or 64 000 data points. The analysis of such information is not feasible without the aid of powerful computational tools. To deal with large NMR data sets, data mining/machine learning techniques have been adopted to build descriptive, predictive, and classification models. Combined with chemometrics techniques, these are thought to be a suitable approach to gain insight into the most important features associated with chemical composition, harvest season, and geographic origin of propolis samples produced in the Santa Catarina state of southern Brazil.

Herein, a novel metabolomics data analysis workflow for propolis samples from southern Brazil using chemometrics and machine learning algorithms, including methods to estimate variable importance in classification, is proposed. A multidimensional metabolomics data set (59 samples × 34 032 variables) of Brazilian propolis NMR spectra was collected and analyzed. It has long been known that the chemical composition of propolis might be strongly influenced by environmental factors peculiar to the sites of collection of a given region of production, as well as by harvest season effects. The underlying hypothesis of this work considers that the huge diversity of plant species derived from climatic and soil conditions found in propolis production regions in the Santa Catarina state, southern Brazil, will lead to quite heterogeneous chemical profiles of the resulting propolis samples.

Such a scenario is critical regarding the need of standardization of the chemical composition of propolis extracts for use as dietary supplements. The search for a single and homogeneous chemical profile of propolis seems to be unfeasible once it is produced worldwide, making it hypervariable in a chemical sense. Indeed, it is well known that the chemical composition of propolis is strongly influenced by environmental factors (e.g., flora, bee genotype, and climate conditions) that vary according to the producing sites worldwide, rendering it difficult to obtain homogeneous samples for industrial applications. Thus, a more realistic and useful strategy to better understand and explore propolis as a source of important bioactive compounds regards a regional (and/or seasonal) analysis of its chemical constituents. For instance, the chemical composition of type 6 propolis, from the Atlantic rainforest in the state of Bahia (northeastern region of Brazil), is distinct from the other known types of propolis, mainly due to the absence of flavonoids and the presence of other nonpolar, long-chain compounds.[28]

One could speculate that by detecting a particular chemical profile associated with a given region of production or harvest season, certain types of propolis could be applied in a dedicated manner to the production of dietary supplements with specific health claims. In this sense, an NMR-based metabolomics analytical platform seems interesting to gain insights as to the chemical heterogeneity of propolis samples associated with production regions and harvest seasons. This will require, however, the aid of powerful bioinformatics tools to extract relevant information associated with metabolite signature identification, as an indispensable tool to characterize propolis samples in a rational basis, as further described by the workflow presented herein.

■ RESULTS AND DISCUSSION

For the purposes of this study, a multidimensional metabolomics data set of Brazilian propolis NMR spectra was collected and analyzed. Figure 1a shows two typical ^{1}H NMR spectra of propolis originated from the highlands region of Santa Catarina state, collected in summer and autumn of 2010. Partial assignments of certain resonances to some of the identified compounds are depicted in panels (b) and (c) for the aromatic and aliphatic regions, respectively. Inspection of the NMR spectra allows one to obtain information regarding chemical heterogeneity of the sample as highlighted for the discrepant set of resonances at the aliphatic and aromatic regions. However, visual analysis of the NMR spectra is not effective to extract all the relevant information from the data set for sample classification, hence rendering the use of bioinformatic tools compulsory.

Univariate Analysis for Seasonal Effects. Univariate analysis methods are the most commonly used for exploratory analysis, providing a preliminary overview about features potentially significant in discriminating the seasonal effects under study. In this study, one-way ANOVA and Tukey's HSD *post hoc* test were used for multigroup analysis, i.e., samples collected in the distinct seasons, and the top 20 important features (resonances) identified are shown in Table 1.

Data from Table 1 reveal that some compounds with anomeric structural moieties seem to have a significant effect on the discrimination of propolis samples over the seasons, since the six main features selected by univariate analysis occur at the anomeric region of the ^{1}H NMR spectra (4.50−5.50 ppm).

Further analysis of the one- and two-dimensional (TOCSY and HSQC) NMR spectra, taking into account the set of features selected, allowed us to putatively identify the metabolic signature of relevant metabolites in propolis samples and, additionally, some phenolic compounds usually found in that raw material (Table 2). Over the past years, our research group has built a 1D NMR spectroscopic data bank of relevant metabolites for both natural products and biochemical studies. Thus, the resonances collected in this work were first assigned to the compounds by visual inspection of the ^{1}H NMR spectra with respect to the information available in that data bank. Such an approach has led to the reduction of the eventual errors added to the interpretation and assignments of NMR spectra, since the hardware setup and the conditions for the sample spectra acquisition (e.g., solvent, pH, temperature) have been kept the same. Eventual ambiguities due to overlapped signals were resolved through analysis of the 2D NMR spectra, as well as searching relevant literature information,[29−32] NMR-based metabolomics databases, e.g., Human Metabolome Database (http://www.hmdb.ca), Biological Magnetic Resonance Data Bank (http://www.bmrb.wisc.edu/metabolomics/metabolomics_standards.html), Madison Metabolomic Consortium Database (http://mmcd.nmrfam.wisc.edu), Spectral Database for Organic Compounds (http://riodb01.ibase.aist.go.jp/sdbs/cgi-bin/cre_index.cgi), Sigma-Aldrich Spectral View Database of Standards, and available tools for metabolomics analysis such as MetaboMiner (http://wishart.biology.ualberta.ca/metabominer/index.html), which includes a comprehensive 2D spectroscopic reference library. Importantly, most of the identified metabolites match the resonance assignments when compared to those of standard compounds available in our NMR spectroscopic database.

Table 1. Important Features Selected by One-Way ANOVA (*p*-Value Threshold 0.05) from the ^{1}H NMR Data Set of Propolis Samples from Southern Brazil, Collected in the Autumn (au), Winter (wi), Spring (sp), and Summer (sm) in 2010

resonances (δ ppm ^{1}H)	*p*-value	−log 10(*p*)	FDR[a]	Tukey's HSD
4.66	9.58×10^{-26}	25.0	2.39×10^{-23}	sm-au; sp-sm; wi-sm
4.58	3.38×10^{-17}	16.5	4.21×10^{-15}	sm-au; sp-sm; wi-sm
4.55	6.09×10^{-14}	13.2	5.06×10^{-12}	sm-au; sp-au; wi-au; sp-sm; wi-sm
4.63	1.04×10^{-13}	13.0	6.50×10^{-12}	sm-au; sp-sm; wi-sm
4.71	2.08×10^{-13}	12.7	1.04×10^{-11}	sm-au; sp-sm; wi-sm
4.50	2.64×10^{-13}	12.6	1.10×10^{-11}	sp-au; wi-au; sp-sm; wi-sm
4.08	1.22×10^{-12}	11.9	4.33×10^{-11}	sp-au; wi-au; sp-sm; wi-sm
4.45	2.45×10^{-12}	11.6	7.23×10^{-11}	sp-au; wi-au; sp-sm; wi-sm
4.17	2.61×10^{-12}	11.6	7.23×10^{-11}	sp-au; wi-au; sp-sm; wi-sm
4.31	4.23×10^{-12}	11.4	1.05×10^{-10}	sp-au; wi-au; sp-sm; wi-sm
4.53	1.04×10^{-11}	11.0	2.36×10^{-10}	sm-au; sp-au; wi-au; sp-sm; wi-sm
4.02	6.61×10^{-11}	10.2	1.37×10^{-9}	sp-au; wi-au; sp-sm; wi-sm
4.38	8.03×10^{-11}	10.1	1.54×10^{-9}	sp-au; wi-au; sp-sm; wi-sm
4.05	1.50×10^{-10}	9.82	2.68×10^{-09}	sp-au; wi-au; sp-sm; wi-sm
4.28	2.62×10^{-10}	9.58	4.35×10^{-9}	sp-au; wi-au; sp-sm; wi-sm
4.25	3.87×10^{-10}	9.41	5.80×10^{-09}	sp-au; wi-au; sp-sm; wi-sm
4.34	3.96×10^{-10}	9.40	5.80×10^{-9}	sp-au; wi-au; sp-sm; wi-sm
4.20	1.34×10^{-9}	8.87	1.85×10^{-08}	sp-au; wi-au; sp-sm; wi-sm
4.13	4.54×10^{-9}	8.34	5.95×10^{-8}	sp-au; wi-au; sp-sm; wi-sm
4.74	2.70×10^{-8}	7.57	3.26×10^{-7}	sm-au; sp-sm; wi-sm

[a] FDR: false discovery rate determines adjusted *p*-values for each Tukey test performed.

