A Machine Learning and Chemometrics Assisted Interpretation of Spectroscopic Data – A NMR-Based Metabolomics Platform for the Assessment of Brazilian Propolis

Marcelo Maraschin[1,3], Amélia Somensi-Zeggio[1], Simone K. Oliveira[1], Shirley Kuhnen[1], Maíra M. Tomazzoli[1], Ana C.M. Zeri[2], Rafael Carreira[3], and Miguel Rocha[3,*]

[1] Plant Morphogenesis and Biochemistry Laboratory Federal University of Santa Catarina, Florianópolis, SC, Brazil
[2] National Laboratory of Bioscience, Campinas, SP, Brazil
[3] CCTC, School of Engineering, University of Minho, Campus Gualtar, Braga, Portugal
mrocha@di.uminho.pt

Abstract. In this work, a metabolomics dataset from ^{1}H nuclear magnetic resonance spectroscopy of Brazilian propolis was analyzed using machine learning algorithms, including feature selection and classification methods. Partial least square-discriminant analysis (PLS-DA), random forest (RF), and wrapper methods combining decision trees and rules with evolutionary algorithms (EA) showed to be complementary approaches, allowing to obtain relevant information as to the importance of a given set of features, mostly related to the structural fingerprint of aliphatic and aromatic compounds typically found in propolis, e.g., fatty acids and phenolic compounds. The feature selection and decision tree-based algorithms used appear to be suitable tools for building classification models for the Brazilian propolis metabolomics regarding its geographic origin, with consistency, high accuracy, and avoiding redundant information as to the metabolic signature of relevant compounds.

Keywords: Supervised classification techniques, evolutionary algorithms, Random Forest, PLS-DA, wrapper methods, NMR-based metabolomics.

1 Introduction

One and two dimensional NMR spectroscopy (1D-, 2D-NMR) has increasingly been used for complex matrix analysis such as plant extracts and biofluids in metabolomics studies. From a ^{1}H-NMR spectrum, a set of peaks, or features, indicative of the metabolite signatures and chemical composition of the sample is obtained and may be used as a basis to build descriptive and predictive models (e.g. for classification tasks). In this context, feature selection may be employed to improve classification accuracy or

* Corresponding author.

S. Tetsuo et al. (Eds.): PRIB 2012, LNBI 7632, pp. 129–140, 2012.

aid model explanation by establishing a subset of class discriminating features. Factors such as experimental noise and threshold selection may adversely affect the set of selected features. Furthermore, the high dimensionality and multi-collinearity inherent to ^{1}H-NMR signals may increase discrepancies between the set of features retrieved and those required to provide a complete explanation of metabolite signatures. Thus, previously to classification of metabolomics data, it is interesting to perform descriptive studies, e.g. using principal component analysis (PCA) [1].

Discriminant analyses such as soft independent modelling by class analogy (SIMCA), support vector machine (SVM), partial least squares discriminant analysis (PLS-DA), and more recently random forests (RF) have also been used within the metabolomics domain.

Feature selection may be employed to improve a classification model in terms of generalization, performance, and accuracy by eliminating non-informative features, as well as to gain deeper insights into the rationale underlying class divisions 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. However, features selected on the basis of classification accuracy, i.e. features that are sufficient to separate classes, may not to always be the best approach due to the redundancy of information. This is typically found in high dimensional NMR-based metabolomics studies, where a metabolite may be represented by one or more spectral features as only a part of the metabolite signature identification may be enough to provide a perfect classification model.

In this work, to overcome such constraints we have adopted an approach where accuracy based approaches are complemented with feature selection methods less prone to the bias effects of multi-collinear features, including those based on variable influence on the projection (VIP) values, derived from PLS-DA and variable importance produced by a RF classifier. Indeed, contrarily to PLS-DA, RF is a non-parametric technique unaffected by feature scale so that the techniques seem to be somewhat complementary.

