RESEARCH ARTICLE

REVISED Functional group and diversity analysis of BIOFACQUIM: A Mexican natural product database [version 2; peer review: 3 approved]

Norberto Sánchez-Cruz, B. Angélica Pilón-Jiménez, José L. Medina-Franco

Department of Pharmacy, School of Chemistry, National Autonomous University of Mexico, Mexico City, Mexico City, 04510, Mexico

v2 **First published:** 10 Dec 2019, **8**(Chem Inf Sci):2071
https://doi.org/10.12688/f1000research.21540.1
Latest published: 08 Jun 2020, **8**(Chem Inf Sci):2071
https://doi.org/10.12688/f1000research.21540.2

Abstract

Background: Natural product databases are important in drug discovery and other research areas. An analysis of its structural content, as well as functional group occurrence, provides a useful overview, as well as a means of comparison with related databases. BIOFACQUIM is an emerging database of natural products characterized and isolated in Mexico. Herein, we discuss the results of a first systematic functional group analysis and global diversity of an updated version of BIOFACQUIM.

Methods: BIOFACQUIM was augmented through a literature search and data curation. A structural content analysis of the dataset was performed. This involved a functional group analysis with a novel algorithm to automatically identify all functional groups in a molecule and an assessment of the global diversity using consensus diversity plots. To this end, BIOFACQUIM was compared to two major and large databases: ChEMBL 25, and a herein assembled collection of natural products with 169,839 unique compounds.

Results: The structural content analysis showed that 15.7% of compounds and 11.6% of scaffolds present in the current version of BIOFACQUIM have not been reported in the other large reference datasets. It also gave a diversity increase in terms of scaffolds and molecular fingerprints regarding the previous version of the dataset, as well as a higher similarity to the assembled collection of natural products than to ChEMBL 25, in terms of diversity and frequent functional groups.

Conclusions: A total of 148 natural products were added to BIOFACQUIM, which meant a diversity increase in terms of scaffolds and fingerprints. Regardless of its relatively small size, there are a significant number of compounds and scaffolds that are not present in the reference datasets, showing that curated databases of natural products, such as BIOFACQUIM, can serve as a starting point to increase the biologically relevant chemical space.

Open Peer Review

Reviewer Status ✓ ✓ ✓

	Invited Reviewers		
	1	**2**	**3**
version 2 (revision) 08 Jun 2020	✓ report		
version 1 10 Dec 2019	? report	✓ report	✓ report

1. **Johannes Kirchmair**, University of Vienna, Vienna, Austria
2. **W. Patrick Walters**, Relay Therapeutics, Cambridge, USA
3. **Trong Tran**, University of the Sunshine Coast, Maroochydore, Australia

Any reports and responses or comments on the article can be found at the end of the article.

Keywords
Consensus Diversity Plot, compound databases, data mining, diversity, natural products, functional groups, in silico

This article is included in the Chemical Information Science gateway.

This article is included in the Mathematical, Physical, and Computational Sciences collection.

Corresponding authors: Norberto Sánchez-Cruz (norberto.sc90@gmail.com), José L. Medina-Franco (medinajl@unam.mx)

Author roles: Sánchez-Cruz N: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Software, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Pilón-Jiménez BA**: Investigation, Methodology, Writing – Original Draft Preparation, Writing – Review & Editing; **Medina-Franco JL**: Conceptualization, Funding Acquisition, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by a Consejo Nacional de Tecnología (CONACyT) scholarship number 335997 (NS-C). The work was also supported by the program NUATEI (Nuevas Alternativas para el Tratamiento de Enfermedades Infecciosas), Instituto de Ciencias Biomédicas, UNAM. The authors are grateful for the computational resources granted by Dirección General de Cómputo y de Tecnologías de Información y Comunicación (DGTIC), project grant LANCAD-UNAM-DGTIC-335 that allows to use the supercomputer Miztli at UNAM.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Sánchez-Cruz N, Pilón-Jiménez BA and Medina-Franco JL. **Functional group and diversity analysis of BIOFACQUIM: A Mexican natural product database [version 2; peer review: 3 approved]** F1000Research 2020, **8**(Chem Inf Sci):2071 https://doi.org/10.12688/f1000research.21540.2

First published: 10 Dec 2019, **8**(Chem Inf Sci):2071 https://doi.org/10.12688/f1000research.21540.1

REVISED **Amendments from Version 1**

We thank all three reviewers for the constructive comments and suggestions to improve the study and the manuscript. In the revised version, we repeated the analysis standardizing the chemical structures using MolVS. The standardization procedure is described in the revised "Methods, Databases and curation" section; the code was made available at GitHub. To visualize the chemical space of BIOFACQUIM and reference compounds databases (entire sets and drug-like subsets) we employed the recently published method TMAP (Tree Manifold Approximation and Projection) using Morgan fingerprints with radius 2. An analysis of the distribution of the fraction of carbon atoms that are sp3 hybridized was added as a new subsection "Molecular complexity". Further details of the BIOFACQUIM database were added (source of compounds and names of the peer-reviewed journals where the compound information was retrieved from). The English language of the entire manuscript was revised.

Any further responses from the reviewers can be found at the end of the article

Introduction

Natural product-based drug discovery continues to be an important part of drug discovery. Recently, the synergy between natural product research with molecular modeling and chemoinformatics is gaining importance, speeding up the drug discovery process[1,2]. As part of these synergistic efforts, curated databases of natural products have an important role as they are major tools for data mining, hypothesis generation, and starting points of virtual screening. There are several databases of natural products in the public domain as reviewed recently[3]. Our research group has reported initial efforts to assemble a database of natural products from Mexico called BIOFACQUIM[4]. As part of that work, scaffold content and chemical space diversity were examined. However, detailed functional group (FG) content analysis, which has been proven to be valuable to characterize compound databases[5], in particular from natural sources[6], has not been reported for BIOFACQUIM. One of the main reasons is that most of the currently available software employed for identification of functional groups rely on a predefined set of substructures, even when it has been established that one of the major features that discriminate natural products from synthetic compounds are their unique functional groups.

Herein, we report a functional group content analysis of an updated version of BIOFACQUIM. We employed a validated and novel algorithm that identifies all functional groups in a molecule. As part of the analysis and to compare the results of BIOFACQUIM we also discuss the functional group contents of other large and related databases in the public domain, namely ChEMBL 25[7] and a herein assembled collection of natural products (NPs) with 169,839 compounds.

Methods

Databases and data curation

As described elsewhere, the first version of BIOFACQUIM was developed as a proof-of-concept database applying several filters to include compounds[4]. Briefly, the database was focused on natural products published between 2000 and 2018 by research groups in a major Mexican institution in eight indexed journals: *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 additional criteria for inclusion of compounds and to increase the quality and reliability of the contents of the database, the procedure for the isolation, purification, and characterization of the natural product should have been described in the article. In this work, we expanded the contents of the BIOFACQUIM database to further explore the diversity of natural products from Mexico.

