Peer-reviewed version available at Biomolecules 2019, 9, 31; doi:10.3390/biom9010031

BIOFACQUIM: A Mexican compound database of natural products

B. Angélica Pilón-Jiménez, Fernanda I. Saldívar-González, Bárbara I. Díaz-Eufracio, José L. Medina-Franco*

Department of Pharmacy, National Autonomous University of Mexico, Mexico City, Mexico.

*Correspondence: medinajl@unam.com.mx (Medina-Franco); Tel.: +5255-5622-3899

Abstract: Compound databases of natural products have a major impact on drug discovery projects and other areas of research. The number of databases in the public domain with compounds from natural origin is increasing. Several countries have initiatives in place to construct and maintain compound databases that are representative of their diversity. Examples are Brazil, France, Panama and recently Vietnam. Herein, we discuss the first version of BIOFACQUIM, a novel compound database with natural products isolated and characterized in Mexico. We discuss its construction, curation, and a complete chemoinformatic characterization of the content and coverage in chemical space. It is reported the profile of physicochemical properties, scaffold content, and diversity, as well as structural diversity based on molecular fingerprints. BIOFACQUIM is freely available.

Keywords: chemical space; chemical data set; chemoinformatics; consensus diversity plot; drug discovery; molecular diversity; visualization

1. Introduction

In light of big data, the significance of compound databases in drug discovery projects is continuously increasing. In fact, compound databases and chemical data sets are a centerpiece in pharmaceutical companies and other academic and government research centers [1]. In addition to compound databases, natural products have been a major resource in drug discovery [2,3]. As reviewed elsewhere, there are several drugs recently approved for clinical use that are natural products or are synthetic analogues of hit compounds initially identified from natural sources. A notable example is the fungi metabolite migalastat (Galafold®) approved in 2018 for the treatment of the Fabry disease [4]. Not unsurprisingly, natural product-based drug discovery is being coupled with other major drug discovery strategies such as high-throughput screening and virtual screening. This synergy has boosted that over

Peer-reviewed version available at Biomolecules 2019, 9, 31; doi:10.3390/biom9010031

the past years, natural products are gaining attention again in the scientific community to address novel and/or difficult molecular targets, for instance, epigenetic targets [5,6].

Several compound databases of natural products have been constructed, curated and often maintained by academic and other non-for-profit research groups. Notable examples are the Universal Natural Product Database [7] and Traditional Chinese Medicine TCM database@Taiwan [8]. As reviewed recently [4], there are other compound databases that collect natural products from specific geographical areas and countries such as $NuBBE_{DB}$ from natural products from Brazil [9]. Recently it was released to the public VIETHERB: A Database for Vietnamese Herbal Species [10]. Other databases of natural products are discussed elsewhere [11-13]. Despite the fact that Mexico also has a large biodiversity, there are limited efforts to assemble a compound database of natural products. One example is UNIQUIM recently reviewed in [11].

The objective of this work is to introduce BIOFACQUIM as one of the first compound databases from natural products isolated and characterized in Mexico. We discuss the assembly of the first version of this chemical data set along with a chemoinformatic characterization of molecular diversity, scaffold content and coverage in chemical space. The compound database is freely available, and it is part of an ongoing effort to build, update and maintain a compound database representative of the biodiversity from Mexico.

2. Materials and Methods

2.1. BIOFACQUIM database

The database of natural products was assembled from a literature search. For the construction of the first version of BIOFACQUIM it was searched on Scopus database (www.scopus.com) the keywords “natural products” and “School of Chemistry of the National Autonomous University of Mexico (FQ, UNAM)”. This search led to a list of scientific papers and researchers that work with natural products. Eight journals were selected on which they had contributed the most thus far: Journal of Ethnopharmacology, Natural Products Research, Journal of Agricultural and Food Chemistry, Journal of Natural Products, Planta Medica, Phytochemistry, Natural Product Letters, and Molecules. As part of the search, three filters were used for the selection of the articles in each journal. The first one was the

search by institution (FQ, UNAM), the second one was the search by publication year (2000-2018), and the last filter was the detailed analysis of the articles to look if they had the procedure for the isolation, purification and characterization of the compounds from natural products.

With the module 'Wash' of the program Molecular Operating Environment (MOE) version 2018 [14], the database was curated. This was done to normalize and collect the most relevant information of the molecules. The data curation involved elimination of salts, adjustment of the protonation states, optimization of the geometry by energy minimization and elimination of the duplicated molecules. Default settings of the 'Wash' module were used.

2.2. Reference data sets

In order to characterize the diversity of BIOFACQUIM and explore its coverage in chemical space, seven compound databases of broad interest in drug discovery were used as reference. The structure files used in this work were taken from previous comparisons and chemoinformatic analysis of natural products [15]. The structures of the reference compound were curated using the same procedure described to prepare BIOFACQUIM. Table 1 summarizes the reference databases and the number of compounds. Of note, the reference collections include seven data sets of natural products.

Table 1. Reference databases [15] used in this work to compare BIOFACQUIM

Database	Size[a]
Approved drugs	1806
Cyanobacteria metabolites	473
Fungi metabolites	206
Marines	6253
MEGx	4103
Semi-syntetics (NATx)	26318
$NuBBE_{DB}$	2214

[a] Number unique compounds after data curation.

2.3. Molecular properties of pharmaceutical relevance

The curated BIOFACQUIM database was characterized calculating six physicochemical properties of therapeutic interest, namely; molecular weight (MW), octanol/water partition coefficient (SLogP),

topological surface area (TPSA), number of rotatable bonds (RB), number of H-bond donor atoms (HBD) and number of H-bond acceptor atoms (HBA). The statistical analysis was done with the program DataWarrior [16] calculating the mean, median and standard deviation of the calculated properties. Based on these statistics BIOFACQUIM was further compared with other natural products databases (NuBBE$_{DB}$, Cyanobacteria, Fungi, Marines and MEGx), with approved drugs and semisynthetic compounds (NATx) (Table 1).

