RESEARCH ARTICLE

REVISED Analysis of a large food chemical database: chemical space, diversity, and complexity [version 2; referees: 3 approved]

J. Jesús Naveja [1,2], Mariel P. Rico-Hidalgo [2], José L. Medina-Franco [2]

[1]PECEM, Faculty of Medicine, Universidad Nacional Autónoma de México, Mexico City, 04510, Mexico
[2]Department of Pharmacy, School of Chemistry, Universidad Nacional Autónoma de México, Mexico City, 04510, Mexico

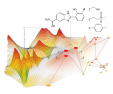
v2 **First published:** 03 Jul 2018, **7**(Chem Inf Sci):993 (doi: 10.12688/f1000research.15440.1)
Latest published: 10 Aug 2018, **7**(Chem Inf Sci):993 (doi: 10.12688/f1000research.15440.2)

Abstract

Background: Food chemicals are a cornerstone in the food industry. However, its chemical diversity has been explored on a limited basis, for instance, previous analysis of food-related databases were done up to 2,200 molecules. The goal of this work was to quantify the chemical diversity of chemical compounds stored in FooDB, a database with nearly 24,000 food chemicals.
Methods: The visual representation of the chemical space of FooDB was done with ChemMaps, a novel approach based on the concept of chemical satellites. The large food chemical database was profiled based on physicochemical properties, molecular complexity and scaffold content. The global diversity of FooDB was characterized using Consensus Diversity Plots.
Results: It was found that compounds in FooDB are very diverse in terms of properties and structure, with a large structural complexity. It was also found that one third of the food chemicals are acyclic molecules and ring-containing molecules are mostly monocyclic, with several scaffolds common to natural products in other databases.
Conclusions: To the best of our knowledge, this is the first analysis of the chemical diversity and complexity of FooDB. This study represents a step further to the emerging field of "Food Informatics". Future study should compare directly the chemical structures of the molecules in FooDB with other compound databases, for instance, drug-like databases and natural products collections. An additional future direction of this work is to use the list of 3,228 polyphenolic compounds identified in this work to enhance the on-going polyphenol-protein interactome studies.

Keywords

ChemMaps, chemical space, chemoinformatics, consensus diversity plots, diversity, FooDB, Foodinformatics, in silico



This article is included in the Chemical Information Science gateway.

Open Peer Review

Referee Status: ✓ ✓ ✓

	Invited Referees 1	2	3
REVISED version 2 published 10 Aug 2018	✓ report		
version 1 published 03 Jul 2018	? report	✓ report	✓ report

1 **Piotr Minkiewicz**, University of Warmia and Mazury in Olsztyn, Poland

2 **Khushbu Shah**, Duquesne University, USA
Kramer Levin Naftalis Frankel LLP, USA

3 **Rachelle J. Bienstock**, RJB Computational Modeling LLC, USA

Discuss this article

Comments (0)

Corresponding author: José L. Medina-Franco (medinajl@unam.mx)

Author roles: Naveja JJ: Conceptualization, Formal Analysis, Investigation, Methodology, Writing – Original Draft Preparation, Writing – Review & Editing; **Rico-Hidalgo MP**: Formal Analysis, Investigation, Writing – Original Draft Preparation, Writing – Review & Editing; **Medina-Franco JL**: Conceptualization, Methodology, Project Administration, Resources, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by a Consejo Nacional de Tecnología (CONACyT) scholarship [622969] (JJN). Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT) Grant [IA203018] from the Universidad Nacional Autónoma de México (JLMF).
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: Naveja JJ, Rico-Hidalgo MP and Medina-Franco JL. **Analysis of a large food chemical database: chemical space, diversity, and complexity [version 2; referees: 3 approved]** *F1000Research* 2018, **7**(Chem Inf Sci):993 (doi: 10.12688/f1000research.15440.2)

First published: 03 Jul 2018, **7**(Chem Inf Sci):993 (doi: 10.12688/f1000research.15440.1)

REVISED **Amendments from Version 1**

We thank the reviewers for the valuable comments and suggestions. We addressed all the comments of Piotr Minkiewicz emphasizing on the novelty, implications and future directions of this work. In the revised version of the manuscript the three suggested references were added and discussed accordingly. It is now mentioned that the findings of this work agree with the results of Lacroix S. *et al.* and the list of polyphenolic compounds made available in this work can further complement the works of Jensen K. *et al.* (2014 and 2015). In the revised manuscript we also acknowledged the optional suggestions of Khushbu Shah. The rationale behind the selection of the three version of the data sets was added. It was also acknowledged as a future work, the suggestion of conducting a systematic analysis of the functional groups in the acyclic compounds of FooDB.

