Talanta 204 (2019) 739–746

Contents lists available at ScienceDirect

Talanta

journal homepage: www.elsevier.com/locate/talanta

Call it a “nightshade”—A hierarchical classification approach to identification of hallucinogenic Solanaceae spp. using DART-HRMS-derived chemical signatures

Samira Beyramysoltan, Nana-Hawwa Abdul-Rahman, Rabi A. Musah*

Department of Chemistry, State University of New York at Albany, 1400 Washington Ave, Albany, NY, 12222, USA

ARTICLE INFO

Keywords:
Hierarchical classification
Psychoactive plants
Seed species identification
Metabolome profiling
Direct analysis in real time-mass spectrometry
Chemometrics

ABSTRACT

Plants that produce atropine and scopolamine fall under several genera within the nightshade family. Both atropine and scopolamine are used clinically, but they are also important in a forensics context because they are abused recreationally for their psychoactive properties. The accurate species attribution of these plants, which are related taxonomically, and which all contain the same characteristic biomarkers, is a challenging problem in both forensics and horticulture, as the plants are not only mind-altering, but are also important in landscaping as ornamentals. Ambient ionization mass spectrometry in combination with a hierarchical classification workflow is shown to enable species identification of these plants. The hierarchical classification simplifies the classification problem to primarily consider the subset of models that account for the hierarchy taxonomy, instead of having it be based on discrimination between species using a single flat classification model. Accordingly, the seeds of 24 nightshade plant species spanning 5 genera (i.e. *Atropa*, *Brugmansia*, *Datura*, *Hyocyamus* and *Mandragora*), were analyzed by direct analysis in real time-high resolution mass spectrometry (DART-HRMS) with minimal sample preparation required. During the training phase using a top-down hierarchical classification algorithm, the best set of discriminating features were selected and evaluated with a partial least square-discriminant analysis (PLS-DA) classifier to discriminate and visualize the data. The method yields species identity through a class hierarchy, and reveals the most significant markers for differentiation. The overall accuracy of the approach for species identification was 95% and 96% using 100X bootstrapping validation and test samples respectively. The method can be extended for the rapid identification of an infinite number of plant species.

1. Introduction

The import of plants as reservoirs of useful compounds is exemplified in part by the observation that ~50% of the drugs approved for clinical use over the last 30 years are either directly from or are semi-synthetic derivatives of molecules from plants [1]. While plant-inspired medicines are often manufactured by synthetic methods, there remain a number of natural products in current clinical use whose synthesis by laboratory methods remains economically unfeasible, and in such cases, they are still isolated from plant tissue. For this reason, the ability to readily detect such compounds, track their occurrence in different parts of the plant, and perform comparative analysis of multiple species that contain the compounds of interest, remains a high priority.

Analysis of the alkaloids in Solanaceae family plants serves as a case in point. Several genera within this family contain species that produce the clinically-relevant and structurally-related tropane alkaloids atropine and scopolamine. These include *Atropa*, *Brugmansia*, *Datura*, *Hyocyamus* and *Mandragora*. Plant species represented by these genera, such as *Atropa belladonna* (aka deadly nightshade), *Brugmansia suaveolens*, *Datura stramonium* (aka Jimson weed), *Hyocyamus niger* (aka henbane) and *Mandragora officinarum* (aka mandrake) have been known since ancient times and have been referenced in the literature of Socrates, Aristotle, Hippocrates, Theophrastus, Avicenna [2–4], Dioscorides [4] and Pliny the Elder. These plants were used to induce sleep [5–7], as ingredients in the first surgical anesthetics [8], as aphrodisiacs, as beauty aids (e.g. for the dilation of pupils to enhance the appearance of the eyes) and even to commit murder, among many other uses. The advent of modern chemistry revealed atropine and scopolamine to be major components responsible for many of their therapeutic

[*] Corresponding author.
E-mail address: rmusah@albany.edu (R.A. Musah).

https://doi.org/10.1016/j.talanta.2019.06.010
Received 7 May 2019; Received in revised form 31 May 2019; Accepted 3 June 2019
Available online 04 June 2019

effects, and showed them to be anticholinergics. While the primary use of these plants in ancient times was medicinal and as a tool of "witchcraft" [9–11], recent years have witnessed exploitation of their mind-altering characteristics for recreational use [12–15] and in criminal activity (e.g. robberies [16], rendering victims in a conscious but compliant state during sexual assault [12,17,18], poisoning). However, the plants are also of horticultural importance, as the flowers of some species, such as several in the *Datura* and *Brugmansia* genera, are highly prized for their large inverted trumpet shapes.

Both atropine and scopolamine are used clinically [19], and much of their production relies on isolation from plant tissue because synthetic approaches remain cost prohibitive. While the compounds themselves are controlled substances in many countries, the plants from which they are derived usually are not. The alarming increase in the use of the plants to "legally" induce altered states of consciousness has been noted by the United Nations Office on Drugs and Crime, which has designated plants such as *Datura stramonium*, as "plants of concern". Because of their ubiquity as ornamentals, reservoirs of clinically important drugs, and use as hallucinogens and as poisons, Solanaceae genus products appear in diverse forms that can be very difficult to identify, particularly in the absence of prior knowledge of their source. For example, the seeds of multiple species that are used in horticulture and for the preparation of psychoactive brews look almost identical in several cases. Since atropine and scopolamine appear prominently throughout the plant matrix, species attribution based on the mere observation of these compounds in a sample is not definitive. Identification is even more difficult if the seeds or other plant material has been processed, because this makes determination of the plant from which it was derived nearly impossible to trace. Since the genomes of the majority of these plants have not been mapped, DNA analysis cannot usually be used as a means of identification.

In principle, it should be possible to determine the species from which a given Solanaceae genus product is derived based on its small molecule profile, since this would be expected to be defined by its unique genome. Recently, Liu et al. demonstrated that chemometric processing of *Brassicacea* spp. volatiles could be used to infer species identity [20]. Similarly, Lesiak et al. showed that statistical analysis processing of the mass spectrometric profiles of members of the *Datura* genus could be used to accurately determine species identity [21]. However, species attribution in the case of complex plant matrices derived from multiple genera, and which include numerous species that all contain the same characteristic biomarkers, is a significantly more complex problem, the solution to which has remained elusive. Tropane alkaloid-containing Solanaceae species plants represent such an example, since there are multiple genera and many more species, all containing atropine and scopolamine. Given that the agricultural, medicinal and forensic importance of the species within the constituent genera are intimately tied to species identity, it is highly desirable to be able to accomplish (in a single experiment, ideally): (1) determination of the presence of compounds of interest; (2) ability to distinguish between plant materials that contain identical biomarkers but which represent different species; (3) monitor the content of biomarkers of interest rapidly and in real-time; and (4) confirm species identity. Additional desirable attributes of such a method include rapid sampling with minimal to no sample preparation. While a number of methods exist for the accomplishment of a subset of these [22–27], an approach that would also enable simultaneous species identification continues to be a formidable challenge.