2.4. Scaffold content

Scaffold content analysis enable to identify the most frequent scaffolds in compound data sets and, in this work, compare the scaffolds that contain approved drugs with those that have natural products. Scaffold content analyses also enable to identify potential novel scaffolds. The most frequent core molecular scaffolds of BIOFACQUIM were computed using the definition of Bemis and Murcko [17] where the core scaffold is obtained by systematically removing the side chains of the compounds. The most frequent scaffolds in BIOFACQUIM where compared with data from the literature (*vide infra*).

2.5. Chemical space: visual representation

In order to generate a visual representation of the chemical space of BIOFACQUIM, two visualization methods were used: principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE). PCA reduces data dimension by geometrically projecting them onto lower dimensions called principal components (PCs). The first PC is chosen to minimize the total distance between the data and its projection on the PC and to maximize the variance of the projected points. t-SNE is a nonlinear dimension reduction, where Gaussian probability distributions over high-dimensional space are constructed and used to optimize a Student t-distribution in low-dimensional space. The low-dimensional space maintains the pairwise similarity to the high-dimensional space, leading to a clustering on the embedding space without losing significant structural information. Further details of each visualization method of the chemical space are discussed elsewhere [18,19]. In this work, for t-SNE subsets of compounds were retrieved from large reference data sets (Table 1); namely: 40% of the Marine, MEGx and NUBBE$_{DB}$ data sets (2501, 1641 and 886 compounds, respectively). For NATx

and approved drugs, 1000 molecules were used. For cyanobacteria metabolites and fungi data sets the entire databases were employed (473 and 206 compounds, respectively).

2.6. "Global" diversity: Consensus diversity analysis

Since the chemical diversity strongly depends on the structure representation, it is recommended to consider multiple representations for a global or complete assessment. To this end, Consensus Diversity (CD) Plots have been proposed as simple two-dimensional graphs that enable the comparison of the diversity of compound data sets using four sets of structure representations [20]; typically, molecular fingerprints, scaffolds, molecular properties and number of compounds. CD Plots have been used to compare the diversity of natural product and other compound data sets [21]. Briefly, in a typical CD plot the scaffold and fingerprint diversity are represented along the Y- and X- axis, respectively. The diversity based on whole molecular properties of pharmaceutical interest are represented with a continuous color scale and the number of compounds is mapped into the plot using different sizes of data points. Further details are provided elsewhere [20]. To generate the CD plot of this work, for the Y-axis we used the area under the cyclic system recovery curve [22]. For the X-axis, we employed the median of the fingerprint-based diversity computed with MACCS keys (166-bits) and the Tanimoto coefficient. Both are established and representative metrics of the scaffold and fingerprint-based diversity, respectively. Subsets of compounds were retrieved from large reference data sets, (Table 1) considering the size of the databases; for NATx, Marines, MEGx, $NuBBE_{DB}$ and approved drugs 2000, 1500, 1000, 800 and 700 molecules were used respectively. For cyanobacteria metabolites and fungi data sets the entire databases were employed (473 and 206 compounds, respectively).

3. Results and Discussion

First, we present the results of the construction of the first version of BIOFACQUIM database followed by a first chemoinformatic characterization in terms of physicochemical properties, scaffold content, diversity and coverage in chemical space.

3.1. BIOFACQUIM database

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 November 2018 doi:10.20944/preprints201811.0627.v1

Peer-reviewed version available at Biomolecules 2019, 9, 31; doi:10.3390/biom9010031

As described in the Materials and Methods section, after the first survey in Scopus with the names of the researchers of the FQ, UNAM, it was applied three filters on the eight journals selected. Each of the 92 scientific papers selected were analyzed individually to extract the information of the natural products. BIOFACQUIM contains the following information: identification number (ID), compound name, SMILES, reference, publication year, kingdom (Plantae or Fungi), genus, and species of the natural product. The current and first version of BIOFACQUIM has 423 compounds. It should be noted that 316 compounds were isolated from 49 different *genus* of plants, 98 were isolated from 19 genus of fungi, and 9 compounds were isolated from Mexican propolis (sticky dark-colored hive product collected by bees from living plant sources). Figure 1 shows the distribution of compounds per year reported since the year 2000 as contained in the first version of the chemical data set.

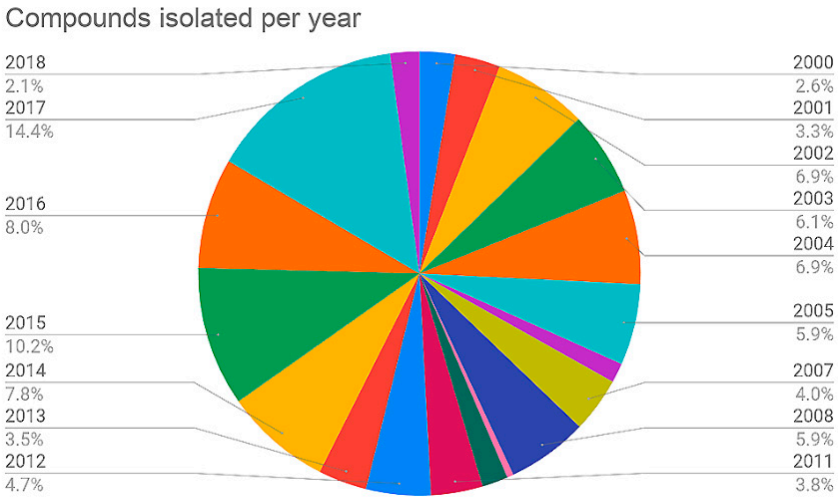
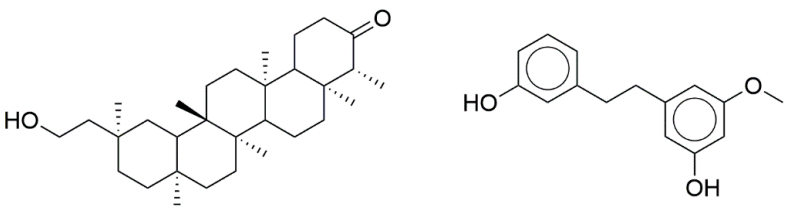


Figure 1. Distribution of compounds reported since 2000 as contained in the first version of BIOFACQUIM.

