https://dx.doi.org/10.21577/0103-5053.20210097

J. Braz. Chem. Soc., Vol. 32, No. 11, 2061-2070, 2021

Article

Secondary Metabolites Extracted from Annonaceae and Chemotaxonomy Study of Terpenoids

Renata P. B. Menezes, [a] ***Zoe Sessions,*** [b] ***Eugene Muratov,*** [b] ***Luciana Scotti*** [a] ***and Marcus T. Scotti*** *,[a]

[a]*Programa de Pós-Graduação de Produtos Naturais e Sintéticos Bioativos, Universidade Federal da Paraíba, 58051-900 João Pessoa-PB, Brazil*

[b]*Laboratory for Molecular Modeling, UNC Eshelman School of Pharmacy, University of North Carolina at Chapel Hill, 27599 North Carolina, USA*

The Annonaceae family of plants is one of the most anatomically and structurally uniform families. Chemotaxonomy is a common practice to determine the chemical patterns within these families at different phylogenetic levels. The aim of this study was to build a dataset of all the secondary metabolites isolated within the Annonaceae family and to perform the respective chemotaxonomic analysis using self-organizing maps (SOMs). This dataset is composed of 5321 botanical occurrences and 1860 unique molecules present in all subfamilies of the Annonaceae. Diterpenes account for 366 unique compounds and 533 botanical occurrences seen in both Annonoideae and Malmeoideae subfamilies. The Annoneae, Xylopieae and Miliuseae tribes had the highest number of botanical occurrences and were therefore selected for the analysis. Molecular descriptors of the diterpenes and their respective botanical occurrences were used to generate the SOMs. These SOMs demonstrated clear and indicative tribe separations, with a match rate higher than 70%. Our results corroborate with the morphological and molecular data. These models can be used to predict the phylogenetic location of certain diterpenes and to accelerate the research of Annonaceae secondary metabolites and their biological potentials.

Keywords: Annonaceae, secondary metabolites, diterpenes, chemotaxonomy, self-organization maps

Introduction

The Annonaceae family was first described by Antoine Laurent de Jussieu in 1789 and is known for its striking anatomical and structural uniformity. The family is very consistent morphologically, with a unique primitive group of angiosperms providing easy identification.[1-4]

Two recent studies relevantly discuss the phylogenetic classification of the Annonaceae family. The first study carried out by Chatrou *et al*.[5] used eight plastid markers and representatives of 94 genera to formally and scientifically classify the Annonaceae into four subfamilies: Anaxagoreoideae, Ambavioideae, Annonoideae and Malmeoideae. The two largest subfamilies, Annonoideae and Malmeoideae, were divided into 14 tribes. The second study was conducted by Guo *et al*.,[6] and considered the phylogenetics of the Annonaceae based on a super matrix of eight chloroplast loci and 749 accessions representing 705 species (29% of ca. 2,400 species of 105 genres; 98% of 107 genres currently accepted). This matrix included almost four times more species as well as representatives of 15 additional genera compared to the first large study of phylogenetic importance by Chatrou *et al*.[5]

In addition to rebuilding the most comprehensive Annonaceae evolutionary tree, Guo *et al*.[6] also determined the phylogenetic position of five genera, *Bocageae*, *Boutiquea*, *Cardiopetalum*, *Duckeanthus* and *Phoenicanthus*, that were not included in any previous phylogenetic reconstruction. Their work assessed the monophyletic status and phylogenetic relationships within each major clade highlighting possible non-monophylides of genera and evaluating alternative resolutions for nomenclatural problems. Additionally, they identified and discussed unresolved problems such as the phylogenetic location and taxonomy of two genera, *Froesiodendron* and *Melodorum*, which have not yet been sampled. Finally, they provided

*e-mail: mtscotti@gmail.com

an updated view of the genera currently recognized in the family using their wealth of species.

Overall, Guo *et al.*[6] reorganized the phylogenetics and taxonomy of Annonaceae and concluded their study stating that the family contains four subfamilies, 15 tribes, 107 genera and 2400 species.

Annonaceae are very important economically given the multitude of ways the derivatives are used; the fruits are used in cooking and the production of ropes, the great diversity of chemical compounds shown to have pharmacological activities inspire new medicines, and the wood that is both light and durable.[7-9] These chemical compounds, also known as secondary metabolites, have great structural diversity in this family and represent many chemical classes including but not limited to alkaloids, terpenes, acetogenins, and steroids.[10-12]

One of the most common classes of Annonaceae is the terpenes. Terpenes are a very diverse class of substances and in addition to their important natural defense mechanisms in plants, terpenes display several therapeutic uses for humans.[9,13]

In the natural biosynthetic route, terpenes are formed from isoprene units, which are considered the basic units for the formation of both terpenes and steroids. Subclasses of terpenes include monoterpenes (two isoprene units, 10 carbons in their structure), sesquiterpenes (15 carbons), diterpenes (20 carbons) and triterpenes (30 carbons).[9,14]

The information gathered from chemical structures of both different species and genera has been and continues to be used in chemotaxonomy, that is, to determine the chemical phylogenetic patterns of a given family.[15-17] For chemotaxonomy studies, it is common practice to use machine learning with either supervised or unsupervised algorithms. A few examples of these machine learning techniques include neural networks (NN), support vector machine (SVM) and k-nearest neighbors (k-NN).[15-17]

Self-organizing maps (SOMs), which were developed by Kohonen,[18] are the main algorithm used in this study. A SOM is an unsupervised neural network that recognizes patterns and performs groupings based on exploratory analysis of the input data to generate non-linear relationships.[18-20] The SOM learning phase is competitive as there is no convergence or minimization criteria, and it works with a defined number of iterations and weight adjustments. In addition, each variable is mapped in a finite space of neurons organized in a typically two-dimensional arrangement (Kohonen map).[19-21]

In order to generate the SOM model, the model must first be trained on a portion of the established data previously separated for training. Then, the second set called the test set evaluates the training of the model. Using the results from the test set evaluation, we then isolate models capable of correctly mapping the test set, since the test data instances are not present in the training data.[20-24]

Vesanto *et al.*[25] created a unified distance matrix (U-matrix) that uses Euclidean distances to further analyze the SOM. In this matrix, it is possible to better visualize the possible groupings of the analyzed data.[25-27]

The goal of this study is to compile and integrate secondary metabolites isolated from Annonaceae into one curated dataset and to perform a chemotaxonomic analysis of diterpenes.