An approach proposed to circumvent these challenges is direct analysis in real time-high resolution mass spectrometry (DART-HRMS). This method has the advantage of requiring little sample preparation, and a mass spectrum can be acquired in under 1 min. Furthermore, it has been previously shown that the overall chemical fingerprint acquired by DART-HRMS can be used for species identification [21].The work described herein tested 24 atropine and scopolamin-containing plant seeds spanning 5 genera. We demonstrate that their DART-HRMS-derived chemical profiles exhibit intraspecies similarities and interspecies differences, and that the mass spectra can be processed using a statistical analysis workflow, to furnish not only genus, but also species identity. In the course of this work, a robust mass spectrometric database of atropine/scopolamine containing plant species was built, and it is shown that it can be used to rapidly and reliably screen direct analysis in real time-high resolution mass spectrometry data generated from seed unknowns, to accurately determine species identity. Furthermore, the samples can be analyzed directly with no pretreatment required. Species-level classification was accomplished by: (1) design of a top-down hierarchical classification approach for the training of a number of multi-class models, with consideration of the hierarchical relationships between samples at the genus and species levels; and (2) discovery of species-specific diagnostic masses that were optimal in enabling discrimination between classes. The proposed approach can be readily extended for the rapid identification of an infinite number of plant species.

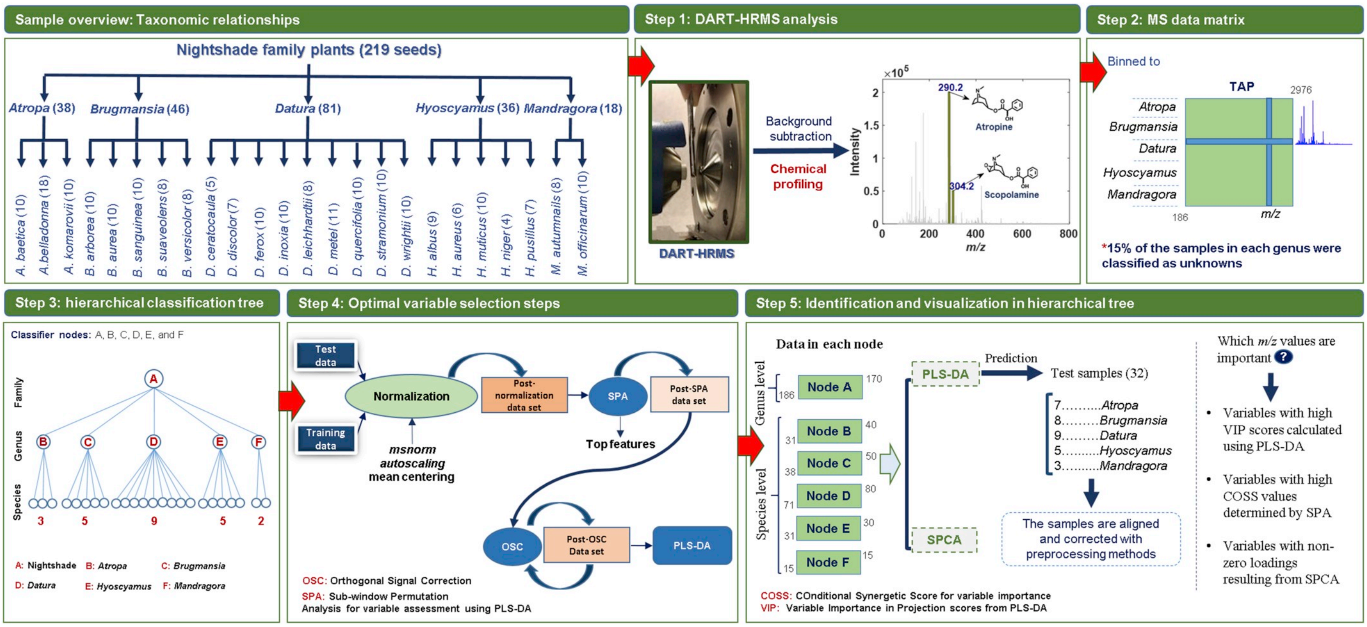
2. Materials and methods

2.1. Materials

Samples were comprised of 219 seeds from nightshade family plants. The multiple vendors from whom the samples were acquired are listed in Table S1. The genera and species studied are listed in Scheme 1 (see Sample overview-Taxonomic relationships), which also provides an outline of the approach taken to accomplish species classification. The following 5 genera were represented: *Atropa* (38 samples), *Brugmansia* (46 samples), *Datura* (81 samples), *Hyoscyamus* (36 samples), *and Mandragora* (18 samples). The number of species analyzed within each genus were as follows: *Atropa* (3 species): *A. baetica*, *A. belladonna* and *A. komarovii*; *Brugmansia* (5 species): *B. arborea*, *B. aurea*, *B. sanguinea*, *B. suaveolens* and *B. versicolor*; *Datura* (9 species): *D. ceratocaula*, *D. discolor*, *D. ferox*, *D. inoxia*, *D. leichhardtii*, *D. metel*, *D. quercifolia*, *D. stramonium* and *D. wrightii*; *Hyoscyamus* (5 species): *H. albus*, *H. aureus*, *H. muticus*, *H. niger* and *H. pusillus*; *Mandragora* (2 species): *M. autumnalis* and *M. officinarum*.

2.2. Instrumentation

Mass spectral data were collected in positive-ion mode over the mass range *m/z* 40–1100 using a DART-SVP ion source (IonSense, Saugus, MA, USA) coupled to a JEOL AccuTOF mass spectrometer (JEOL USA, Peabody, MA, USA) with a resolving power of 6000 FWHM. The helium gas flow rate for the DART ion source was set to 2.0 L/min, and the gas heater temperature and the DART ion source grid voltage were set to 350 °C and 50 V respectively. The optimal setting for the ring lens, orifice 1, orifice 2, and peaks voltages were 5, 20, 5 and 400 V respectively. Polyethylene glycol 600 (PEG) (Sigma-Aldrich, Burlington, MA, USA) was used as a mass callibrant. For DART-HRMS analysis, each seed was divided into four segments using a razor blade and each of the pieces was sampled directly by presenting it to the open-air space between the ion source and mass spectrometer inlet using a pair of stainless steel tweezers. The generated DART-HRMS spectra for each sample represent the average of the spectra of the four seed segments. Step 1 in Scheme 1 shows a representative DART-HRMS spectrum of a seed sample. The Mass spectral data were stored in text format after performing data processing steps including background subtraction, mass calibration with PEG 600 and peak centroiding using TSSPro3 software (Schrader Analytical Labs, Detroit, MI, USA). The raw spectral data, in the form of two column tables comprised of *m/z* and relative intensity values respectively for the detected peaks, were imported into MATLAB 9. 3. 0, R2017b Software (The MathWorks, Natick, MA, USA) for multivariate statistical analysis.