Chemical analyses of aqueous alcoholic extracts of Brazilian propolis have previously permitted identification of a number of compounds by using LC-MS and GC-MS techniques. Some compounds found in this work have been previously identified in that complex matrix,[18−33] but it seems to be the first time that some (γ-aminobutyric acid, citric acid, and amino acids, for example) were identified in propolis by NMR techniques. Currently, more than 300 compounds have been identified in several types of propolis worldwide, and the phenolics seem to be the most abundant metabolites.[34] Indeed, as phenolic acids and their derivatives, flavonoids, diterpenoids, and triterpenoids are commonly found in propolis, since they are constituents of resinous substances collected by honeybees from plant sources, the occurrence of monosaccharides (e.g., D-glucose and D-fructose) and disaccharides (e.g., sucrose) has been thought to originate from nectar and honey, as well as from plant mucilages collected by honeybees, implying the latter to be an additional propolis source.[32,35−38]

Of interest, the most important feature selected by one-way ANOVA (Table 1) from the ^{1}H NMR data set of propolis refers to β-D-glucose (4.66 ppm, d, 1β-CH, $J = 7.40$ Hz). Analysis of the spectroscopic profiles revealed that propolis samples harvested in the winter did not show typical resonances associated with a monosaccharide moiety, indicating that propolis produced at lower temperatures differs in its monosaccharidic composition with respect to those produced in the summer and spring. Other features relevant to discriminate the propolis samples over the harvesting seasons have been associated with the metabolites L-ascorbic acid (4.50 ppm, d, 4-CH, $J = 1.65$ Hz), β-D-fructose (4.08 ppm, m, 3-CH; 4.05 ppm, m, 5-CH), tartaric acid (4.31 ppm, s), L-alanine (4.20 ppm, q, 2-CH, $J = 7.21$ Hz), and lactic acid (4.13 ppm, q, 2-CH, $J = 6.95$ Hz).

By expanding the number of top features selected by the univariate statistical methods applied to the ^{1}H NMR data set (data not shown) to 50, significant differences (Tukey test, $p < 0.05$) were detected in the chemical profiles. Indeed, the southern region of Brazil shows a well-defined pattern of seasonal effects compared to other regions, e.g., northern and northeastern Brazil. This issue is a matter of concern as one aims to investigate eventual discrepancies in the chemical composition of propolis samples collected throughout the year in southern Brazil.

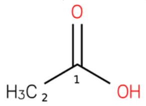
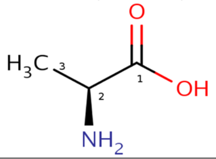
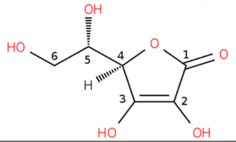
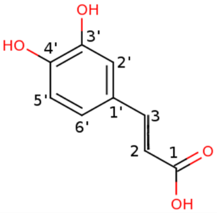
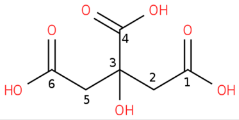
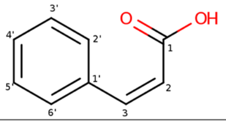
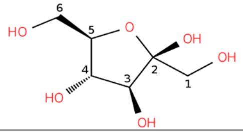
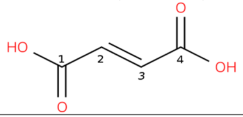
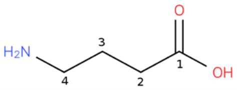
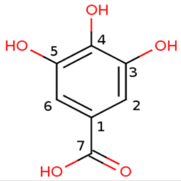
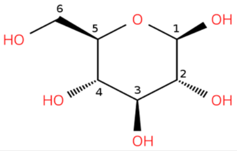
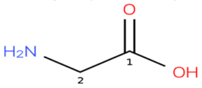
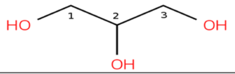
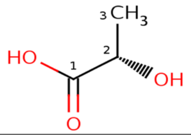
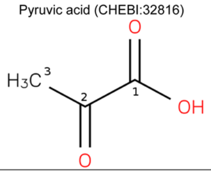
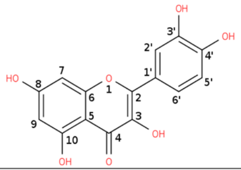
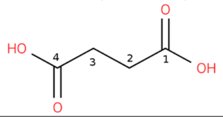
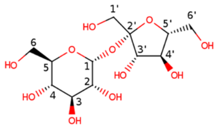
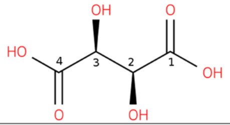
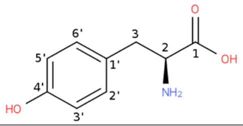
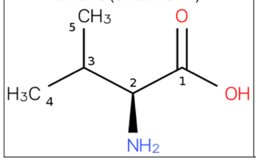
Several studies have found that any meaningful effects of the harvest season on the chemical profiles of Brazilian propolis have been predominantly quantitative. However, one should bear in mind that propolis samples in those studies[9,39−44] originated from a single site (i.e., Botucatu County, São Paulo state) in southeastern Brazil and the analytical tools used, GC-FID and GC-MS, for chemically profiling the samples differ from the analytical techniques and the experimental approach adopted here, where samples originated from many regions. Despite being a more sensitive and selective technique than NMR spectroscopy, gas chromatography is typically a more time-consuming and sample-destructive technique. Besides, it is also possible to combine different NMR approaches (i.e., 1D and 2D) in the same experiment, allowing the metabolic profiling of samples in a more versatile way.

Discriminant Analysis for Seasonal Effects. The aforementioned results indicate a clear seasonal effect on the chemical composition of propolis (Table 1). These findings may be used as a basis to build descriptive and predictive models and prompted a further investigation of the set of features that might be employed to correctly classify propolis samples according to their harvest season, taking into account their chemical variability.

Thus, in a second series of experiments, unsupervised and supervised multivariate statistical methods were applied to the ^{1}H NMR data set to build descriptive and classification models that could extract latent information on interest from the spectroscopic data. Thus, the first stage of these experiments involved the reduction of the dimensionality of the ^{1}H NMR data set by PCA. PC1 and PC2 afforded only 30.7% of the explained variance, and this descriptive model was not effective in providing a clear separation for the samples into the different seasons.

