Article

Proof-of-Principle of rTLC, an Open-Source Software Developed for Image Evaluation and Multivariate Analysis of Planar Chromatograms

Dimitri Fichou, Petar Ristivojevi#, and Gertrud Elisabeth Morlock

Anal. Chem., **Just Accepted Manuscript** • DOI: 10.1021/acs.analchem.6b04017 • Publication Date (Web): 15 Nov 2016

Proof-of-Principle of rTLC, an Open-Source Software Developed for Image Evaluation and Multivariate Analysis of Planar Chromatograms

Dimitri Fichou[a], Petar Ristivojević[a,b] and Gertrud E. Morlock[a,*]

[a]Chair of Food Science, Institute of Nutritional Science, and Interdisciplinary Research Center (IFZ), Justus Liebig University Giessen, Heinrich-Buff-Ring 26-32, 35392 Giessen, Germany

[b]On leave from the Innovation Center of the Faculty of Chemistry, University of Belgrade, PO Box 51, 11158 Belgrade, Serbia

*Corresponding author. Tel.: +49-641-99-39141; fax: +49-641-99-39149; E-mail address: Gertrud.Morlock@uni-giessen.de (G. E. Morlock).

ABSTRACT: High-performance thin-layer chromatography (HPTLC) is an advantageous analytical technique for analysis of complex samples. Combined with multivariate data analysis, it turns out to be a powerful tool for profiling of many samples in parallel. So far, chromatogram analysis has been time-consuming and required the application of at least two software packages to convert HPTLC chromatograms into a numerical data matrix. Hence, this study aimed to develop a powerful, all in one open-source software for user-friendly image processing and multivariate analysis of HPTLC chromatograms. Using the caret package for machine learning, the software was set up in the R programming language with an HTML-user interface created by the shiny package. The newly developed software, called rTLC, is deployed online and instructions for direct use as web application, and in case required, for local installation are available on GitHub. rTLC was created especially for routine use in planar chromatography. It provides the necessary tools to guide the user in a fast protocol to the statistical data output (*e. g.*, data extraction, preprocessing techniques, variable selection and data analysis). rTLC offers a standardized procedure and informative visualization tools that allow the user to explore the data in a reproducible and comprehensive way. As proof-of-principle of rTLC, German propolis samples were analyzed using pattern recognition techniques, principal component analysis, hierarchic cluster analysis and predictive techniques, such as random forest and support vector machines.

KEYWORDS: High-performance thin-layer chromatography; Multivariate analysis; Chemometrics; Open-source software, R programming language; Caret package

INTRODUCTION: Natural extracts may contain thousands of individual compounds, and the majority of these are present in low concentrations down to the trace level. Though it is challenging, it is important to obtain reliable fingerprints that represent sound profiles of physiologically active compounds[1]. Its simplicity, cost-effective operation and the possibility of simultaneous analysis of up to 20 samples in parallel makes high-performance thin-layer chromatography (HPTLC) a technique of choice in herbal and food analysis[2, 3]. For evaluation, the HPTLC fingerprint of a complex sample is visually compared to that of a certified reference sample or to marker compounds being characteristic for the respective sample. The main disadvantage of such a manual pattern recognition technique and its visual comparison is its subjectivity, and it highly depends on the analyst's perception. Hence, hyphenation of HPTLC with high-sophisticated multivariate techniques provides objective fingerprints, mainly based on mathematical models[4,5]. As HPTLC chromatograms contain hundreds of pixels, this multidimensionality is used to extract a maximum of information out of the chromatograms[4]. For example, pattern recognition techniques can recognize chemical compound patterns, identify characteristic marker compounds as well as classify unknown samples according to their biological activity.

Though increasing, there are still a limited number of research papers on the combination of HPTLC with multivariate data analysis. Most of these are based on the investigation of propolis, herbal samples, biopolymers and microalgae[6-16]. Although propolis is one of the most investigated honeybee product, the separation of its complex phenolic compound composition is still challenging analysts. After derivatization with Neu's reagent and detection at UV 366 nm, phenolic components showed differently colored bands. Such colorful HPTLC chromatograms are highly appropriate input data for evaluation by multivariate data analysis. There exists a wide

range of derivatization reagents with different specificity and capability of detection. The resulting characteristically colored bands generate different profiles on the red, green and blue (RGB) channels. Thus, derivatization reagents can substantially influence the separation performance and data evaluation[6].

Contrary to other chromatography techniques, such as high performance liquid chromatography (HPLC) and gas chromatography (GC) which offer a direct export of data for further multivariate analysis, images of HPTLC chromatograms must first be converted to a numerical data matrix. Various software, toolboxes and algorithms have been applied for image processing and multivariate analysis of HPTLC chromatograms so far (Table 1)[6-16]. Such packages lack in domain-specific functionality, which results in a manual, lumbering and time-consuming pipeline of the data handling. The user is forced to open, process and save the data through different software packages and toolboxes to perform the analysis[6-16].

For the first time, we describe and introduce rTLC in this study. It is a newly developed open-source web application for image processing and multivariate analysis of HPTLC chromatograms. The focus is laid on the different possibilities and advantages of the application, such as a fast and simple image processing workflow and application of a range of chemometric techniques suited for planar chromatography. One driving force for developing rTLC was to provide users with a unique solution to analyze HPTLC data. The access to a simple and accurate open-source web application, instead of purchasing a number of licenses, was another impetus. Many useful features for the analysis of HPTLC data, such as line profile of target compounds, band comparison, signal preprocessing as well as comma separated value (CSV) export for analysis on other platforms were integrated. Pattern recognition techniques such as principal component analysis (PCA), hierarchical cluster analysis (HCA) and heat map are applicable on

separate channels (RGB and gray scale) or in combination. Prediction techniques such as random forest (RF), linear discriminant analysis (LDA), support vector machine (SVM), partial least square (PLS) and classification and regression tree (CART) analysis were integrated as well. The increasing number of publications in the field of planar chromatography hyphenated with multivariate analysis motivated to redesign software and add many new tools. This makes rTLC suitable for a wide range of applications in herbal, food and environmental science.