PLS extracts the set of latent variables which model the data, but which are also correlated to the class membership vector. Once a PLS model has been built the influence of individual features is captured by measuring the VIP scores derived from the PLS coefficients for the optimal set of features. After that, features are ranked by these scores and selected considering the choice of an appropriate threshold (usually $\alpha \geq 1$), a step that may greatly affect the set of retrieved features. Finally, PLS-DA is also a scale dependent technique as the choice of scaling factor affects the features selected [2].

In its turn, RF is a classification technique based on growing many classification trees, in which feature values are used to build a model that enables the classification of unlabeled samples. RF allows assigning importance values to features resulting from their influence on the classification accuracy of the forest, aiding feature selection, and allowing gaining further insights into the data. The importance of a particular feature is determined by randomly permuting the feature over samples in each tree's 'out-of-bag' test set, followed by the reclassification of the samples using the RF. Such a calculation approach is advantageous for feature selection because it covers both the impact of each feature individually and its multivariate interactions with other features. Besides, as RF is a decision tree-based technique it also deals well with

differently scaled features [3], a relevant trait for NMR-based metabolomics where the peaks vary greatly in intensity.

An alternative approach for feature selection is the use of wrapper methods. In wrapper approaches, the feature selection processes are performed by optimization algorithms that search the space of possible subsets of attributes, to find the best alternative. These approaches train the classifier with a subset of the available attributes and estimate its generalization error. These methods are dependent on the classifier that is used. Indeed, there is no guarantee that an optimal subset of attributes chosen for one classifier will be the optimal one when used with another algorithm.

The wrapper approach followed in this work is based on two components: the use of classifiers implemented by the open-source data mining software Weka [4] for the inner layer (decision trees and rule set induction methods will be used), and the use of Evolutionary Algorithms (EAs) as the optimization engine. Together, these techniques may allow extracting relevant features from a given dataset, minimizing the redundant information as to metabolite signature identification. This work aimed at proving the later assertion as our scientific hypothesis, using a high dimensional, multi-collinear metabolomics dataset (80 samples x 81675 variables) of Brazilian propolis NMR spectra as a study model.

Propolis has been chosen because it has long been recognized as a useful source of valuable compounds for human health, but due to its huge chemical heterogeneity, the production of standardized and homogeneous extracts is a difficult task. This is due to the fact that chemical characterization and standardization of propolis extract is technically tedious, time expensive, and non-cost effective as one adopt traditional analytical selective techniques such as high performance liquid chromatography. Besides, the effect of flora composition on the propolis' chemical profile is well known and considering the huge biodiversity of plant species found in some producer regions [5], e.g., Atlantic Rainforest in Santa Catarina State, southern Brazil, one could expect a high chemical heterogeneity among samples from distinct geographic regions where propolis has been collected; an important underlying assumption addressed in this study.

On the other hand, over the past years nuclear magnetic resonance (NMR) spectroscopy has been recognized as a powerful tool as one aims at characterizing chemically complex matrices. Indeed, NMR spectroscopy is a fast, robust, and non-selective analytical technique able to detect virtually any molecule in a solution, given a minimum value of concentration (detection limit, ug/ml). However, the amount of information afforded by NMR analysis is huge as a typical high dimensional ^{1}H-NMR spectrum easily contains 32.000 or 64.000 data points. The analysis of such an amount of information is unthinkable without the aid of powerful computational tools, but one should bear in mind this scenario for metabolomics studies.

In order to deal with large NMR datasets, data mining techniques have been adopted to build descriptive and predictive models. Here, machine learning and chemometrics techniques are thought to be a suitable approach to gain insights as to important spectroscopic features associated to the chemical composition and geographic origin of propolis produced in Santa Catarina state, southern Brazil. For that, emphasis will be given to accurate feature selection and classification techniques in order to avoid retrieving redundant information (i.e., multi-collinear features) and overfitting in classification models by using prominent machine learning algorithms.