See referee reports

Introduction

Despite the high relevance of food chemicals in many areas including nutrition, disease prevention, and broad impact in the food industry, the chemical space and diversity of food chemical databases (Minkiewicz *et al.*, 2016) has been quantified on a limited basis. Previous efforts include the analysis and comparison of about 2,200 Generally Recognized as Safe (GRAS) flavoring substances (discrete chemical entities only) with compound databases relevant in drug discovery and natural product research e.g., drugs approved for clinical use, compounds in the ZINC database, and natural products from different sources (Burdock & Carabin, 2004; González-Medina *et al.*, 2016; González-Medina *et al.*, 2017; Martinez-Mayorga *et al.*, 2013; Medina-Franco *et al.*, 2012; Peña-Castillo *et al.*, 2018). Other food-related chemical databases, comprising around 900 compounds, were analyzed by Ruddigkeit and J.-L. Reymond (Ruddigkeit & Reymond, 2014). The limited quantitative analysis of food chemicals has been in part due to the scarce availability of food chemical databases in the public domain. A major exception, however, is FooDB a large database with more than 20,000 food chemicals (The Metabolomics Innovation Centre, 2017). To date, it is the most informative public repository of food compounds.

As part of a continued effort to characterize the chemical contents and diversity of food chemicals (González-Medina *et al.*, 2016; Martinez-Mayorga & Medina-Franco, 2009; Medina-Franco *et al.*, 2012), herein we report a quantitative analysis of the chemical space and chemical diversity of FooDB. Widely characterized compound databases such as GRAS, approved drugs and screening compounds used in drug discovery projects were employed as references. We used well-established and novel (but validated) chemoinformatic methods to analyze compound collections. Although most of these approaches are commonly used in drug discovery, this and previous works show they can be readily applied for food chemicals (Peña-Castillo *et al.*, 2018). Thereby this study represents a contribution to further advance the emerging field of Foodinformatics (Martinez-Mayorga & Medina-Franco, 2014).

Methods

Databases and data curation

Four chemical databases were homogeneously curated and analyzed, namely: FooDB version 1.0 (accessed November, 2017) (The Metabolomics Innovation Centre, 2017), drugs approved for clinical use available in DrugBank 5.0.2. (Law *et al.*, 2014), GRAS (Burdock & Carabin, 2004), and a random subset of drug-like natural products from ZINC 12 (Irwin & Shoichet, 2005), of a size comparable to FooDB. The GRAS and DrugBank sets used in this work also have been used as reference in other comparative studies (Medina-Franco *et al.*, 2012). The random set from ZINC was employed just as reference and other random sets from ZINC could be used. Compounds from all databases were washed and prepared using Wash MOE 2017 node in KNIME version 3.5.3 (Berthold *et al.*, 2008). Briefly, the washing protocol implemented in MOE included removing salts and neutralizing the charges in the molecules. The largest fragments were kept and duplicates in each dataset deleted. Table 1 summarizes the databases and sizes after data preprocessing.

Chemical space visualization

The visual representation was generated with ChemMaps, a novel method for large chemical space visualizations (Naveja & Medina-Franco, 2017). Briefly, ChemMaps is able to generate two- and three-dimensional representations of the chemical space based. It uses as input the pairwise chemical similarity computed using fingerprints data. This approach exploits the 'chemical satellites' concept (Oprea & Gottfries, 2001), i.e., molecules whose similarity to the rest of the molecules in the database yield sufficient information for generating a visualization of the chemical space. Further details of ChemMaps are described elsewhere (Naveja & Medina-Franco, 2017).

Physicochemical properties

Six physicochemical properties (PCP) were calculated with RDKit KNIME nodes version 3.4, namely: SlogP (partition coefficient), TPSA (topological polar surface area), AMW (atomic mass weight), RB (rotatable bonds), HBD (hydrogen bond donors) and HBA (hydrogen bond acceptors). For the analysis reported in this short communication, these properties were selected based on their broadly extended use for cross-comparison

Table 1. Compound databases analyzed in this work.

Database	Size[a]
FooDB	23,883
GRAS	2,244
DrugBank	8,748
Natural products in ZINC (drug-like random subset)	24,000

[a]Number of compounds after data curation

GRAS: Generally Recognized as Safe

of compound databases of biological relevance. However, additional properties can be calculated.

Molecular complexity

Fraction of sp^3 carbons and number of stereocenters were computed for FooDB as measures of structural complexity. Despite the fact that there are several other measures, these two are straightforward to interpret, easy to calculate and are becoming standard to make cross comparisons among databases (Méndez-Lucio & Medina-Franco, 2017). As described in the Results and Discussion section, the computed values for FooDB were compared to literature data already reported for the reference data sets.

Scaffold content

The term "molecular scaffold" is employed to describe the core structure of a molecule (Brown & Jacoby, 2006). Different approaches have been proposed to consistently obtain a molecule's scaffold *in silico*. In this work, scaffolds were generated under the Bemis-Murcko definition using the RDKit nodes available in KNIME (Bemis & Murcko, 1996). Bemis and Murcko define a scaffold as "the union of ring systems and linkers in a molecule", i.e., all side chains of a molecule are removed.