Figure 2 shows the chemical structures of representative compounds from the first version of BIOFACQUIM and further discussed below.



FQNP329 FQNP130

Figure 2. Selected compounds contained in BIOFACQUIM.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 November 2018 doi:10.20944/preprints201811.0627.v1

Peer-reviewed version available at Biomolecules 2019, 9, 31; doi:10.3390/biom9010031

155 *3.2. Molecular properties*

156 Figure 3 shows box plots of the distribution of the six calculated physicochemical properties (*vide supra*)
157 calculated for BIOFACQUIM. For comparison, the box plots also include the distribution of the same
158 properties of the seven reference data sets that were retrieved from the literature [15]. The three main
159 molecular properties of size, flexibility, and molecular polarity are described by MW; RB; and SlogP,
160 TPSA, HBA, and HBD, respectively. In the plots, the boxes enclose the data points with values within
161 the first and third quartile; the line that divides the box denote the median of distributions, and the lines
162 above and below indicate the upper and lower adjacent values. The red asterisks indicate the data
163 points with values beyond the upper and lower adjacent values. Summary statistics are presented at
164 the bottom of the box plots.

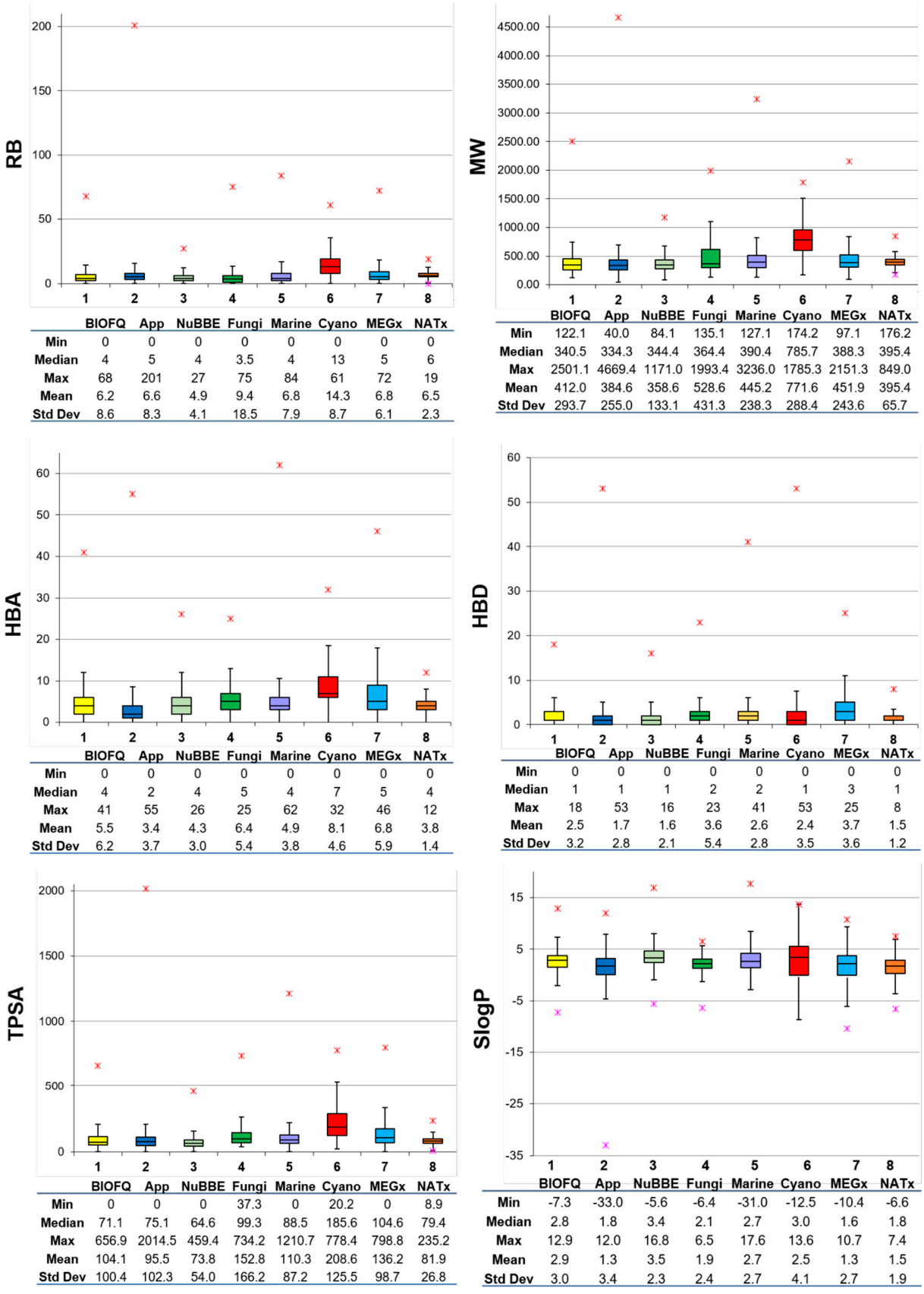
165 According to Figure 3, based on the mean of RB, BIOFACQUIM compounds have comparable
166 flexibility to approved drugs. The figure also shows that, except for Cyanobacteria metabolites, all
167 databases have a median up to 5 rotatable bonds (including approved drugs). The median and mean
168 MW of BIOFACQUIM are 340.5 and 412 g/mol, respectively. Notably, BIOFACQUIM and $NuBBE_{DB}$ have
169 the most similar MW profile as compared to drugs. BIOFACQUIM has a median of 4 HBA, the same
170 number that $NuBBE_{DB}$ and Marine's data sets. Furthermore, BIOFACQUIM has a very similar profile of
171 HBA as compared to MEGx. Comparing HBD, BIOFACQUIM, $NuBBE_{DB}$, NATx and Cyanobacteria have
172 the same median values with similar profile to approved drugs although with higher standard deviation
173 than approved drugs. Regarding TPSA, compounds in BIOFACQUIM are those that share the closest
174 values to the approved drugs. It should be noted that the Cyanobacteria metabolites set has the largest
175 distribution and the highest mean values of TPSA, being the double of the mean of the approved drugs.
176 The distribution of SlogP values indicates that, overall, natural products are slightly more hydrophobic
177 than approved drugs.