2 Methods

2.1 Propolis Sample Preparation and NMR Spectroscopy

In autumn, 2010, propolis samples (n=16) were collected from each of the five geographic regions (East, Central, Highlands, North, and West) of Santa Catarina State, southern Brazil. The lyophilized ethanolic extracts (2g/10 ml, EtOH 70%, v/v) were added of 700 µl of CD_3OD, centrifuged (5 krpm/10min), and transferred to 5 mm NMR tubes. The propolis ^{1}H-NMR spectra were acquired on a Varian Inova 600 MHz NMR spectrometer by collecting (time domain) 32,000 data points (32 scans, acquisition time = 4s, delay time = 2s, recycle time = 6s, 25ºC) over a spectral window of 8000 Hz, and water signal suppression. The recycle time was considered of sufficient length (e.g. 3 T_1 s) to avoid significant (<10%) peak saturation. Prior to Fourier transformation (FT), the 1D FIDs were zero-filled to 64K data points and a line broadening factor of 0.5 Hz was applied. A routine implemented in the ACD/NMR processor software (v.12.01) consisting of phasing, baseline correction, and calibration ($TSP_{\delta H}$ 0.00ppm) was used for processing all the ^{1}H-NMR spectra. Each relevant peak, i.e., selected feature, in the spectrum was integrated using a quantitation script of the Quanalyst tool of ACD/NMR processor software.

2.2 Metabolomics Data Processing

From the processed full spectra dataset (0.80 – 13.00 ppm) a peak list was extracted using a two-column comma separated values format, where the first column indicates peak position (ppm) and the second one represents peak intensities. A set of 80 samples was used, containing a total of 81675 peaks with an average of 425.4 peaks per sample. Peak alignment grouped proximal peaks together according to their position using a moving window of 0.03ppm and a step of 0.015ppm. Peaks of the same group were aligned to their median positions across all samples and those detected in very few samples (< 50% in both classes) were excluded. Besides, the missing and zero values were replaced with a value of 0.00005, the half of the minimum positive values in the original data, assuming to be the detection limit. Indeed, most missing values are caused by low abundance metabolites with contents lower than the detection limit.

In order to identify and remove variables that are unlikely to be of use when modeling data, a filtering protocol was applied based on interquantile range, affording a 5% reduction in features. No phenotype information was used in the filtering process, allowing the result to be used in any downstream analysis. Such processing step is strongly recommended for datasets with large number of variables (> 250) containing much noise [6] as typically found in NMR-based metabolomics analysis. Taking into account the very distinct orders of magnitude of the variables, quantile normalization within replicates of the dataset was performed [7].

2.3 Statistical and Machine Learning Data Analysis

In order to extract latent information from the ^{1}H-NMR dataset, classification models were built by applying supervised classification and feature selection methods. The MetaboAnalyst 2.0 tool provides a framework for conducting analyses over metabolomics

datasets and was used to perform PLS-DA and RF analysis [8]. The wrapper approach was implemented by combining classifiers from the Weka open-source data mining software (v. 3.6.6) [4] and EAs were implemented using the Java open-source library JECoLi (http://darwin.di.uminho.pt/jecoli).

The methods used in this work are described in detail next:

PLS-DA: PLS is a supervised method that uses multivariate regression techniques to extract via linear combination of original variables (X) the information that can predict the class membership (Y). To assess the significance of class discrimination, a permutation test is performed. In each permutation, a PLS-DA model is built between the data (X) and the permuted class labels (Y) using the optimal number of components determined by cross-validation for the model based on the original class assignment. Further variable importance in projection (VIP), a weighted sum of squares of the PLS loadings taking into account the amount of explained Y-variation in each dimension was measured for purpose of calculation of the feature importance.

The PLS regression was performed using the *plsr* function provided by R *pls* package. The classification and cross-validation were performed using the corresponding wrapper function of the *caret* package [9, 10].

Random Forest (RF): Random Forest is a supervised learning algorithm suitable for high dimensional data analysis. It uses an ensemble of classification trees, each of which is 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 OOB (out-of-bag) error, variable importance measure, and outlier measures. During tree construction, about one-third of the instances are left out of the bootstrap sample. This OOB data is then used as test sample to obtain an unbiased estimate of the classification error. Variable importance is evaluated by measuring the increase of the OOB error when it is permuted. The outlier measures are based on the proximities during tree construction. RF analysis was performed using the *randomForest* package for R [11].