Results and Discussion

We collected and processed all Web of Science-indexed research papers published between 1970 and 2019 to create a database of secondary metabolites isolated from Annonaceae, except for the acetogenin class that is exclusive to this family. As seen in Figure 1, the interest in studying the Annonaceae plants has grown over time. One explanation for this growth is the abundant and diverse biological activity of the Annonaceae that comes from the structural diversity of the secondary metabolites. Alkaloids, for example, exhibited a wide variety of pharmacological activities and have been clinically studied for the treatment of cancer, Parkinson's disease, cardiovascular diseases, and various viral infections.[1-4,8,28,29]

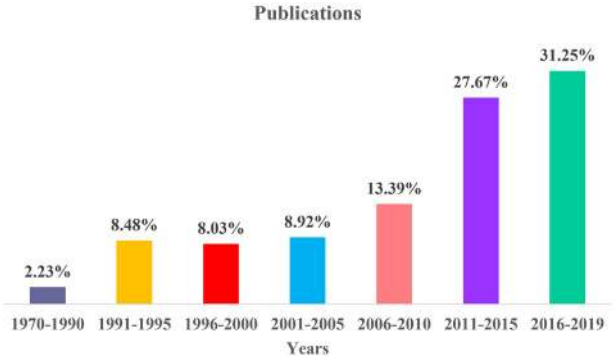

Figure 1. Distribution of published phytochemical studies of the Annonaceae plant family over time.

Our database consisted of 5321 botanical occurrences and 1860 unique molecules present in all subfamilies, 12 tribes, 64 genera and 380 species of the Annonaceae. Terpenes and alkaloids are the largest classes present in these plants (Figure 2).

It is important to note that although Annonaceae has 107 genera and 2400 species, only a small percentage of them have been studied chemically and therefore our database was considered comprehensive.

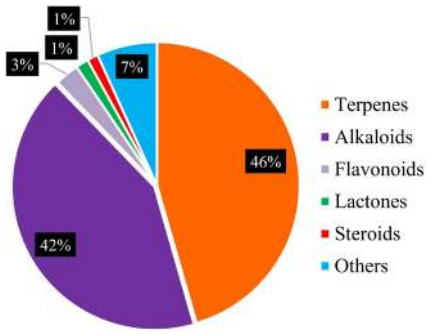

Figure 2. Secondary metabolite classes isolated from the Annonaceae family.

The alkaloids present in the Annonaceae are isoquinolines but the biosynthetic origins of the main nuclei occurring in the Annonaceae are the simple isoquinoline, proaporphine, aporphine, benzylisoquinoline, protoberberine, and phenanthrene.[30,31]

Terpenes, the second most common class in Annonaceae, occur in all subclasses (mono-, di-, sesqui-, and triterpenes), with the diterpenes being the most abundant. The most frequent diterpenes are kaurene, trachylobane, labdane, and atisane, wherein kaurane is the most common. Figure 3 shows the skeletons of some of the most present alkaloids and diterpenes in this family.

Once the database was compiled and the classes and skeletons of the secondary metabolites most present in Annonaceae were identified, the chemotaxonomic analysis was performed.

Chemotaxonomy is defined as a taxonomic classification method based on the chemical similarity of compounds identified in the organisms/plants being classified.[32] Thus, we sought to investigate chemical molecules that serve as taxonomic markers of the Annonaceae.

Given the assortment of the secondary metabolites collected, the terpenes were selected for the chemotaxonomic studies because they were the predominant class (46% of metabolites). As mentioned earlier, terpenes can be classified into mono-, di-, sesqui-, and triterpenes.

Among these four subclasses, about 50% of the terpenes were diterpenes. Annonaceae diterpenes have promising anti-inflammatory activity, making compounds of this class excellent candidates for clinical trials in anti-inflammatory therapy.[33]

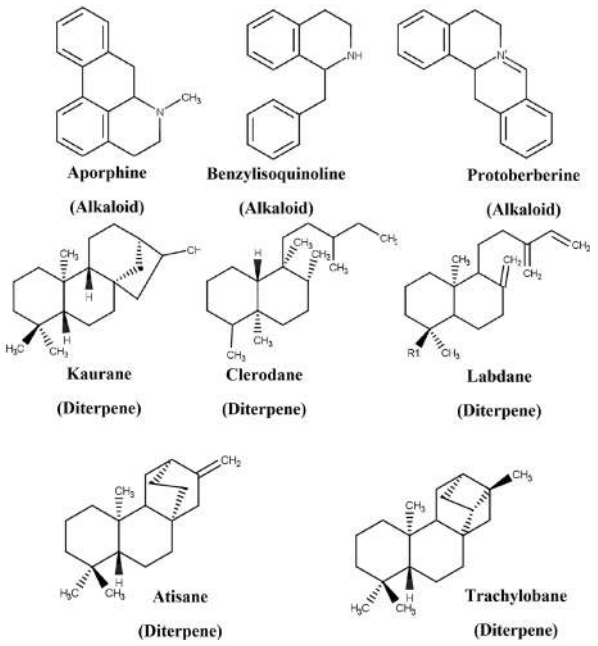

Figure 3. Skeletons of the most abundant alkaloids and diterpenes in the Annonaceae family.

Diterpenes represented a total of 366 unique chemical structures and 533 botanical occurrences; a botanical occurrence indicates that the compounds are present in several species.