Global diversity

The so-called "global diversity" (or total diversity) of FooDB was assessed and compared to other reference collections using a consensus diversity plot (González-Medina *et al.*, 2016). As described recently, a consensus diversity plot simultaneously represents, in two-dimensions, four diversity criteria: structural (based on pairwise molecular fingerprint similarity values), scaffolds (using Murcko scaffolds computed as described in the Scaffold content section), physicochemical properties (based on the six properties described in Physicochemical properties section), and database size (the number of compounds) (González-Medina *et al.*, 2016). The structural diversity of each data set is represented on the X-axis and was defined as the median Tanimoto coefficient of MACCS keys fingerprints. The scaffold diversity of each database is represented on the Y-axis and was defined as the area under the corresponding scaffold recovery curve, a well-established metric to measure scaffold diversity (Medina-Franco *et al.*, 2009). The diversity based on PCP was defined as the Euclidean distance of six auto-scaled properties (SlogP, TPSA, AMW, RB, HBD, and HBA - *vide supra*) and is shown as the filling of the data points using a continuous color scale. The relative number of compounds in the data set is represented with a different size of the data points (smaller data sets are represented with smaller data points).

Results and discussion

Visual representation of the chemical space

Chemical space of FooDB in comparison with the compounds of the three reference databases is visualized in Figure 1. The figure also shows the individual comparisons of FooDB with GRAS, DrugBank and natural products subset from ZINC, respectively. As shown in Figure 1a, the coverage of chemical space of FooDB is quite large as compared to other datasets. Most GRAS compounds lie within the chemical space framed by FooDB (Figure 1b): indeed, 1,193 compounds (53% of GRAS) are structurally identical between the two databases. Hence, FooDB largely contains and upgrades structural information from GRAS. There is significant overlap with approved drugs (Figure 1c) and natural products from ZINC with FooDB (Figure 1d).

Distribution of physicochemical properties

Figure 2 shows the boxplots for the distribution of PCP in all the four databases. For better visualization, the outliers above or below the median +/- 1.5 interquartile range are omitted. As expected, due to the large structural diversity, distribution of PCP in FooDB is broad, in many cases overcoming even approved drugs. For most properties, except RB, several compounds in FooDB share the properties of drugs, and drug-like natural products in ZINC. The comparable physicochemical properties between compounds from FooDB and DrugBank encourages additional systematic investigations for bioactivity of food components. Of course, during this search one needs to consider that compounds with similar properties may have different activity profile. In turn, GRAS consists mostly of small-sized compounds. Table S1 (Supplementary File 1) summarizes the statistics for FooDB and other reference collections.

Molecular complexity

For FooDB, the fraction of sp^3 carbons (mean: 0.62; standard deviation: 0.28) and the number of stereocenters (mean: 4.7; standard deviation: 7.1) indicated a high structural complexity. For comparison, it has reported that the mean of the fraction of sp^3 carbons for approved drugs, compounds in the clinic and a general screening collections of organic compounds is 0.47, 0.41 and 0.32, respectively (González-Medina *et al.*, 2016; Lovering *et al.*, 2009). Moreover, the reported mean of the fraction of sp3 carbons for natural products collections ranges between 0.41 and 0.58 (for natural products in ZINC and Traditional Chinese Medicine (López-Vallejo *et al.*, 2012). The complexity of compounds in FooDB is comparable to molecules in GRAS (mean: 0.63; standard deviation: 0.28) (González-Medina *et al.*, 2016).

Scaffold content

Figure 3 shows the frequency of the most common scaffolds in FooDB. Many compounds are acyclic (32%), followed by monocyclic compounds with a benzene (6%), cyclohexene (2%) and tetrahydropyran (1%) as a core structure. The benzene ring is the most common core scaffold in chemical databases used in drug discovery (Bemis & Murcko, 1996; Singh *et al.*, 2009; Yongye *et al.*, 2012). Many of the most frequent scaffolds in FooDB are also common in other compound databases of natural products (González-Medina *et al.*, 2017). In a follow-up work, it will be interesting to explore the type of functional groups commonly present in the acyclic structures of FooDB.