Wrapper Approach - Weka Classifiers and Evolutionary Algorithms: An EA is used to evolve the best set of attributes for the classification task, using a set-based representation to encode each solution. Regarding the reproduction operators, two types were used: crossover and mutation. The crossover operator used was inspired on uniform crossover and works as follows: the genes that are present in both parent sets are kept in both offspring; the genes that are present in only one of the parents are sent to one of the offspring, selected randomly with equal probabilities. Regarding mutation, the random mutation operator was deployed, replacing a gene in the set by a random value in the allowed range. Both reproduction operators are used with equal probabilities to create new solutions. The operators are implemented taking into consideration the need to comply with the constraints imposed by the minimum and maximum set size and also to avoid repeated elements in the sets. In the experiments reported in this work, the minimum size is always set to 1 and the maximum size to 10.

The selection procedure is a tournament scheme with k=2. In each generation, 50% of the individuals are kept from the previous generation and 50% are bred by the application of the reproduction operators. An elitism value of 1 is used, allowing the best individual of the population to be always kept. The EA's population size is set to 100 and the termination criterion was defined based on a maximum of 100 generations. The EA was executed 30 times for each case.

Each solution in the EA is evaluated by retrieving the attributes encoded in its genome and building classifiers based solely on those attributes. These classifiers are built and evaluated resorting to Weka and therefore it is easy to select different classifiers implementing distinct data mining algorithms. In this work, we used J48, a classification decision tree induction method based on the well known C4.5 algorithm and JRip, a rule set induction method inspired in the RIPPER algorithm. The fitness function of each solution is computed calculating an accuracy estimation of the classifier, obtained by performing 5-fold cross-validation over the available dataset.

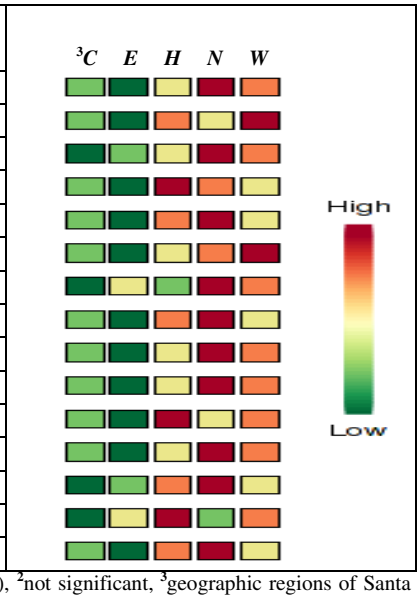
3 Results

The dataset for this classification task includes 80 samples with five classes, one per each geographic region. The dataset is balanced since there are 16 samples for each class. The aim is to classify samples regarding their geographic region.

Previously to PLS-DA and RF analyses, a descriptive model was built based on the calculation of the principal components (PCAs) for the ^{1}H-NMR dataset as previously suggested [1]. PC1 and PC2 afforded for 89.9% of the explained variance of the data, but a clear discrimination was not achieved as the samples spread over the PC1 and PC2 axes. These findings prompted us to adopt a classification model in order to gain insights as to the relevant features associated to an eventual discrimination according to the propolis sample chemical composition and its geographic origin. In order to extract relevant but not redundant information we applied PLS-DA and RF to the propolis metabolomics dataset.

A model was built by performing PLS-DA that was able to identify important features to predict the propolis sample classification by measuring the variable importance in projection (VIP) as shown in Table 1. The most important fifteen ^{1}H-NMR resonances, i.e. features, were identified by PLS-DA and most of them (10) resulting from aliphatic compounds, as five features were associated to anomerical ones. Among other, the features detected by PLS-DA were mostly assigned to chemical groups of the alkane moiety (e.g., C-CH_2-C, 1.30-1.33 ppm; C-CH-C, 1.21 and 1.47 ppm) or acetyl group ($COCH_3$, 2.07-2.37 ppm) [12, 13, 14] of fatty acids and waxes commonly found in propolis. One-way ANOVA followed by the *post-hoc* Tukey test of the ^{1}H -NMR dataset confirmed the significance of most of the selected features.