EXPERIMENTAL SECTION

Set-Up of the Open-Source Web Application. The rTLC application is written with the R programming language[17]. R is an open-source language and environment for statistical computing and graphics. A key feature of R lies in its community of sharing users, who contribute to the extension of the language via packages, allowing others to use their work. rTLC uses in particular the shiny package to create an HTML based user interface[18] and the caret package for machine learning[19]. This way, the application was deployed online and is directly accessible via a modern internet browser having internet connection. As it is a web application, the user needs not to install software. Direct use of rTLC (http://shinyapps.ernaehrung.uni-giessen.de/rtlc), and in case required, instructions for local installation are available on GitHub: https://github.com/DimitriF/rTLC-apps.

Example Data Set. A given sample set was used as proof-of-principle of the newly developed software. 106 samples of German propolis obtained from the Apicultural State Institute (Stuttgart, Germany) were analyzed in a previous study[20, 21]. The resulting 7 chromatograms in the JPEG format were manually labeled before the statistical analysis, leading to the assignment

of 37 blue-type and 69 orange-type samples of German propolis. The rTLC parameters set are discussed subsequently.

RESULTS AND DISCUSSIONS

rTLC, the newly developed open-source web application for image processing and multivariate analysis of HPTLC chromatograms, is introduced for the first time. The simple and streamlined workflow (Figure 1) provides the necessary tools to reproducibly guide the user in a fast protocol to the statistical data output. For regular cases, the evaluation of HPTLC chromatograms took only few minutes. The proof-of-principle was demonstrated via a German propolis data set, which was also made available as demonstration file in the rTLC software. Thus, the user is able to follow and reproduce the results reported below.

Data Input. The user had to upload two files in rTLC to provide an appropriate data set for image evaluation and multivariate data analysis: (1) HPTLC chromatograms which contain the independent variables and (2) a batch file which contains the dependent variables about each sample on the plates, such as classes, botanical and geographical origin. Information on the experimental conditions is necessary to automatically extract each chromatogram from the HPTLC plate, *e.g.* the distances used during sample application and chromatography. rTLC supports the upload of commonly used image formats such as jpeg, tiff and png. The software computes the horizontal mean for each pixel of the chromatogram on the RGB channels as well as the gray scale, which is the mean of those three channels. At the end of this step, the data are in the form of a 3D array with samples as rows, R_F as columns and channels as layers (Figure 2). rTLC provides tools for line profiles of target compounds, comparison between tracks, pattern identification as well as identification of characteristic chemical and biological markers. The

profile comparison of RGB channels as well as gray scale helps to find similarities and dissimilarities between samples before and after signal preprocessing.

Data Preprocessing. Recently, preprocessing methods used in HPTLC fingerprinting were discussed[22]. Among others, the appearance of a non-homogeneous background after derivatization, an increased noise level and band shifts are caused by variation in mobile and vapor phase composition, humidity, temperature, operator handling and instrumental instability. Thus, warping techniques are recommended to mitigate such experimental drawbacks[4, 23]. Two peak alignment procedures were integrated into the rTLC software and are available to correct inter- and intra-plate band shifts[24]: (1) parametric time warping and (2) dynamic time warping. Further integrated options for data preprocessing such as denoising, normalization, and baseline removal aimed at improving the quality of the data set. The software provides the Savitzky-Golay and median filter, which are denoising/filtering methods commonly used in preprocessing of HPTLC chromatograms[4, 5]. The baseline removal process was found to be mandatory in almost all cases[4], whereas good statistical models were also obtained without baseline correction[7]. Hence, it is recommended to compare results with and without baseline correction. Also, a normalization step is not mandatory and there is no consensus when it is obligatory - sometimes it makes the results better, sometimes even worse. The preferred method of signal normalization is the standard normal variate (SNV) method. Finally, rTLC provides auto-scaling and mean centering to transform variables in the same unit[5, 6, 24]. The selection and need for preprocessing tools depends on the project and may be chosen by the users to obtain ready-to-use data for statistical analysis.

Variable Selection. HPTLC chromatograms provide a high number of variables for the given, often limited number of available samples. There are several approaches regarding the nature of

used variables for multivariate analysis. Important variables that contain information for the aimed classification should be kept, whereas variables encoding the noise and/or with no discriminating power should be removed[25]. For this purpose, rTLC provides options for careful selection of variables for a specific channel or all channels together. The statistical analysis part also informs on this selection, which can be optimized to keep the important information only.

Exploratory Statistics. The user is only working with a data matrix, *i.e.* with samples as rows and variables as columns; with this, it is possible to compute pattern recognition techniques such as PCA, HCA and heatmap. For each of these techniques, informative visualization tools are available that illustrate the data in various perspectives and allow the user to highlight patterns by comparing the results with a chosen column of the batch file. For both, beginners and experienced R-users, an editor is available and can be used for other types of techniques or custom-made plots.

Predictive Statistics. With the same matrix as mentioned before, this feature allows the user to train a predictive model, used for the subsequent prediction of the properties of new samples. There are two main techniques in predictive statistics, *i.e.* classification and regression; both are available in the software. Before the training, the data set is split into training and test set to produce a true validation set and avoid overfitting. The application uses the caret package[19] of the R language to tune a model and choose the optimal parameters for a given algorithm. The available predictive techniques are LDA, PCA (regression only), PLS (regression only), RF, CART as well as SVM with linear and polynomial kernel.

A model will be trained for each value of a grid, automatically created but editable, and the parameters which give the best validation result will be kept for the final model. The choice of the best set of parameters is made according to a cross-validation procedure; available

procedures are k-fold cross validation, bootstrapping and leave-one-out-cross-validation. A summary metric must be chosen to select the best model. For regression, the statistical parameters can be expressed by root mean squared error (RMSE) or R^2. For classification, accuracy, kappa, sensitivity or sensibility are available as summary metrics.