These 533 botanical occurrences are distributed in two subfamilies, Annonoideae and Malmeoideae, which are the largest subfamilies of the Annonaceae and are distributed in 8 tribes, 13 genera and 50 species. The phylogenetic classification of the Annonaceae family proposed by Guo *et al.*[6] was utilized.

The three tribes with the highest number of botanical occurrences and molecules were then selected for the self-organizing neural maps, as the high number of diterpenes allows for the recognition of chemical pattern among the tribes. These tribes were Annoneae, Xylopieae and Miliuseae, and Table 1 contains the botanical characteristics and quantities of the selected molecules.

The genera represented in each selected tribe are: *Annona* (Annoneae), *Xylopia* (Xylopieae), *Polyalthia*, *Pseudouvaria*, *Piptostigma* and *Greenwayodendron* (Malmeoideae). Malmeoideae is the most studied genera of these tribes.

Table 1. Botanical characteristics and occurrences of the diterpenes of the tribes Annoneae, Xylopieae and Miliuseae

Tribe	Subfamily	Genus	Species	Diterpenes	Occurrences
Annoneae	Annonoideae	1	11	150	179
Xylopieae	Annonoideae	1	14	179	241
Miliuseae	Malmeoideae	4	13	95	101
					Total: 521

For the 521 molecules of the three selected tribes, molecular descriptors were calculated using the DRAGON 7.0 software,[34] which has 5270 descriptors organized in 30 logic blocks. From these three blocks of descriptors, 60 molecular descriptors were selected to consider ring descriptors, functional groups, and fragments of central atoms.

The botanical occurrences were classified in the three selected tribes and the values of the 60 molecular descriptors were used as input data in the SOM Toolbox software.[25] The self-organized matrix of diterpenes was then generated, classified into the three aforementioned tribes according to the chemical similarity between them. Then, the classification generated was compared with the phylogenetic classification proposed by Guo *et al.*[6] The phylogenetic classification of Guo *et al.*[6] can be seen in Figure 4.

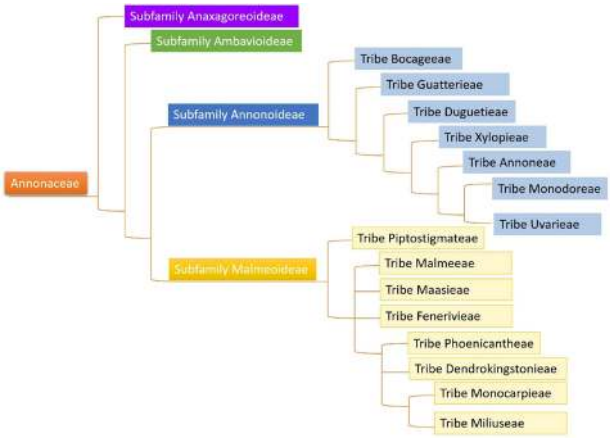

Figure 4. Phylogenetic diagram of the Annonaceae family (adapted from Guo *et al.*[6]).

In the generated maps, the hit rate using the two types of DRAGON 7.0 descriptors was > 77%. Thus, the 5-fold validation was performed for the generated SOM model, in which the diterpenes were divided into five training groups and five test groups, always maintaining the proportion of molecules from the three tribes (Annoneae, Xylopieae and Miliuseae). The results of the validation are described in Table 2.

Table 2, like Table 3, also describes the accuracy values for each training and test. Accuracy provides us with information about the overall performance of the model, indicating the overall hit rate. The values of this metric vary between 0 and 1, and the closer to 1 it indicates that the model is getting more correct in its classification of molecules in terms of their tribes, that is, correctly classifying a molecule of the Annoneae tribe in the Annoneae tribe. Models with an accuracy greater than 0.70 are already considered models of excellent performance.[24]

After analyzing Table 2, it is observed that the hit rate was overall > 70%, with the best hit rate of 95% for the Miliuseae tribe. The average hit rate of the test sets was 80% and is very close to the average hit rate for the training, which was 83%, revealing not only the good predictive power of the model, but that the model is robust. The applicability domain was also analyzed and was > 99% of the predictions of the test sets.

To verify the tribes dependence on chemical similarity and the ability to separate them accordingly, chemotaxonomy analysis was performed using other machine learning algorithms such as the support vector machine (SVM) and the k-nearest neighbors' algorithm k-NN, in addition to neural maps generated using the fingerprint descriptors calculated by the DRAGON 7.0 software. The results are shown in Table 3 for this SOM analysis of the Annoneae, Xylopieae and Miliuseae tribes and like those in Table 2, the hit rates are excellent.

To visualize the generated SOM, we utilize a U-matrix and display it alongside a principal component analysis (PCA) which was developed from the correlation matrix of the database used in the generation of SOM. PCA is measured using eigenvectors with higher eigenvalues. In

Table 2. Accuracy statistics of the training and tests groups of the 5-fold cross-validation of the self-organizing map from the Annoneae, Xylopieae and Miliuseae tribes

Tribe	Training 1	Training 2	Training 3	Training 4	Training 5	Average
Annoneae	0.89	0.90	0.80	0.80	0.78	0.83
Miliuseae	0.90	0.86	0.86	0.90	0.90	0.88
Xylopieae	0.77	0.76	0.85	0.86	0.85	0.82
Accuracy	0.83	0.83	0.83	0.84	0.84	0.83
Tribe	Test 1	Test 2	Test 3	Test 4	Test 5	Average
Annoneae	0.83	0.70	0.77	0.86	0.70	0.77
Miliuseae	0.90	0.94	0.85	0.95	0.80	0.88
Xylopieae	0.71	0.89	0.83	0.67	0.85	0.79
Accuracy	0.79	0.83	0.81	0.79	0.78	0.80

Table 3. Summary of test averages corresponding to 5-fold cross-validation using the different machine learning algorithms and self-organizing map (SOM) with the fingerprint descriptors for the Annoneae, Xylopieae and Miliuseae tribes

Tribe	SOM molecular descriptors average	SOM fingerprint descriptors average	SVM average	k-NN average
Annoneae	0.766	0.77	0.70	0.70
Miliuseae	0.88	0.92	0.81	0.85
Xylopieae	0.79	0.89	0.87	0.81
Accuracy	0.80	0.85	0.80	0.78

SVM: support vector machine; k-NN: k-nearest neighbors' algorithm.

the projection of the PCA, the neighboring map units are connected by lines to make the visualization of the data on the map more clear and defined. The PCA performed has an explained variance of 37.04%, that is, using only two variables it is possible to visualize one third of the entire variance.