The second version of BIOFACQUIM was assembled using the same methodology described to develop the first version[4] extending the date of publication to 2019. To achieve the objective of being representative of Mexico, one additional criterion was considered, including only compounds collected in Mexico at any of its institutions (universities, research laboratories and research centers). For the new version of the database, the same procedure for the curation was performed[4], using Molecular Operating Environment (MOE) software, although this procedure can be performed using open source software, such as MolVS and RDKit. The updated and curated version of BIOFACQUIM contains 531 compounds.

Table 1 summarizes the information of BIOFACQUIM and other major compound databases used in this work as reference: ChEMBL 25 as a representative example of the biologically

Table 1. Compound databases analyzed in this work and summary statistics of their diversity.

Database	Size (compounds)	Median similarity (MACCS keys - 166 bits)	Median similarity (Morgan2 - 1024 bits)	Mean distance (PCP)	Scaffold diversity (AUC)	Scaffold diversity (F_{50})
BIOFACQUIM V1	403	0.457	0.123	3.648	0.725	0.165
BIOFACQUIM V2	503	0.446	0.119	3.319	0.710	0.171
Natural products	168,030	0.422	0.111	3.775	0.830	0.032
ChEMBL 25	1,667,509	0.382	0.117	2.187	0.809	0.057

PCP: physicochemical properties; AUC: area under the cyclic system retrieval curve.

tested chemical space with 1,667,509 unique compounds; and a collection of known natural products with a total of 168,030 molecules. The reference natural product collection was assembled from three general and publicly available natural products databases: the Universal Natural Products Database (UNPD)[8], the Natural Products Atlas[9] and Natural Products in PubChem Substance Database[10]. The data sets were curated using the same procedure. Briefly, compounds were standardized and those consisting of multiple components were split and the largest component was retained. Compounds consisting of any element other than H, B, C, N, O, F, Si, P, S, Cl, Se, Br and I, as well as compounds with valence errors, were removed from the data set. The remaining compounds were neutralized and reionized to subsequently generate a canonical tautomer. Finally, canonical simplified molecular-input line-entry system (SMILES) (ignoring stereochemistry information) were generated as molecular representation and duplicate structures in the context of each database were removed. The entire process was performed by using the functions Standardizer, LargestFragmentChoser, Uncharger, Reionizer and TautomerCanonicalizer implemented in the molecule validation and standardization tool MolVS for the open source cheminformatics toolkit RDKit. The code is available at GitHub (https://github.com/DIFACQUIM/IFG_General).

Databases overlap

Overlap of BIOFACQUIM with the databases selected as reference was assessed in terms of three different structural levels: compounds, scaffolds, and functional groups. Compound overlap was determined in terms of canonical SMILES. For scaffold comparison we use the definition proposed by Bemis and Murcko[11] as implemented in RDKit, while for functional group overlap we selected the recently published definition and implementation suggested by Ertl[5]. For each structural level, we identified the unique structures belonging to each dataset as well as those belonging to two or three of them.

Functional group analysis

For the functional group content analysis we selected the algorithm recently described by Ertl[5], which is able to identify all functional groups in a molecule based on an iterative marching through its atoms. In short, the proposed algorithm identifies all heteroatoms in a molecule, all atoms connected by multiple bonds as well as the atoms in oxirane, aziridine, and thiirane rings. Afterwards, all connected atoms are joined together to form a functional group. Single aromatic heteroatoms are retained only if they are connected to an additional aliphatic functionality. Finally, a generalization scheme is applied in which for a defined list of common FGs, information about the parent carbon is retained (e.g. to differentiate between alcohols and phenols) as well as hydrogen atoms (e.g. to differentiate between aldehydes and ketones). The method is fully described in 5. An open source version of this algorithm is available for Python (https://github.com/rdkit/rdkit/tree/master/Contrib/IFG); however, it does not cover the generalization scheme proposed originally. To this end and based on the code available, we implemented with RDKit a fragmentation approach considering keeping the parent carbon and hydrogen atoms proposed originally, were the remaining carbon atoms are replaced by dummy atoms. This implementation works over a SMILES string and returns a list with the canonical SMILES of the FGs identified in the molecule. The code is freely available at GitHub (https://github.com/DIFACQUIM/IFG_General). After determining the FGs content of the different datasets, we compare the proportion of the most frequent FGs at each library.

Complexity analysis

The fraction of carbon atoms that are sp^3 hybridized (F-sp3) is a common metric to quantify molecular complexity[12]. Higher values for this descriptor have been associated to an improved binding selectivity of compounds[13]. In order to compare the complexity of BIOFACQUIM with the data sets selected as reference, we computed the F-sp3, as implemented in RDKit, for all compounds in the three data sets and compared its distribution among libraries.

Chemical space visualization

In order to generate a visual representation of the chemical space covered by the analyzed databases, we selected a recently proposed method, named TMAP[14] (Tree Manifold Approximation and Projection). This method enables the visualization of up to millions of data points with high dimensionality as a two-dimensional tree and has shown to be better suited than t-distributed Stochastic Neighbor Embedding[15] (t-SNE) and Uniform Manifold Approximation and Projection[16] (UMAP) for the exploration of large datasets. TMAP consists in four phases: (I) the input data are indexed in an local sensitive hashing forest data structure, using l prefix trees and d hash functions in encoding the data, (II) an undirected weighted c-approximate k-nearest neighbor graph (c-k-NNG) is constructed from the indexed data points with the Jaccard distances between vertices used as edges weights, (III) a minimum spanning tree (MST) is constructed for the weighted c-k-NNG using Kruskal's algorithm and (IV) a layout for the resulting MST is constructed by using a spring-electrical model layout algorithm with multilevel multipole-based force approximation as provided by the modular C++ library, open graph drawing framework[17] (OGDF).

For the description of compounds, we selected Morgan fingerprints with radius 2 (Morgan2, 1024-bits) as implemented in RDKit. For the generation of the TMAP, the input data was encoded by 1024 hash functions and indexed using 64 prefix trees. The weighted c-k-NNG was built using the 5 nearest neighbors and a factor of 20 for the LSH forest query algorithm. For the layout generation a node size of 0.01 was selected while all the remaining parameters were set to default. These calculations were done using the TMAP python package and all the charts were generated using the matplotlib library[18].