Thus, we adopted a classification model to gain insight into the relevant features associated with an eventual discrimination according to the propolis chemical composition. To extract relevant nonredundant information, we applied PLS-DA to the propolis metabolomics data set. A first estimation of the error, for different numbers of components, has shown that 19 PCs lead to

Table 2. ^{1}H and ^{13}C NMR Chemical Shifts (600 MHz, CD_3OD), Proton Multiplicity, and Coupling Constants (*J*, Hz) for Assigned Compounds Found in Brazilian Propolis Produced in Southern Brazil in 2010

Compound	δ_{ppm} ^{1}H (multiplicity, assignment, *J* in Hz)	δ_{ppm} ^{13}C
Acetic acid (CHEBI:15366) 	2.03 (s, 2-CH_3)	24.05 (C-2)
L-Alanine (CHEBI:16977) 	1.58 (d, 3-CH_3, 7.16); 4.20 (q, 2-CH, 7.21)	19.03 (C-3), 51.10 (C-2)
L-Ascorbic acid (CHEBI:29073) 	3.72 (m, 6-CH_2); 4.50 (d, 4-CH, 1.65)	63.30 (C-6), 77.80 (C-4)
Caffeic acid (CHEBI:36281) 	6.28 (d, 2-CH, 15.81); 6.96 (dd, 2'-CH, 7.81, 1.93); 9.17 (s, 3'-OH); 9.51 (s, 4'-OH)	115.18 (C-2), 146.10 (C-3'), 148.88 (C-4')
Citric acid (CHEBI:30769) 	2.74/2.95 (d, 2, 5-CH_2, 15.14)	43.55 (C-2, 5)
Cinnamic acid (CHEBI:27386) 	6.52 (d, 2-CH, 15.80); 7.67 (m, 2', 6'-CH); 7.59 (d, 3-CH, 15.80); 7.42 (m, 3', 4', 5'-CH)	126.79 (C-3'), 115.53 (C-2), 128.55 (C-2'), 143.55 (C-3)
β-D-Fructose (CHEBI:28645) 	3.57 (m, 1-CH_2); 3.78 (dd, 6-CH_2, 12.72, 1.05); 3.90 (dd, 4-CH, 9.82, 3.42); 4.05 (m, 5-CH); 4.08 (m, 3-CH)	62.65 (C-1), 74.55 (C-4)
Fumaric acid (CHEBI:18012) 	6.53 (s, 2, 3-CH)	135.11 (C-2,3)
GABA (γ-Aminobutyric acid, CHEBI:16865) 	1.90 (q, 3-CH_2); 2.29 (t, 2-CH_2, 7.35); 3.01 (t, 4–CH_2, 7.57)	26.77 (C-3)
Gallic acid (CHEBI:30778) 	7.03 (s, 2, 6-CH)	111.10 (C-2, 6)
β-D-Glucose (CHEBI:15903) 	3.22 (t, 2-CH, 9.42); 3.44 (dd, 5-CH, 2.32, 5.92); 4.66 (d, 1-CH, 7.40)	77.25 (C-2), 77.30 (C-5), 99.10 (C-1)
L-Glycine (CHEBI:15428) 	3.44 (s, 2-CH_2)	42.10 (C-2)
Glycerol (CHEBI:17754) 	3.54 (m, 1-CH_2); 3.72 (t, 2-CH, 6.51); 3.63 (m, 3-CH_2)	65.40 (C-1.3), 74.97 (C-2)
L- Lactic acid (CHEBI:422) 	1.33 (d, 3-CH_3, 6.98); 4.13 (q, 2-CH, 6.95)	22.90 (C-3), 71.10 (C-2)
Pyruvic acid (CHEBI:32816) 	2.32 (s, 3-CH_3)	29.00 (C-3)
Quercetin (CHEBI:16243) 	9.55 (s, 4'-OH); 9.17 (s, 3-OH); 7.71 (d, 2'-CH, 2.10); 6.46 (d, 8-CH, 1.81), 6.20 (d, 6-CH, 1.81)	144.91 (C-3', 5'), 136.09 (C-3), 115.05 (C-2', 6'), 98.99 (C-6)
Succinic acid (CHEBI:15741) 	2.38 (s, 2, 3-CH_2)	33.89 (C-2, 3)
Sucrose (CHEBI:17992) 	5.42 (d, 1-CH, 3.91)	93.07 (C-1)
Tartaric acid (CHEBI:26849) 	4.31 (s, 2, 3-CH)	76.57 (C-2, 3)
L-Tyrosine (CHEBI:17895) 	6.88 (d, 3', 5'-CH, 8.50); 7.16 (d, 2', 6'-CH, 8.50)	118.89 (C-6), 133.80 (C-5)
L-Valine (CHEBI:16414) 	1.02 (d, 5-CH_3, 7.05); 2.29 (m, 3-CH)	20.75 (C-5), 31.20 (C-3)

the best accuracy in the cross-validation procedure, around 87.7%, with a Kappa statistical value of 0.83.

The 3D scores scatter plot shown in Figure 2 provides an overview of the separation of the classes using only the first three

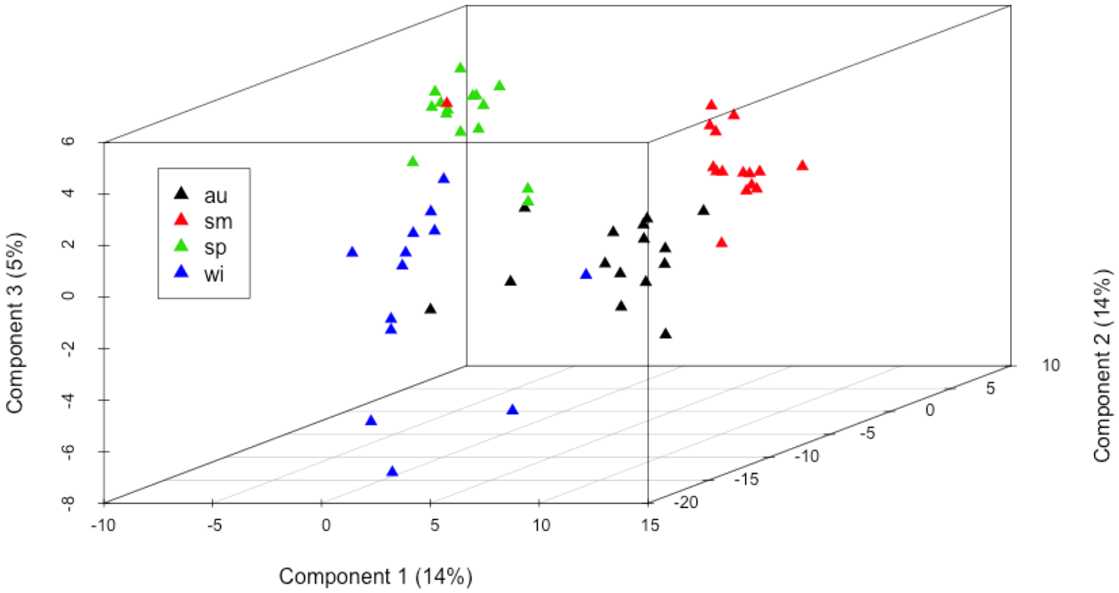


Figure 2. 3D score plot for the selected PCs calculated by PLS-DA from the propolis ^{1}H NMR data set.