178 Taken together the results of the distribution of properties, it can be concluded that the current
179 version of BIOFACQUIM is, in general, most similar to $NuBBE_{DB}$ and Fungi data sets. This outcome is
180 in agreement with the findings that while assembling BIOFACQUIM and analyzing in detail the source
181 papers, it turned out that compounds were mostly isolated from plants and fungi.

182

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 November 2018 doi:10.20944/preprints201811.0627.v1

Peer-reviewed version available at Biomolecules 2019, 9, 31; doi:10.3390/biom9010031



RB	BIOFQ	App	NuBBE	Fungi	Marine	Cyano	MEGx	NATx
Min	0	0	0	0	0	0	0	0
Median	4	5	4	3.5	4	13	5	6
Max	68	201	27	75	84	61	72	19
Mean	6.2	6.6	4.9	9.4	6.8	14.3	6.8	6.5
Std Dev	8.6	8.3	4.1	18.5	7.9	8.7	6.1	2.3

MW	BIOFQ	App	NuBBE	Fungi	Marine	Cyano	MEGx	NATx
Min	122.1	40.0	84.1	135.1	127.1	174.2	97.1	176.2
Median	340.5	334.3	344.4	364.4	390.4	785.7	388.3	395.4
Max	2501.1	4669.4	1171.0	1993.4	3236.0	1785.3	2151.3	849.0
Mean	412.0	384.6	358.6	528.6	445.2	771.6	451.9	395.4
Std Dev	293.7	255.0	133.1	431.3	238.3	288.4	243.6	65.7

HBA	BIOFQ	App	NuBBE	Fungi	Marine	Cyano	MEGx	NATx
Min	0	0	0	0	0	0	0	0
Median	4	2	4	5	4	7	5	4
Max	41	55	26	25	62	32	46	12
Mean	5.5	3.4	4.3	6.4	4.9	8.1	6.8	3.8
Std Dev	6.2	3.7	3.0	5.4	3.8	4.6	5.9	1.4

HBD	BIOFQ	App	NuBBE	Fungi	Marine	Cyano	MEGx	NATx
Min	0	0	0	0	0	0	0	0
Median	1	1	1	2	2	1	3	1
Max	18	53	16	23	41	53	25	8
Mean	2.5	1.7	1.6	3.6	2.6	2.4	3.7	1.5
Std Dev	3.2	2.8	2.1	5.4	2.8	3.5	3.6	1.2

TPSA	BIOFQ	App	NuBBE	Fungi	Marine	Cyano	MEGx	NATx
Min	0	0	0	37.3	0	20.2	0	8.9
Median	71.1	75.1	64.6	99.3	88.5	185.6	104.6	79.4
Max	656.9	2014.5	459.4	734.2	1210.7	778.4	798.8	235.2
Mean	104.1	95.5	73.8	152.8	110.3	208.6	136.2	81.9
Std Dev	100.4	102.3	54.0	166.2	87.2	125.5	98.7	26.8

SlogP	BIOFQ	App	NuBBE	Fungi	Marine	Cyano	MEGx	NATx
Min	-7.3	-33.0	-5.6	-6.4	-31.0	-12.5	-10.4	-6.6
Median	2.8	1.8	3.4	2.1	2.7	3.0	1.6	1.8
Max	12.9	12.0	16.8	6.5	17.6	13.6	10.7	7.4
Mean	2.9	1.3	3.5	1.9	2.7	2.5	1.3	1.5
Std Dev	3.0	3.4	2.3	2.4	2.7	4.1	2.7	1.9

Figure 3. Box plots for the physicochemical properties of BIOFACQUIM (BIOFQ) and reference data sets (Table 1). The boxes enclose data points with values within the first and third quartile. The red asterisks indicate outliers.

3.3. Scaffold content

Figure 4 shows the 27 most populated molecular scaffolds in BIOFACQUIM that include half (50.6%) of the 423 compounds that make up the database. Other than benzene that is also highly frequent in several other compound databases [21], the second most frequent scaffold is a flavan-related scaffold (5 %) followed by 1,3-benzodioxole and dibenzyl core scaffolds (2.4 %). Interestingly, these last three frequent scaffolds in BIOFACQUIM are not the most frequent in other databases of natural products [15].