Figure 5 shows the U-matrix of the generated SOM where we can see a chemical pattern separating the three tribes Annoneae (blue), Xylopieae (red) and Miliuseae (green), which are best observed in the principal component analysis chart (PCA).

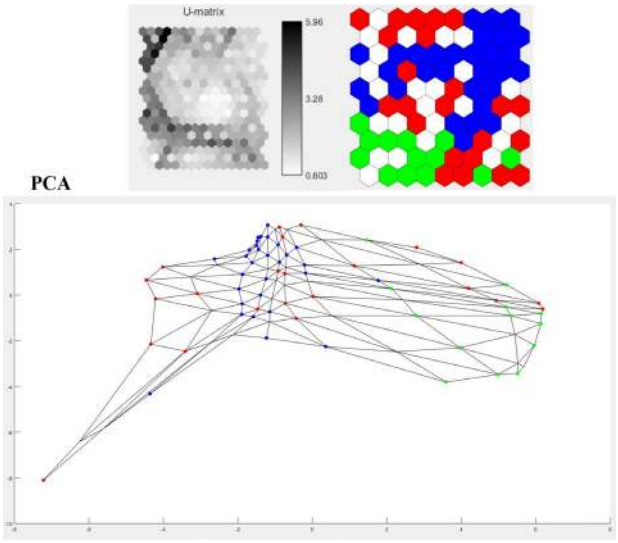

Figure 5. Visualization of the SOM of Annonaceae diterpenes data. In the upper corner we have the U-matrix. The left U-matrix does not identify the tribes while the right U-matrix identifies the tribes by color; Annoneae is blue, Xylopieae is red, and Miliuseae is green. The values shown on the scale between the two U-matrices represent the values of the molecular descriptors of the diterpenes, varying between 0.603 and 5.96. These values were used to group the diterpenes by tribes. At the bottom, we have the PCA projection of the SOM measured by its two eigenvectors with higher eigenvalues. The tribes were plotted using the same identification colors as the U-matrix.

We can see that the Miliuseae tribe, despite having the fewest number of diterpenes and, consequently, the fewest botanical occurrences, was the tribe with the best hit rates (greater than 85% in all algorithms and different descriptors in SOM) and is more structurally distant from the Annoneae and Xylopieae tribes, corroborating Guo's[6] phylogenetic classification, seen in Figure 4.

Annoneae and Xyopieae are part of the same subfamily, Annonoideae, explaining the proximity of the two tribes in the SOM, while Miliuseae is part of the Malmeoideae subfamily, and is therefore further away. When observing the diterpenes present in the tribes present in the SOM (Figure 6), we can see that each tribe has a higher frequency of a certain subtype of diterpene. The subtypes present in the Annoneae and Xylopieae tribes, although different, maintain a certain chemical similarity in their skeletons, explaining once again the approximation of these two tribes in the SOM.

Figure 6 shows some of the isolated diterpenes in each of the analyzed tribes, focusing on the most frequent skeletons identified from each tribe. The Miliuseae tribe has a clerodane subclass of diterpenes. The clerodane diterpene is able to undergo structural changes and generate some subtypes,[35] and the kolava subtype is present in the Miliuseae tribe. The Annoneae and Xylopieae tribes have kaurane and trachylobane diterpenes, respectively. Although different, these subclasses have similarities in their chemical skeletons, even further supporting the closeness of the two tribes in the SOM.

The most significant descriptors in the separation of each cluster (each tribe in SOM) are represented in Figure 7. For the Annoneae tribe, the descriptors that presented a high value were (*i*) NROH, which describes hydroxyl groups (OH) linked to aliphatic groups, (*ii*) nOHp descriptor that points to primary alcohols, (*iii*) C-006, which indicates CH_2 carbons attached to a radical and that radical attached to an OH, and (*iv*) the descriptor O-056 that describes the alcohol function. Thus, these descriptors report that the diterpenes of this tribe are distinguished by the large number of hydroxyls in their chemical structure (Figure 6).

For the Xylopieae tribe, the most representative descriptors were nCIR, which indicates the number of circuits (rings/cycles connected to each other) present in the molecule, the RFD descriptor of ring melting density,