Global diversity

The "global" or total diversity of the datasets was analyzed through the Consensus Diversity (CD) Plot[19]. A CD Plot is a two-dimensional representation of compound datasets based on four different and complementary diversity criteria: molecular fingerprints, molecular scaffolds, physicochemical properties (PCP), and size. Fingerprint-based diversity of each dataset is represented in the X-axis, while scaffold-based diversity is

represented in the Y-axis, PCP-based diversity is represented as the filling of the data points using a continuous color scale and the size of the data set is represented with the size of the data points.

For this work, scaffold diversity was assessed as the area under the cyclic system retrieval curve and the fraction of chemotypes that covers 50% of the dataset (F_{50}). The median of the lower triangle from the pairwise similarity matrix computed as the Tanimoto coefficient of both MACCS keys (166-bits) fingerprint and Morgan2 (1024-bits), were used as molecular fingerprint-based diversity. For PCP-based diversity, six molecular properties of pharmaceutical interest were computed for each unique compound, being averaged molecular weight (AMW) partition coefficient octanol/water (SlogP), number of hydrogen bond donors (HBD), number of hydrogen bond acceptors (HBA), number of rotatable bonds (RB), and topological polar surface area (TPSA); PCP-based diversity was measured as the mean distance of the lower triangle of the pairwise distance matrix computed as the Euclidean distance of those PCPs scaled (mean 0 and unit variance). The number of compounds in each dataset was selected as measure of the size-based diversity. PCPs were calculated as implemented in RDKit and the CD Plot was constructed using R.

Results and discussion

Update of BIOFACQUIM

As described in the Methods section, the updated BIOFACQUIM database contains the chemical structure of 531 compounds, all collected from Mexico. As with the first version[4], each molecule is annotated with information of the chemical structure, the original source of the information (Digital Object Identifier, DOI, to reference paper), kingdom, genus, and species of the organism from which the natural product was isolated, place of collection (city and state), and activity value of the reported biological activity. From the original dataset containing 423 compounds, 40 were discarded since they were not collected in Mexico, which means an increase of 148 unique compounds compared to the previous release of the database. The sources of the 531 compounds are distributed as follows: 406 from plants, 97 from fungus, 15 from propolis and 13 from marine animals.

Database overlap

To assess the chemical space not covered by ChEMBL and NPs but by BIOFACQUIM, we characterized the structural content of the three datasets in terms of unique compounds, scaffolds and functional groups and determined the overlap among them. Figure 1 depicts Venn diagrams showing the overlap

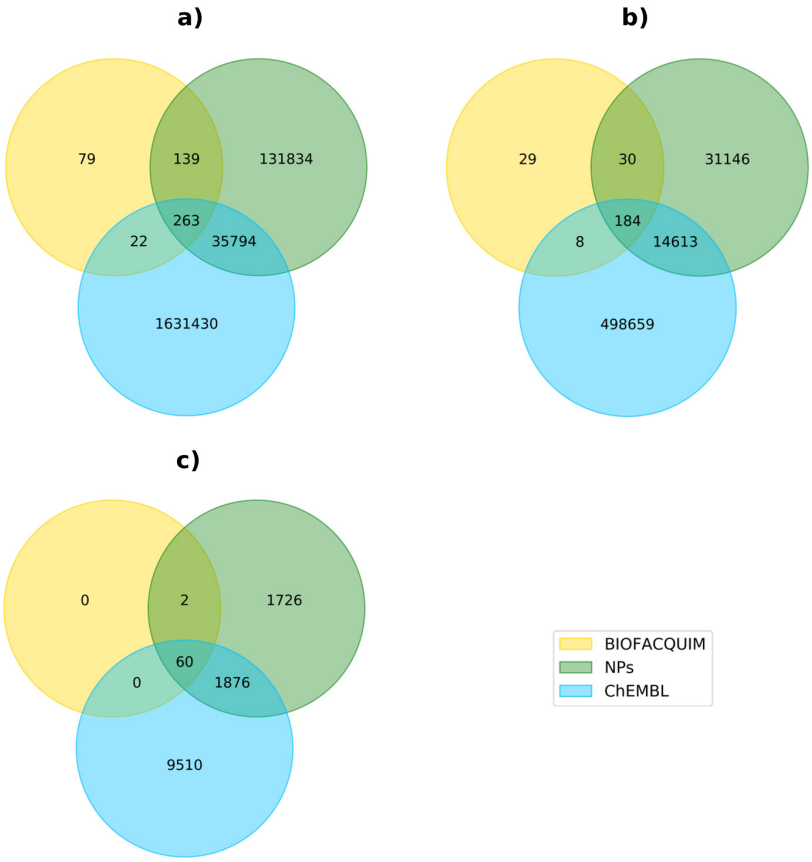

Figure 1. Overlap between BIOFACQUIM, reference natural products (NPs) and ChEMBL. Datasets content was analyzed in terms of (**a**) Compounds, (**b**) Scaffolds and (**c**) Functional Groups.

among those datasets. It should be noted that despite its small size in comparison with ChEMBL and NPs, 15.7% of compounds present in BIOFACQUIM have not been reported in those other major datasets, as well as 11.6% of its scaffolds. Figure 2 shows the structure of representative scaffolds identified in at least 2 compounds from BIOFACQUIM, all of them associated to compounds isolated from different plants. Scaffold **a** associated to hexasaccharides of convolvulinic and jalapinolic acids isolated from *Ipomoea purga*[20], scaffold **b** corresponding to glycosides of 4-phenylcoumarin isolated from *Exostema caribaeum*[21], scaffold **c** representing karwinaphthopyranones A2 and B3 isolated from *Karwinskia parvifolia*[22] and scaffold **d** associated to batatins X and XI isolated from *Ipomoea batatas*[23]. Another remarkable observation is the fact that most of the overlap of BIOFACQUIM with the other datasets involves NPs, either alone or in combination with ChEMBL, being 79.9%, 85.3%, and 100% for compounds, scaffolds and FGs, respectively.

Functional group analysis

A systematic analysis of functional groups was carried out over BIOFACQUIM and the two datasets selected as reference. 62, 3664, and 11446 functional groups were identified in BIOFACQUIM, NPs and ChEMBL datasets, respectively (the overlap between them is shown in Figure 1). From the total number of functional groups present in each dataset, only 12, 15, and 22 were present in at least 1% of the corresponding library (19.4%, 0.4% and 0.2%, respectively) while 30, 1879, and 5212 (48.4%, 51.3% and 45.5%, respectively) were singletons. This result is consistent with the typical power law observed in other databases[6]. The most frequent FGs present in BIOFACQUIM are oxygen-containing FGs, being the phenolic hydroxyl group (46.1%), followed by ether (41.4%), alcohol hydroxyl group (38.4%), alkene (28.6%) and ester (26.8%), which although in a different order, are the most frequent FGs in the herein assembled NPs collection and in other natural product libraries[6]. This is in contrast to ChEMBL in which only ether is part of the most frequent FGs while the rest of them are nitrogen containing FGs and halogens. The complete results of the FGs found of the datasets is included as *Extended data* (Supplementary File 1).