Recently, Schneider *et al.* published an analysis on the selectivity of Bemis-Murcko scaffolds based on public bioactivity data available in ChEMBL (Schneider & Schneider, 2017). 78 of the

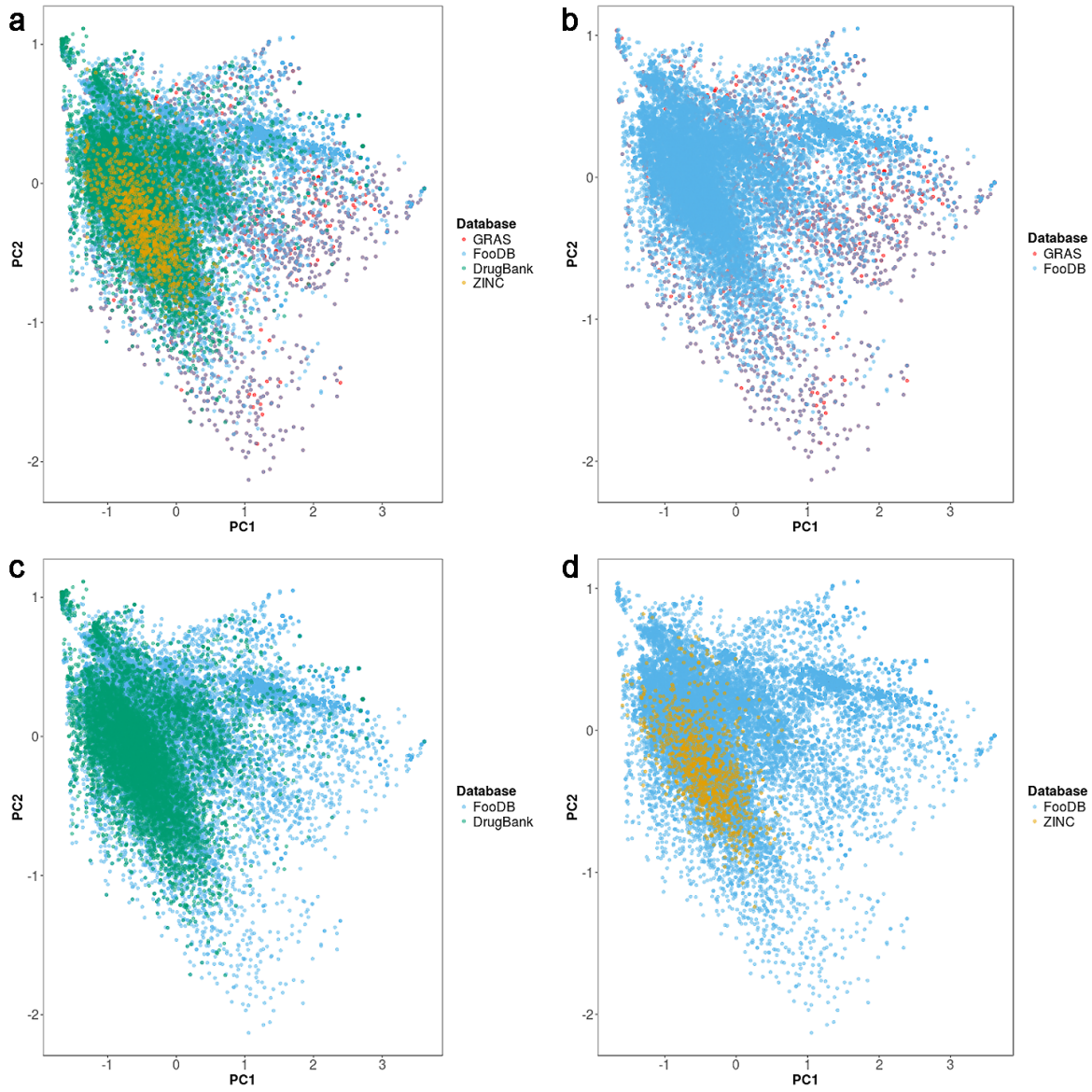


Figure 1. Representation of the chemical space of FooDB. The visual representation was generated with ChemMaps (Naveja & Medina-Franco, 2017). **a**) Comparison of FooDB with three reference collections. Panels **b**–**d**) show comparisons of FooDB with individual data sets.

585 scaffolds reported therein were present in FooDB. The list of the 78 matching scaffolds, along with the original statistics calculated by Schneider *et al.*, is made available as Dataset 1 (Naveja *et al.*, 2018a). Of note, the three most frequent scaffolds in FooDB (benzene, cyclohexane and tetrahydropyran, with more than 300 compounds - Figure 3) are matching scaffolds. Interestingly, the mean *Information content* (I) value of all 585 Schneider's scaffolds is 2.8 (sd= 0.6), while the subset of the 78 scaffolds also present in FooDB has a mean I value of only 2.1 (sd = 0.7). Lower I values point towards more promiscuous scaffolds (Schneider & Schneider, 2017), an expected finding given the nature of the database. As example, Table S2 (Supplementary File 1) shows and discusses briefly the statistics for the three most frequent matching scaffolds.