Scheme 1. Workflow devised for the analysis and identifictaion of plant species in the nightshade family.

2.3. Statistical analysis

Multivariate statistical analysis methods were applied to the mass spectral data acquired from plant samples to achieve discrimination between species and reveal the presence of diagnostic markers. The overall approach is outlined in Steps 2 through 5, presented in Scheme 1. These steps are described below:

Step 2: Following DART-HRMS analysis (i.e. Step 1), a subset of the data, comprised of 186 randomly selected sample spectra (85% of the total number) were used for development of the training model, and the remaining 32 samples (i.e. 15%) were subsequently used to test the developed approach. Within the training subset, the number of samples from each species was based on the relative proportions of the species that were represented in a given genus. For example, 31 samples were selected from a total of 38 *Atropa* seeds for the training data. The mass spectra of the training set were aligned in a matrix according to common m/z values (i.e. binned-see Step 2). In this process, the optimal bin width was determined to be within ± 15 mmu of the observed mass, and the relative abundance threshold was set to 0.2% of the maximum intensity. This furnished a matrix of dimension 186 (i.e. number of samples) by 2976 (number of m/z values), and which represented all of the tropane alkaloid-containing plant (TAP) species studied.

Step 3: Hierarchical classification simplfies the classification problem such that the primary consideration is the subset of classification models that account for the hierarchy taxonomy, instead of being based on specie discrimination accomplished through use of a flat classification model [28]. Therefore, inspired by the taxonomic relationships between samples in terms of genera within a family and species within each genus, a “top-down” tree structure-hierarchical classification workflow, which enabled visualization, discrimination and prediction of sample classes was used. This was accomplished using an in-house written MATLAB program applied to the matrix generated in Step 2. As illustrated in Scheme 1, Step 3, the classification tree was designed to have two discrimination levels (one for genus and one for species). Its first node, termed “A” represents the Nightshade (i.e. Solanaceae) family level. From it, the nodes of the first level of discrimination, representative of the various genera, were derived (i.e. B, C, D, E and F, denoting *Atropa*, *Brugmansia*, *Datura*, *Hyocyamus* and *Mandragora* respectively). Within these, the species classes could be observed and distinguished in the second level of discrimination.

Step 4: Information on the attributes within the dataset that enabled accomplishment of the clustering observed by hierarchical clustering analysis (HCA), was extracted through a series of iterative operations summarized in Scheme 1, Step 4. These operations were performed on the data representative of each node, in order to further reduce the dimensionality of the data to contain the best features that enabled discrimination between genera and species. These in turn could then be used to classify sample unknowns. First, each node dataset was normalized, with the optimal normalization method being determined through subjection of the data to *msnorm*, autoscaling and mean centering approaches. Following determination of the best normalization method in each case, outliers were detected by performing principal component analysis (PCA), and assessing the PCA results by using the Hotelling's T-squared test on the first two PCs. Best feature selection from the resulting reduced data matrices was accomplished by sub-window permutation analysis (SPA) [29], using PLS-DA as the underlying classifier. In SPA, the number of Monte Carlo simulations was set to 500, with the training dataset (~85% of samples) and m/z values (150) being sampled in each step. The resulting matrices (representing the post-SPA dataset) were analyzed by cross validated PLS-DA to find the optimal number of latent variables (LVs) required to create a PLS-DA model. The remaining 15% of samples (i.e. the validation dataset) were permuted and analyzed by PLS-DA to investigate the prediction performance of the individual variables. The root mean squared error (RMSE) of prediction values that were observed in this process were recorded and used to calculate p-values and the conditional synergetic score (COSS) value (COSS = − log10 (p)) for variable importance, with higher scores reflecting higher variable importance. A variable with $p \leq 0.05$ has a COSS of ≥1.3. The training dataset was further reduced by using those variables with the higher COSSs. It should be noted that to select the optimum number of the best features revealed by SPA, m/z values sorted in order of descending COSS value were analyzed in an iterative manner, starting from 10, and increasing by increments of 10, to a total of 300. Finally, to eliminate variance that was not correlated to a discriminative response (i.e. in order to address the possibility that the eliminated variance might be correlated with within class variances, or represent a proportion of the background that was corrected in earlier steps), orthogonal signal correction (OSC) [30] was implemented. This operation was also performed iteratively, in that from

1 up to 4 (up to the maximum number of classes in classification problems with lower than 5 classes) within each node were analyzed to find the optimum number to use. For example, orthogonal components in the range of 1 through 3 were examined for node B data, while from 1 to 4 were analyzed for node D data. The results of the iterative operations performed in Step 4 were 6 matrices representing nodes A, B, C, D, E and F, of dimensions 186 × 170; 31 × 40; 38 × 50; 71 × 80; 31 × 30; and 15 × 15 respectively. It should be noted that for each iteration, assessment of the best pre-processing approach was validated by 5-fold cross validation, and determination of the accuracy of class prediction using test dataset samples (i.e. the subset of 15% of the samples that were reserved for testing the models).

Step 5: The reduced data matrices generated in Step 4 served as the input for Step 5. Two operations were performed on this data: PLS-DA and sparse principle component analysis (SPCA). The former was performed in order to accomplish sample classification and prediction. The number of LVs for PLS-DA was set to the number of components that were found to be optimal. SPCA [31] was used to aid visualization of the best discriminating features in each node. It uses a lasso (elastic net) to produce modified principal components with sparse loadings. In summary, to reveal the *m/z* values with the most potential to serve as markers enabling distinctions between classes to be made, PLS-DA along with the results of exploration by SPCA were considered for each node. Then, to rank the identified potential features (i. e. *m/z* values), calculated COSS (conditional synergetic score for variable importance) values from SPA in Step 4, as well as variable importance in projection (VIP) scores from PLS-DA in Step 5, were considered.

3. Results and discussion

Accomplishment of discrimination between species of tropane alkaloid containing seeds using DART-HRMS data was of interest in this study. The 219 seeds of species in five genera, namely *Atropa*, *Brugmansia*, *Datura*, *Hyoscyamus*, and *Mandragora* were investigated. Although mass measurements were made in the range *m/z* 40–1100, the data actually used were in the *m/z* range of 40-700, since the masses above 700 were not informative (i.e. they did not enhance the predictive capacity of the model). Fig. 1Panels I through V displays averaged composite DART mass spectra of the species within the indicated genera (i.e. *Atropa*, *Brugmansia*, *Datura*, *Hyoscyamus* and *Mandragora* respectively). At first glance, the spectra representing the pairs *Hyoscyamus/Mandragora* (Panels IV and V), and *Brugmansia/Datura* (Panels II and III) appear very similar. In each of the panels, the most prominent *m/z* values are labeled and include 124.11, 193.06 and 290.17 for *Atropa*; 144.09, 158.11, 174.11 and 304.15 for *Brugmansia*; 142.12, 174.11, 290.17 and 304.15 for *Datura*; 127.04, 281.24, 290.17 and 298.27 for *Hyoscyamus* and 61.03, 90.08, 96.04, 127.04, 142.12, 145.05, 281.24, 290.17 and 298.27 for *Mandragora*. With the exception of *m/z* 61.03 which was absent in *Datura*, the five genera contained (to differing extents) all of the aforementioned prominent peaks.