Table 1. Important features (^{1}H resonances) ranked according to the VIP score calculated by PLS-DA of propolis samples. The colored boxes on the right indicate the relative concentrations of the corresponding metabolite in each studied group, i.e., geographic region.

Resonances (δH ppm)	**VIP score**	***p*-value[1] (-log 10)**	
1.30	4.57	47165e-14	
2.29	3.08	2.9594e-10	
1.33	2.88	n.s[2]	
5.26	2.78	5.8257e-19	
1.21	2.74	1.3077e-17	
4.84	2.63	8.9478e-10	
1.47	2.43	5.9462e-14	
4.81	2.21	6.9457e-10	
5.07	2.16	3.7371e-18	
2.12	2.11	2.4206e-09	
4.87	2.09	1.0308e-10	
2.07	2.02	3.1639e-10	
1.80	1.87	1.1753e-17	
1.53	1.82	n.s	
2.37	1.81	n.s	

[1] One-way ANOVA and *post-hoc* Tukey test ($p<0.05$), [2]not significant, [3]geographic regions of Santa Catarina state (southern Brazil): **C** = central, **E** = east, **H** = highlands, **N** = north, and **W** = west.

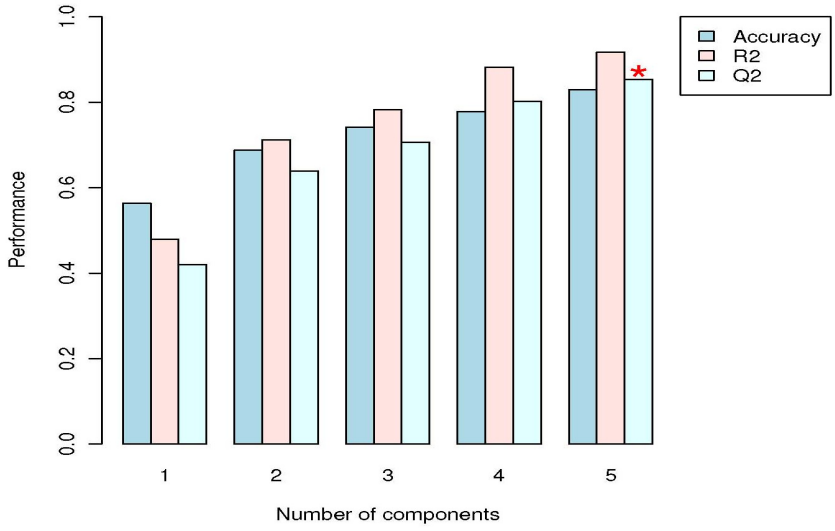


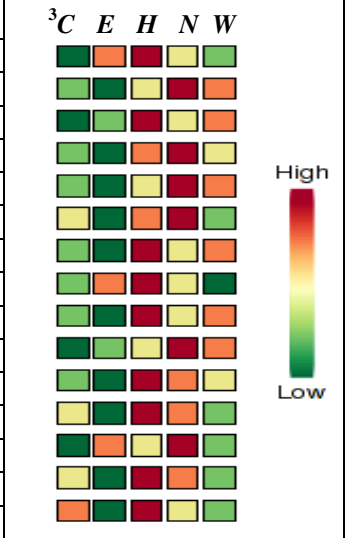
Fig. 1. Performance of the PLS-DA model classification using different numbers of components. The red asterisk indicates the best classifier.

The quantitative measure of the performance for PLS-DA classification model given by the R2, Q2, and accuracy values showed to be higher than 83% for those statistics and reveals a good performance of the method (Fig. 1).

In a second set of experiments, the non parametric RF analysis was applied to the ^{1}H-NMR dataset allowing to selecting extra and non-redundant features for an accurate classification of Brazilian propolis according the geographic origin (Table 2). Eleven out of the top fifteen features identified by RF analysis occur in the spectral window of aliphatic compounds, corroborating the PLS-DA findings, but expanding the metabolite signatures associated to the selected features. Indeed, features associated to saturated (C-CH_3, C-CH-C, C-CH_2-C) and unsaturated (=C-CH_3) alkyl and acetyl ($COCH_3$) groups were predominantly identified by the RF supervised learning algorithm. Preliminary analysis of the features selected by RF, PLS-DA, and also 2D-NMR experiments (data not shown) suggests the presence of long chain fatty acids in propolis samples such as arachidonic, oleic, stearic, and palmitic/palmitoleic acids, associated to the resonances at 1.30, 1.64, 2.04, and 2.76 ppm, for instance [14].