Different output tools are available to explore the result, such as confusion matrix of the test set, prediction table and model summary. Also here, an editor is available to produce custom-made plots. At the end of this step, a model file can be downloaded and used in other sessions to predict the properties of new samples.

Proof-of-Principle of rTLC. HPTLC chromatograms contain comprehensive information regarding the polarity, chemical, and spectral properties of individual compounds in a sample. As a case study, HPTLC chromatograms of German propolis samples were used to illustrate the practical application of the rTLC software. The HPTLC chromatograms of propolis showed a complex mixture of phenolic compounds, and thus, were highly appropriate input data to demonstrate the performance and power of rTLC. Visual comparison of the respective HPTLC chromatograms and RGB channels (Figure 3A) revealed a difference in the chemical composition of the two types of German propolis. The blue type of propolis had several blue bands at R_F values around 0.2, 0.3 and 0.6 (Figure 3B). The orange type of propolis showed a rich phenolic profile and contained characteristic orange and yellow bands in the R_F range of 0.1-0.5, and high fluorescent blue bands in the R_F range of 0.5-0.8 (Figure 3C). Next, two unsupervised techniques (PCA and HCA) and two supervised techniques (RF and SVM) were selected to illustrate the capabilities of multivariate analysis by rTLC. Parametric time warping (aligned to the first sample), SNV and mean centering were used as preprocessing step.

Unsupervised Techniques. Commonly used pattern recognition techniques[5, 6] as PCA and HCA, are performed by rTLC in a fast and simple way. PCA was applied on the data set for the RGB channels as well as on the grayscale image. The variable of interest was class as color assignment (labelling and different symbols were not chosen). The blue channel (Figure 4A) and grayscale data (Figure 4D) with a R_F range of 0-1 as variable selection showed the best discrimination between the two sorts of German propolis samples and their statistical performances were discussed subsequently.

In case of the blue channel data, PCA resulted in a five-component model, explaining 78.41% of the total variance. PC1 described 40.99%, while PC2 explained 15.34% of the total variance (Figure 4A). The most influential phenolic compounds were identified using the loading plots. For PC1, the compounds at R_F 0.04, 0.38, 0.53, 0.66 and 0.98 had positive contributions while the compounds at R_F 0.29, 0.58 and 0.77 had negative contributions (Figure 4B). For PC2, the compounds at R_F 0.27, 0.52, 0.63 and 0.82 had positive contributions, while the compounds at R_F 0.06, 0.30, 0.36, 0.56 and 0.72 had negative contributions (Figure 4C).

In the case of the grayscale image, the total variance explained by the first three PCs was 59.66% (PC1: 32.58%, PC2: 15.61%, and PC3: 11.45%) (Figure 4D). The discrimination between the two types of propolis samples is mainly driven by the first component. For PC1, positive influences were found at R_F 0.06, 0.34, 0.39, 0.53 and 0.66 and negative ones at R_F 0.28, 0.58 and 0.79 (Figure 4E). For PC2, positive influences were observed at R_F 0.04, 0.36, 0.57 and 0.73 and negative ones at R_F 0.27, 0.52, 0.65 and 0.84 (Figure 4F). Once those influential R_F values are known, the researcher can apply other analytical techniques or refer to the literature to identify such discriminatory compounds.

Cluster analysis is an often used classification technique. This algorithm performs a hierarchical cluster analysis using the distance between samples. At the beginning, each sample is assigned to its own cluster, iteratively, the closest clusters are joined together and the distances between clusters are recomputed, continuing until there is only one cluster. The simplest and most intuitive way to mathematically define the similarity between objects is based on the Euclidean distance. rTLC provides several routes to define the similarity between objects. According to the blue channel and grayscale data, there was a good discrimination between the orange- and blue-type propolis samples, which was in agreement with PCA[5-7]. For the blue channel, 'class bis' (x-labelling and color), Euclidean distance, ward method and a cluster number of 3 were chosen (2 clusters for gray scale).

In the dendrogram of the blue channel data, there were three clusters (Figure 5A). The first cluster had a distance of 49 and was mainly composed of orange samples, whereas the second cluster had a distance of 25 and was dominated of blue-type propolis samples and the third cluster had a distance of 28 and consisted mainly of the latter samples. The dendrogram obtained for the grayscale data showed two clusters (Figure 5B). The first cluster contained almost all blue samples, while the second cluster consisted mostly of orange samples, the distance were respectively 57 and 62. The few blue-type propolis samples grouped into the orange-type cluster differed in their patterns compared to the other blue-type propolis samples. These samples can be considered as a mixture of both types of propolis due to the natural variation in the chemical composition. For such cases, it has to be proven that the variations in the experimental condition had been removed during the preprocessing step, as far as possible.

Supervised Techniques. In supervised techniques, a set of data describing objects of known features is used to construct a training set that is used to predict those features for new samples

then. Supervised techniques were applied in a wide range of chromatographic, spectrophotometric and sensorial data, for quantification, fingerprinting, authentication and detection of adulteration of food and herbal products[25]. The feature can be discrete, like the geographical or botanical origin, or continuous like the concentration of a target molecule in the investigated samples.

As a first step of the supervised procedure, the data were split between training and test set. Secondly, preprocessing techniques were applied on the training and test set. Note for normalization, that the mean centering and standard deviation of the training set is used to standardize the test set to avoid overfitting[25]. After the following variable selection, prediction models were built using the training set for each row of the tuning grid and each step of the cross validation procedure. Afterwards, the best parameters were selected and the final model was trained with those parameters on the entire training set. Lastly, the reliability of the model was evaluated using the test set. Two powerful supervised algorithms were selected to present this feature: RF and SVM with linear kernel. Like for PCA and HCA, the following preprocessing was used: parametric time warping, SNV and mean centering. In each case, the ratio of training to test set was 3:1 and the cross validation method was 5-fold cross validation with total accuracy as summary metric of choice for the selection of the best model. The outcome was studied for each of the three channels and the grayscale image. In all cases, the prediction efficiency was high and demonstrated the power of the technique to reproduce human decisions.