Comphlexity analysis

As described in the Methods section, molecular complexity of BIOFACQUIM, NPs and ChEMBL data sets were assessed through the F-sp3 of its compounds. Figure 3 depicts letter plots for the distribution of this descriptor among data sets. This shows that NPs is the more complex data set overall regarding to this metric, with a median of 0.60 and a long tail towards low values. In contrast, ChEMBL is the less complex data set, with a median F-sp3 of 0.31 and a long tail towards high values. BIOFACQUIM is in an intermediate position with a median of 0.42 and a more symmetrical distribution, which is consistent with its small size in comparison to the other sets and its major overlap with NPs.

Chemical space visualization

For visualization of the chemical space of the datasets compared in this work, we built a TMAP as described in the Methods section. This method allows the representation of large datasets as a two-dimensional tree. TMAP shows the relationships among subsets of data points and data points itself as branches and sub-branches, so similar compounds and clusters tend to be close in the final representation even if the tree edges are not included, for that reason all charts in this work do not show the tree edges. Figure 4 shows a visual representation of the chemical space of the three datasets analyzed in this work, including the whole datasets and drug-like subsets. Drug-like compounds for each subset were defined those complying with the Lipinski "rule of 5"[24] and Veber criteria[25] (AMW ≤ 500, -1 ≤ SlogP ≤ 5, HBA ≤ 10, HBD ≤ 5, RB ≤ 10 and TPSA ≤ 140). For this plot, in order to better illustrate the unique compounds present in BIOFACQUIM, all compounds belonging to more

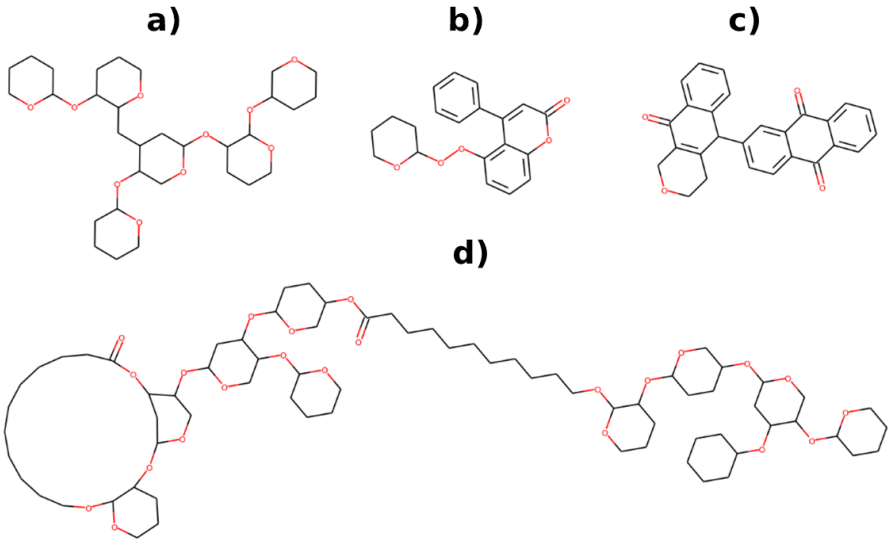

Figure 2. Representative unique scaffolds from BIOFACQUIM.

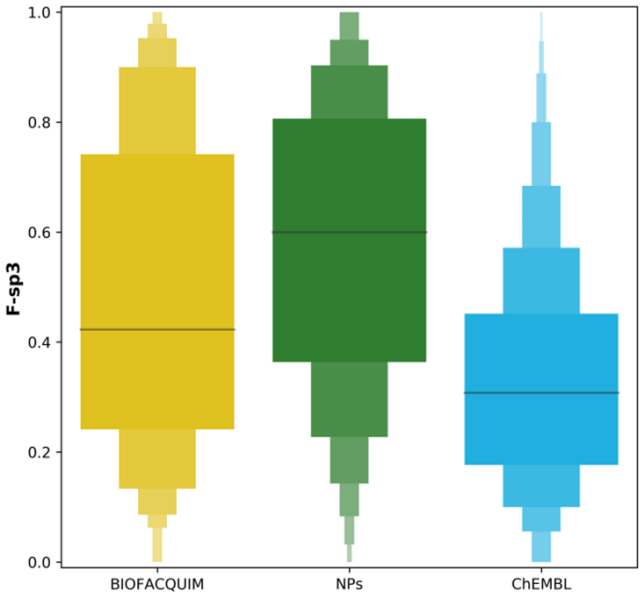

Figure 3. Distribution of F-sp3 for BIOFACQUIM, NPs, and ChEMBL.

than one library were assigned to a single one: ChEMBL if they belong to this dataset, NPs if they belong to this dataset but not to ChEMBL and BIOFACQUIM if they were unique for this library. Figure 4 shows that the chemical space covered by the analyzed datasets is practically defined by ChEMBL and highly focused in the upper right section of the plot, meaning that the biologically relevant space does not cover evenly the available chemical space. It is also shown that NPs cover, in a sparser manner, the same space as ChEMBL. Unique compounds from BIOFACQUIM on the other hand are distributed only in the less populated regions of the space, and even in the outer region of the plot, which implies the presence of few similar compounds in the other datasets. All these observations are equally applicable for the drug-like subsets from the original data sets, which represent 44.3%, 48.4% and 69.1% of BIOFACQUIM, NPs, and ChEMBL, respectively.

Global diversity

In order to compare the chemical diversity of the current version of BIOFACQUIM with the previous one and the two datasets selected as reference, we employed a CD Plot. Figure 5 shows the plot comparing the diversity of all datasets considering four different criteria: scaffolds in the *y*-axis, molecular fingerprints