Table 2. Significant features (^{1}H resonances) ranked by the mean decrease in classification accuracy when permuted by RF analysis. The colored boxes on the right indicate the relative effect of the corresponding metabolite in each group of propolis in study, according to their regions of production.

Resonances (δH ppm)	**Mean decrease Accuracy**	***p*-value[1] (-log 10)**
2.04	0.035	3.1456e-22
1.08	0.034	3.2706e-21
2.46	0.028	9.7457e-25
1.18	0.026	3.7861e-16
1.30	0.024	4.7165e-14
2.56	0.023	3.9608e-16
5.07	0.021	3.7371e-18
2.61	0.020	7.9958e-11
2.49	0.019	6.5947e-20
1.64	0.018	n.s[2]
5.20	0.017	5.6611e-21
6.16	0.016	3.4739e-11
2.76	0.015	n.s
2.98	0.014	3.8818e-18
6.10	0.013	3.6287e-19



[1] One-way ANOVA and *post-hoc* Tukey test ($p<0.05$), [2]not significant, [3]geographic regions of Santa Catarina state (southern Brazil): **C** = central, **E** = east, **H** = highlands, **N** = north, and **W** = west.

The confusion matrix revealed a quite interesting performance of the RF supervised learning algorithm, since the classification error found for the predicted class and actual class was zero. Besides, the descriptive model based on the univariate

statistics one-way ANOVA followed by the *post-hoc* Tukey test corroborate thirteen out of the top fifteen features selected by RF analysis. Indeed, only two features (1.30 ppm and 5.20 ppm) were simultaneously detected by both PLS-DA and RF, characterizing redundant information.

PLS-DA and RF methods were also able to reveal distinct effects of the selected features regarding the geographic origin of propolis samples (Tables 1 and 2). A quantitative approach was applied to the PLS-DA selected features by calculating the values of their absolute integral (data not shown). Differences in relative concentrations of the corresponding metabolite in each studied group (geographic origin) were detected for all the features, adding extra information to the classification model. Thus, for example, the propolis samples originated from the east region of Santa Catarina state were characterized for their lower content of metabolites comparatively to samples from the other studied regions.

Finally, it is worth mentioning that propolis is a complex matrix well known for its phenolic constituents so that the most interesting spectral windows are 5.50-8.25 ppm, containing mainly the aromatic compound signals, and 8.25-13.00 ppm, where the carbonylic and carboxylic proton signals are found. However, features belonging to those spectral regions did not influence the classification by PLS-DA and RF analysis.

The following task involved the validation of the wrapper approach described in section 2.3. In this study, the coupling of J48, a decision tree-inducing algorithm, and JRip, a rule set induction method, to EA as an optimization engine, in a wrapper approach allowed to identify a certain number of features over the ^{1}H-NMR spectral window as shown in Table 3.

Table 3. J48-EA and JRip-EA wrapper performances and the 15-top ^{1}H-NMR resonances identified taking into account a calculation for 5 and 10 features. The EA was executed 30 times for each case and the prediction accuracy of the classifiers was evaluated using 5-fold cross-validation.

Wrappers	Features	Mean fitness (%)	Standard deviation	Mean cross-validation accuracy (%)	Standard deviation	Resonances (δH ppm)
J48-EA	5	99.84	0.34	93.70	2.61	2.04, **5.61**, **6.64**, 0.89, 4.97, **5.37**, **5.29**, **7.21**, 3.10, **7.05**, **7.08**, 1.27, 2.12, 4.84, **7.62**
	10	99.67	0.19	94.27	2.80	**5.61**, 1.08, 2.04, **5.20**, **7.05**, 2.49, 1.27, **6.49**, 4.00, 1.50, **7.62**, **8.07**, **7.12**, 2.10, **5.10**
JRip-EA	5	99.67	0.57	92.09	2.92	1.64, 3.63, **6.46**, 2.04, **6.64**, **6.72**, **6.79**, **8.07**, **6.25**, 0.88, **5.17**, 1.02, **6.16**, **9.18**, **6.82**
	10	99.90	0.21	92.77	2.82	2.12, 1.08, **5.29**, 1.86, **7.05**, **7.08**, **6.79**, **5.79**, 1.56, 0.81, 2.76, **6.46**, 2.04, 1.60, **8.07**