Polyphenols. Since polyphenols are an important class of compounds in food chemistry (Rasouli *et al.*, 2017), we investigated and quantified the amount of polyphenols in FooDB. Polyphenols are well-known antioxidants, which may play a role in the prevention of several diseases including type 2 diabetes, cardiovascular diseases, and some types of cancer (Neveu *et al.*, 2010). In this line, it is known that oxidative/nitrosative stress

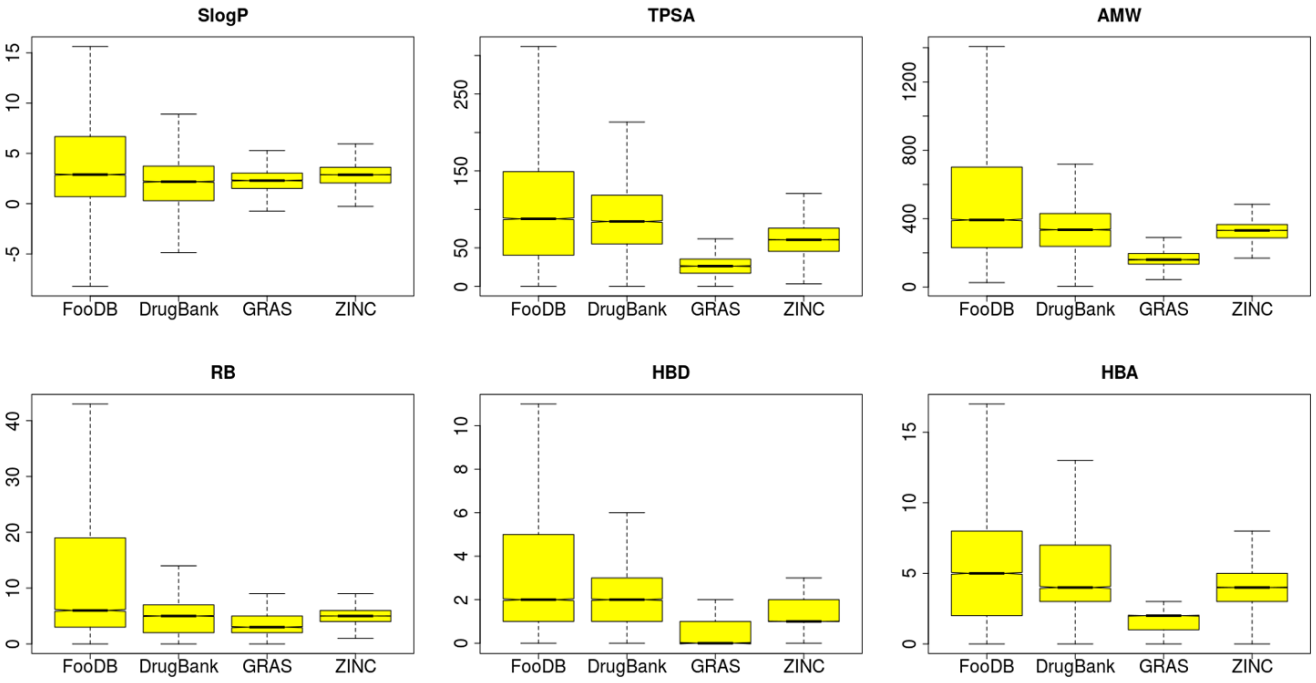


Figure 2. Distribution of physicochemical properties. Box plots of the distribution of six physicochemical properties of FooDB and reference data sets. SlogP (partition coefficient), TPSA (topological polar surface area), AMW (atomic mass weight), RB (rotatable bonds), HBD (hydrogen bond donors) and HBA (hydrogen bond acceptors).

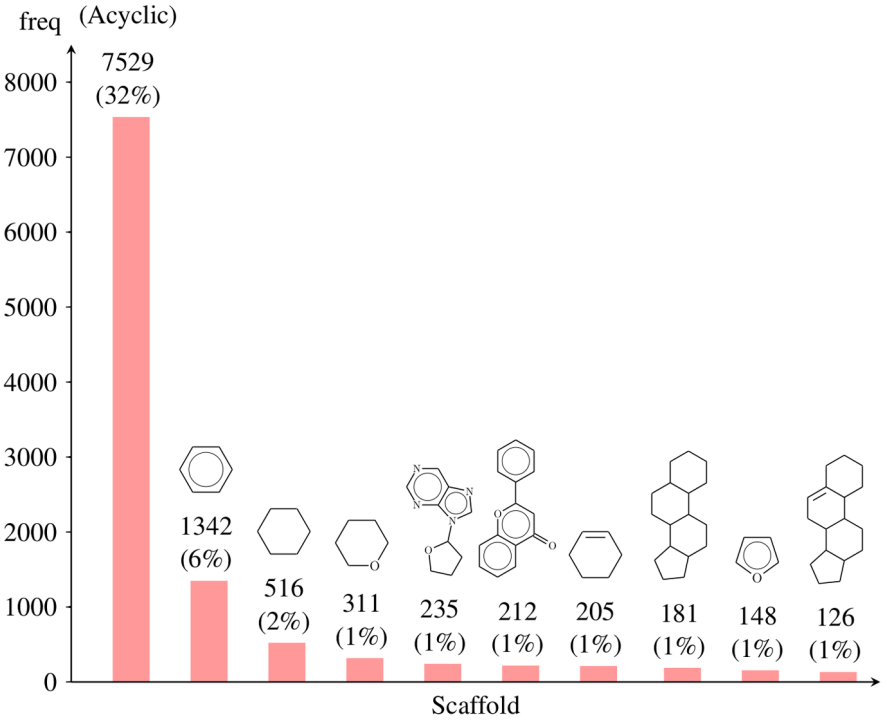


Figure 3. Frequency of the ten most common scaffolds in FooDB.

