SciForum
MOL2NET

Application of Self-Organizing Maps generated from Molecular Descriptors of diterpenoids in Chemotaxonomy Studies of Lamiaceae Family

Andreza Cavalcanti [1,*], Marcelo Silva [1] , Vicente Costa [1], Renata Barros [1], Luciana Scotti [1], Josean Tavares [1] and Marcus Scotti [1]

[1] Program of Natural and Synthetic Bioactive Products (PgPNSB), Health Sciences Center, Federal University of Paraíba, João Pessoa-PB, Brazil; E-mail: andreza.jp.pb@gmail.com

***** Author to whom correspondence should be addressed; E-Mail: andreza.jp.pb@gmail.com; Tel.: 55-83-98713-6155.

Received: / Accepted: / Published:

Abstract: Lamiaceae is the largest family-level clade of the order Lamiales and comprises approximately 295 genera and 7775 species, presenting cosmopolitan distribution. It is estimated that in Brazil there are 36 genus and 490 species. Lamiaceae is classified into 10 subfamilies that present a large variety of secondary metabolites, among them diterpenes are commonly reported for this family. These diterpenes can be used in the chemotaxonomy of this family, because they have stable and quite diversified structures, being found in several species of the Lamiaceae family. Thus, the objective of this study is to classify two subfamilies of Lamiaceae based on the identification of diterpenes and their respective botanical occurrences available in our internal database (www.sistematx.ufpb.br), using descriptors calculated by DRAGON 7.0 software. The 3551 botanical occurrences and their 119 descriptors obtained from molecular fragments were used as input data in SOM Toolbox 2.0 (Matlab) to generate a self-organizing map (SOM), allowing to classify two subfamilies: Lamioideae (L) and Scutellarioideae (S). Therefore, the results obtained by the chemotaxonomic study corroborate with the phylogenetic classification based on the DNA that was proposed by Li et al., 2016.

Keywords: Lamiaceae; diterpenes; chemotaxonomy

1. Introduction

Currently Lamiaceae is the largest family clade of the order Lamiales and comprises approximately 295 genus and 7775 species, presenting cosmopolitan distribution. It is estimated that in Brazil there are 36 genus and 490 species. Lamiaceae is classified in ten subfamilies (Ajugoideae, Lamioideae, Nepetoideae, Prostantheroideae, Scutellarioideae,

Symphorematoideae, Viticoideae Cymarioideae, Peronematoideae and Premnoideae) and two genus (Callicarpa and Tectona) that are not assigned to a subfamily (Figure 1) [1]. This family presents a large variety of secondary metabolites, among which diterpenes are commonly reported for this family. These diterpenes can be used in the chemotaxonomy of this family, since they have stable and quite diversified structures, being found in several species of the Lamiaceae family [2,3].

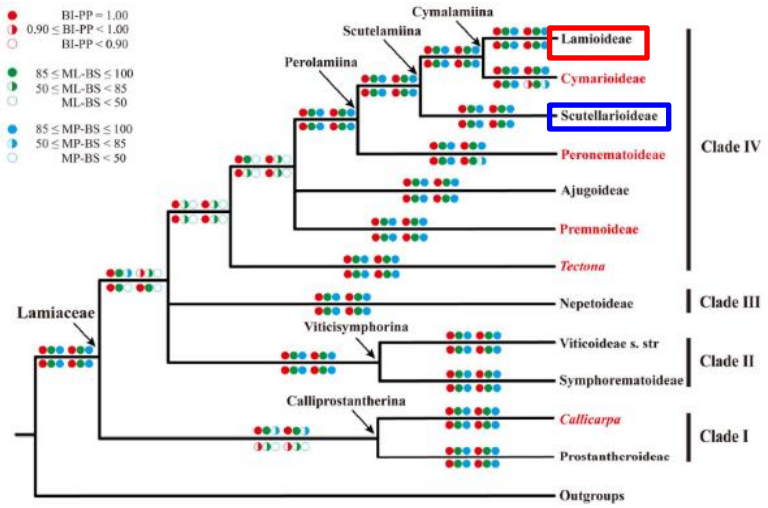

Figure 1. Phylogenetics of Lamiaceae from analyzes of the DNA dataset [1].