Though RF has rarely been used as multivariate tool in food and herbal research so far, there are several benefits that could make the RF algorithm an appropriate supervised tool in HPTLC analysis: it can be used (1) when there are much more variables than observations, (2) for two- or multi-classification and (3) for a good predictive performance, even when most of the predictive

variables are noise, and thus a preselection of variables is not required. As another benefit, this algorithm does not need standardization. The RF classifier needs optimization for two parameters to generate a prediction model: the number of classification trees desired (ntree) and the number of variables (mtry) which are used for tree growing in each tree. The accepted default values for those two parameters are 500 for ntree, and sqrt (mall) for mtry, whereby mall is the total number of variables in the original data set. The most important parameter, mtry, can be optimized with the caret package, in contrast to the ntree parameter. This optimization led to more accurate models[26, 27, 28]. By the way, the option PLS resulted in an equivalent outcome to RF and SVM.

For all channels, the accuracy of classification of the training set was 100%. Those models were clearly overfitted and this outcome must not be taken into account to judge a model. The confusion matrix was obtained for each channel on the test set and during the cross validation (Table 2 A). The green channel showed a good accuracy for cross validation and for the test set. For the blue channel data and gray scale image, the comparison between cross validation and test set showed more consistency, which was in accordance to PCA and HCA. Detailed statistical parameters for the blue channel showed the performance of the model according to different metrics (Table 2 B). The importance of the variables for the RF algorithm trained on the blue channel is evident (Figure 6). In contrast to the variables highlighted in the loading plots of the PCA, the model resulted in other variables to discriminate the two types of propolis.

The SVM algorithm separates the classes by an optimal hyper plane that maximizes the distances between classes by defining boundaries for the closest classes (support vectors) from the margins of the class. This way, SVM minimizes the training error with regard to the separation of the considered classes by using the least complex boundaries out of all possible ones. The optimal

hyper plane is obtained by an interactive algorithm that minimizes an error function that contains a parameter (penalty error) to control the complexity of the model and to avoid overfitting[24, 25]. Even if the results for each channel and the grayscale image (Table 3 A) were comparable with the RF results, this algorithm performed slightly worse in particular on the cross-validation data set. The tuning step chose values of 0.25 for cost and 2 for gamma, except for the blue channel where the optimum cost was 0.5. Detailed statistical parameters for the grayscale showed the performance of the model according to different metrics (Table 3 B).

CONCLUSIONS

According to our knowledge, there was no dedicated all-in-one software for a streamlined image evaluation and multivariate analysis of HPTLC chromatograms. The newly developed rTLC application was designed as user-friendly open-source software to ease fingerprint comparisons. New perspectives and conclusions on the data set are supported by a wide range of visualization tools, owed to high plotting capabilities of the R software. A great step forward was achieved by a substantial reduction of the analysis time. rTLC solved the supervised and unsupervised data handling within few minutes, whereas the current practice needs several hours using at least two different software packages. To the best of our knowledge, rTLC is the most concise tool available for application of different pattern recognition and prediction techniques for HPTLC chromatograms. On the one hand, the open-source asset of this application may attract users for the powerful combination of HPTLC and multivariate analysis. On the other hand, it may encourage the users to contribute to this technology through feedback, discussing ideas and adding new functionalities to the software.

AUTHOR INFORMATION

Corresponding Author

*E-mail: Gertrud.morlock@uni-giessen.de Tel. +49-641-99-39141. Fax +49-641-99-39149.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

Thank is owed to the Ministry of Education, Science and Technological Development of the Republic of Serbia, grant No. 172017 for financial support of P.R.

References

(1) Joshi D.D., *Herbal Drugs and Fingerprints*, Springer: New Dehli, India, 2012. DOI: 10.1007/978-81-322-0804-4.

(2) Krüger S.; Morlock G.E. In *Instrumental Thin-Layer Chromatography, Handbooks in Separation Science*; Poole C., Ed; Elsevier Science: Amsterdam, 1st edition, 2014; pp. 409-429. DOI:10.1021/ac00073a721.

(3) Morlock G.E.; Schwack W. *J. Chromatogr. A* **2010**, *1217*, 6600–6609. DOI: 10.1016/j.chroma.2010.04.058.

(4) Komsta Ł. *Chromatogr. Res. Int.* **2012**, *Article ID 893246,* 1–5. DOI: 10.1155/2012/893246.

(5) Milojković-Opsenica D.; Ristivojević P.; Andrić F.; Trifković J. *Chromatographia* **2013**, *76*, 1239–1247. DOI: 10.1007/s10337-013-2423-9.

(6) Ristivojević P.; Andrić F; Trifković J.; Vovk I.; Stanisavljević L.Ž.; Tešić Ž.L.; Milojković-Opsenica D. *J. Chemom.* **2014**, *28*, 301–310. DOI: 10.1002/cem.2592.

(7) Morlock G.E.; Ristivojević P.; Chernetsova E.S. *J. Chromatogr. A* **2014**, *1328*, 104–112. DOI: 10.1016/j.chroma.2013.12.053.

(8) Sârbu C.; Moţ A.C. *Talanta* **2011,** 85, 1112–1117. DOI: 10.1016/j.talanta.2011.05.030.

(9) Sagi S.; Avula B.; Wang Y.H.; Zhao J.; Khan I.A. *J. Sep. Sci.* **2014**, *37*, 2797–2804. DOI: 10.1002/jssc.201400646.

(10) Tian R.T.; Xie P.S.; Liu H.P. *J. Chromatogr. A* **2009**, *1216*, 2150–2155. DOI: 10.1016/j.chroma.2008.10.127.

(11) Tang T.X.; Guo W.Y.; Xu Y.; Zhang S.M.; Xu X.J.; Wang D.M.; Zhao Z.M.; Zhu L.P.; Yang D.P. *Phytochem. Anal.* **2014**, *25*, 266–272. DOI: 10.1002/pca.2502.