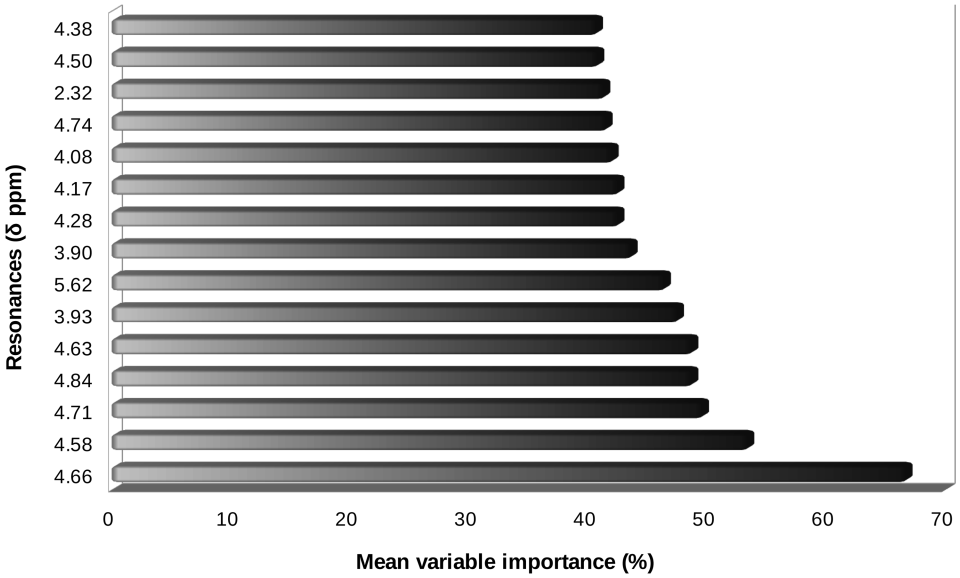


Figure 3. Significant features (^{1}H NMR resonances, ppm) identified by the PLS-DA classification model of propolis samples and their variable importance on the projection scores (VIP). Variables are ranked by mean importance.

components of PLS-DA's best model. The propolis sample classification according to the harvest seasons, afforded by the PLS-DA model, discriminates the samples convincingly. The confusion matrix of the best performing model is given as Supporting Information.

After the PLS-DA model has been built, the influence of individual features was captured by measuring the variable importance on the projection (VIP) derived from the PLS-DA coefficients. The top 15 features selected by the PLS-DA classification algorithm are shown in Figure 3.

Again, as identified by one-way ANOVA and Tukey test analyses, most of the main features accounting for the classification model shown in Figure 3 are resonances occurring at the anomeric region of the ^{1}H NMR spectra (e.g., 4.50, 4.74, 4.63, 4.84, 4.71, 4.58, and 4.66 ppm). Further analysis by visual inspection of the 1D and 2D NMR (TOCSY and HSQC) spectra allowed assignment of resonances to β-D-glucose (4.66 ppm, d, 1β-CH, J = 7.40 Hz), β-D-fructose (3.90 ppm, dd, 4-CH, J = 9.82, 3.42 Hz; 4.08 ppm, m, 3-CH), pyruvic acid (2.32 ppm, s), and L-ascorbic acid (4.50 ppm, d, 4-CH, J = 1.65 Hz). Since these metabolites have shown importance as eventual chemical signatures of propolis samples collected in different seasons as indicated by one-way ANOVA e by PLS-DA, in a follow-up approach qNMR experiments were performed to assess the contents of relevant compounds for propolis sample discrimination. The quantitative data are shown in Table 3, revealing the discrepancy of the samples collected over the seasons in terms of contents of certain metabolites.

The amounts of β-D-glucose differed mostly between the colder seasons with respect to the hotter ones, with virtually no β-D-glucose been found in propolis samples collected in winter. L-Ascorbic acid was also identified as varying in its content, as winter propolis showed ca. 1.6 times lesser amounts compared to the other seasons. Similarly, reduced concentrations of pyruvic,

Table 3. Metabolic Profile and Concentrations (mg/mL) of Metabolites Determined by qNMR in Propolis Samples According to the Season of Harvest and the Agroecological Region of Production in Santa Catarina State, Southern Brazil

	season				agroecological regions		
metabolites	au	wi	sp	sm	plain	plateau	highlands
β-D-glucose	0.013	0.000	0.433	0.493	0.183	0.208	0.335
β-D-fructose	10.690	8.920	8.876	5.421	2.152	7.657	7.817
L-alanine	6.113	4.302	6.876	7.181	3.429	6.560	7.206
L-glycine	0.188	0.114	0.095	0.233	0.147	0.151	0.208
lactic acid	1.546	0.750	0.686	1.357	0.531	1.050	1.634
tartaric acid	0.037	0.034	0.089	0.085	0.039	0.059	0.093
acetic acid	0.727	0.316	0.464	0.820	0.181	0.624	0.854
citric acid	0.431	0.301	0.657	0.575	0.129	0.600	0.486
fumaric acid	0.485	0.552	0.407	0.252	0.219	0.420	0.698
pyruvic acid	1.030	0.657	1.123	1.063	0.272	1.109	1.079
succinic acid	2.133	1.793	1.428	1.680	0.198	2.148	2.150
L-ascorbic acid	4.301	2.653	4.310	4.290	2.839	3.644	5.805
ferulic acid	0.164	0.036	0.075	0.142	0.179	0.070	0.144
caffeic acid	0.031	0.001	0.010	0.023	0.006	0.020	0.022

acetic, citric, ferulic, and caffeic acids were also observed in the winter-collected propolis. Other important features discriminating propolis samples identified by PLS-DA were assigned to β-D-fructose, with lower amounts being detected in the propolis harvested in the summer.

When building classification models, one should bear in mind that PLS-DA is a scale-dependent technique, as the choice of scaling factor affects the features selected.[45] Such a trait might add some constraints to NMR-based metabolomics data where usually the peaks vary greatly in intensity so that the classification models obtained might be unsatisfactory. Taking into account such an assumption, we applied a decision tree-based technique that deals well with differently scaled features[46] to the ^{1}H NMR data set, i.e., RF analysis. This was conducted following the same approach as before for the PLS-DA. The confusion matrix for the RF best model with an overall accuracy in the cross-validation of 82.6% (Kappa statistic of 0.76) is presented as Supporting Information.

The RF analysis allowed the selection of extra and nonredundant features for an accurate classification of Brazilian propolis (Figure 4). Similarly to one-way ANOVA and PLS-DA results, some important features selected by RF analysis belong to the anomeric region of the ^{1}H NMR spectra (4.66, 4.58, 4.71, 4.63, and 4.50 ppm), while new features (e.g., 5.99, 6.03, 5.56, and 6.28 ppm) have also been identified. Besides, nine out of the top 15 features (60%) identified by RF analysis corroborate the PLS-DA findings by VIP measurements. Further analysis of 1D and 2D NMR spectra allowed structural assignment of a number of compounds, confirming the chemical signatures previously found (Table 2) such as β-D-glucose (4.66 ppm, d, 1β-CH, J = 7.40 Hz), pyruvic acid (2.32 ppm, s), L-ascorbic acid (4.50 ppm, d, 4-CH, J = 1.65 Hz), L-alanine (4.20, q, 2-CH, J = 7.21 Hz), β-D-fructose (4.08 ppm, m, 3-CH), and caffeic acid (6.28 ppm, d, 2-CH, J = 15.81 Hz). The L-alanine contents showed a more pronounced discrepancy between the winter and summer propolis samples, again indicating the effect of temperature on the chemical composition of the raw material produced in southern Brazil.

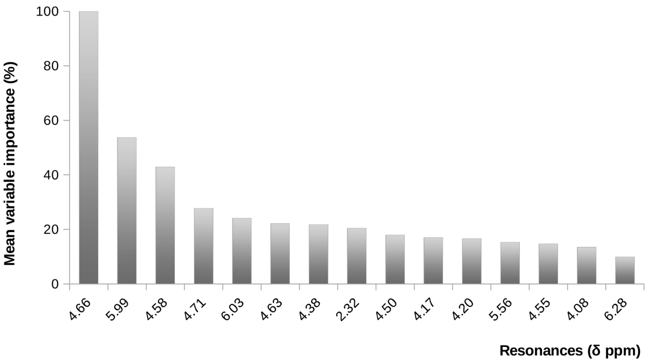


Figure 4. Significant features (^{1}H resonances, ppm) ranked by the mean decrease in classification accuracy from RF analysis.