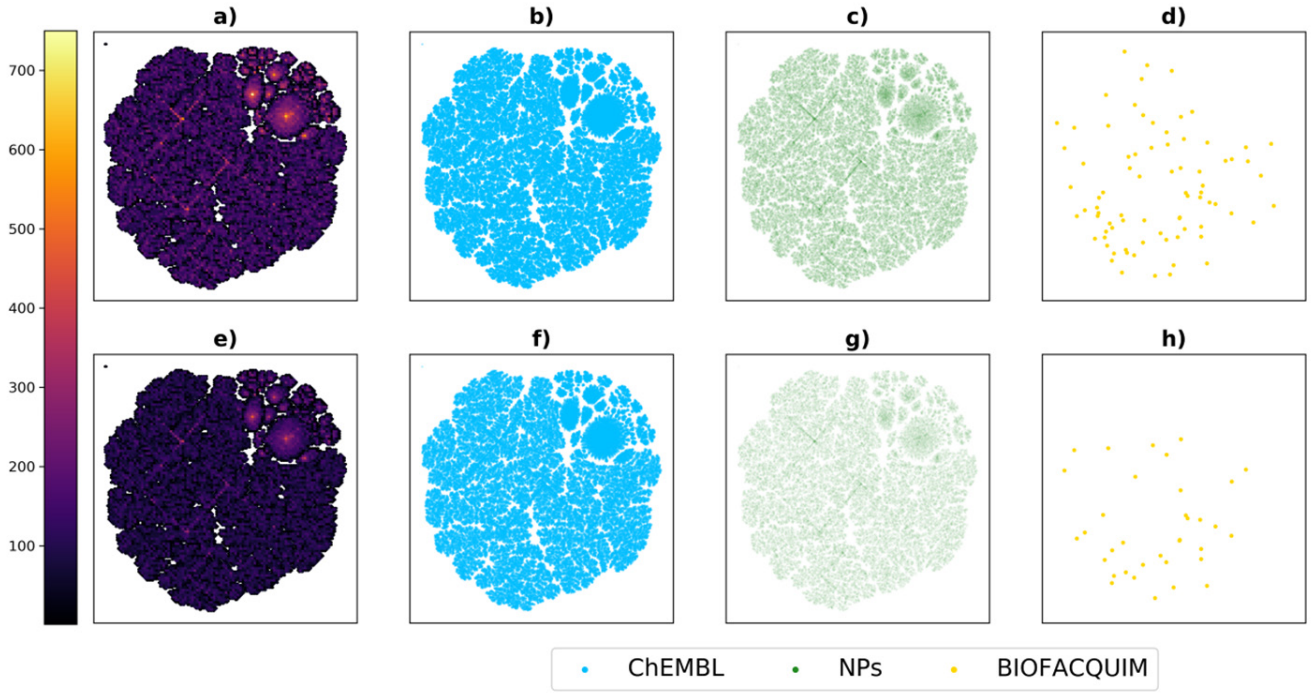

Figure 4. TMAP visualization of the chemical space covered by BIOFACQUIM. Comparison of BIOFACQUIM with two reference datasets. Panels **a**–**d** show all compounds in the datasets, panels **e**–**h** show drug-like compounds only. Panels **a** and **f** show the distribution of compounds from the three data sets among the chemical space in a continuous color scale.

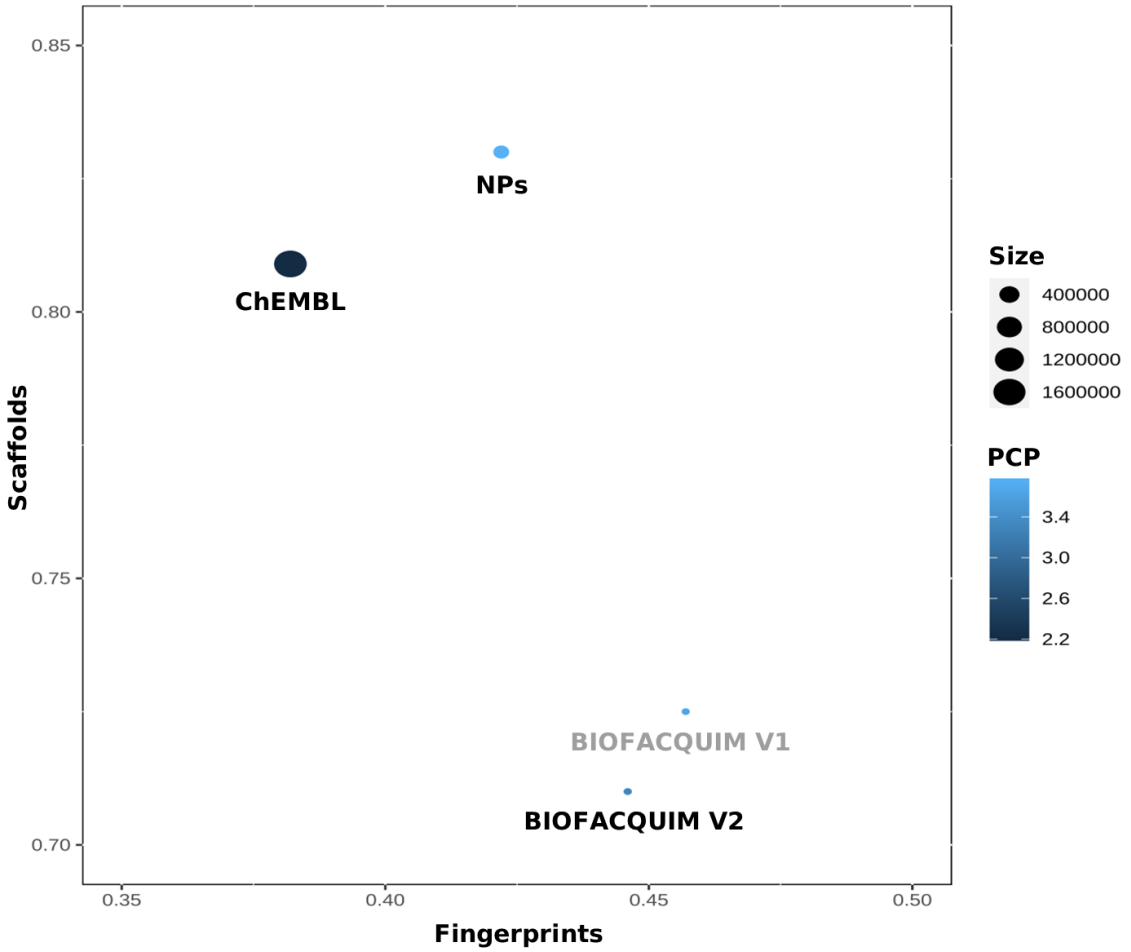

Figure 5. Consensus diversity plot of BIOFACQUIM.

in the *x*-axis, physicochemical properties as the filling of the data points in a continuous color scale, and number of compounds as the data points size. This comparison shows the relatively small size of BIOFACQUIM in comparison with the reference datasets. As compared to the previous release of BIOFACQUIM, the current version has increased its diversity in terms of scaffolds and fingerprints but decreased in terms of physicochemical properties. Also, it is shown that its diversity in terms of molecular fingerprints and physicochemical properties, although not the greatest ones of the three datasets, are closer to the ones for NPs, contrary to scaffolds, in which is the most diverse.