Contrarily to PLS-DA and RF supervised learning algorithms, the wrapper algorithms selected features that spread over all the ^{1}H-NMR spectral regions, but a predominance of meaningful resonances associated to the aromatic ring moiety of metabolites i.e., 5.50-8.25 ppm, could be detected, typically suggesting an important effect of, e.g., (poly)phenolic compounds in the classification models. Furthermore, it is also possible to notice some redundant information given by both wrapper methods.

The J48/JRip-EA wrapper methods showed to complement RF and PLS-DA since important features addressing the occurrence of phenolic compounds were identified, even suggesting the occurrence of phenolic acids (gallic – 7.05 ppm, *singlet*; *t*-cinnamic – 6.49, *duplet*; hydrocinnamic – 2.50 ppm, *triplet* and 7.12 ppm, *multiplet*; and caffeic – 8.07 ppm, *singlet*, 7.08 ppm, *double duplet*, 6.82 ppm, *singlet,* 6.79 ppm, *duplet*), as well as the tentatively assigned flavone apigenin (6.16 ppm-*duplet*, 6.46 ppm-*duplet*, and 6.72 ppm-*singlet*) [15] in the studied propolis. In fact, in this regard the application of the wrapper algorithms to the propolis metabolomics dataset expanded the possibilities of detecting relevant metabolite signatures typically found in that complex matrix. Such findings were further confirmed by reverse-phase high performance liquid chromatography coupled to a UV-visible detector (data not shown).

Besides, similarly to PLS-DA and RF analysis, among the significant top fifteen features identified by J48/JRip-EA wrapper methods a series of resonances (0.88, 1.27, 1.60, 2.04, 2.12, 2.76, and 5.29 ppm) associated to metabolite signatures of the, e.g., alkane moiety (C-CH_3, 0.88-1.02 ppm; C-CH_2-C, 1.27-1.30 ppm) and acetyl group ($COCH_3$, 2.04-2.37 ppm) [12, 13, 14] of monosaturated or unsaturated fatty acids was found in propolis samples. Finally, the presence of the nucleoside uridine in the samples is inferred as meaningful for the classification model, since typical resonances at 5.61 ppm-*duplet*, 5.37 ppm-*duplet*, and 3.63 ppm-*double duplet* were identified by the J48/JRip-EA algorithms and further confirmed by 2D-NMR (TOCSY and HSQC experiments).

The wrapper models showed very high mean fitness (≥ 99%) and prediction accuracy (≥92%) on the cross-validation studies. The validation of each final solution was conducted by doing an independent validation procedure, performing a 10 times 5-fold cross-validation process, using the set of selected features coming from the EA's best solution. Such a finding is worth mentioning taking into account the effect of the EA as optimization engine in controlling overfitting in classification-tree models. It is quite interesting to notice that the performance of the classifiers is quite acceptable with only 5 features, showing the ability of the classifiers to provide high accuracy models with a very limited set of features.

Taken together, the several test domains performed by running the J48/JRip-EA interfaces showed to be effective for feature selection and to develop a classification model tree with high prediction accuracy and consistency.

4 Conclusions

The selected classification methods PLS-DA, RF and the wrapper methods J48/EA and JRip/EA based on machine learning and feature selection appear usable tools for building classification models for the Brazilian propolis metabolomics, with high prediction accuracy.

PLS-DA, RF and J48-EA/JRip-EA analyses of the NMR-based propolis metabolomics dataset showed to be complementary approaches by retrieving and expanding the set of class discriminating features and by adding relevant information for the identification of the class determining metabolites. This allowed further elucidation of the system under investigation in regards to the metabolite signature of important compounds, i.e., chemical fingerprint, and geographic origin of Brazilian propolis.