Representative mass spectra of the 24 species analyzed in this study are shown in Fig. S1, and they illustrate that *m/z* 290.2 and 304.2 (for protonated atropine and scopolamine respectively) are prominent in most samples. Atropine was the base peak in the *Atropa* genus, and it was also very prominent in *Datura*, *Hyoscyamus* and *Mandragora* seeds. Scopolamine was well-represented in *Datura*, *Brugmansia* and *Hyoscyamus* genera. Confirmation of the presence of scopolamine and atropine was accomplished through in source collision induced dissociation (CID) experiments as previously described [21].

As outlined in Scheme 1-Step 2, a data matrix of dimension 186 × 2976 (termed “Tropane Alkaloid Plants” and abbreviated “TAP”) was generated to begin development of a model to accomplish species discrimination. The first number in the matrix refers to the number of spectra (i.e. observations) and the second, the number of *m/z* values (i.e. variables). The matrix was subjected to HCA, and the resulting “top-down” hierarchical classification tree is presented in Scheme 1-Step 3. Using HCA, the data were explored to assess the similarities between the represented genera and species (i.e. the hierarchical relationships within the family members) based on the DART-HRMS data. The species spectral replicates were scaled using the *msnorm* function in MATLAB, and then averaged to create a new matrix with dimension 24 × 2976. HCA was applied to the first principal component (PC) resulting from PCA. This PC accounted for ~50% of the observed variance. The plant species clustered within subgroups based on the *Euclidean* distance using the unweighted average distance (UPGMA) method. Fig. 2 illustrates the resulting dendrogram which shows clustering as a function of species and provides an indication of the relative closeness/relatedness of the genera. Two main clades were computed (labeled 1 and 2) and they show a major branch point between *Mandragora* and *Hyoscyamus* on the one hand, and *Atropa*, *Datura* and *Brugmansia* respectively on the other. Cluster 1 is divided into the two subgroups labeled 3 and 4 for *Mandragora* and *Hyoscyamus* respectively.

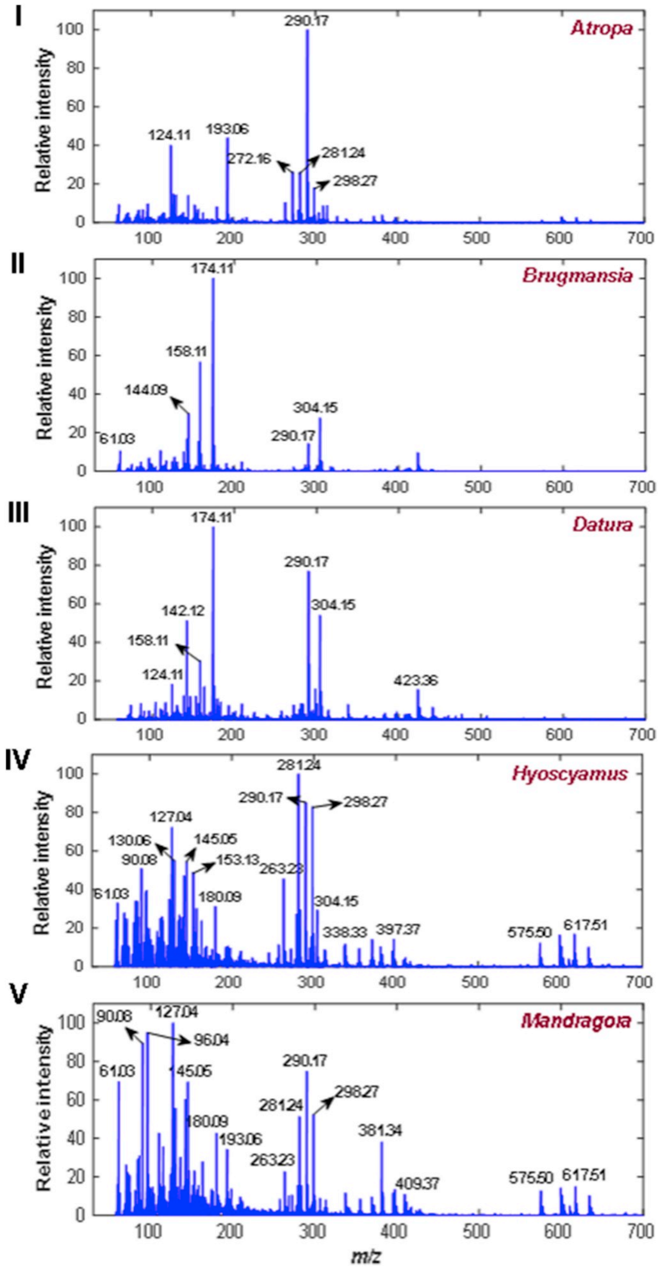


Fig. 1. Representative DART-HRMS averaged spectra of the species within the five studied genera. *Atropa* (I); *Brugmansia* (II); *Datura* (III); *Hyoscyamus* (IV); and *Mandragora* (V).

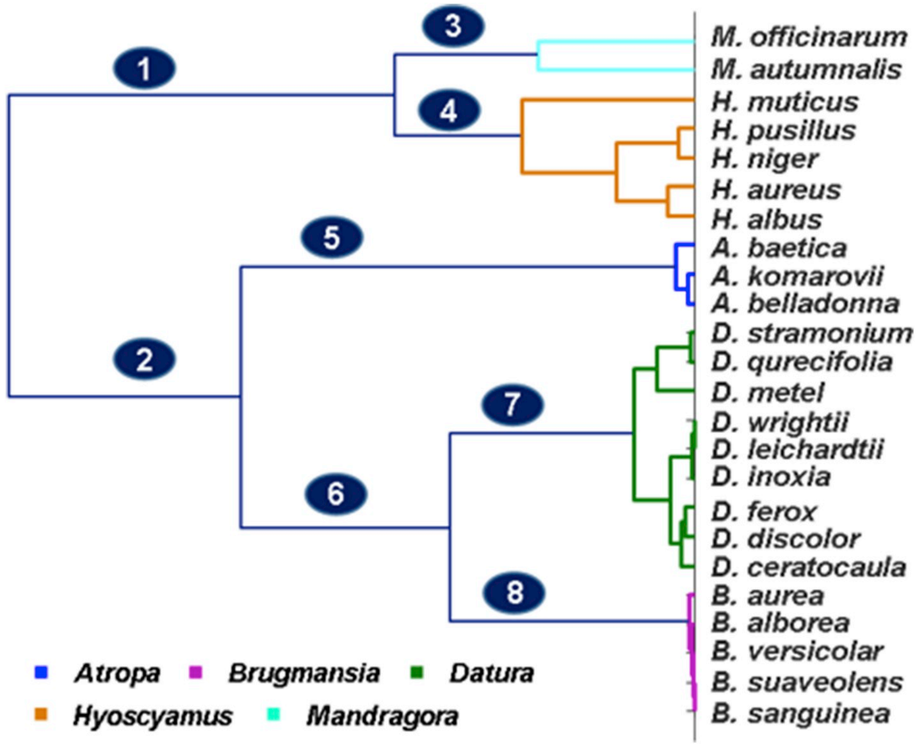


Fig. 2. Dendrogram representation of HCA results based on the first principal component from PCA of the combined species matrix (24 × 230). Labels show the two main clusters and six sub-clusters. The two main clusters contain genera *Mandragora* and *Hyoscyamus* on the one hand, and *Atropa*, *Datura* and *Brugmansia* on the other.