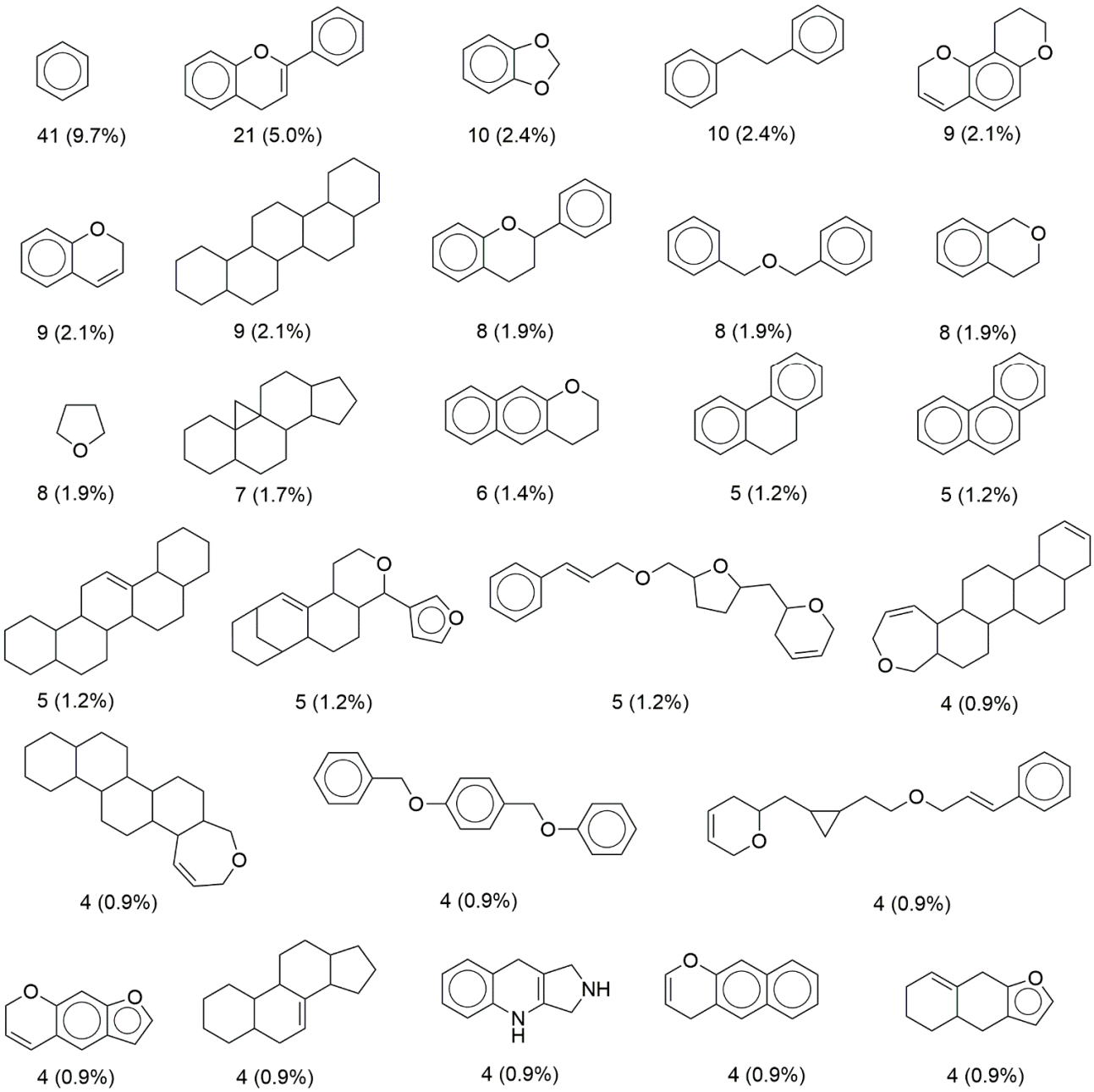


Figure 4. Most frequent scaffolds in BIOFACQUIM. The frequency and percentage are shown. The 27 scaffolds shown in the figure contain half of the total compounds in the database (50.6%).

196 *3.4. Chemical space*

197 As explained in the Materials and Methods section, a visual analysis of the chemical space of
198 BIOFACQUIM was done with two visualization methods, PCA and t-SNE. The visual representation with
199 PCA was based on physicochemical properties while the visualization with t-SNE was based on
200 molecular topological fingerprints.

201

202 *3.4.1. Visual representation based on properties*

203 Using the program KNIME [23], we did a visual comparison of the chemical space of BIOFACQUIM and
204 the reference databases. We used the node “Normalizer” which gives a linear transformation of all
205 values such that the minimum and maximum of each database. Then PCA was applied to reduce the
206 dimensionality of the six calculated physicochemical properties and then compare BIOFACQUIM with
207 the reference collections (*vide supra*, Table 1).

208 Figure 5 shows a visual representation of the property-based chemical space. Table S1 in the
209 Supplementary Material summarizes the corresponding loadings and eigenvalues for the first three PC.
210 The first two PCs capture 84 % of the variance while the first three recover 92 % of variance. Table S1
211 shows that for first PC, the larger loadings correspond to SlogP, followed by RB, whereas for the second
212 PC the largest loading corresponds to HBD.

213 The visual representation of the chemical space in Figure 5 indicates that some of the natural
214 products compounds occupy the same space as the already approved drugs. It also shows that there
215 are molecules in BIOFACQUIM and the Marine set that cover neglected regions of the currently drug–
216 like chemical space. Finally, Figure 5 suggest that BIOFACQUIM shares the chemical space of almost
217 all Fungi and $NuBBE_{DB}$.

218

219

220

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 November 2018 doi:10.20944/preprints201811.0627.v1

Peer-reviewed version available at Biomolecules 2019, 9, 31; doi:10.3390/biom9010031

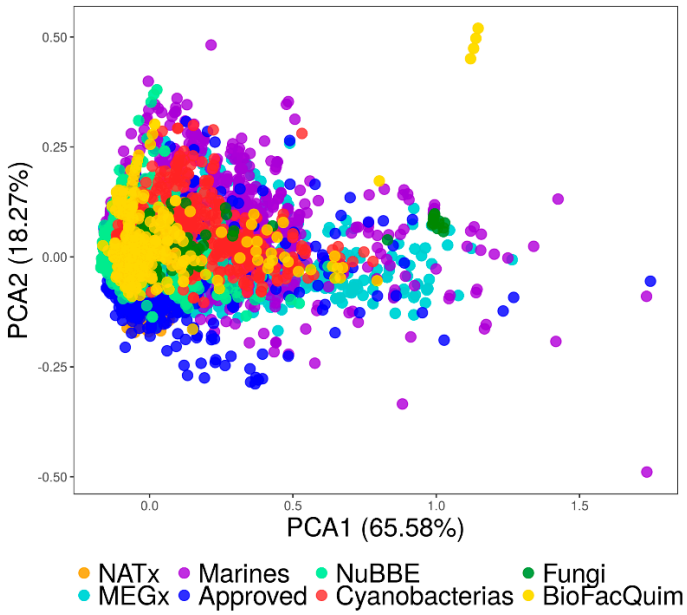


Figure 5. Visual representation of the chemical space based on physicochemical properties of eight data sets. BIOFACQUIM (423 compounds, yellow); Fungi metabolites (206 compounds, green); Cyanobacteria metabolites (473 compounds, red); $NuBBE_{DB}$ (2214 compounds, light green); NATx (26318 compounds, orange); MEGx (4103 compounds, blue); Marine metabolites (6253 compounds, lilac); FDA Approved drugs (1806 compounds, dark blue).