The molecular fingerprint diversity of each data set is represented on the *x*-axis and was defined as the median Tanimoto coefficient of MACCS keys (166-bits) fingerprint. The scaffold diversity of each database is represented on the *y*-axis and was defined as the area under the corresponding cyclic system retrieval curve. The diversity based on physicochemical properties (PCP) was defined as the mean euclidean distance of six scaled physicochemical properties (SlogP, TPSA, AMW, RB, HBD, and HBA) and is shown as the filling of the data points using a continuous color scale. The number of compounds is represented by the size of the data points.

Conclusions

The current version of BIOFACQUIM involved the addition of 148 natural products. This was reflected in a diversity increase based on both scaffolds and molecular fingerprints. It was shown that in terms of diversity, structural content overlap and complexity, BIOFACQUIM is more similar to the assembled set of natural products than to the set of biologically tested compounds. The herein reported chemoinformatic study revealed that 44.3% of the unique compounds contained in BIOFACQUIM are focused in the drug-like space in terms of physicochemical properties. Interestingly, despite the fact of its relative small size, there were identified a significant number of compounds and scaffolds (79 and 29, respectively) that were not present in the

two large sets used as reference, showing that curated databases of natural products, such as BIOFACQUIM, can serve as a starting point for the study and increase of the biologically relevant chemical space.

Data availability

Underlying data

Figshare: BIOFACQUIM_V2. http://doi.org/10.6084/m9.figshare.11312702

This file contains the chemical structures of 531 compounds in SDF format, alongside ID number, compound name, simplified molecular input line entry system, literature reference, kingdom, genus, species, geographical location and biological activity.

Extended data

Figshare: Supporting information for "Functional group and diversity analysis of BIOFACQUIM: A Mexican natural product database". http://doi.org/10.6084/m9.figshare.11312735

This project contains the following extended data:

- Supplementary File 1. File with summary results of the functional group analysis.

Acknowledgments

We thank Roberto Hernández for his valuable contribution toward the update of BIOFACQUIM.

References

1. Kinghorn AD, Falk H, Gibbons S, *et al.*: eds. **Progress in the Chemistry of Organic Natural Products 110: Cheminformatics in Natural Product Research.** Cham: Springer International Publishing. 2019; **110**.
Publisher Full Text
2. Medina-Franco JL: **New Approaches for the Discovery of Pharmacologically-Active Natural Compounds.** *Biomolecules.* 2019; **9**(3): 115.
PubMed Abstract | Publisher Full Text | Free Full Text
3. Chen Y, Garcia de Lomana M, Friedrich NO, *et al.*: **Characterization of the Chemical Space of Known and Readily Obtainable Natural Products.** *J Chem Inf Model.* 2018; **58**(8): 1518–1532.
PubMed Abstract | Publisher Full Text
4. Pilón-Jiménez BA, Saldívar-González FI, Díaz-Eufracio BI, *et al.*: **BIOFACQUIM: A Mexican Compound Database of Natural Products.** *Biomolecules.* 2019; **9**(1): 31.
PubMed Abstract | Publisher Full Text | Free Full Text
5. Ertl P: **An Algorithm to Identify Functional Groups in Organic Molecules.** *J Cheminform.* 2017; **9**(1): 36.
PubMed Abstract | Publisher Full Text | Free Full Text
6. Ertl P, Schuhmann T: **A Systematic Cheminformatics Analysis of Functional Groups Occurring in Natural Products.** *J Nat Prod.* 2019; **82**(5): 1258–1263.
PubMed Abstract | Publisher Full Text
7. Gaulton A, Hersey A, Nowotka M, *et al.*: **The ChEMBL database in 2017.** *Nucleic Acids Res.* 2017; **45**(D1): D945–D954.
PubMed Abstract | Publisher Full Text | Free Full Text
8. Gu J, Gui Y, Chen L, *et al.*: **Use of Natural Products as Chemical Library for Drug Discovery and Network Pharmacology.** *PLoS One.* 2013; **8**(4): e62839.
PubMed Abstract | Publisher Full Text | Free Full Text
9. van Santen JA, Jacob G, Singh AL, *et al.*: **The Natural Products Atlas: An Open Access Knowledge Base for Microbial Natural Products Discovery.** *ACS Cent Sci.* 2019; **5**(11): 1824–1833.
PubMed Abstract | Publisher Full Text | Free Full Text
10. Ming H, Tiejun C, Yanli W, *et al.*: **Web Search and Data Mining of Natural Products and Their Bioactivities in PubChem.** *Sci China Chem.* 2013; **56**(10): 1424–1435.
PubMed Abstract | Publisher Full Text | Free Full Text
11. Bemis GW, Murcko MA: **The properties of known drugs. 1. Molecular frameworks.** *J Med Chem.* 1996; **39**(15): 2887–2893.
PubMed Abstract | Publisher Full Text
12. Lovering F, Bikker J, Humblet C: **Escape From Flatland: Increasing Saturation as an Approach to Improving Clinical Success.** *J Med Chem.* 2009; **52**(21): 6752–6756.
PubMed Abstract | Publisher Full Text
13. Clemons PA, Bodycombe NE, Carrinski HA, *et al.*: **Small Molecules of Different Origins Have Distinct Distributions of Structural Complexity That Correlate With Protein-Binding Profiles.** *Proc Natl Acad Sci.* 2010; **107**(44): 18787–18792.
PubMed Abstract | Publisher Full Text | Free Full Text
14. Probst D, Reymond JL: **Visualization of Very Large High-Dimensional Data Sets as Minimum Spanning Trees.** *J Cheminform.* 2020; **12**(1): 12.
Publisher Full Text
15. van der Maaten L, Hinton G: **Visualizing Data Using T-SNE.** *J Mach Learn Res.* 2008; **9**: 2579–2605.
Reference Source
16. McInnes L, Healy J, Melville J: **UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction**. 2018; arXiv:1802.03426v2.
Reference Source
17. Chimani M, Gutwenger C, Junger M, *et al.*: **The Open Graph Drawing Framework (OGDF).** *Handb Graph Draw Vis.* 2013; 543–570.
Reference Source
18. Hunter JD: **Matplotlib: A 2D Graphics Environment.** *Comput Sci Eng.* 2007; **9**(3): 90–95.
Publisher Full Text
19. González-Medina M, Prieto-Martínez FD, Owen JR, *et al.*: **Consensus Diversity Plots: a Global Diversity Analysis of Chemical Libraries.** *J Cheminform.* 2016; **8**: 63.
PubMed Abstract | Publisher Full Text | Free Full Text
20. Pereda-Miranda R, Fragoso-Serrano M, Escalante-Sanchez E, *et al.*: **Profiling of the Resin Glycoside Content of Mexican Jalap Roots with Purgative Activity.** *J Nat Prod.* 2006; **69**(10): 1460–1466.
PubMed Abstract | Publisher Full Text
21. Pérez-Vásquez A, Castillejos-Ramírez E, Cristians S, *et al.*: **Development of a UHPLC-PDA Method for the Simultaneous Quantification of 4-phenylcoumarins and Chlorogenic Acid in *Exostema Caribaeum* Stem Bark.** *J Nat Prod.* 2014; **77**(3): 516–520.
PubMed Abstract | Publisher Full Text
22. Rojas-Flores C, Rios MY, López-Marure R, *et al.*: **Karwinaphthopyranones From the Fruits of *Karwinskia Parvifolia* and Their Cytotoxic Activities.** *J Nat Prod.* 2014; **77**(11): 2404–2409.
PubMed Abstract | Publisher Full Text
23. Rosas-Ramírez D, Pereda-Miranda R: **Batatins VIII-XI, Glycolipid Ester-Type Dimers From *Ipomoea Batatas*.** *J Nat Prod.* 2015; **78**(1): 26–33.
PubMed Abstract | Publisher Full Text
24. Lipinski CA, Lombardo F, Dominy BW, *et al.*: **Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings.** *Adv Drug Deliv Rev.* 2001; **46**(1–3): 3–26.
PubMed Abstract | Publisher Full Text
25. Veber DF, Johnson SR, Cheng HY, *et al.*: **Molecular Properties That Influence the Oral Bioavailability of Drug Candidates.** *J Med Chem.* 2002; **45**(12): 2615–2623.
PubMed Abstract | Publisher Full Text