Cluster 2 is comprised of the two subgroups 5 (involving the genus *Atropa*) and 6, which is split into subgroups 7 and 9 representing the *Datura* and *Brugmansia* genera respectively. The emergence of these groupings reveals that the chemical profiles of the studied plants are dissimilar enough to enable classification not only at the genus level, but also in terms of species. Therefore, the TAP matrix was introduced into the "top-down" hierarchical classification tree to characterize the best pre-processing steps (as presented in Scheme 1-Step 4) and the discriminating set of features for classification with PLS-DA in the 6 classifier nodes labeled A, B, C, D, E, and F in Scheme 1. The nodes A, B, C, D, E and F specify the classification of 5 genera in the nightshade family, namely species within the *Atropa, Brugmansia*, *Datura*, *Hyoscyamus*, and *Mandragora* respectively. The suitability of the steps applied to each node, was confirmed by 5-fold venetian blind cross validation, and assessment of the accuracy of the test set predictions.

For normalization, scaling the intensities of the peaks in every spectrum to the maximum intensity of the base peak using the *msnorm* function of MATLAB was found to work for the data in all nodes except the *Brugmansia* case (node C). For this node, the best normalization method was found to be *autoscaling*. Examination of the data by PCA combined with Hotelling's T-squared test revealed no outliers. Furthermore, it was determined by OSC that the optimum number of orthogonal components was 1 for nodes A, B, D, E and F, and 4 for node C. The application of SPA to each node provided a ranking of the variables (i. e. *m/z* values) that were of significance in enabling classification using the PLS-DA model. From this, 170, 40, 50, 80, 30 and 15 variables were found to be the most effective *m/z* values for the classifications in nodes A, B, C, D, E and F respectively. These preprocessing steps resulted in extraction of the most informative data to yield matrices with dimensions of 186 × 170, 31 × 40, 38 × 50, 71 × 80, 31 × 30 and 15 × 15 for nodes A, B, C, D, E, and F respectively. The number of latent variables required to build the PLS-DA models were found to be 7, 3, 4, 11, 5, and 2 in A, B, C, D, E, and F nodes respectively. The PLS-DA models explained 85% of the MS spectral data variance and 84% of the response-variance in node A; 97% and 90% in node B; 50% and 75% in node C; 97% and 82% in node D; 96% and 87% in node E; and 91% and 98% in node F. The 5-fold cross validation result details (accuracy and error rate) of the PLS-DA models are presented in Table S2, while the merits of the discrimination model for each class (i.e. sensitivity, specificity, and precision) are displayed in Table S3. The PLS-DA scores (with a specific color used to define each class) and loadings plots (which indicate the *m/z* values that correspond with the loadings coordinates) are illustrated in Figs. S2–S7 for nodes A through F respectively. The figures made discrimination visually apparent. For example in Fig. S2, the *Atropa*, *Mandragora*, *Hyoscyamus, Brugmansia* and *Datura* genera are discriminated from each other based on the first two latent variables. To examine the overall accuracy of the trained top-down hierarchical classification algorithm, a 100X bootstrapping of a random sampling of the training set was performed. In each repetition of the bootstrapping, 158 out of 186 samples were randomly resampled to train the model, and the model was then applied to test the remaining samples (i. e. 28). The results were integrated and used to compute the hierarchy performance characteristics. The overall accuracy was determined to be 95% for species specification using the hierarchical classification tree, while the accuracy was 86% for a flat PLS-DA classification for the 24 species. The validation result details and the merits of the species identification (i. e. sensitivity, specificity, and precision) are displayed in Table 1. The test samples (i. e. 32 in number) comprised of 7 *Atropa*, 8 *Brugmansia*, 9 *Datura*, 5 *Hyoscyamus* and 3 *Mandragora* species, were used to test the workflow strategy. The prediction results of the test set are presented in Table 2. At the genus level, all 32 samples were predicted correctly. At the species level, all samples in the *Atropa*, *Hyoscyamus* and *Mandragora* genera were correctly predicted using the corresponding nodes. One *Brugmnasia* sample (*B. suaveolens*) and one *Datura* sample (*D. stramonium*) were not identified correctly. Overall, these results illustrate that the model has the ability to predict species identity from seeds with high accuracy.

The six extracted data sets were analyzed using SPCA to visualize the structure within each node. SPCA computes the sparse loadings with many values equal to zero while incorporating distinguishing structural information between classes. The sparse loadings simplify the interpretation of the principal components based on a subset of variables. The two first loading vectors contain non-zero values equal to 20

Table 1
The merits of the 100X bootstrap validation of the trained top-down hierarchical classification tree used for species identification (i.e. overall accuracy, sensitivity, specificity, and precision).

Overall accuracy: 0.95

Species	Classification model performance		
	Sensitivity	Specificity	Precision
A. baetica	0.99	0.99	0.90
A. belladonna	0.98	1	1
A. komarovii	1	1	0.99
B. arborea	1	1	1
B. aurea	0.88	1	1
B. sanguinea	1	1	1
B. suaveolens	1	1	1
B. versicolor	1	1	1
D. ceratocaula	0.81	1	1
D. discolor	1	1	0.98
D. ferox	1	1	1
D. inoxia	0.76	1	0.90
D. leichhardtii	1	1	1
D. metel	1	0.98	0.79
D. quercifolia	0.90	0.99	0.86
D. stramonium	0.96	0.99	0.89
D. wrightii	0.90	1	0.96
H. albus	0.96	1	1
H. aureus	1	0.99	0.87
H. muticus	0.97	1	0.99
H. niger	1	1	0.98
H. pusillus	0.83	1	1
M. autumnalis	1	1	1
M. officinarum	0.97	1	1

Table 2
PLS-DA model prediction results for the test samples (i.e. sample unknowns). The true (i.e. correct) genus and species labels are listed in the first two columns, and the genus and species-level predictions are listed in the last two columns. The test samples (32 in total) were randomly selected observations, with the number of samples of each species (indicated in parentheses) being reflective of the proportion of that species that was represented in the total number of samples analyzed in this study.