has a pivotal role in pathophysiology of neurodegenerative disorders and other kinds of disease (Ebrahimi & Schluesener, 2012). Polyphenols have been demonstrated to elicit several biological effects in *in vitro* and *ex vivo* tests (Del Rio *et al.*, 2010; Scalbert *et al.*, 2005).

The molecular structure of polyphenols includes at least two phenolic groups, or one biphenol, and up to any additional number of OH substitutions in aryl rings. They may be classified by their structure in two major groups: flavonoids and non-flavonoids (phenolic acid derivatives) (Del Rio *et al.*, 2013). Some polyphenols, such as quercetin, are found in all plant products, whereas others are specific to particular foods. In many cases, food contain complex mixtures of polyphenols, which are often poorly characterized (Manach *et al.*, 2004).

Polyphenols are also a common chemical motif among natural products, and they are often associated to promiscuity (Tang, 2016). In this work it was found that 3,228 (13.5%) compounds in FooDB are polyphenolic. The list of all 3,228 polyphenolic compounds is made available as Dataset 2 (Naveja *et al.*, 2018b). This set of polyphenols is larger than the 502 polyphenols from food indexed in Phenol-Explorer (Neveu *et al.*, 2010). For comparison, all the reference databases used in this work contained less polyphenols than FooDB. GRAS, ZINC and DrugBank contained 15 (0.6%), 24 (0.1%) and 325 (3.7%) polyphenols, respectively. The large list of polyphenols identified from FooDB is larger than the list of 1,395 polyphenols identified and used in the recent work of Lacroix *et al.* (Lacroix *et al.*, 2018) that was retrieved from Phenol-Explorer and the Dictionary of Natural Products. Indeed, the list of 3,228 polyphenolic compound made available in this work can be used to augment the already extensive polyphenol-protein interactome work of Lacroix *et al.* (Lacroix *et al.*, 2018).

Global diversity

Since the diversity of compound data sets depend on the molecular representation (Sheridan & Kearsley, 2002), a global assessment of the diversity of FooDB was analyzed using different criteria: molecular fingerprints, scaffolds, physicochemical properties and number of compounds. The four criteria were analyzed in an integrated manner through a Consensus Diversity Plot generated as described in the Global diversity section of the Methods. The Consensus Diversity Plot in Figure 4 shows that FooDB has about average diversity both by fingerprints and relatively low diversity by scaffolds.