3.4.2. Visual representation based on molecular fingerprints

Figure 6 shows a visual representation of the chemical space of BIOFACQUIM based on topological fingerprints using t-SNE (see the Materials and Methods). Figure 6a compare BIOFACQUIM with all other reference data sets. Figure 6b shows a comparison of BIOFACQUIM with approved drugs. Figure 6a shows three main groups or clusters of in which all the databases have compounds. The clusters indicate that the visualization method and the fingerprints can distinguish three major core structures that would have detailed variations in the structure. Figure 6b indicates that there are compounds in BIOFACQUIM with a high structural similarity to approved drugs. Notable examples are the compounds FQNP329 (chemical structure in Figure 2), that is similar to ethinylestradiol (App_75), and FQNP130 to choline (App_878). Other comparisons with t-SNE are shown in Figure S3 in the Supplementary Material.

Based on the assessment of the chemical space in particular the position of BIOFACQUIM relative to other reference libraries in chemical space, it can be concluded that the compounds in BIOFACQUIM are very similar to drugs based on: physicochemical properties (PCA) and structural fingerprints (t-SNE). Therefore, the chemical space analysis further supports the use of BIOFACQUIM in drug discovery projects.

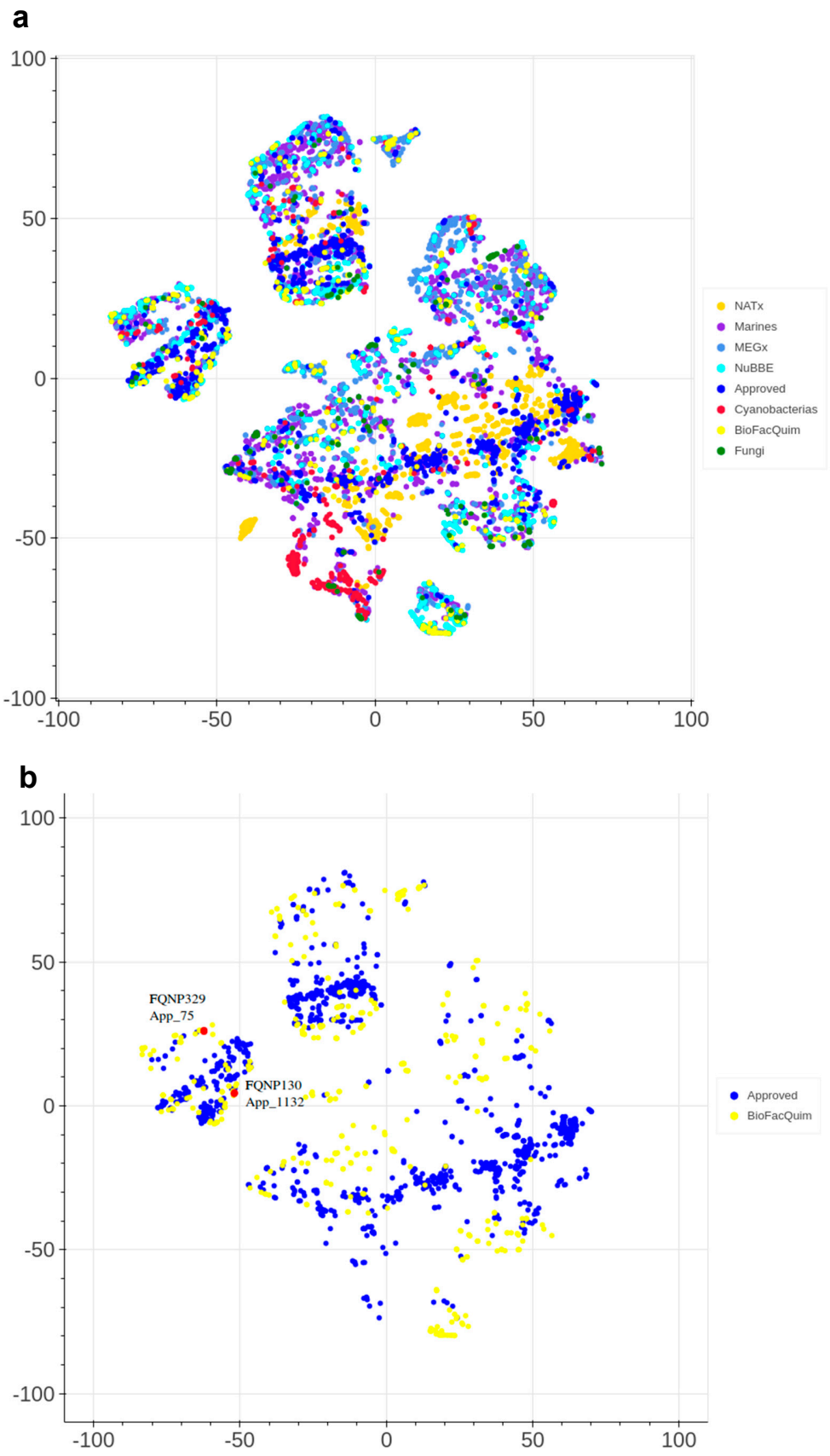


Figure 6. Visual representation of the chemical space of BIOFACQUIM compared with: **a**) All reference data sets; **b**) Approved drugs. The visualization was generated using t-SNE based on topological fingerprints.

3.5. "Global" diversity: Consensus diversity analysis

As elaborated in the Materials and Methods section, a CD plot was used to compare the diversity of BIOFACQUIM with the diversity of the reference data sets based on molecular fingerprints, scaffolds, and whole (physicochemical) properties. Figure 7 shows the CD plot representing on the X-axis the MACCS keys/Tanimoto similarity. Here, lower values indicate larger fingerprint-based diversity (further details of the fingerprint-based diversity assessment are presented in Figure S1 in the Supplementary Material). The Y-axis of the CD plot represents the scaffold diversity where lower values (of the area under the scaffold recovery curve – see Table S2 in the Supplementary Material) indicate higher scaffold diversity. The property-based diversity of BIOFACQUIM and each database was calculated as the Euclidean distance of the scaled properties. The values were represented on the CD Plot color the data points using a continuous color scale: darker color represents lower diversity while lighter color represents higher diversity. Finally, the relative size of the databases is represented with different point sizes where smaller data points indicate data sets with less number of molecules. The CD plot in Figure 7 shows that BIOFACQUIM and cyanobacteria are found in the area that represents a low diversity of both scaffold and fingerprints. This may be attributed to the fact that this is the first version of the database. Regarding the diversity based on physicochemical properties, it is observed that cyanobacteria metabolites have a larger diversity (e.g., lighter blue data point in Figure 7) as compared to BIOFACQUIM. This is consistent with the analysis of the box plots discussed in section 3.2. Figure 7 also indicates that approved drugs are in the region of the plot that represents a high diversity of scaffolds and fingerprints. The large scaffold and fingerprint-based diversity of approved drugs is consistent with previous reports [20, 21].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 November 2018 doi:10.20944/preprints201811.0627.v1