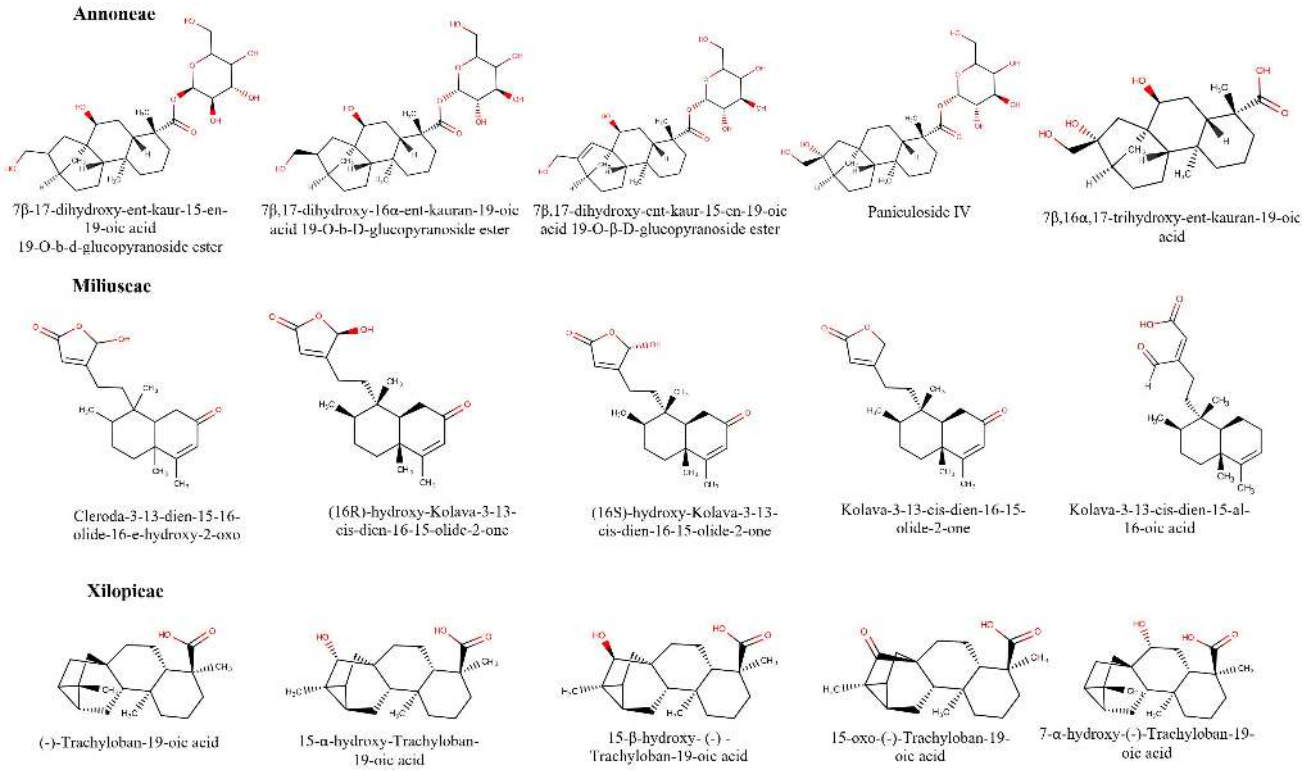

Figure 6. Diterpenes of the most present subtypes in the Annoneae, Xylopieae, and Miliuseae tribes.

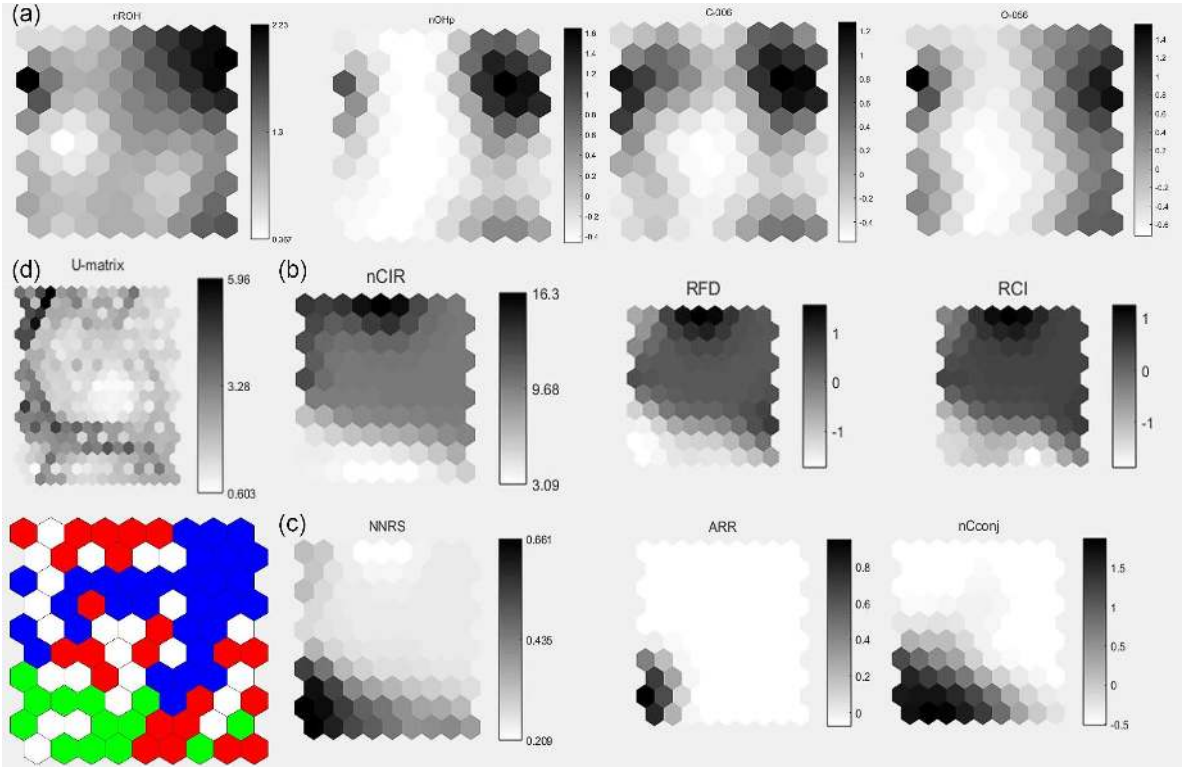

Figure 7. The most significant descriptors for the Annoneae, Xylopieae and Miliuseae tribes. In (a) we have the U-matrix for the four most significant descriptors in the grouping of the diterpenes of the Annoneae tribe. In (b), the U-matrix is shown for the three most significant descriptors of the Xylopieae tribe. (c) Shows the U-matrix for the three most significant descriptors of the Miliuseae tribe. Finally, in (d) we have the U-matrix of the self-organizing map generated in the study, with the upper U-matrix not identifying the tribes and the lower U-matrix identifying the tribes by color; Annoneae is blue, Xylopieae is red, and Miliuseae is green.

and the RCI descriptor that provides information about the ring complexity of the molecule. These descriptors point to the presence of molecules with a large number of interconnected rings/cycles; as seen in Figure 6, the diterpenes of this tribe have many interconnected rings/cycles and a certain degree of complexity.