Acknowledgments. The authors are indebted to CAPES, CNPq, FAPESC, and National Laboratory of Bioscience.

The work is partially funded by ERDF - European Regional Development Fund through the COMPETE Programme (operational programme for competitiveness) and by National Funds through the FCT (Portuguese Foundation for Science and Technology) within projects ref. COMPETE FCOMP-01-0124-FEDER-015079 and PEst-OE/EEI/UI0752/2011. RC's work is funded by a PhD grant from the Portuguese FCT (ref. SFRH/BD/66201/2009).

References

1. Raamsdonk, L.M., Teusink, B., Broadhurst, D., Zhang, N., Hayes, A., Walsh, M.C., Berden, J.A., Brindle, K.M., Kell, D.B., Rowland, J.J., Westerhoff, H.V., van Dam, K., Oliver, S.G.: A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. Nature Biotechnology 19, 45–50 (2001), doi:10.1038/83496
2. van den Berg, R.A., Hoefsloot, H.C., Westerhuis, J.A., Smilde, A.K., van der Werf, M.J.: Centering, scaling and transformations: improving the biological information content of metabolomics data. BMC Genomics 7, 142–147 (2006), doi:10.1186/1471-2164-7-142
3. Weljie, A., Newton, J., Mercier, P., Carlson, E., Slupsky, C.: Targeted profiling: quantitative analysis of 1HNMR metabolomics data. Analytical Chemistry 78, 4430–4442 (2006), doi:10.1021/ac060209g
4. Witten, I.H., Frank, E., Hall, M.A.: Data Mining: Practical Machine Learning Tools and Techniques, 3rd edn. Morgan-Kaufmann, Burlington (2011)
5. Watson, D.G., Peyfoon, E., Zheng, L., Lu, D., Seidel, V., Johnston, B., Parkinson, J.A., Fearnley, J.: Application of principal components analysis to ^{1}H-NMR data obtained from propolis samples of different geographical origin. Phytochemical Analysis 17, 323–331 (2006), doi: 10.1002.pca
6. Hackstadt, A.J., Hess, A.M.: Filtering for increased power for microarray data analysis. BMC Bioinformatics 10, 11–23 (2009), doi:10.1186/1471-2105-10-11
7. Brodsky, L., Moussaie, A., Shahaf, N., Aharoni, A., Rogachev, I.: Evaluation of peak picking quality in LC-MS metabolomics data. Analytical Chemistry 15, 9177–9187 (2010), doi:10.1021/ac101216e
8. Xia, J., Psychogios, N., Young, N., Wishart, D.S.: MetaboAnalyst: a web server for metabolomic data analysis and interpretation. Nucleic Acids Research 37(Web Server issue), W652–W660 (2009), doi:10.1093/nar/gkp356
9. Wehrens, R., Mevik, B.H.: Pls: partial least squares regression (PLSR) and principal component regression (PCR) (2007), R package version 2.1-0
10. Kuhn, M., Wing, J., Weston, S., Williams, A.: Caret: classification and regression training (2008), R package version 3.45
11. Liaw, A., Wiener, M.: Classification and regression by random Forest (2002), R News

12. Leyden, D.E., Cox, R.H.: Analytical Applications of NMR. John Wiley & Sons, New York (1977)
13. Waterman, P.G., Mole, S.: Analysis of Plant Metabolites. Blackwell Scientific Publications, London (1994)
14. Fan, T.W.M., Lane, A.N.: Structure-based profiling of metabolites and isotopomers by NMR. Progress in Nuclear Magnetic Resonance Spectroscopy 52, 69–117 (2008), doi:10.1016/j.pnmrs.2007.03.002
15. Bertelli, D., Papotti, G., Bortolotti, L., Marcazzanb, G.L., Plessia, M.: ^{1}H-NMR simultaneous identification of health-relevant compounds in propolis extracts. Phytochemical Analysis 23, 260–266 (2011), doi:10.1002/pca.1352