Propolis is a complex matrix well known for its phenolic constituents, so that the most interesting ^{1}H NMR spectral window is the 5.50−8.25 ppm region, which contains mainly the aromatic compound signals, and 8.25−13.00 ppm, where the hydroxycarbonyl proton signals are found. However, features belonging to those spectral regions did not influence significantly the classification models built by PLS-DA, while in the case of RF a few phenolic constituents appear as relevant. Collectively, the data of metabolite concentrations from Table 3 suggest that propolis samples collected over the seasons are distinct regarding their metabolic profiles. In principle, one could argue that this is because in southern Brazil the seasons are well defined, with their own climatic traits that strongly modulate the flora surrounding the beehives, i.e., the source of resinous material and the chemical composition thereof.

Analysis of the Effect of Agroecological Regions. Aiming to further validate the machine learning and chemometrics metabolomics approach adopted to build descriptive and classification models of Brazilian propolis, a third series of experiments were performed taking into account the NMR data set of propolis samples produced in the regions of interest, e.g., plain, plateau, and highlands, taken here as the classes for classification. We have assumed that discrepancies in propolis chemical composition, resulting from environmental factors distinct among the production regions, might meaningfully affect the performance of the models.

By performing PLS-DA analysis, the method was able to identify important features to predict the propolis sample region by measuring the variable importance as shown in Figure 5. Seven out of the 15 most important ^{1}H NMR resonances identified by PLS-DA result from compounds with an aliphatic

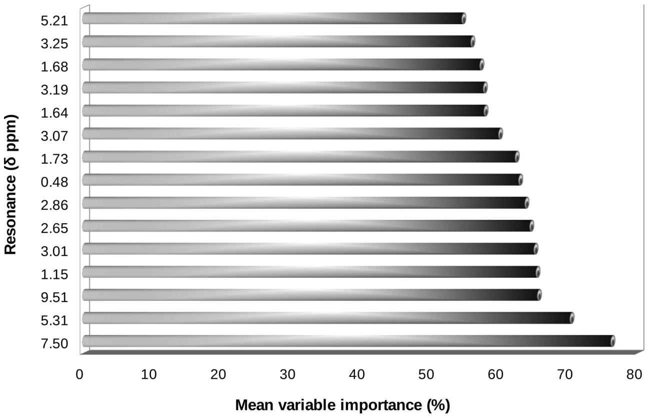


Figure 5. Important features (^{1}H NMR resonances, ppm) ranked according to the VIP score calculated by PLS-DA analysis of propolis samples.

chain, some of them eventually assigned to chemical groups of the alkyl moiety (e.g., 1.64, 1.68, and 1.73 ppm, C−C−CH_2−C and 5.31 ppm of CH_2 of C-9, C-10)[29,31,47] of fatty acids (e.g., palmitic, oleic, stearic, and arachidonic acids) and waxes commonly found in propolis. Two features referred to metabolites occurring in the anomeric region of the ^{1}H NMR spectra, and two were detected in the aromatic region. Collectively, these features suggest the presence of ferulic acid (7.50 ppm, d, 3-CH, *J* = 15.95 Hz), caffeic acid (9.51 ppm, s, 4-′OH), and pinostrobin (2.86 ppm, dd, 3-CH, *J* = 17.1, 3.05 Hz and 3.25 ppm, dd, 3-CH, *J* = 17.15, 12.90 Hz), as previously reported in propolis extracts,[48] as well as *α*-D-glucose (5.21 ppm, d, 1*α*-CH, *J* = 3.80 Hz), and the structures of all compounds were tentatively assigned. The PLS-DA best classification model showed a prediction accuracy of about 79% in the cross validation (with 4 PCs).

The quantification of ferulic acid through qNMR experiments (Table 3) detected amounts more than 2 times higher in propolis samples produced in the plain and highland agroecological regions compared to the plateau. Differences even larger were found for the caffeic acid contents among the geographical regions investigated, but in this case lower concentrations were observed in samples originating from the plain sites. Indeed, this region was characterized by the lowest amounts of all the metabolites investigated, except ferulic acid. In contrast, propolis from the highlands contained higher amounts for most of the compounds relevant for sample discrimination.

In a follow-up set of experiments, RF analysis was applied to the ^{1}H NMR data set allowing the selection of extra and nonredundant features for an accurate classification (Figure 6). A significant number of ^{1}H NMR resonances selected by RF analysis were in accordance with the PLS-DA (six out of 15 resonances are the same, i.e., 40%). Interestingly, features occurring at the anomeric and aromatic regions were more frequent compared to the PLS-DA model. Indeed, resonances from the aromatic moieties (e.g., 6.13, 6.17, 6.75, 6.96, and 9.51 ppm) and from the anomeric region (5.17, 5,21, 5.28, and 5.31 ppm) were predominantly identified by the RF algorithm and tentatively assigned to the phenolic compounds pinostrobin (6.17 ppm, d, 8-CH, *J* = 2.31 Hz; 3.25 ppm, dd, 3-CH, *J* = 17.15, 12.90 Hz) and caffeic acid (6.13 ppm, d, 2-CH, *J* = 16.10 Hz; 6.75 ppm, d, 5′-CH, *J* = 7.95 Hz; 6.96 ppm, dd, 2′-CH, *J* = 7.81, 1.93 Hz; 9.51 ppm, s, 4-′OH), as well as to *α*-D-glucose (5.21 ppm, d, 1*α*-CH, *J* = 3.80 Hz). The mean variable with the highest importance detected, i.e., the resonance at 2.86 ppm (dd, 3-CH, *J* = 17.1, 3.05 Hz, aliphatic region) was associated with pinostrobin. Further MS and MS/MS experiments (data not shown) corroborate the NMR results indicating the presence of the phenolic compounds pinostrobin (*m*/*z* 270.08) and caffeic acid (*m*/*z* 180.05) in propolis samples.[32] ^{1}H NMR and MS reference data of pinostrobin are provided in Supporting Information. The RF supervised learning algorithm obtained a classification accuracy of 75%, not being able to improve the results of the PLS-DA.

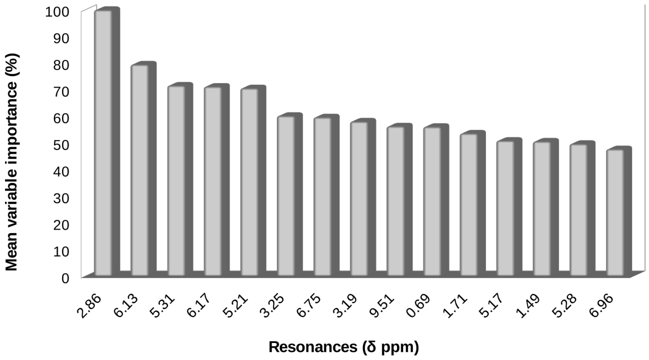


Figure 6. Significant features (^{1}H NMR resonances, ppm) ranked by the variable importance in classification accuracy by RF analysis.

In both models, the classes where the majority of the classification errors occur are the ones with a reduced number of samples, leading us to believe that the improvement of the results would be possible with the collection of further data.

Pinostrobin amounts were determined in the propolis samples with differences having been detected regarding the sites of production. A range from 2.60 to 3.10 mg pinostrobin/mL has been found in the samples produced in the plain and plateau regions, as the propolis from the highlands contained a much lower amount, i.e., 0.090 mg pinostrobin/mL. The sites of production located in the highlands seemed to be more influenced by climatic factors over the seasons, as pinostrobin was not detected in winter-collected samples. Importantly, peak overlapping hampered accurate quantification of pinostrobin for a certain number of samples.