For the Miliuseae tribe, the descriptors with the highest values were nConj, a descriptor that expresses the presence of non-aromatic C conjugates (sp^2), NNRS, the normalized number of ring system, which accounts for both the ratio between the number of ring systems (NRS) and the cyclomatic number (nCIC, discriminates cyclic compounds from acyclics) to provide information related to the presence of aromatic rings in the chemical structure, and lastly the ARR descriptor. The ARR, aromatic ratio, is the ratio of the number of aromatic bonds to the total number of bonds in the molecule. These descriptors reveal that the diterpenes of this tribe have an aromatic ring and conjugated non-aromatic bonds, which can also be seen in Figure 6.

An article by Scotti *et al*.,[15] constructed a SOM with nuclear magnetic resonance (NMR) data of 118 diterpenes from three genera of the Annonaceae, the genera *Xylopia*, *Polyalthia* and *Annona*. The SOM was able to separate the diterpenes of the three genera with the NMR data and specific chemical displacement values of ^{13}C were observed for the skeletal carbons of each type of diterpenes of each genus. Kauranes skeletons were found for *Annona*, while trachylobans were found for *Xylopia* and clerodanes were found for *Polyalthia*.

Review papers concerning the *Annona* genus and some of its species have suggested that *ent*-kauranes are the most abundant diterpenes.[36-38] A review by Barbosa and Vega,[9] highlights that diterpenes are the second most common class of secondary metabolites in species of the *Xylopia* genus, with kaurane, labdane, atisane and trachylobane diterpenes being the most frequent. Of these, trachylobanes are considered as chemotaxonomic markers of *Xylopia* as they are the most abundant in *Xylopia* and are difficult to find elsewhere in Annonaceae.[9,39]

The four genera selected from the Miliuseae tribe are those with the most phytochemical studies, with the *Polyalthia* and *Pseuduvaria* genera being the most chemically and biologically studied of the tribe. As in the other genera, there are studies in the literature that show that the most isolated diterpenes of *Polyalthia* and *Pseuduvaria* species are clerodanes.[40-43]

Conclusions

The literature corroborates the information obtained in this study. In this way, this study of Annonaceae diterpenes establishes a way to separate the Annoneae, Xylopieae and Miliuseae tribes in accordance with the family's morphological and taxonomic separation. This phenomenon makes it possible to predict the location of a certain diterpene in the Annoneae, Xylopieae and Miliuseae tribes of the Annonaceae and to search for these secondary metabolites and their biological potentials more effectively.

Methodology

Construction of the Annonaceae database

The articles used for the construction of the database were selected by means of an electronic search in the Web of Science research base, and were composed of studies and literature reviews on secondary metabolites isolated in plants of the Annonaceae. The following terms were used in the search for scientific articles: "Annonaceae", "secondary metabolites", "terpenes", "alkaloids", "flavonoids". All secondary metabolites, the species from which they were isolated, and the geographic locations will be registered on the SISTEMATX[44] web tool and developed by the Chemistry Laboratory of the Postgraduate Course on Natural and Bioactive Synthetic Products.[45]

Obtaining structures in three dimensions of compounds

For all structures, SMILES codes were used as input data for Marvin v. 19.27.0.[46] It was also used the Standardizer software[47] which made it possible to convert the various chemical structures into personalized canonical representations. This standardization is extremely important to create libraries of consistent compounds, in addition to canonizing the structures, adding hydrogens, aromatizing molecules, generating the 3D structures, and saving the compounds in SDF format.

Obtaining the molecular descriptors

Molecular descriptors are used to calculate the physicochemical properties of the molecules of each set of molecules. To obtain the molecular descriptors, the DRAGON 7.0 program[34] was used.

The DRAGON 7.0 software[34] can calculate 5270 molecular descriptors, covering several approaches. These molecular descriptors are arranged in 30 logic blocks.[34] Of the 30 blocks of molecular descriptors available in the Dragon 7.0 software,[34] only the ring descriptors, functional groups, and fragments of central atoms blocks were selected.

Pre-processing of data

In this step, the variables/descriptors were selected. This selection tactic is used to identify those descriptors that are most important for the grouping of the diterpenes and in this case were mostly related to the tribes. The selection of descriptors is an important step that must be carried out before the generation of the model, since it is useful for reducing the dimensionality of the data, helping to obtain a generic and not over-adjusted model, reducing computational cost, simplifying extraction processes and transformation of data, and further simplifying the presentation and demonstration of data.[48] In short, this step helps to reduce overfitting, increases the accuracy of the model, and reduces training time.

The pre-treatment criteria removed descriptors that had equal values in the series, ones that only a different value, and ones that had a correlation greater than 0.99. The majority of descriptors end up being removed, as many were inter-correlated, such that the independent variable remained the most correlated with the dependent variable.

Self-organizing maps (SOMs)

For the realization of the neural maps, the selection of molecular descriptors was performed for the bank of isolated molecules of the Annonaceae. The functional group, central atom, and ring descriptors were selected. Then, the constant variables for each block of descriptors and those with a different value in the series were excluded.

The molecular descriptors selected were analyzed with SOMs in Matlab 6.5 and SOM Toolbox 2.0.[25,26,49] The SOM Toolbox tool is a set of Matlab functions that can be used for the elaboration and implementation of neural networks, since it contains functions for the creation, visualization, and analysis of self-organizing maps. The data set was presented to the network before any adjustments were made. Subsequently, the data group was partitioned according to the regions of the weight vectors of the map, in each training stage. Then, the correct prediction of these sets and the total correct predictions of the compounds were evaluated. In the most relevant models, the set was divided into training and test sets to assess the forecasting capacity. Training and test performance were assessed by calculating the proportion of the number of samples correctly classified by SOM. For each map, 5 cross-validations were performed, being partitioned into 80% training and 20% testing. In the SOM, sites containing molecules for each descriptor were identified to highlight existing chemical patterns.