In the search for these secondary metabolites, we can use dereplication tools, which is the access to characteristics of molecules already reported in the literature and which are available in virtual databases. These databases can then provide information on compounds such as biological, biogeographical, and taxonomic data, or the presence of a certain compound (new or known) in other individuals of the same species, genus, subfamily, and family. However, we currently have a web interface that is SISTEMAT X web (https://sistematx.ufpb.br/), which provides a database of secondary metabolites, presenting a wealth of information to the scientific community about the products (SMILES code), relative mass, exact mass, name of the compound as well as specific information for taxonomic classification (from family to species) and the location of species from which the compounds were isolated [4].

There is a great diversity of molecular descriptors, which can be distinguished by the physical-chemical findings or the specific mathematical tools that are used for their calculation, such as the DRAGON 7.0 program [5]. We use methodologies that detect chemical clusters and patterns, such as artificial neural networks (RNAs) that are not restricted to linear correlations and are able to consider nonlinear data correlations, they can be used efficiently for modeling, prediction and classification The RNA architecture often used for pattern recognition and classification is the self-organizing map (SOM) that can map multivariate data into a two-dimensional grid, grouping similar patterns close to each other [6,7].

The objective of this study is to classify two subfamilies of Lamiaceae based on the identification of diterpenes and their respective botanical occurrences available in our internal database (www.sistematx.ufpb.br) using descriptors calculated by DRAGON 7.0 software [5]. With the Matlab software [8], the chemical patterns were recognized and analyzed from unsupervised neural networks along with the Self Organizing Map (SOM) to create the maps.

2. Results and Discussion

From the data collected from the botanical occurrences of the diterpenes obtained from the Lamiaceae family, 119 molecular descriptors were generated for each diterpene molecule, using

the DRAGON 7.0 software, and the self-organized matrix for each molecule could be calculated by dividing the data into groups according to similarity. It was possible to observe in Table 1, the success rates of diterpene occurrences of the Lamioideae (L) and Scutellarioideae (S) subfamilies belonging to the Lamiaceae family according to Li et al., 2016 [1]. Results of the analysis: 832 occurrences and 800 hits, showing a total hit of 96%.

Table 1. SOM hit rate of subfamilies Lamioideae (L) and Scutellarioideae (S).

Subfamilies	Nº of hits	Nº of occurrences	% of hits
L	534	551	97
S	266	281	95
Total	800	832	96

Figure 2 shows the Self-Organizing Map and some molecular descriptors generated from the diterpenes of the Lamioideae (L) and Scutellarioideae (S) subfamilies, which are used in the study of Lamiaceae chemotaxonomy.

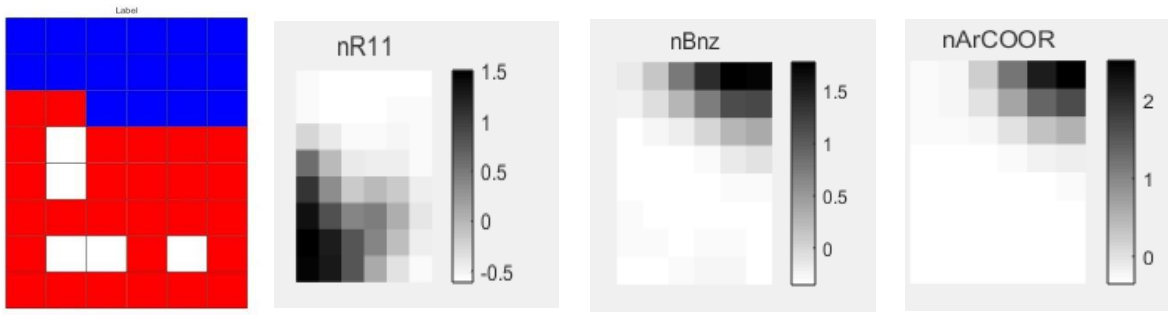

Figure 2. Self-organizing map obtained with the diterpenes of the subfamilies Lamioideae (red) and Scutellarioideae (blue) and generated descriptors: nR11, nBnz and nArCOOR.