Peer-reviewed version available at Biomolecules 2019, 9, 31; doi:10.3390/biom9010031

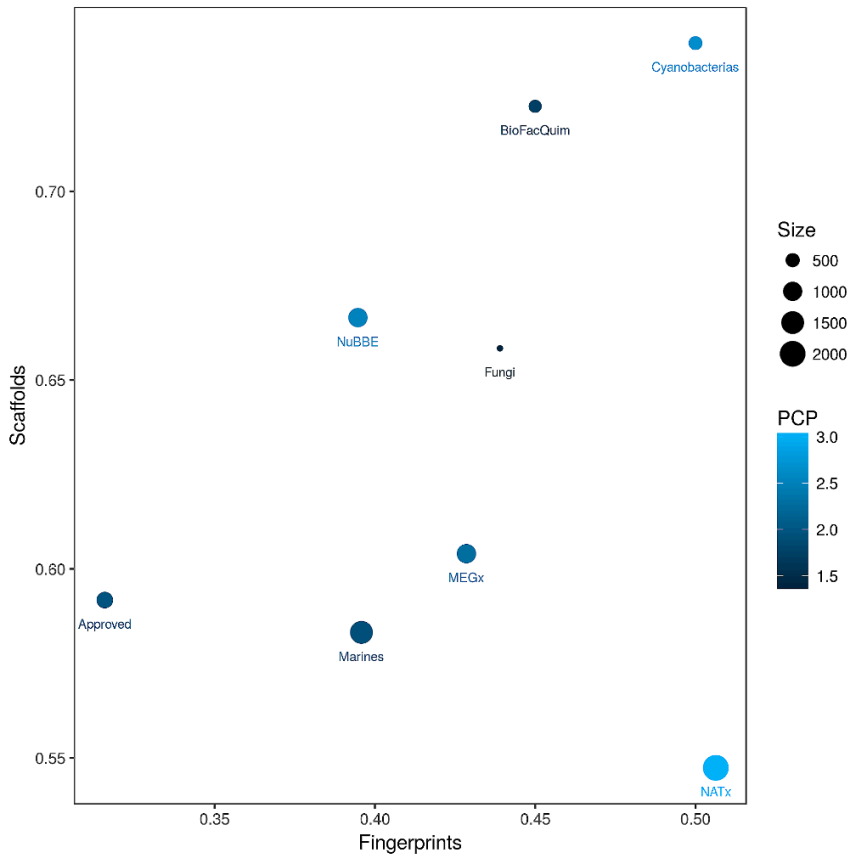


Figure 7. Consensus Diversity Plot comparing the global diversity of BIOFACQUIM with other natural products databases. The structural diversity (fingerprint diversity) was calculated with the median Tanimoto coefficient of MACCS keys fingerprints is plotted on the X axis. The scaffold diversity of each database was defined as the area under the curve (AUC) of the respective scaffold recovery curves, and it is represented on the Y axis. The diversity based on physicochemical properties (PCP) was calculated with the Euclidean distance of six scaled properties (SlogP, TPSA, MW, RB, HBD and HBA) and is shown in a color scale. The distance is represented with a continuous color scale from light blue (more diverse) to dark blue (less diverse). The relative size of the data set is represented with the size of the data point: smaller data points indicate compound data sets with fewer molecules.

4. Conclusions

BIOFACQUIM is a compound database of natural products from Mexico being constructed, curated and maintained by an academic group. The first and current version of BIOFACQUIM herein described has 423 compounds reported over the past 10 years at the School of Chemistry of the National Autonomous University of Mexico (UNAM). The compound database contains the chemical name, SMILES notation, reference, kingdom (Plantae or Fungi), genus, and species of the natural product. The chemoinformatic characterization and analysis of the coverage and diversity of BIOFACQUIM in chemical space suggests

that they have a broad coverage overlapping with regions in drug like chemical space. The analysis also indicated that there are compounds in BIOFACQUIM with chemical structures very similar to drugs approved for clinical use and could, based on the similarity principle, be of pharmaceutical interest. Similar to other natural product databases BIOFACQUIM can be used in virtual screening to identify potential lead compounds or starting points for additional optimization. BIOFACQUIM is freely accessible through the web-site of D-TOOLS (www.difacquim.com/d-tools/) [24].

One of the major perspectives of this work already in progress is augmenting the size of BIOFACQUIM by expanding the search to other universities and research centers in Mexico, increasing the number of years and number of scientific peer-reviewed journals covered. A second major perspective of this work is to develop a searchable interface that will be called “BIOFACQUIM Explorer”. The interface is under construction and will be released to the public in due course.

Supplementary Materials: The following are available online. **Table S1.** Loadings for the first three principal components of the property space of eight databases; **Figure S1**. Distribution of the pairwise similarity values calculated for BIOFACQUIM and the reference data sets computed with MACCS keys (166-bits) and the Tanimoto coefficient; **Table S2**. Statistics of the cyclic system recovery curves for BIOFACQUIM and the reference data sets; **Figure S2**. Visual representation of the chemical space of BIOFACQUIM generated with t-SNE.

Author Contributions: Conceptualization, BAP-J, FIS-G, JLM-F; Methodology, BAP-J, FIS-G, BID-E, Formal analysis, BAP-J, BID-E, Writing, editing, BAP-J, JLM-F; Funding acquisition, JLM-F.

Funding: This research was supported by the *Programa de Apoyo a la Investigación y el Posgrado* (PAIP) grant 5000-9163, Facultad de Química, UNAM.