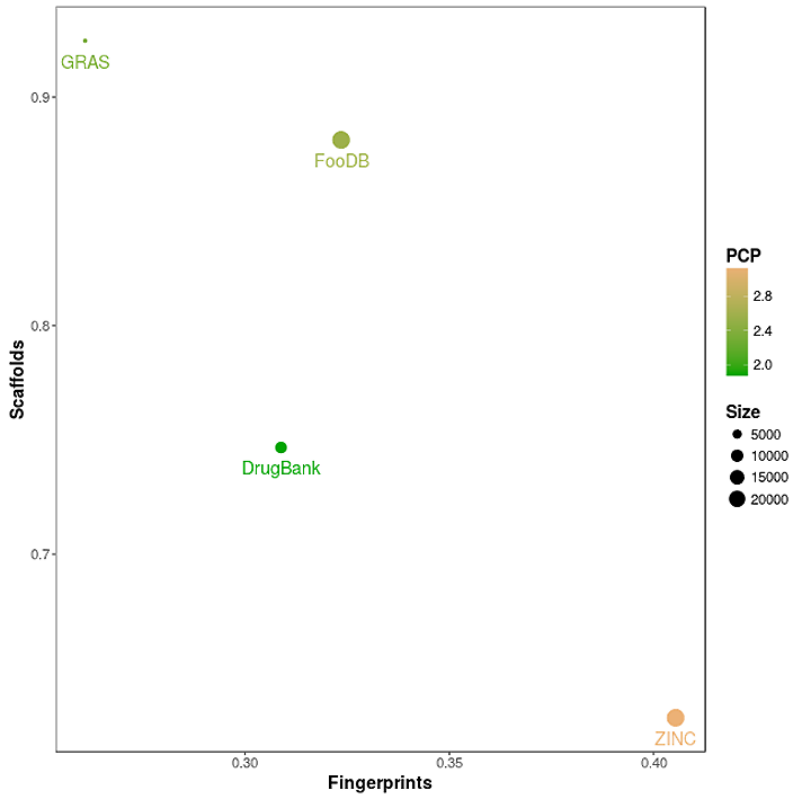


Figure 4. Consensus Diversity Plot of FooDB and reference data sets. The structural diversity of each data set is represented on the X-axis and was defined as the median Tanimoto coefficient of MACCS keys fingerprints. The scaffold diversity of each database is represented on the Y-axis and was defined as the area under the corresponding scaffold recovery curve. The diversity based on physicochemical properties (PCP) was defined as the Euclidean distance of six auto-scaled properties (SlogP, TPSA, AMW, RB, HBD, and HBA) and is shown as the filling of the data points using a continuous color scale. The relative number of compounds is represented with a different size of the data points (smaller data sets are represented with smaller data points).

Although PCP (represented with the color of the data points) are extremely diverse, structural motifs seem to reappear with slight variations. Figure 4 shows the overall large fingerprint and scaffold diversity of approved drugs (e.g., data points towards the lower left region of the plot). Similarly, the relative global diversity of GRAS i.e., high fingerprint diversity but low scaffold diversity (e.g., upper left region of the plot), is consistent with previous comparisons of these compounds with other reference data sets (González-Medina *et al.*, 2016; Medina-Franco *et al.*, 2012).

Dataset 1. Schneidermatch.sdf. This file contains the list of the 78 matching scaffolds in SDF format, along with the original statistics calculated by Schneider *et al.*

http://dx.doi.org/10.5256/f1000research.15440.d209071

No special software is required to open the SDF files. Any commercial or free software capable of reading SDF files will open the data sets supplied

Dataset 2. FooDBpolyphenols.sdf. This file contains 3,228 polyphenolic compounds available in FooDB, in SDF format

http://dx.doi.org/10.5256/f1000research.15440.d209072

No special software is required to open the SDF files. Any commercial or free software capable of reading SDF files will open the data sets supplied

Conclusions

FooDB is a novel, large and diverse library containing information of more than 23,000 compounds found in food. To date, it is the most informative public resource of food compounds. Visual representation of the chemical space revealed that FooDB largely contains and upgrades structural information from GRAS. Indeed, most of GRAS is contained in FooDB. Compounds in FooDB have a large diversity of physicochemical properties. The distributions of most physicochemical properties of FooDB compounds overlap with those of approved drugs and natural products in ZINC. GRAS mostly contains small-sized compounds. The global diversity indicates that FooDB has a large structural diversity as measured by molecular fingerprints, though it has relatively low scaffold diversity. One third of the compounds in FooDB are acyclic. The most frequent cyclic scaffolds are monocyclic. Of note, polyphenols represent a large fraction of FooDB. The list of 3,228 polyphenolic compounds identified in this work to enhance the on-going polyphenol-protein interactome studies. Analysis of the chemical complexity revealed that compounds in FooDB are more complex than approved drugs and natural products and have complexity comparable to GRAS compounds. A next step of this work is to compare the chemical space of FooDB with that of natural products from different sources, e.g., plants, terrestrial, cyanobacteria. A second suggested future study is to perform the virtual screening of FooDB across a range of targets, for instance, the increasingly important epigenetic targets (Naveja & Medina-Franco, 2018). Virtual screening can be done using multiple methods, for instance, using similarity searching. In this case one needs to consider, however, the potential presence of activity cliffs i.e., compounds with similar structure but different activity (Stumpfe *et al.*, 2014). The goal of such study would be to identify systematically dietary components that may be participating in epigenetic regulatory processes (Martinez-Mayorga *et al.*, 2013). These efforts are ongoing in our group and will be reported in due course. Other perspective of this work is integrating the knowledge of FooDB with other large databases with the aim of identifying food-disease associations and food-drug interactions such as the works previously published by Jensen *et al.* (Jensen *et al.*, 2014; Jensen *et al.*, 2015).

Data availability

Dataset 1: (Schneidermatch.sdf). **This file contains the list of the 78 matching scaffolds in SDF format**, along with the original statistics calculated by Schneider *et al.* No special software is required to open the SDF files. Any commercial or free software capable of reading SDF files will open the data sets supplied. 10.5256/f1000research.15440.d209071 (Naveja, *et al.*, 2018a)

Dataset 2: (FooDBpolyphenols.sdf). **This file contains 3,228 polyphenolic compounds available in FooDB, in SDF format**. No special software is required to open the SDF files. Any commercial or free software capable of reading SDF files will open the data sets supplied. 10.5256/f1000research.15440.d209072 (Naveja *et al.*, 2018b)

Competing interests

No competing interests were disclosed.

Grant information

This work was supported by a Consejo Nacional de Tecnología (CONACyT) scholarship [622969] (JJN). Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT) Grant [IA203018] from the Universidad Nacional Autónoma de México (JLMF).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

The authors thank Karina Martínez-Mayorga, Andrea Peña-Castillo and Nicole Trujillo for rich discussions and valuable insights.