3. Materials and Methods

A database of diterpene molecules isolated from the Lamiaceae family was constructed, and all the structural data and respective botanical occurrences were added to Sistemat X Web (http://sistematx.ufpb.br). There were 3551 botanical occurrences extracted from 402 species, 58 genera and seven subfamilies (described in the clade) of the family Lamiaceae. For all structures, SMILES codes were used as input data for Marvin, ChemAxon (http://www.chemaxon.com/). Then Standardizer software (http://www.chemaxon.com/) was used to convert the various chemical structures into custom canonical representations, add hydrogens, aromatize, generate 2D and save the compounds in SDF format.

Afterwards, the two-dimensional (2D) structures of the compounds were used as input data in the DRAGON 7.0 program [5], which resulted in molecular descriptors to predict biological and physicochemical properties of the database molecules. Allowing a chemotaxonomic analysis between two of the seven subfamilies of the Lamiaceae family, using molecular descriptors and unsupervised neural networks
These descriptors were used as input data in the SOM Toolbox 2.0 (Matlab) [8], a program that separates the relevant descriptors and their respective maps, obtaining the location of the molecules with higher and lower values for each descriptor. In the self-organizing matrix, it was possible to observe the location of the sites assigned to the molecules for each descriptor, to relate the similarities between the different types of diterpenes.

4. Conclusions

The Self-Organizing Map obtained separated the subfamilies Lamioideae (L) and Scutellarioideae (S) from the family Lamiaceae, using the molecular descriptors. Thus, the similar compounds were grouped in relation to the molecular fragments, which were later labeled according to the botanical occurrence in these subfamilies. Therefore, the SOMs that were generated enable the use of this tool in the search for diterpenes with potential biological activity according to the respective taxonomic information.

Acknowledgments

National Council for Scientific and Technological Development (Conselho Nacional de Desenvolvimento Científico e Tecnológico) CNPq

References

1. Li, B., Cantino, P. D., Olmstead, R. G., Bramley, G. L. C., Xiang, C.-L., Ma, Z.-H., Tan, Y.-H., Zhang, D.-X. A large-scale chloroplast phylogeny of the Lamiaceae sheds new light on its subfamilial classification. *Scientific Reports*, 6:34343, 1-18, 2016.
2. Lemes, G. F.; Ferri, P. H.; Lopes, M. N. CONSTITUINTES QUÍMICOS DE *Hyptidendron canum* (Pohl ex Benth.) R. Harley (LAMIACEAE). *Química Nova*, 34(1), 39-42, 2011.
3. Falcão, D. Q.; Fernandes, S. B. O.; Menezes, F. S. Triterpenos de *Hyptis fasciculata* Benth. *Revista Brasileira de Farmacognosia*, 13 (supl.), 81-83, 2003.
4. Scotti, M.; Herrera-Acevedo, C.; Oliveira, T.; Costa, R.; Santos, S.; Rodrigues, R.; Scotti, L.; Da-Costa, F. SistematX, an Online Web-Based Cheminformatics Tool for Data Management of Secondary Metabolites. *Molecules*, 23(1), 103, 2018.
5. Kode, S.R.L. Dragon (Software for Molecular Descriptor Calculation) version 7.0, 2016, Available online: <http://chm.kode-solutions.net>. accessed on August 9, 2018.
6. Zupan, J., Gasteiger, J. Neural Networks in Chemistry and Drug Design, 2nd ed.; Wiley-VCH: Weinheim, Germany, 1999.
7. Kohonen, T. *Self-Organizing Maps*, 1st ed.; Springer: Berlin, Germany, 2001.
8. Vesanto, J., Himberg, J., Alhoniemi, E., Parhankangas, J. Self-organizing map in Matlab: the SOM Toolbox. Proceedings of the Matlab DSP Conference, Espoo, Finland, 35-40, 1999.