(12) Ogegbo O.L.; Eyob S.; Parmar S.; Wang Z.T.; Bligh S.W.A. *Anal. Methods* **2012**, *4*, 2522–2527. DOI: 10.1039/c2ay25373a.

(13) Xie P.S.; Sun S.; Xu S.; Guo L. *J. Chromatogr. Sep. Tech.* **2014**, *5,* 249–258. DOI: 10.4172/2157-7064.1000249.

(14) Zarzycki P.K.; Zarzycka M.B.; Clifton V.L.; Adamski J.; Głód B.K. *J. Chromatogr. A* **2011**, *1218*, 5694–5704. DOI: 10.1016/j.chroma.2011.06.065.

(15) Milojković Opsenica D.; Ristivojević P.; Trifković J.; Vovk I.; Lušić D.; Tešić Ž. *J. Chromatogr. Sci.* **2016,** *54*, 1077–1083. DOI: 10.1093/chromsci/bmw024.

(16) Morlock G.E.; Ristivojević P. *Food Hydrocolloids*, in print. DOI: 10.1016/j.foodhyd.2016.10.005.

(17) R Core Team. *R: A language and environment for statistical computing;* R Foundation for Statistical Computing: Vienna, Austria, 2016. https://www.R-project.org/.

(18) Chang W.; Cheng J.; Allaire J.; Xie Y.; McPherson J. *Shiny: Web application framework for r*; 2016. https://CRAN.R-project.org/package=shiny.

(19) Kuhn M. *J. Stat. Softw.* **2008**, *28*, 1–26. DOI: 10.18637/jss.v028.i05.

(20) Kunz N.; Scholl I.; Schroeder A.; Morlock G.E. In *58th Annual convention of the Association of the German Bee Research Institutes*, Berlin, Germany, March 29–31, 2011; Poster P 21.

(21) Morlock G.E.; Scholl I.; Kunz N.; Schroeder A. *CAMAG Bibliogr. Service CBS* **2013**, *111*, 13-15.

(22) Ristivojević P.; Morlock G.E. *J. Planar Chromatogr.--Mod. TLC* **2016**, *29*, 310-317. DOI: 10.1556/1006.2016.29.4.10

(23) Wong K.H.; Razmovski-Naumovski V.; Li K.M.; Li G.Q.; Chan K. *J. Pharm. Biomed. Anal.* **2014**, *95*, 11–19. DOI: 10.1016/j.jpba.2014.02.007.

(24) Mohd K.S.; Azemin A.; Hamil M.S.R.; Bakar A.R.A.; Dharmaraj S.; Hamdan M.R.; Mohamad H.; Mat N.; Ismail Z. *Asian J. Pharm. Clin. Res.* **2014**, *7*, 110–116.

(25) Berrueta L.A.; Alonso-Salces R.M.; Heberger K. *J. Chromatogr. A* **2007**, *1158*, 196–214. DOI: 10.1016/j.chroma.2007.05.024.

(26) Breiman L. *Mach. Learn.* **2001**, *45*, 5–32. DOI: 10.1023/A:1010933404324.

(27) Ai F.F.; Bin J.; Zhang Z.M.; Huang J.H.; Wang J.B.; Liang Y.Z.; Yu L.; Yang Z.Y. *Food Chem.* **2014**, *143*, 472–478. DOI: 10.1016/j.foodchem.2013.08.013.

(28) Svetnik V.; Liaw A.; Tong C.; Culberson J.C.; Sheridan R.P.; Feuston B.P. *J. Chem. Inf. Comp. Sci.* **2003**, *43*, 1947–1958. DOI: 10.1021/ci034160g.

(29) Chang C.-C.; Lin C.-J. *ACM Trans. Intell. Syst. Technol.* **2011**, *2*, 27:1–27:27. DOI: 10.1145/1961189.1961199.

(30) Hamber V.; Irish H. *J. Chem. Educ.* **2007**, *84*, 842—847. DOI: 10.1021/ed084p842.

Table 1

Overview of publications related to HPTLC and multivariate analysis.

No.	Samples	Multivariate techniques	Software	Ref.
1	Herbs	PCA, Artificial neural network (ANN)	Matlab R2007 (MathWorks, Natick, MA, USA)	10
2	Propolis	PCA, HCA, Partial least square-discriminant analysis (PLS-DA)	Matlab R2011a (MathWorks, Natick, MA, USA), PLS toolbox version 6.2.1 (Eigenvector Research Incorporated, Manson, WA, USA) Image J1.48c version (Research Services Branch, National Institute of Mental Health, Bethesda, MD, USA.)	6
3	Herbs	PCA, PLS-DA Orthogonal PLS-DA (O-PLS DA)	SIMCA-P+ Version 12 (Umetrics AB, Umea, Sweden), VideoScan (CAMAG. Muttenz, Switzerland)	12
4	Propolis	PCA, HCA, LDA	Matlab R2011a (MathWorks, Natick, MA, USA), PLS toolbox version 6.2.1 (Eigenvector Research Incorporated, Manson, WA, USA) SPSS Version 21 (BM Corporation, Armonk, NY, USA), LIBSVM Version 3.16[29]	7
5	Herbs	K-nearest neighbors Classification and regression tree (CART) Successive projection algorithm-linear discriminant analysis (SPA-LDA) PCA-discriminant analysis (PCA-DA) Support vector machine-discriminant analysis (SVM-DA), PLS-DA	Matlab R2012b (MathWorks, Natick, MA, USA) PLS toolbox version 7.3.1 (Eigenvector Research Incorporated, Manson, WA, USA) SPA toolbox 1.0(Homemade programs written in Matlab) Classification toolbox version 2.0 (Milano Chemometrics and QSAR Research Group, Milano, Italy)	21
6	Herbs	PCA	XLSTAT (Addinsoft, New York, NY, USA)	9
7	Herbs	PCA	Origin pro (OriginLab, Northampton, MA, USA)	13
8	Propolis	PCA, HCA	TLC Analyzer[30]	8
9	Propolis	Similarity analysis, HCA K-means clustering, ANN, SVM	Self-programmed software Xe2 IDE (Embarcadero, San Francisco, CA, USA), SPSS Version 21 (IBM Corporation, Armonk, NY, USA),LIBSVM Version 3.16[29]	11
10	Biopolymers	PCA, HCA	Matlab R2011a (MathWorks, Natick, MA, USA), PLS toolbox version 6.2.1 (Eigenvector Research Incorporated, Manson, WA, USA) Image J1.48c version (Research Services Branch, National Institute of Mental Health, Bethesda, MD, USA.)	16

Table 2

RF algorithm model: confusion matrix for the test set of blue- and orange-type propolis samples as well as cross validation set for each channel (A) and detailed summary metrics for the blue channel on the three data sets (accuracy, sensitivity and specificity; B).