Test sample true label		Test sample predicted label	
Genus	Species	Genus level	Species level
Atropa (7[a])	*A. beatica (2)*	*Atropa*	True
	A. belladonna (3)		True
	A. belladonna (2)		True
Brugmansia (8)	*B. arborea (2)*	*Brugmansia*	True
	B. aurea (2)		True
	B. sanguinea (2)		True
	B. suaveolens (1)		False
	B. versicolor (1)		True
Datura (9)	*D. discolor (1)*	*Datura*	True
	D. ferox (1)		True
	D. inoxia (1)		True
	D. leichhardtii (1)		True
	D. metel (2)		True
	D. quercifolia (1)		True
	D. stramonium (1)		False
	D. wrightii (1)		True
Hyoscyamus (5)	*H. albus (1)*	*Hyoscyamus*	True
	True		True
	H. aureus (1)		True
	H. muticus (1)		True
	H. niger (1)		True
	H. pusillus (1)		True
Mandragora (3)	*M. autumnalis (1)*	*Mandragora*	True
	M. officinarum (2)		True

[a] The numbers within parentheses indicate the number of each genus and species in the test samples.

and 10 in node A; 3 and 4 in node B; 25 and 25 in node C; 15 and 8 in node D; 10 and 4 in node E; and 4 and 4 in node F. The two first computed principal components explained 43, 61, 25, 71, 57, and 42% of the variance for the data representative of nodes A, B, C, D, E, and F respectively. Fig. 3 illustrates the corresponding SPCA scores plots for PCs 1 and 2. In addition, the SPCA scores and loadings bar plots are presented in Figs. S8–S13, Panels A and B. These illustrations provide a way to visualize the discrimination structure between classes, and reveal which variables are most discriminative.

3.1. Determination of diagnostic markers

The results show that DART-HRMS-derived chemical signatures can provide adequate information for classification and identification of the genera and species of the psychoactive seeds studied here. However, an important question that remained to be investigated was which *m/z* values (i. e. which molecular components) in the DART-HRMS-derived metabolome profile were responsible for discrimination. To answer this question, the variable importance of *m/z* values that resulted from the application of SPCA and PLS-DA for each node of the hierarchical classification tree were examined. The weighted relative importance of *m/z* values in terms of SPCA can be extracted from the computed loadings values, which is shown in Panel B of Figs. S8–S13. Table S4 shows the variables with calculated PLS-DA VIP scores of > 1. In addition, the corresponding COSS values (resulting from SPA coupled with PLS-DA) are displayed for each variable. The informative variables revealed by SPA were considered to be those with COSS values of > 1.3.

The data, rendered as heatmaps representing the *m/z* values associated with VIP scores of > 1, are displayed in Fig. 4, along with horizontal and vertical dendrograms representing the correlation between *m/z* values and sample identification, respectively. Dendrograms were computed by HCA based on *Euclidean* distance and the UPGMA method. In each of the plots in Fig. 4, the horizontal axis shows the discriminative *m/z* values, and the vertical axis shows the classes. Colors (red, green and black) are reflective of the relative intensities of the indicated *m/z* values, with red conoting a high value, green a low value and black, an intensity of zero. From this rendering, the m/z values that are important in enabling a given species to be distinguished from others is visually apparent. For example, *m/z* 174.1148 is relatively intense (as indicated by the red color corresponding to this mass) for *Datura* and *Brugmansia*, but is of much lower intensity in the other species (as indicated by the green color) (Fig. 4A). This indicates that it is important for enabling discrimination between *Datura* and *Brugmansia* from the other genera. The heat map structure for the *Atropa* genus (Fig. 4B) shows that while *m/z* 193.0501 and 272.1601 are absent in *A. baetica* (indicated by the black color), they are present with high intensities in the two other species. On the other hand, *m/z* 281.2442 and 298.2717 are of high intensity in *A. baetica* (red color), but low intensity for the other two Atropa spsecies. Thus, it is clear that these four masses are most impactful in enabling discrimination between *A. baetica* and the other species (i.e. *A. komarovii* and *A. belladonna*). On the other hand, *m/z* 127.0407, 145.0506 and 153.1258 were important in *A. komarovii* discrimination. Fig. 4C shows the 4 main clusters of variables (indicated in red) that define the relationship structure between the species in the *Brugmansia* genus. For the *Datura* case (Fig. 4D) the pattern differences were a consequence of specific masses that were highly correlated to each individual species. Within the *Hyocyamus* genus (Fig. 4E), it was apparent that *H. niger* could be separated from the other species based on *m/z* 281.2442. For *Mandragora* (Fig. 4F), the *m/z* values 381.3498, 290.1726, 145.0506, 130.0521 and 31.0306 had a positive impact on differentiating between both *M. officinarum* and *M. autumnalis*. *H. niger* can be separated from other species based on *m/z* 281.2442 (Fig. 4F). The *m/z* values 290.1726, 145.0506, 130.0521, 381.3498 and 61.0306 had a positive impact of differentiating between both *M. officinarum* and *M. autumnalis* in the *Mandragora* genus (Fig. 4E).