Acknowledgements: BAP-J is grateful for the support given by the subprogram 127 “Basic Training in Research” of the School of Chemistry, UNAM. FIS-G and BID-E are thankful to *Consejo Nacional de Ciencia y Tecnología, Mexico* (CONACyT) for scholarships number 629458 and 620289, respectively. Discussions with Oscar Palomino-Hernández to implement t-SNE are acknowledged.

Conflict of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

References

1. Miller, M.A. Chemical database techniques in drug discovery. *Nat Rev Drug Discov* **2002**, *1*, 220-227.
2. Newman, D.J. From natural products to drugs. *Phys Sci Rev* **2018**, in press. DOI:10.1515/psr-2018-0111.
3. Newman, D.J.; Cragg, G.M. Natural products as sources of new drugs from 1981 to 2014. *J. Nat. Prod.* **2016**, *79*, 629-661.
4. Saldívar-González, F.I.; Pilón-Jiménez, B.A.; Medina-Franco, J.L. Chemical space of naturally occurring compounds. *Phys Sci Rev* **2018**, in press. Doi: 10.1515/psr-2018-0103.
5. Saldívar-González, F.I.; Gómez-García, A.; Chávez-Ponce de León, D.E.; Sánchez-Cruz, N.; Ruiz-Rios, J.; Pilón-Jiménez, B.A.; Medina-Franco, J.L. Inhibitors of DNA methyltransferases from natural sources: A computational perspective. *Front. Pharmacol.* **2018**, *9*, 1144.
6. Thomford, N.; Senthebane, D.; Rowe, A.; Munro, D.; Seele, P.; Maroyi, A.; Dzobo, K. Natural products for drug discovery in the 21st century: Innovations for novel drug discovery. *Int J Mol Sci* **2018**, *19*, 1578.
7. Gu, J.; Gui, Y.; Chen, L.; Yuan, G.; Lu, H.-Z.; Xu, X. Use of natural products as chemical library for drug discovery and network pharmacology. *PLoS One* **2013**, *8*, e62839.
8. Chen, C.Y.-C. TCM database@Taiwan: The world's largest traditional chinese medicine database for drug screening in silico. *PLoS One* **2011**, *6*, e15939.
9. Pilon, A.C.; Valli, M.; Dametto, A.C.; Pinto, M.E.F.; Freire, R.T.; Castro-Gamboa, I.; Andricopulo, A.D.; Bolzani, V.S. Nubbedb: An updated database to uncover chemical and biological information from brazilian biodiversity. *Sci Rep* **2017**, *7*, 7215.
10. Nguyen-Vo, T.-H.; Le, T.Q.M.; Pham, D.T.; Nguyen, T.D.; Le, P.H.; Nguyen, A.D.T.; Nguyen, T.D.; Nguyen, T.-N.N.; Nguyen, V.A.; Do, H.T., *et al.* VIETHERB: A database for vietnamese herbal species. *J. Chem. Inf. Model.* **2018**, in press. DOI: 10.1021/acs.jcim.1028b00399.
11. Medina-Franco, J.L. Discovery and development of lead compounds from natural sources using computational approaches. In *Evidence-based validation of herbal medicine*, Mukherjee, P., Ed. Elsevier: 2015; pp 455-475.
12. Chun-Wei, T. Public databases of plant natural products for computational drug discovery. *Curr. Comput. Aided Drug Des.* **2014**, *10*, 191-196.

13. Chen, Y.; Garcia de Lomana, M.; Friedrich, N.-O.; Kirchmair, J. Characterization of the chemical space of known and readily obtainable natural products. *J. Chem. Inf. Model.* **2018**, *58*, 1518-1532.
14. Molecular Operating Environment (MOE), version 2018.08, Chemical Computing Group INC., Montreal, Quebec, Canada. Available at http://www.chemcomp.com
15. Saldívar-González, F.I.; Valli, M.; Andricopulo, A.D.; Vanderlan da Silva Bolzani; Medina-Franco, J.L. Chemical diversity of nubbe database: A chemoinformatic characterization. *J. Chem. Inf. Model.* **2018**, (revision submitted).
16. Sander, T.; Freyss, J.; von Korff, M.; Rufener, C. Datawarrior: An open-source program for chemistry aware data visualization and analysis. *J. Chem. Inf. Model.* **2015**, *55*, 460-473.
17. Bemis, G.W.; Murcko, M.A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1996**, *39*, 2887-2893.
18. van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn Res.* **2008**, *9*, 2579-2605.
19. Osolodkin, D.I.; Radchenko, E.V.; Orlov, A.A.; Voronkov, A.E.; Palyulin, V.A.; Zefirov, N.S. Progress in visual representations of chemical space. *Exp. Opin. Drug Discov.* **2015**, *10*, 959-973.
20. González-Medina, M.; Prieto-Martínez, F.D.; Medina-Franco, J.L. Consensus diversity plots: A global diversity analysis of chemical libraries. *J. Cheminf.* **2016**, *8*, 63.
21. Naveja, J.; Rico-Hidalgo, M.; Medina-Franco, J. Analysis of a large food chemical database: Chemical space, diversity, and complexity. *F1000Research* **2018**, *7*, 993.
22. Medina-Franco, J.L.; Martínez-Mayorga, K.; Bender, A.; Scior, T. Scaffold diversity analysis of compound data sets using an entropy-based measure. *QSAR Comb. Sci.* **2009**, *28*, 1551-1560.
23. Berthold, M.R.; Cebron, N.; Dill, F.; Gabriel, T.R.; Kötter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B. Knime: The konstanz information miner. In Data analysis, machine learning and applications: Proceedings of the 31st annual conference of the gesellschaft für klassifikation e.V., albert-ludwigs-universität Freiburg, March 7–9, 2007, Preisach, C.; Burkhardt, H.; Schmidt-Thieme, L.; Decker, R., Eds. Springer Berlin Heidelberg: Berlin, Heidelberg, 2008; pp 319-326.
24. Naveja, J.J.; Oviedo-Osornio, C.I.; Trujillo-Minero, N.N.; Medina-Franco, J.L. Chemoinformatics: A perspective from an academic setting in Latin America. *Mol. Divers.* **2018**, *22*, 247-258.