The high dimensionality and multicollinear nature of NMR-based metabolomics data provide significant challenges for feature selection, metabolite annotation, and sample classification. These issues, combined with other factors such as experimental noise, scaling, and threshold selection, may lead to the omission of features relevant to class explanation. To address this issue, a set of feature selection and classification methods was applied to a metabolomics data set to eliminate irrelevant and redundant information regarding the seasonal and geographical origin effects on the chemical composition and classification of Brazilian propolis samples.

The selected classification methods based on machine learning and feature selection appear to be effective for building classification models for Brazilian propolis, with prediction accuracy always above 75% and reaching nearly 90% in the best case. The results of the classifiers were slightly better in the discrimination of the collection seasons, with less convincing results in the discrimination of the regions, mainly due to the imbalance in the number of samples per class.

PLS-DA was able to reach better results comparatively to RF, which could be explained by the collinear nature of the data. When considering the identification of biomarkers for sample discrimination, PLS-DA and RF were complementary approaches by retrieving and expanding the set of discriminating features and by adding relevant information for the identification of the class-determining metabolites. Important features have been identified that address in a more rational basis the quantitative analysis of the assigned metabolites and enable the differentiation of the various propolis samples as to their contents of important bioactive compounds. This allowed further elucidation of the investigated system regarding the metabolic signature of important compounds, i.e., chemical fingerprint, harvest seasons, and regions of production of Brazilian propolis.

Finally, pattern recognition techniques of chemically complex matrices have been claimed to be of interest for selecting samples typically characterized by their high heterogeneity as shown herein for Brazilian propolis. In this context, and in connection with the development of a medicine/dietary supplement with a specific pharmacological activity, the classification methods applied to the NMR data set of propolis permit the selection of a specific set of propolis samples from a certain geographical region/season according to the pertinence of their chemical profiles or target compound (e.g., ferulic acid, pinostrobin) as to the biological activity of interest. Such an approach might eventually optimize the development of the medicine/dietary supplement in a more rational manner, by targeting a certain number of selected samples to a dedicated preclinical assay. It is worth mentioning that a similar approach can be adopted for selecting propolis samples as one aims at development of a dietary supplement with a certain nutritional effect. Propolis samples from elsewhere might be routinely assayed as to their metabolic profiles, target compounds, chemical signatures, and biomarkers when necessary, in quality control pipelines through the NMR-based metabolomics data analysis workflow described herein, making it relevant for biotechnological purposes.

■ EXPERIMENTAL SECTION

General Experimental Procedures. EtOH of analytical grade was purchased from Sync (São Paulo, Brazil). Ultrapure H_2O was obtained through a reverse-osmosis system (Permution E-10, Curitiba, Brazil). The methanol-d_4 was purchased from TediaBrazil (Rio de Janeiro, Brazil), and sodium 3-(trimethylsilyl)propionate-d_4 (TSP-d_4) and deuterium chloride solution (DCl 35 wt % in D_2O, 99 atom % D) were obtained from Sigma-Aldrich (Saint Louis, MO, USA).

Samples. Propolis samples were collected in the autumn (au), winter (wi), spring (sp), and summer (sm) of 2010 from *Apis mellifera* hives located in southern Brazil (Santa Catarina state). A total of 59 samples were collected, divided by the collection seasons as follows: sm, 16 samples; au and sp, 15 samples each; wi, 13 samples. The southern region in Brazil typically presents well-marked seasons regarding temperature, photoperiod, rainfall, and sunlight intensity, i.e., a set of ecological factors that modulate the regional flora and accordingly the source of plant resins collected by honeybees for propolis production.

The apiaries were divided into three agroecological regions—plain, plateau, and highlands—derived from their temperature and altitude conditions. The number of samples per region was unequal (plain, 11; plateau, 36; and highlands, 12), due to the number of available apiaries in each.

Propolis was collected preferentially at noon from a special woody frame collector developed in southern Brazil, containing an open central rectangular window where the bees were induced to deposit the propolis to seal the opening. The collectors were inserted under the hive's cover, and the propolis was extracted by carefully scratching the inner opening of the rectangular frame collectors, followed by the removal of any eventual debris of wood and bees.

The propolis samples (~100 g/sample) were stored at −20 °C for further analysis. Prior to extraction, the samples were added to liquid N_2, ground, and homogenized. Afterward, to the samples (2 g) was added 10 mL of an 80% EtOH (v/v) solution and extracted for 1 h at room temperature, with continuous stirring and protection from light. The aqueous−EtOH extract was filtered on a cellulose membrane under reduced pressure, the volume was made up to 10 mL with 80% EtOH (v/v) solution, and the solution was frozen at −80 °C and freeze-dried.

1D and 2D NMR Spectroscopy. To the lyophilized samples were added 700 μL of methanol-d_4 containing 0.024% of sodium 3-(trimethylsilyl)propionate-d_4 (TSP-d_4) as internal standard for solubilization, the solution was centrifuged (5000 rpm/5 min), and 650 μL of the supernatant was collected and transferred to 5 mm NMR tubes for further 1D and 2D NMR analyses. The pH of the samples was adjusted to 3.48 with a deuterium chloride solution (35 wt % in D_2O, 99 atom % D).

A Varian Inova 600 MHz NMR spectrometer, operating at 599.89 MHz for ^{1}H and at 150.85 MHz for ^{13}C NMR, using VNMRJ software, was used for the NMR experiments. The chemical shifts are expressed in δ (parts per million) referenced to the TSP peak at $\delta(^1H)$ 0.00 ppm and $\delta(^{13}C)$ 0.00 ppm. The 2D total correlation spectroscopy (TOCSY, ^{1}H/^{1}H NMR) and (HSQC) ^{1}H/^{13}C NMR experiments were performed using conventional sequence pulses with H_2O suppression of −4.87 ppm (watergate, wgtocsy, and gChsqc).

All experiments were performed at 300 K with no spinning. The high-resolution ^{1}H NMR spectra were measured at 599.89 MHz, by collecting 32K data points (time domain), with a 10 μs (90°) rf pulse, acquisition time 5.0 s, delay time 1.0 s, mixing time 100 ms for saturation of H_2O, 4 dummy scans, and 32 total scans. The spectral width was 9330 Hz, and digital resolution was ±0.3228 Hz/point. The Varian Inova 600 MHz NMR spectrometer was set up for an automatic receiver gain adjustment. The function GAIN was set to "*N*", i.e., not used, the acquisition was started, a number of trial FIDs were recorded to determine the best gain value, and then the acquisition began. For the quantification of the metabolites (qNMR) all spectra were recorded as the sum of 32 free induction decays into 32K complex data points, at 300 K, using a spectral width of 9330 Hz (native resolution = 0.323 Hz/point), acquisition time of 1.67 s, relaxation delay (*D*1) of 7.00 s (total recycling time 8.67 s), and pulse duration of 10 μs (90°).

For the TOCSY spectra the acquisition parameters were as follows: number of scans, 16; dummy scans, 4; time domain, 1024 (F2) and 521 (F1) data points; spectral width, 7992.19 Hz in F2 and 7984.38 Hz in F1; digital resolution, 1.1120 Hz; acquisition time, 0.1280 s (F2) and 0.0320 s (F1); mixing time, 100 ms; delay time 0.5 s; total acquisition time, 8 h and 11 min.

The acquisition parameters for the HSQC spectra, acquired with inverse detection and ^{13}C NMR decoupling during acquisition, were as follows: number of scans, 32; dummy scans, 4; 1024 and 512 data points (time domain) in F2 (^{1}H) and F1 (^{13}C), respectively; spectral width, 7992.19 Hz (F2) and 21080.12 Hz (F1); digital resolution, 1.7206 Hz (F2) and 200.27 Hz (F1); acquisition time, 0.1280 s (^{1}H) and 0.0121 s (^{13}C); delay time, 0.5 s; delay time, 60.5 ms; total acquisition time, 11 h and 23 min.