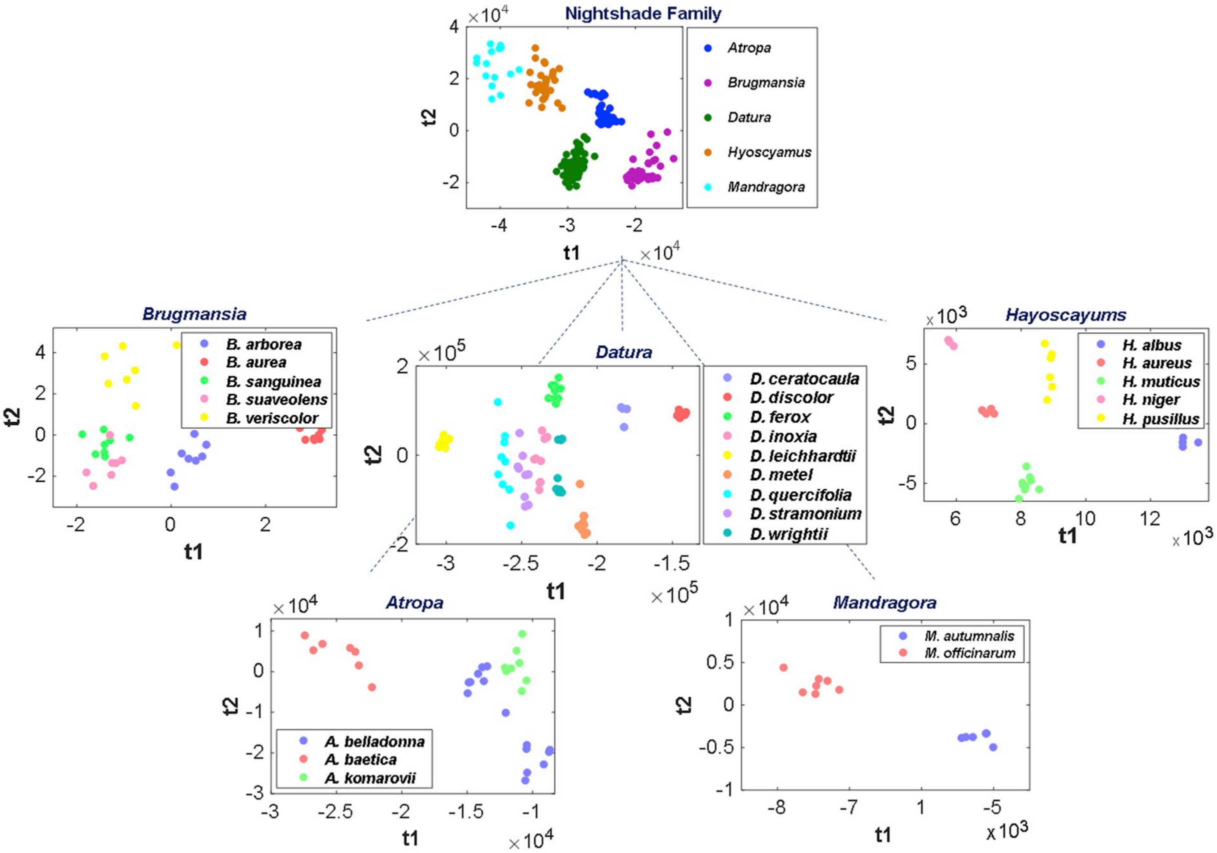


Fig. 3. Data visualization in each top-down hierarchical classification tree node using SPCA. The scores plots represent the first two principal components. The samples corresponding to each class are specified by color.

It is noteworthy that no *m/z* values unique to a given species were revealed by in the DART-HRMS analysis. Nevertheless, it is possible, based on the observation of species-specific small-molecule signatures, to distinguish between species. The results of PLS-DA and SPCA were 80% similar in revealing the features that were important in enabling discrimination between features.

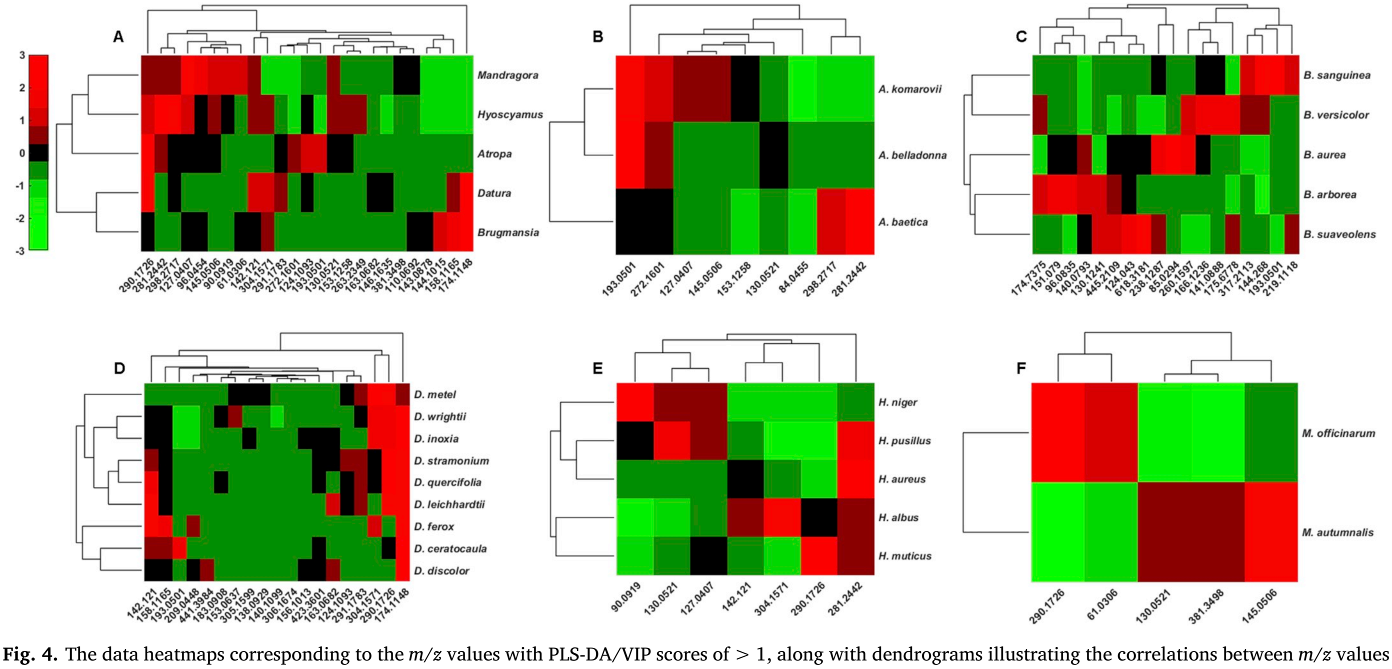


Fig. 4. The data heatmaps corresponding to the *m/z* values with PLS-DA/VIP scores of > 1, along with dendrograms illustrating the correlations between *m/z* values and samples. Dendrograms were calculated by HCA based on *Euclidean* distance and the UPGMA method. The relative intensities of for the represented *m/z* values were scaled such that the mean is 0 and the standard deviation is 1. Each heat map corresponds with features associated with a single node of the hierarchical classification tree as follows: (A) Tropane alkaloid-containg plant (TAP), (B) *Atropa*, (C) *Brugmansia*, (D) *Datura*, (E) *Hyocyamus*, and (F) *Mandragora*.

4. Conclusions

We demonstrate that DART-HRMS data in combination with a hierarchical classification model is an effective approach for identification of species of psychoactive plants in the nightshade family. The general method results in the ability to determine species identity within a few minutes, as opposed to rearing seeds to maturity to base species attribution on the gross morphological features of the reproductive parts (which can take years to accomplish). Here, the method was applied to differentiation of the seeds of multiple plant species of related Nightshade (Solanacea) family plants that have the common alkaloid biomarkers atropine and scopolamine. Using the metabolome profiles of the seeds that were furnished by DART-HRMS analysis, a top-down hierarchical classification method was developed and applied to discriminate and readily visualize the differences between species. In contrast to the approach of discriminating using a "flat" classification model, the hierarchical algorithm simplified the classification problem to the several discrimination models that revealed hierarchical relationships. By the 100X bootstrap validation method, this increased the accuracy of species identification from 84% to 95%. Here, the best pre-processing steps and the best set of features were selected in each classification node using the PLS-DA model.

Competing interest statement

There are no conflicts to declare.

Acknowledgements

This project was supported by grants to RAM from the United States National Science Foundation (#1429329) and the United States National Institute of Justice (award number 2015-DN-BX-K057). The opinions, findings, and conclusions or recommendations expressed are those of the authors and do not necessarily reflect those of the Department of Justice. Both grants are gratefully acknowledged.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.talanta.2019.06.010.