A				**Test set**			**Cross validation**		
	Channel	**Optimum mtry**		**Blue-type**	**Orange-type**	**Accuracy**	**Blue-type**	**Orange-type**	**Accuracy**
	Red	2	Blue	6	6	0.7857	14	11	0.8077
			Orange	0	16		4	49	
	Green	15	Blue	9	3	0.8571	16	9	0.8590
			Orange	1	15		2	51	
	Blue	2	Blue	9	3	0.8929	17	8	0.8333
			Orange	0	16		5	48	
	Gray	2	Blue	9	3	0.8929	16	9	0.8590
			Orange	0	16		2	51	

B	**RF model parameters**	**Accuracy**	**Sensitivity**	**Specificity**
	Training set	1.0000	1.0000	1.0000
	Test set	0.8929	1.0000	0.8421
	Cross validation	0.8333	0.7727	0.8571

Table 3

SVM model with linear kernel: Confusion matrix for the test set of blue- and orange-type propolis samples and cross-validation set for each channel (A) and summary metrics for the grayscale image data on the three data sets (accuracy, sensitivity and specificity; B).

A	Optimum			Test set			Cross validation		
	Channel	**Cost**	**Gamma**	**Type**	**Blue-type**	**Orange-type**	**Accuracy**	**Blue-type**	**Orange-type**	**Accuracy**
	Red	0.25	2	Blue	6	6	0.7143	19	6	0.7692
				Orange	2	14		12	41	
	Green	0.25	2	Blue	8	4	0.8214	18	7	0.8077
				Orange	1	15		8	45	
	Blue	0.5	2	Blue	9	3	0.8214	17	8	0.7564
				Orange	2	14		11	42	
	Gray	0.25	2	Blue	8	4	0.8517	18	7	0.8590
				Orange	0	16		4	49	

B	SVM model parameters	Accuracy	Sensitivity	Specificity
	Training set	1.0000	1.0000	1.0000
	Test set	0.8571	1.0000	0.8000
	Cross validation	0.8590	0.8182	0.8750

List of figures

Figure 1

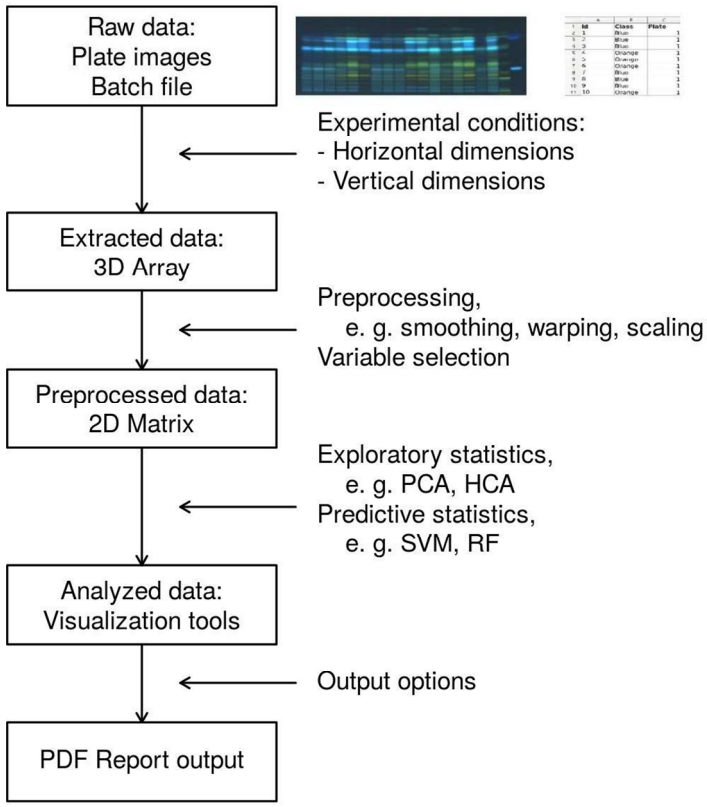

Raw data:
Plate images
Batch file
Experimental conditions:
- Horizontal dimensions
- Vertical dimensions
Extracted data:
3D Array
Preprocessing,
e. g. smoothing, warping, scaling
Variable selection
Preprocessed data:
2D Matrix
Exploratory statistics,
e. g. PCA, HCA
Predictive statistics,
e. g. SVM, RF
Analyzed data:
Visualization tools
Output options
PDF Report output