NMR Data Processing. The ^{1}H NMR data set was collected and processed using a routine procedure implemented in ACD/NMR processor software (Advanced Chemistry Development, release 12.01) consisting of Fourier transforming the 32K data points as the signal-to-noise ratio of the spectra was improved by multiplying each free induction decay with an additional exponential factor corresponding to 0.3 Hz. The resulting ^{1}H NMR spectroscopic profiles were automatically phased (Ph0 and Ph1), manually baseline corrected, and referenced to the internal standard (TSP, δ_{1H} 0.00 ppm), and the relevant spectroscopic information was extracted as a peak intensity list considering a signal−noise ratio for detecting peaks higher than 5.0. Resonances at 3.29−3.31 and 4.85−5.00 ppm, containing the methanol-d_4 and H_2O signals and the internal standard used (TSP) to calibrate and normalize each ordinate allowing quantitative comparison of spectra, were removed from the data set for further analysis. TOCSY and HSQC spectra were processed by applying a squared sine function and squared sine constant in both the F1 and F2 dimension after Fourier transforming the data set and calibration. For the qNMR experiments, FIDs were processed before the Fourier transformation using an exponential multiplication window function with a line-broadening factor of 0.5 Hz, permitting a compromise between both sensitivity (threshold signal-to-noise ratio = 5) and resolution. All spectra were zero filled to 128K data points Quantitative analyses of the identified metabolites were performed after integration of the well-separated resonances relative to the integration of the internal standard TSP peak at $\delta(^1H)$ 0.00 ppm (0.024%) as follows: β-D-glucose (4.66 ppm), β-D-fructose (3.57 ppm), L-alanine (1.58 ppm), L-glycine (3.44 ppm), lactic acid (1.33 ppm), tartaric acid (4.31 ppm), acetic acid (2.03 ppm), citric acid (2.74 ppm), fumaric acid (6.53 ppm), pyruvic acid (2.32 ppm), succinic acid (2.38 ppm), L-ascorbic acid (4.50 ppm), ferulic acid (7.50 ppm), and caffeic acid (9.17 ppm). For the peak integration of the target compounds spectroscopic regions were defined for each of the metabolites and the integral area of these regions was calculated using scripts created and run in ACD/NMR processor software (Advanced Chemistry Development, release 12.01).

Statistical Analysis and Chemometrics. From the processed full spectra data set (0.20−13.00 ppm), a peak list was extracted into a file (comma separated values format .csv), where the first column indicates peak positions (ppm) and the second one represents peak intensities. A set of 59 samples was used, containing a total of 25 403 peaks with an average of 430.6 peaks per sample. Peak alignment grouped proximal peaks together according to their position, using a moving window of 0.03 ppm. Peaks of the same group were aligned to their median positions across all samples, and those detected in very few samples (<25% of the samples) were excluded. In addition, the missing values were replaced with a value of 0.00005, half of the minimum positive values in the original data, assumed to be the detection limit. Indeed, most missing values are caused by low-abundance metabolites with contents lower than the detection limit. Taking into account the distinct orders of magnitude of the variables, a generalized logarithmic transformation of the data was performed followed by data standardization (mean-centering each variable and dividing by its standard deviation).

The ^{1}H NMR data set was analyzed using statistical univariate techniques, in this case ANOVA and the *post hoc* Tukey test, to detect eventual statistical differences ($p < 0.05$) derived from the effects of propolis harvest seasons or regions on the propolis spectroscopic profiles, i.e., chemical composition of that complex matrix. Then, descriptive and classification models were built by applying unsupervised and supervised classification methods. The first step was to perform PCA over the data set, followed by supervised approaches.

Partial least-squares discriminant analysis is a supervised method frequently used for classification in the metabolomics area,[21,22] being an extension of PCA that takes advantage of class information to maximize the separation between groups of observations. It works by uncovering the latent variables within the data that both model the feature values and separate the sample classes.

In turn, Random Forest is a supervised learning algorithm suitable for multidimensional data analysis. It uses an ensemble of classification trees, each grown by random feature selection from a bootstrap sample at each branch. Class prediction is based on the majority vote of the ensemble. RF also provides other useful information such as variable importance measures. Interestingly, there is a fundamental difference between RF and PLS-DA. Unlike PLS-DA, RF is a nonparametric technique and is unaffected by feature scale. For this reason, these techniques may be seen to be somewhat complementary.

A number of scripts using the R scientific computing system (http://www.r-project.org) were developed for data analysis, which provides a framework for conducting analyses over metabolomics data sets.[49] These scripts were used to perform the analysis, including ANOVA, PCA, PLS-DA, and RF analyses. For PCA, the *prcomp* function was used, which takes advantage of singular value decomposition for the computation. Regarding the machine learning methods (PLS-DA and RF), the package Caret was selected, building a wrapper over a number of different machine learning methods. In both cases, an estimation of the error of the classifier was computed using a 10 times 10-fold cross validation scheme, for different configurations of the internal parameters (for PLS-DA, the number of components was varied from 1 to 20, while for RF the *mtry* parameter, number of variables randomly sampled as candidates at each split, was tuned over 20 different values in the range from 2 to 242). Afterward, for each of the methods, the variable importance is estimated for each possible feature (resonance) and the top features are ranked.

The main analyses and results are given in the Supporting Information. The markdown features of the R language were used to create a data report analysis, including runnable chunks of code, making all the data analysis process fully reproducible (both the source R markdown file and the generated report are given in the online materials).

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jnatprod.5b00315.

> The confusion matrices of the best performing models built (e.g., PLD-DA and RF) and the main analyses and additional results; NMR and MS data of the identified pinostrobin compound; the markdown features of the R language were used to create the data report analysis (in the HTML format) (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*Phone: +55-48-3721-4812. Fax: +55-48-3721-5333. E-mail: m.maraschin@ufsc.br.

Author Contributions

$^\perp$M. Maraschin and A. Somensi-Zeggio contributed equally.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

Financial support for this investigation by National Council for Scientific and Technological Development (CNPq), Coordination for the Improvement of Higher Education Personnel (CAPES), Brazilian Biosciences National Laboratory (LNBio-CNPEM/MCTI), Foundation for Support of Scientific and Technological Research in the State of Santa Catarina (FAPESC), and Portuguese Foundation for Science and Technology (FCT) is acknowledged. The research fellowship granted by CNPq to the first author is also acknowledged. The work was partially funded by a CNPq and FCT agreement through the PropMine grant.

■ REFERENCES

(1) Wollenweber, E.; Hausen, B. M.; Greenaway, W. *Bull. Liaison Groupe Polyphenols* **1990**, *15*, 112−120.

(2) Marcucci, M. C. *Apidologie* **1995**, *26*, 83−99.

(3) Burdock, G. A. *Food Chem. Toxicol.* **1998**, *36*, 347−363.

(4) Banskota, A. H.; Tezuka, Y.; Kadota, S. H. *Phytother. Res.* **2001**, *15*, 561−571.

(5) Yildirim, Z.; Hacievliyagil, S.; Kutlu, N. O.; Aydin, E. N.; Kurkcuoglu, M.; Iraz, M.; Durmaz, R. *Pharmacol. Res.* **2004**, *49*, 287−292.

(6) Vardar-Ünlü, G.; Silici, S.; Ünlü, M. *World J. Microbiol. Biotechnol.* **2008**, *24*, 1011−1017.

(7) Kumazawa, S.; Ueda, R.; Hamasaka, T.; Fukumoto, S.; Fujimoto, T.; Nakayama, T. *J. Agric. Food Chem.* **2007**, *55*, 7722−7725.