References

[1] A. Gurib-Fakim, Medicinal plants: Traditions of yesterday and drugs of tomorrow, Mol. Asp. Med. 27 (1) (2006) 1–93.
[2] E. Aziz, B. Nathan, J. McKeever, Anesthetic and analgesic practices in Avicenna's canon of medicine, Am. J. Chin. Med. 28 (1) (2000) 147–151.
[3] F.S. Haddad, The spongia somnifera, Middle East J. Anesthesiol. 17 (3) (2003) 321–327.
[4] V.A. Peduto, [The mandrake root and the Viennese Dioscorides], Minerva Anestesiol. 67 (10) (2001) 751–766.
[5] A. Stewart, B. Morrow-Cribbs, J. Rosen, Wicked Plants : The Weed that Killed Lincoln's Mother & Other Botanical Atrocities, 1st ed., Algonquin Books of Chapel Hill, Chapel Hill, N.C., 2009.
[6] H.N. Ellacombe, The Plant-Lore & Garden-Craft of Shakespeare, 2nd ed., W. Satchell and Co. etc., London,, 1884.
[7] J. Wynbrandt, The Excruciating History of Dentistry/Toothsome Tales & Oral Oddities from Babylon to Braces, 1st ed., St. Martin's Press, New York, 1998.
[8] A.J. Carter, Narcosis and nightshade, BMJ 313 (7072) (1996) 1630–1632.
[9] M.R. Lee, Solanaceae III, Henbane, hags and hawley harvey crippen, J. R. Coll. Phys. Edinb. 36 (4) (2006) 366–373.
[10] M.R. Lee, Solanaceae IV: Atropa belladonna, deadly nightshade, J. R. Coll. Phys. Edinb. 37 (1) (2007) 77–84.
[11] J.L. Muller, Love potions and the ointment of witches: Historical aspects of the nightshade alkaloids, J. Toxicol. Clin. Toxicol. 36 (6) (1998) 617–627.
[12] J. Saiz, T.D. Mai, M.L. Lopez, C. Bartolome, P.C. Hauser, C. Garcia-Ruiz, Rapid determination of scopolamine in evidence of recreational and predatory use, Sci. Justice 53 (4) (2013) 409–414.
[13] M.M. Glatstein, F. Alabdulrazzaq, F. Garcia-Bournissen, D. Scolnik, Use of physostigmine for hallucinogenic plant poisoning in a teenager: Case report and review of the literature, Am. J. Therapeut. 19 (5) (2012) 384–388.
[14] P. Nikolaou, I. Papoutsis, M. Stefanidou, A. Dona, C. Maravelias, C. Spiliopoulou, S. Athanaselis, Accidental poisoning after ingestion of "aphrodisiac" berries: Diagnosis by analytical toxicology, J. Emerg. Med. 42 (6) (2012) 662–665.
[15] S.P. Spina, A. Taddei, Teenagers with Jimson weed (Datura stramonium) poisoning, CJEM 9 (6) (2007) 467–468.
[16] K.J. Lusthof, I.J. Bosman, B. Kubat, M.J. Vincenten-van Maanen, Toxicological results in a fatal and two non-fatal cases of scopolamine-facilitated robberies, Forensic Sci. Int. 274 (2017) 79–82.
[17] A. Negrusz, R.E. Gaensslen, Analytical developments in toxicological investigation of drug-facilitated sexual assault, Anal. Bioanal. Chem. 376 (8) (2003) 1192–1197.
[18] A. Ardila, C. Moreno, Scopolamine intoxication as a model of transient global amnesia, Brain Cogn. 15 (2) (1991) 236–245.
[19] M. Lochner, A.J. Thompson, The muscarinic antagonists scopolamine and atropine are competitive antagonists at 5-HT3 receptors, Neuropharmacology 108 (2016) 220–228.
[20] Y. Liu, H. Zhang, S. Umashankar, X. Liang, H.W. Lee, S. Swarup, C.N. Ong, Characterization of plant volatiles reveals distinct metabolic profiles and pathways among 12 brassicaceae vegetables, Metabolites 8 (4) (2018).
[21] A.D. Lesiak, R.B. Cody, A.J. Dane, R.A. Musah, Plant seed species identification from chemical fingerprints: A high-throughput application of direct analysis in real time mass spectrometry, Anal. Chem. 87 (17) (2015) 8748–8757.
[22] Z. Long, Y. Zhang, P. Gamache, Z. Guo, F. Steiner, N. Du, X. Liu, Y. Jin, L. Liu, Determination of tropane alkaloids by heart cutting reversed phase - Strong cation exchange two dimensional liquid chromatography, J. Chromatogr. B Analyt. Technol. Biomed. Life Sci. 1072 (2018) 70–77.
[23] J. Marin-Saez, R. Romero-Gonzalez, A. Garrido Frenich, Multi-analysis determination of tropane alkaloids in cereals and solanaceaes seeds by liquid chromatography coupled to single stage Exactive-Orbitrap, J. Chromatogr. A 1518 (2017) 46–58.
[24] H. Chen, J. Marin-Saez, R. Romero-Gonzalez, A. Garrido Frenich, Simultaneous determination of atropine and scopolamine in buckwheat and related products using modified QuEChERS and liquid chromatography tandem mass spectrometry, Food Chem. 218 (2017) 173–180.
[25] E. Aehle, B. Drager, Tropane alkaloid analysis by chromatographic and electrophoretic techniques: An update, J. Chromatogr. B Analyt. Technol. Biomed. Life Sci. 878 (17–18) (2010) 1391–1406.
[26] M. Cirlini, T.M. Demuth, A. Biancardi, M. Rychlik, C. Dall'Asta, R. Bruni, Are tropane alkaloids present in organic foods? Detection of scopolamine and atropine in organic buckwheat (Fagopyron esculentum L.) products by UHPLC-MS/MS, Food Chem. 239 (2018) 141–147.
[27] A. Romera-Torres, R. Romero-Gonzalez, J.L. Martinez Vidal, A. Garrido Frenich, Analytical methods, occurrence and trends of tropane alkaloids and calystegines: An update, J. Chromatogr. A 1564 (2018) 1–15.
[28] A. Freitas, C.N. Silla, A survey of hierarchical classification across different application domains, Data Min. Knowl. Discov. 22 (2011) 31–72.
[29] H.-D. Li, Q.-S. Xu, Y.-Z. Liang, libPLS: An integrated library for partial least squares regression and discriminant analysis, Chemometr. Intell. Lab. Syst. 176 (2018) 34–43.
[30] S. Wold, E. Johansson, M. Cocchi, PLS - Partial Least-squares projections to latent structures, in: H. Kubinyi (Ed.), 3D QSAR in Drug Design: Theory, Methods, and Applications, ESCOM, Leiden, The Netherlands, 1993, pp. 523–550.
[31] K. Sjöstrand, L. Clemmensen, R. Larsen, B. Ersbøll, G. Einarsson, SpaSM-a Matlab toolbox for sparse statistical modeling, J. Stat. Softw. 84 (2018) 1–37.