SVM and k-NN models

Knime 3.6.2 software[50] was used to perform all the following analyzes. The class descriptors and variables were imported from the Dragon 7.0 software[34] and, for each, the data was divided into the "partitioning" node with the "stratified sample" option to create a training set and a set of tests, covering 80 and 20% of the compounds, respectively. Although the compounds were selected at random, the same proportion of active and inactive samples was maintained in both sets. Two models were generated using the support vector machine (SVM) algorithm[51] and the K-nearest neighbors' algorithm (k-NN).[52] An external cross-validation was modeled 5 times.

SVM is a supervised machine learning algorithm that analyzes data and recognizes patterns.[51,53] The parameters selected for the SMV for all the models generated were polynomials, with power 1.0, bias 1.0, and range 1.0.

k-NN consists of instance-based machine learning as the function and is approximated only locally (neighbors) so the entire calculation is postponed until classification.[53,54] It is a technique that gives weight to the contributions of neighbors, so that the closest neighbors contribute more to the average than the more distant ones.[52-54] The parameters selected for the SVM for all the generated models were k = 3.

Acknowledgments

We thank the CNPq for financial support, grant numbers 309648/2019-0 and 431254/2018-4.

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

References

1. Cronquist, A.; *An Integrated System of Classification of Flowering Plants*; Columbia University Press: New York, 1981.
2. Cunha, L. M. A.: *Estudo Fitoquímico e Biológico de Duguetia riparia (Annonaceae)*; MSc Dissertation, Universidade Federal do Amazonas, Manaus, 2009, available at https://tede.ufam.edu.br/handle/tede/3307, accessed in June 2021.
3. Silva, C. A.; Domingues Neta, A. M.; *Biotemas* **2011**, *23*, 69.
4. Doyle, J. A.; Sauquet, H.; Scharaschkin, T.; le Thomas, A.; *Int. J. Plant Sci.* **2004**, *165*, S55.
5. Chatrou, L. W.; Pirie, M. D.; Erkens, R. H. J.; Couvreur, T. L. P.; Neubig, K. M.; Abbott, J. R.; Mols, J. B.; Maas, J. W.; Saunders, R. M. K.; Chase, M. W.; *Bot. J. Linn. Soc.* **2012**, *169*, 5.