(8) Gekker, G.; Hu, S.; Spivak, M.; Lokensgard, J. R.; Peterson, P. K. *J. Ethnopharmacol.* **2005**, *102*, 158−163.

(9) Sforcin, J. M. *J. Ethnopharmacol.* **2007**, *113*, 1−14.

(10) Tan-No, K.; Nakajima, K. T.; Shoii, T.; Nakagawasa, O.; Nijima, F.; Ishikawa, M.; Endo, Y.; Sato, S.; Tadano, T. *Biol. Pharm. Bull.* **2006**, *29*, 96−99.

(11) Awale, S.; Li, F.; Onozuka, H.; Esumi, H.; Tezuka, Y.; Kadota, S. *Bioorg. Med. Chem.* **2008**, *16*, 181−189.

(12) Chen, J.; Long, Y.; Han, M.; Wang, T.; Chen, Q.; Wang, R. *Pharmacol., Biochem. Behav.* **2008**, *90*, 441−446.

(13) Tosi, E. A.; Ré, E.; Ortega, M. E.; Cazzoli, A. F. *Food Chem.* **2007**, *104*, 1025−1029.

(14) Bankova, V. S.; de Castro, S. L.; Marcucci, M. C. *Apidologie* **2000**, *31*, 3−15.

(15) Salatino, A.; Teixeira, E. W.; Negri, G.; Message, D. *eCAM* **2005**, *2*, 33−38.

(16) König, B. *Bee World* **1985**, *66*, 136−139.

(17) Marcucci, M. C.; Bankova, V. S. *Curr. Top. Phytochem.* **1999**, *2*, 115−123.

(18) Park, Y. K.; Alencar, S. M.; Aguiar, C. L. *J. Agric. Food Chem.* **2002**, *50*, 2502−2506.

(19) Cuesta-Rubio, O.; Frontana-Uribe, B. A.; Ramırez-Apan, T.; Cardenas, J. *Z. Naturforsch., C: J. Biosci.* **2002**, *57*, 372−378.

(20) Yang, J.; Xu, G.; Hong, Q.; Liebich, H. M.; Lutz, K.; Schmülling, R. M.; Wahl, H. G. *J. Chromatogr. B: Anal. Technol. Biomed. Life Sci.* **2004**, *813*, 53−58.

(21) Bryan, K.; Brennan, L.; Cunningham, P. *BMC Bioinf.* **2008**, *9*, 470.

(22) Wold, S.; Sjöström, M.; Eriksson, L. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109−130.

(23) Fan, X.; Bai, J.; Shen, P. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* **2006**, *6*, 6081−6084.

(24) Goodacre, R.; Kell, D.; Bianchi, G. *J. Sci. Food Agric.* **1993**, *63*, 297−307.

(25) Holmes, E.; Nicholson, J.; Tranter, G. *Chem. Res. Toxicol.* **2001**, *14*, 182−191.

(26) Hattori, H.; Okuda, K.; Murase, T.; Shigetsura, Y.; Narise, K.; Semenza, G. L.; Nagasawa, H. *Bioorg. Med. Chem.* **2011**, *19*, 5392−5401.

(27) Simmler, C.; Napolitano, J. G.; McAlpine, J. B.; Chen, S. N.; Pauli, G. F. *Curr. Opin. Biotechnol.* **2014**, *10*, 51−59.

(28) Park, Y. K.; Ikegaki, M.; Alencar, S. M.; Moura, F. F. *Honeybee Sci.* **2000**, *21*, 85−90.

(29) Fan, T. W. M.; Lane, A. N. *Prog. Nucl. Magn. Reson. Spectrosc.* **2008**, *52*, 69−117.

(30) Fan, T. W. M. *Prog. Nucl. Magn. Reson. Spectrosc.* **1996**, *28*, 161−219.

(31) Waterman, P. G.; Mole, S. *Analysis of Phenolic Plant Metabolites*; Blackwell Scientific Publications: Oxford, 1994; p 238.

(32) Bertelli, D.; Papotti, G.; Bortolotti, L.; Marcazzanb, G. L.; Plessia, M. *Phytochem. Anal.* **2012**, *23*, 260−266.

(33) Meneguelli, C.; Joaquim, L. S. D.; Félix, J. L. Q.; Somensi, A.; Tomazolli, M.; Silva, D. A.; Berti, F. V.; Velerinho, M. B. R.; Recouvreux, D. O. S.; Zeri, A. C. M.; Maraschin, M. *Microvasc. Res.* **2013**, *88*, 1−11.

(34) Fontana, J. D.; Passos, M.; Santos, M. H. R.; Fontana, C. K.; Oliveria, B. H.; Schause, L.; Pontarolo, R.; Barbirato, M. A.; Ruggiero, M. A.; Lanças, F. M. *Chromatographia* **2000**, *52*, 147−151.

(35) Falcão, S. I.; Vilas-Boas, M.; Estevinho, L. M.; Barros, C.; Domingues, M. R. M.; Cardoso, S. M. *Anal. Bioanal. Chem.* **2010**, *396*, 887−897.

(36) Popova, M.; Trusheva, B.; Antonova, D.; Cutajar, S.; Mifsud, D.; Farrugia, C.; Tsvetkova, I.; Najdenski, H.; Bankova, V. *Food Chem.* **2011**, *126*, 1431−1435.

(37) Hernández, I. M.; Cuesta-Rubio, O.; Fernández, M. C.; Perez, A. R.; Porto, R. M. O.; Piccinelli, A. L.; Rastrelli, L. *J. Agric. Food Chem.* **2010**, *58*, 4725−4730.

(38) Watson, D. G.; Peyfoon, E.; Zheng, L.; Lu, D.; Seidel, V.; Johnston, B.; Parkinson, J. A.; Fearnley, J. *Phytochem. Anal.* **2006**, *17*, 323−331.

(39) Bankova, V.; Boudourova-Krasteva, G.; Popov, S.; Sforcin, J. M.; Funari, S. R. C. *Apidologie* **1998**, *29*, 361−367.

(40) Sforcin, J. M.; Fernandes, A., Jr.; Lopes, C. A. M.; Bankova, V.; Funari, S. R. C. *J. Ethnopharmacol.* **2000**, *73*, 243−249.

(41) Sforcin, J. M.; Fernandes, A., Jr.; Lopes, C. A. M.; Funari, S. R. C.; Bankova, V. *J. Venomous Anim. Toxins* **2001**, *7*, 139−144.

(42) Sforcin, J. M.; Kaneno, R.; Funari, S. R. C. *J. Venomous Anim. Toxins* **2002**, *8*, 19−29.

(43) Sforcin, J. M.; Novelli, E. L. B.; Funari, S. R. C. *J. Venomous Anim. Toxins* **2002**, *8*, 244−254.

(44) Sforcin, J. M.; Orsi, R. O.; Bankova, V. *J. Ethnopharmacol.* **2005**, *98*, 301−305.

(45) Van den Berg, R. A.; Hoefsloot, H. C.; Westerhuis, J. A.; Smilde, A. K.; Van der Werf, M. J. *BMC Genomics* **2006**, *7*, 142.

(46) Weljie, A.; Newton, J.; Mercier, P.; Carlson, E.; Slupsky, C. *Anal. Chem.* **2006**, *78*, 4430−4442.

(47) Leyden, D. E.; Cox, R. H. *Analytical Applications of NMR*; John Wiley & Sons: New York, 1977; p 456.

(48) Papotti, G.; Bertelli, D.; Plessi, M.; Rossi, M. C. *Int. J. Food Sci. Technol.* **2010**, *45*, 1610−1618.

(49) Xia, J.; Psychogios, N.; Young, N.; Wishart, D. S. *Nucleic Acids Res.* **2009**, *37*, W652−W660.