Figure 2

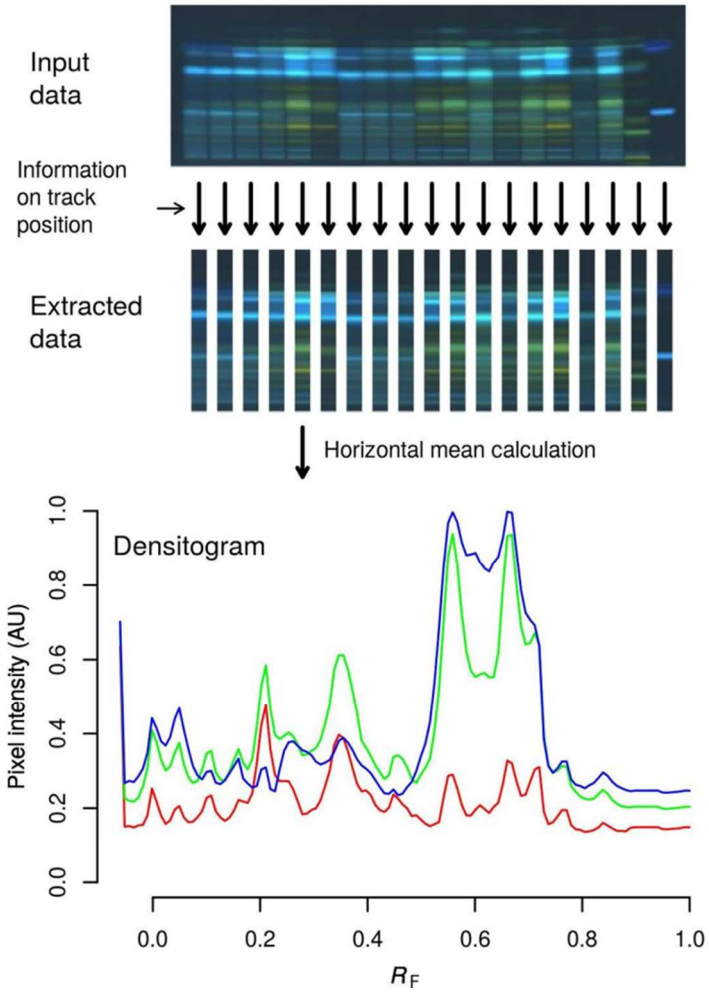

Input data
Information on track position
Extracted data
Horizontal mean calculation
Densitogram
Pixel intensity (AU)
0.0
0.2
0.4
0.6
0.8
1.0
R_F

Figure 3

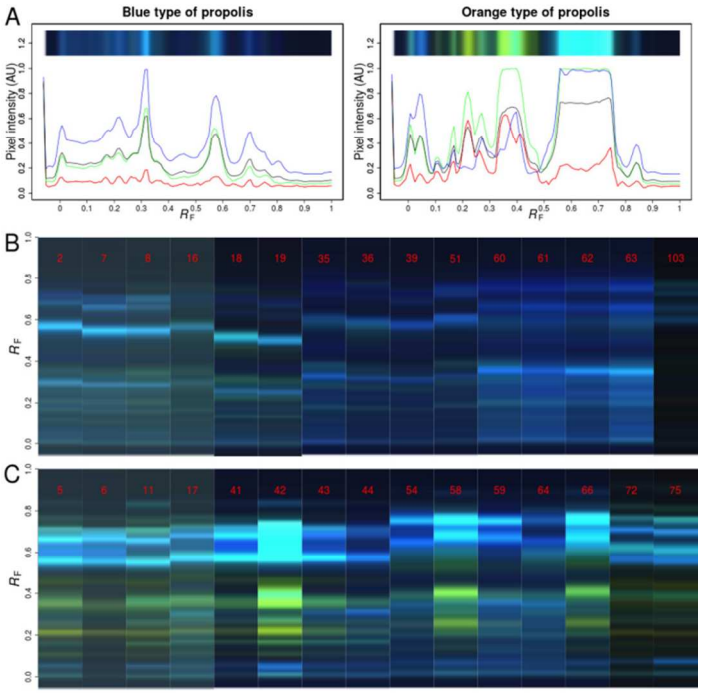

A
Blue type of propolis
Orange type of propolis
Pixel intensity (AU)
R_F
B
2 7 8 16 18 19 35 36 39 51 60 61 62 63 103
C
5 6 11 17 41 42 43 44 54 58 59 64 66 72 75