Open Peer Review

Current Peer Review Status: ✔ ✔ ✔

Version 2

Reviewer Report 11 June 2020

https://doi.org/10.5256/f1000research.27062.r64382

✔ **Johannes Kirchmair**
Department of Pharmaceutical Chemistry, University of Vienna, Vienna, Austria

I have no further comments or requests.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: cheminformatics, natural products research

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 30 January 2020

https://doi.org/10.5256/f1000research.23733.r58743

✔ **Trong Tran**
GeneCology Research Centre, University of the Sunshine Coast, Maroochydore, Qld, Australia

The manuscript by Sánchez-Cruz *et al.* describes the cheminformatic analysis of the updated BIOFACQUIM database consisting of 531 Mexican natural products. Despite its relatively small size, the results supported the potential of using BIOFACQUIM as a good source for virtual screening in drug discovery.

Generally, the study was well designed, and the paper was well-written. However, since stereochemistry plays a significant role in the interactions between small molecules and protein targets, I suggest the authors should add the analysis of the fraction of sp^3-hydridized atoms for the new BIOFACQUIM version and compare that value with those of other databases to assess its 3D unique. Can authors detail the “eight indexed journals” which was referred in the “Methods - Databases and data curation”?
All in all, I feel this manuscript is suitable for indexing.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Natural product chemistry, Drug Discovery.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 01 Jun 2020

José L. Medina-Franco, National Autonomous University of Mexico, Mexico City, Mexico

The manuscript by Sánchez-Cruz *et al.* describes the cheminformatic analysis of the updated BIOFACQUIM database consisting of 531 Mexican natural products. Despite its relatively small size, the results supported the potential of using BIOFACQUIM as a good source for virtual screening in drug discovery.
Generally, the study was well designed, and the paper was well-written. However, since stereochemistry plays a significant role in the interactions between small molecules and protein targets, I suggest the authors should add the analysis of the fraction of sp - hydridized atoms for the new BIOFACQUIM version and compare that value with those of other databases to assess its 3D unique.
RESPONSE: We thank the reviewer for the suggestion. In the revised “Methods” and “Results

and discussions", we added a new section "Complexity analysis" to compare the distribution of the fraction of carbon atoms that are sp3 hybridized. A new figure (3 in the revised version) was added.

Can authors detail the "eight indexed journals" which was referred in the "Methods - Databases and data curation"?
RESPONSE: The names of the 8 journals were added: "Journal of Ethnopharmacology, Natural Products Research, Journal of Agricultural and Food Chemistry, Journal of Natural Products, Planta Medica, Phytochemistry, Natural Product Letters, and Molecules."

Competing Interests: No competing interests were disclosed.

Reviewer Report 24 January 2020

https://doi.org/10.5256/f1000research.23733.r57650

W. Patrick Walters
Relay Therapeutics, Cambridge, MA, USA

This paper describes an analysis of the BIOFACQUIM database of natural products derived from Mexico. The paper is well written and provides a useful overview of a resource that will be of use to medicinal and computational chemists, as well as those involved in research on natural products. The authors should be commended for making their database and associated code available as Open Source.

While the paper is well written, there are a few areas where changes in the text may make the presentation a bit more clear:

- The second sentence in the abstract could be changed to read: "An analysis of its structural content, as well as functional group occurrence, provides a useful overview, as well as a means of comparison with related databases."

- In the first sentence of the "Methods" section, the authors may want to consider changing "done" to "performed".

- In the Introduction, the sentence "As part of that work, it was reported an initial scaffold content, and chemical space diversity, and coverage" could be changed to "As part of that work, scaffold content and chemical space diversity were examined".

- In the first paragraph of the "Results and discussion" section, "place of recollection" would be better stated as "place of collection".

In the methods section, the authors detail their procedure for “cleaning” the databases. It would be useful to others if the source code for these procedures was made available as part of the GitHub repository associated with this paper.
The authors used canonical SMILES to identify exact matches between compounds in different databases. While this is a reasonable approach, canonical SMILES do not standardize tautomers and may miss some duplicates. An alternate approach would be to use InChI keys, which canonicalize tautomeric forms.

The description of the “Consensus Diversity (CD) Plot” in the “Global diversity” section is a bit vague. An additional paragraph describing this method would be useful.
The density of points in Figure 3 obscures many of the points in the plot. The authors may want to consider using a plotting technique like hexagon binning which enables the plotting of large datasets without obscuring points. For example, see these links:

- https://datavizproject.com/data-type/hexagonal-binning/
- https://rdrr.io/cran/som.nn/man/hexbinpie.html

I’m not convinced that plotting molecular weight vs LogP is the best way to represent the diversity and overlap of datasets. An alternative might be to use a technique like t-SNE to project a set of fingerprints into a lower dimensional space: https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html.

It might also be interesting to provide another plot which only shows the molecules from the three datasets studied that are in “drug-like” chemical space rather than the very large range shown in Figure 3.

In summary, this paper provides a valuable resource to researchers working in a number of different areas. Hopefully, these suggestions will provide minor enhancements to the work.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Computational Chemistry, Cheminformatics, Drug Design

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 01 Jun 2020

José L. Medina-Franco, National Autonomous University of Mexico, Mexico City, Mexico

This paper describes an analysis of the BIOFACQUIM database of natural products derived from Mexico. The paper is well written and provides a useful overview of a resource that will be of use to medicinal and computational chemists, as well as those involved in research on natural products. The authors should be commended for making their database and associated code available as Open Source. While the paper is well written, there are a few areas where changes in the text may make the presentation a bit more clear:
RESPONSE: Thank you for the feedback and constructive comments. In the second version of the manuscript we incorporated all your suggestions as described hereunder.