6. Guo, X.; Tang, C. C.; Thomas, D. C.; Couvreur, T. L. P.; Saunders, R. M. K.; *Sci. Rep.* **2017**, *7*, 7323.
7. Tekuri, S.; Pasupuleti, S.; Konidala, K.; Pabbaraju, N.; *J. Complementary Med. Res.* **2019**, *10*, 38.
8. Aminimoghadamfarouj, N.; Nematollahi, A.; Wiart, C.; *J. Asian Nat. Prod. Res.* **2011**, *13*, 465.
9. Barbosa, L. T. C.; Vega, M. R. G.; *Rev. Virtual Quim.* **2017**, *9*, 1712.
10. Salehi, B.; Sharopov, F.; Martorell, M.; Rajkovic, J.; Ademiluyi, A. O.; Sharifi-Rad, M.; Sharifi-Rad, J.; *Int. J. Mol. Sci.* **2018**, *19*, 2361.
11. Chakraborty, P.; *Biochimie Open* **2018**, *6*, 9.
12. Mohammadi, S.; Jafari, B.; Asgharian, P.; Martorell, M.; Sharifi-Rad, J.; *Phytother. Res.* **2020**, *34*, 1556.
13. Viegas Jr., C.; *Quim. Nova* **2003**, *26*, 390.
14. Dewick, P. M.; *Medicinal Natural Products: A Biosynthetic Approach*, 2nd ed.; John Wiley & Sons Ltd: New Jersey, 2002.
15. Scotti, L.; Tavares, J. F.; da Silva, M. S.; Falcão, E. V.; e Silva, L. M.; Soares, G. C. S.; Scotti, M. T.; *Quim. Nova* **2012**, *35*, 2146.
16. Cavalcanti, A. B. S.; Barros, R. P. C.; Costa, V. C. O.; da Silva, M. S.; Tavares, J. F.; Scotti, L.; Scotti, M. T.; *Molecules* **2019**, *24*, 3908.
17. Emerenciano, V. P.; Barbosa, K. O.; Scotti, M. T.; Ferreira, M. J. P.; *J. Braz. Chem. Soc.* **2007**, *18*, 891.
18. Kohonen, T.; *Biol. Cybern.* **1982**, *43*, 59.
19. Kohonen, T.; Saarela, A.; *IEEE Trans. Neural Networks* **2000**, *11*, 574.
20. Kitani, E. C.: *Mapeamento e Visualização de Dados em Alta Dimensão com Mapas Auto-Organizados*; PhD Thesis, Escola Politécnica da Universidade de São Paulo, São Paulo, 2013, available at https://www.teses.usp.br/teses/disponiveis/3/3142/tde-11072014-114804/en.php, accessed in June 2021.
21. Affonso, G. S.: *Mapas Auto-Organizáveis de Kohonen (SOM) Aplicados na Avaliação dos Parâmetros da Qualidade da Água*; MSc Dissertation, Autarquia Associada à Universidade de São Paulo, São Paulo, 2011, available at http://pelicano.ipen.br/PosG30/TextoCompleto/Gustavo%20Sousa%20Affonso_M.pdf, accessed in June 2021.
22. Salzberg, S. L.; *Mach. Learn.* **1994**, *16*, 235.
23. Livingstone, D.; *Data Analysis for Chemists*; Oxford Science Publications: UK, 1995.
24. Barros, R. P. C.: *Triagem Virtual de Metabólitos Secundários com Potencial Atividade Antimicrobiana do Gênero Solanum e Estudo Fitoquímico de Solanum capsicoides All.*; MSc Dissertation, Programa de Pós-Graduação em Produtos Naturais e Sintéticos Bioativos, Universidade Federal da Paraíba, João Pessoa, 2017, available at https://repositorio.ufpb.br/jspui/handle/tede/9069, accessed in June 2021.
25. http://lib.tkk.fi/Diss/2002/isbn951226093X/article2.pdf, accessed in June 2021.
26. Vesanto, J.; Himberg, J.; Alhoniemi, E.; Parhankangas, J.; *SOM Toolbox for Matlab 5*; Helsinki University of Technology, Finland, 2000, available at https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.25.7561&rep=rep1&type=pdf, accessed in June 2021.
27. Arcoverde, G. F. B.; Almeida, C. M.; Ximenes, A. C.; Maeda, E. E.; Araújo, L. S.; *Bol. Ciênc. Geod.* **2011**, *1*, 379.
28. Suedee, A.; Mondranondra, I. O.; Kijjoa, A.; Pinto, M.; Nazareth, N.; Nascimento, M. S. J.; Silva, A. M. S.; Herz, W.; *Pharm. Biol.* **2007**, *45*, 575.
29. Cuendet, M.; Oteham, C. P.; Moon, R. C.; Keller, W. J.; Peaden, P. A.; Pezzuto, J. M.; *Pharm. Biol.* **2008**, *46*, 3.
30. Lúcio, A. S. S. C.; Almeida, J. R. G. S.; da-Cunha, E. V. L.; Tavares, J. F.; Barbosa Filho, J. M.; *Alkaloids: Chem. Biol.* **2015**, *74*, 233.
31. Rabêlo, S. V.: *Revisão de Alcalóides do Gênero Annona, Estudo Fitoquímico e Avaliação da Atividade Biológica de Atemoia (Annona cherimola x Annona squamosa)*; MSc Dissertation, Programa de Pós-Graduação em Recursos Naturais do Semiárido, Universidade Federal do Vale do São Francisco, Petrolina, 2014, p. 234, available at http://www.cpgrnsa.univasf.edu.br/uploads/7/8/9/0/7890742/rab%C3%8Alo_s._v._disserta%C3%87%C3%83o.pdf, accessed in June 2021.
32. Chen, Y.; Zou, C.; Mastalerz, M.; Hu, S.; Gasaway, C.; Tao, X.; *Int. J. Mol. Sci.* **2015**, *16*, 30223.
33. Attiq, A.; Jalil, J.; Husain, K.; *Front. Pharmacol.* **2017**, *8*, 752.
34. *Dragon*, *Software for Molecular Descriptor Calculation*, version 7; Talete srl, 2013, available at http://www.talete.mi.it/, accessed in June 2021.
35. Li, R.; Morris-Natschke, S. L.; Lee, K.-H.; *Nat. Prod. Rep.* **2016**, *33*, 1166.
36. Leite, D. O. D.; Nonato, C. F. A.; Camilo, C. J.; Carvalho, N. K. G.; Nobrega, M. G. L. A.; Pereira, R. C.; Costa, J. G. M.; *Curr. Pharm. Des.* **2020**, *26*, 4056.
37. Ma, C.; Chen, Y.; Chen, J.; Li, X.; Chen, Y.; *Am. J. Chin. Med.* **2017**, *45*, 933.
38. Coria-Téllez, A. V.; Montalvo-Gónzalez, E.; Yahia, E. M.; Obledo-Vázquez, E. N.; *Arabian J. Chem.* **2018**, *11*, 662.
39. Moreira, I. C.; Roque, N. F.; Vilegas, W.; Zalewski, C. A.; Lago, J. H. G.; Funasaki, M.; *Chem. Biodiversity* **2013**, *10*, 1921.
40. Hasan, C. M.; Hossain, M. A.; Rashid, M. A.; *Biochem. Syst. Ecol.* **1995**, *23*, 331.
41. Hasan, C. M.; Islam, M. O.; Rashid, M. A.; *Pharmazie* **1995**, *50*, 227.
42. Gbedema, S. Y.; Bayor, M. T.; Annan, K.; Wright, C. W.; *J. Ethnopharmacol.* **2015**, *169*, 176.
43. Yao, L. J.; Jalil, J.; Attiq, A.; Hui, C. C.; Zakaria, N. A.; *J. Ethnopharmacol.* **2019**, *229*, 303.
44. Scotti, M. T.; Herrera-Acevedo, C.; Oliveira, T. B.; Costa, R. P. O.; Santos, S. Y. K. O.; Rodrigues, R. P.; Scotti, L.; Da-Costa, F. B.; *Molecules* **2018**, *23*, 103.

45. https://sistematx.ufpb.br/, accessed in June 2021.
46. *Marvin*, v. 19.27.0; ChemAxon, Budapest, Hungary, 2019.
47. *Standardizer*, v. 19.27.0; ChemAxon, Budapest, Hungary, 2019.
48. Amaral, F.; *Aprenda Mineração de Dados*, 1st ed.; Alta Books: Rio de Janeiro, RJ, 2016.
49. Todeschini, R.; Consonni, V.; *Molecular Descriptors for Chemoinformatics*; Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, Germany, 2009.
50. Berthold, M. R.; Cebron, N.; Dill, F.; Di Fatta, G.; Gabriel, T. R.; Georg, F.; Meinl, T.; Ohl, P.; Sieb, C.; Wiswedel, B.; *ACM SIGKDD Explor. Newsl.* **2006**, *11*, 58.
51. Mei, H.; Zhou, Y.; Liang, G.; Li, Z.; *Chin. Sci. Bull.* **2005**, *50*, 2291.
52. Cheng, D.; Zhang, S.; Deng, Z.; Zhu, Y.; Zong, M. In *Advanced Data Mining and Applications*; Luo, X.; Yu, J. X.; Li, Z., eds.; Springer: Cham, 2014, p. 499.
53. Baskin, I. I. In *Methods in Molecular Biology*; Walker, J. M.; Rapley, R., eds.; Humana Press Inc.: Totowa, 2018, p. 119.
54. Altman, N. S.; Altman, N. S.; *Am. Stat.* **1991**, *46*, 175.

Submitted: April 1, 2021
Published online: July 7, 2021