Figure 4

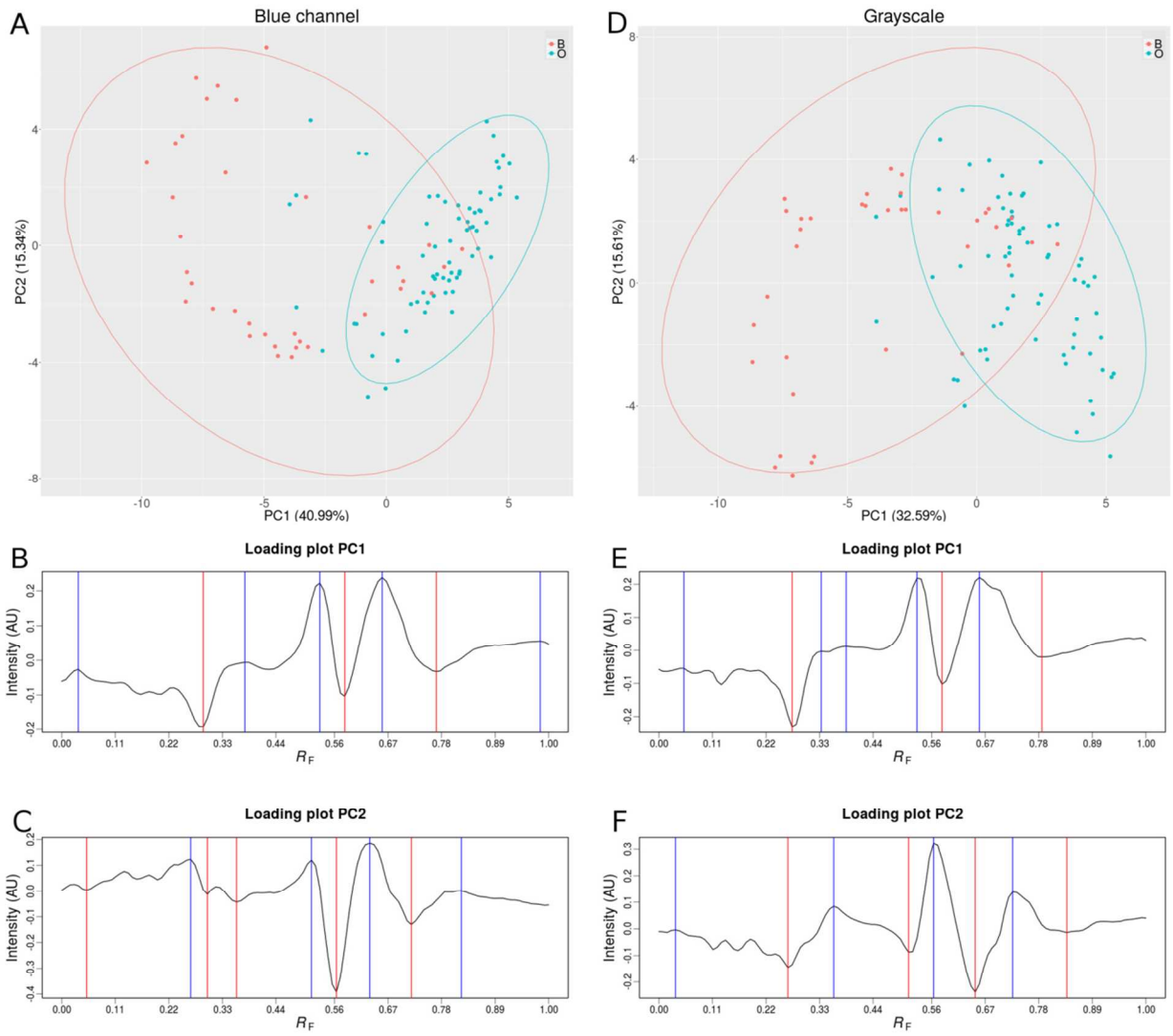

A
Blue channel
B
O
PC2 (15.34%)
PC1 (40.99%)
D
Grayscale
PC2 (15.61%)
PC1 (32.59%)
B
Loading plot PC1
Intensity (AU)
R_F
E
Loading plot PC1
C
Loading plot PC2
F
Loading plot PC2

Figure 5

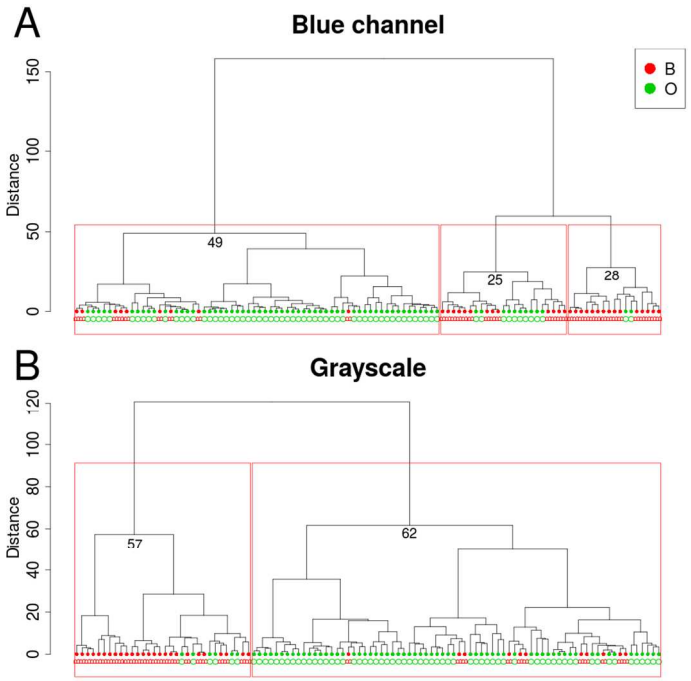
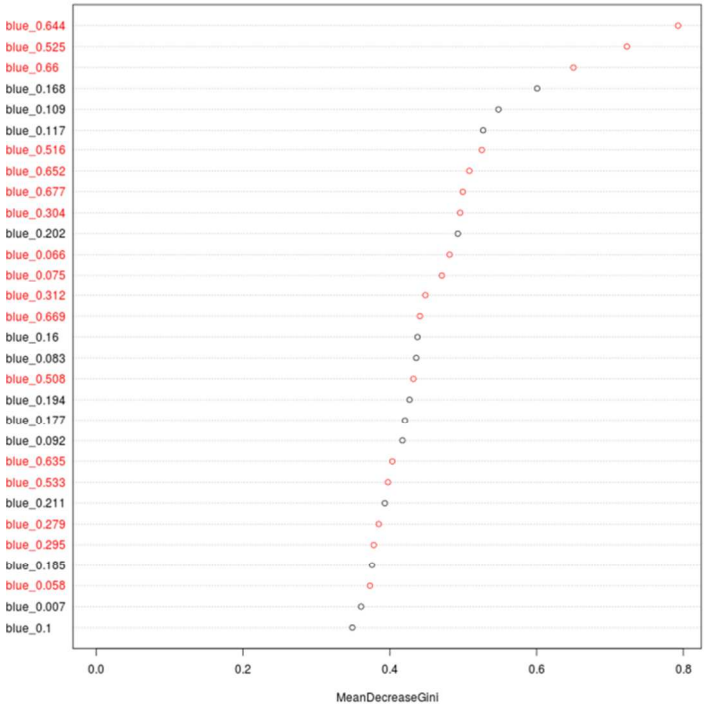

A
Blue channel
B
O
Distance
0
50
100
150
49
25
28
B
Grayscale
Distance
0
20
40
60
80
100
120
57
62

Figure 6


blue_0.644
blue_0.525
blue_0.66
blue_0.168
blue_0.109
blue_0.117
blue_0.516
blue_0.652
blue_0.677
blue_0.304
blue_0.202
blue_0.066
blue_0.075
blue_0.312
blue_0.669
blue_0.16
blue_0.083
blue_0.508
blue_0.194
blue_0.177
blue_0.092
blue_0.635
blue_0.533
blue_0.211
blue_0.279
blue_0.295
blue_0.165
blue_0.058
blue_0.007
blue_0.1
0.0
0.2
0.4
0.6
0.8
MeanDecreaseGini

for TOC only