- The second sentence in the abstract could be changed to read: “An analysis of its structural content, as well as functional group occurrence, provides a useful overview, as well as a means of comparison with related databases."
- In the first sentence of the “Methods” section, the authors may want to consider changing “done” to “performed”.
- In the Introduction, the sentence “As part of that work, it was reported an initial scaffold content, and chemical space diversity, and coverage” could be changed to “As part of that work, scaffold content and chemical space diversity were examined”.
- In the first paragraph of the “Results and discussion” section, “place of recollection” would be better stated as “place of collection”.

RESPONSE: All four changes in the text were done.

In the methods section, the authors detail their procedure for “cleaning” the databases. It would be useful to others if the source code for these procedures was made available as part of the GitHub repository associated with this paper.
RESPONSE: In agreement with the suggestion, the code to prepare the databases was made freely available. In the revised version of the manuscript “Methods, Databases and data curation section” we included the sentence: “The code is available at GitHub (https://github.com/DIFACQUIM/IFG_General).”

The authors used canonical SMILES to identify exact matches between compounds in different databases. While this is a reasonable approach, canonical SMILES do not standardize tautomers and may miss some duplicates. An alternate approach would be to use InChI keys, which canonicalize tautomeric forms.
RESPONSE: To address this point we repeated the analysis standardizing the structures using MolVS that does consider tautomers. The standardization procedure was described in detail in the revised “Methods, Databases and curation” section. All the quantities associated with the new and correct number of unique molecules were revised and the main text and

figures were updated accordingly.

The description of the “Consensus Diversity (CD) Plot” in the “Global diversity” section is a bit vague. An additional paragraph describing this method would be useful.
RESPONSE: We described in more detail the Consensus Diversity (CD) Plots.

The density of points in Figure 3 obscures many of the points in the plot. The authors may want to consider using a plotting technique like hexagon binning which enables the plotting of large datasets without obscuring points. For example, see these links:
https://datavizproject.com/data-type/hexagonal-binning/
https://rdrr.io/cran/som.nn/man/hexbinpie.html
RESPONSE: To address the reviewer’s point, we used the recently published visualization method TMAP (Tree Manifold Approximation and Projection) (J. Cheminform. 2020, 12, 12.). This approach is suited for visualization of very large data sets (up to millions of data points with high dimensionality). Figure 3 was replaced with a TMAP where the 3 data sets are shown in separate panels and a heatmap in the form of hexbin plot for the distribution of compounds is included (Figure 4 in the revised manuscript). A description of TMAP was added to the Methods section citing the original work and the one generated is discussed accordingly in the Results and Discussion section.

I’m not convinced that plotting molecular weight vs LogP is the best way to represent the diversity and overlap of datasets. An alternative might be to use a technique like t-SNE to project a set of fingerprints into a lower dimensional space: https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html.
RESPONSE: Following the recommendation, we now compare the three data sets using TMAPs. For the description of compounds, we selected Morgan fingerprints with radius 2.

It might also be interesting to provide another plot which only shows the molecules from the three datasets studied that are in “drug-like” chemical space rather than the very large range shown in Figure 3.
RESPONSE: In agreement with the suggestion, in the newly added TMAPs (Figure 4 in the revised manuscript) we included a visualization of drug-like subsets of the three databases.

Competing Interests: No competing interests were disclosed.

Reviewer Report 13 December 2019

https://doi.org/10.5256/f1000research.23733.r57653

?

Johannes Kirchmair

Department of Pharmaceutical Chemistry, University of Vienna, Vienna, Austria

Sánchez-Cruz et al. report on the diversity of an extended version of BIOFACQUIM, a database of natural products extracted from organisms endemic to Mexico.

The overall scientific approach is sound. The only technical issue that I believe requires attention is tautomerism, which appears to have been neglected. Consideration of tautomerism will unlikely change the outcomes and conclusions of the statistical analysis but for the functional group analysis it will be important to understand whether the five new functional groups from BIOFACQUIM presented in Figure 2 are indeed unique to that source. Could the authors please use InChI, MolVS or a similar notation/approach for standardizing the molecular (sub-) structures and conduct a tautomer-invariant identification of functional groups?

Language issues are apparent throughout the manuscript and careful proofreading will be required. Some (parts of) statements are unclear, e.g.
"it was reported an initial scaffold content"
"the articles should have described the procedure"
"decided to augment the number of Mexican institutions"
"A similar procedure can be done using RDKit"

Minor comment: Could the authors please clarify in the manuscript the sources of the natural products included in BIOFACQUIM (e.g. plants, bacteria, fungi)?

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: cheminformatics, natural products research

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have

significant reservations, as outlined above.

Author Response 01 Jun 2020

José L. Medina-Franco, National Autonomous University of Mexico, Mexico City, Mexico

The overall scientific approach is sound. The only technical issue that I believe requires attention is tautomerism, which appears to have been neglected. Consideration of tautomerism will unlikely change the outcomes and conclusions of the statistical analysis but for the functional group analysis it will be important to understand whether the five new functional groups from BIOFACQUIM presented in Figure 2 are indeed unique to that source. Could the authors please use InChI, MolVS or a similar notation/approach for standardizing the molecular (sub-) structures and conduct a tautomer-invariant identification of functional groups?
RESPONSE: Following the recommendation, we repeated the analysis reported in the manuscript standardizing the structures of all databases using MolVS. The standardization procedure was described in detail in the revised "Methods, Databases and curation" section and the code was made available at GitHub (https://github.com/DIFACQUIM/IFG_General). All the quantities associated with the new and correct number of unique molecules were revised and the main text and figures were updated accordingly (including figure 2).

Language issues are apparent throughout the manuscript and careful proofreading will be required. Some (parts of) statements are unclear, e.g.
"it was reported an initial scaffold content"
"the articles should have described the procedure"
"decided to augment the number of Mexican institutions"
"A similar procedure can be done using RDKit"
RESPONSE: The statements pointed out were fixed. We also proofread carefully the second version of the manuscript.

Minor comment: Could the authors please clarify in the manuscript the sources of the natural products included in BIOFACQUIM (e.g. plants, bacteria, fungi)?
RESPONSE: In the revised "Results and discussion, Update of BIOFACQUIM section" we included the number of compounds from each source ("...406 from plants, 97 from fungus, 15 from propolis and 13 from marine animals....").

Competing Interests: No competing interests were disclosed.