Supplementary material

Supplementary File 1: File with supporting tables. Table S1: Summary statistics of the distribution of six PCP of FooDB and other reference collections. Table S2: Selected scaffold statistics as reported by (Schneider & Schneider, 2017).

Click here to access the data.

References

Bemis GW, Murcko MA: **The properties of known drugs. 1. Molecular frameworks.** *J Med Chem.* 1996; **39**(15): 2887–93.
PubMed Abstract | Publisher Full Text

Berthold MR, Cebron N, Dill F, *et al.*: **KNIME: The Konstanz Information Miner.** In: Preisach C, Burkhardt H, Schmidt-Thieme L, Decker R, (Eds.), *Data Analysis, Machine Learning and Applications.* Berlin, Heidelberg: Springer Berlin Heidelberg. 2008; 319–326.
Publisher Full Text

Brown N, Jacoby E: **On scaffolds and hopping in medicinal chemistry.** *Mini Rev Med Chem.* 2006; **6**(11): 1217–29.
PubMed Abstract | Publisher Full Text

Burdock GA, Carabin IG: **Generally Recognized as Safe (GRAS): history and description.** *Toxicol Lett.* 2004; **150**(1): 3–18.
PubMed Abstract | Publisher Full Text

Del Rio D, Costa LG, Lean ME, *et al.*: **Polyphenols and health: what compounds are involved?** *Nutr Metab Cardiovasc Dis.* 2010; **20**(1): 1–6.
PubMed Abstract | Publisher Full Text

Del Rio D, Rodriguez-Mateos A, Spencer JP, *et al.*: **Dietary (poly)phenolics in human health: structures, bioavailability, and evidence of protective effects against chronic diseases.** *Antioxid Redox Signal.* 2013; **18**(14): 1818–92.
PubMed Abstract | Publisher Full Text | Free Full Text

Ebrahimi A, Schluesener H: **Natural polyphenols against neurodegenerative disorders: potentials and pitfalls.** *Ageing Res Rev.* 2012; **11**(2): 329–45.
PubMed Abstract | Publisher Full Text

González-Medina M, Owen JR, El-Elimat T, *et al.*: **Scaffold Diversity of Fungal Metabolites.** *Front Pharmacol.* 2017; **8**: 180.
PubMed Abstract | Publisher Full Text | Free Full Text

González-Medina M, Prieto-Martínez FD, Naveja JJ, *et al.*: **Chemoinformatic expedition of the chemical space of fungal products.** *Future Med Chem.* 2016; **8**(12): 1399–412.
PubMed Abstract | Publisher Full Text | Free Full Text

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

Irwin JJ, Shoichet BK: **ZINC--a free database of commercially available compounds for virtual screening.** *J Chem Inf Model.* 2005; **45**(1): 177–82.
PubMed Abstract | Publisher Full Text | Free Full Text

Jensen K, Panagiotou G, Kouskoumvekaki I: **Integrated text mining and chemoinformatics analysis associates diet to health benefit at molecular level.** *PLoS Comput Biol.* 2014; **10**(1): e1003432.
PubMed Abstract | Publisher Full Text | Free Full Text

Jensen K, Ni Y, Panagiotou G, *et al.*: **Developing a molecular roadmap of drug-food interactions.** *PLoS Comput Biol.* 2015; **11**(2): e1004048.
PubMed Abstract | Publisher Full Text | Free Full Text

Lacroix S, Klicic Badoux J, Scott-Boyer MP, *et al.*: **A computationally driven analysis of the polyphenol-protein interactome.** *Sci Rep.* 2018; **8**(1): 2232.
PubMed Abstract | Publisher Full Text | Free Full Text

Law V, Knox C, Djoumbou Y, *et al.*: **DrugBank 4.0: shedding new light on drug metabolism.** *Nucleic Acids Res.* 2014; **42**(Database issue): D1091–7.
PubMed Abstract | Publisher Full Text | Free Full Text

López-Vallejo F, Giulianotti MA, Houghten RA, *et al.*: **Expanding the medicinally relevant chemical space with compound libraries.** *Drug Discov Today.* 2012; **17**(13–14): 718–26.
PubMed Abstract | Publisher Full Text

Lovering F, Bikker J, Humblet C: **Escape from flatland: increasing saturation as an approach to improving clinical success.** *J Med Chem.* 2009; **52**(21): 6752–6.
PubMed Abstract | Publisher Full Text

Manach C, Scalbert A, Morand C, *et al.*: **Polyphenols: food sources and bioavailability.** *Am J Clin Nutr.* 2004; **79**(5): 727–47.
PubMed Abstract | Publisher Full Text

Martinez-Mayorga K, Medina-Franco JL: **Chemoinformatics-applications in food chemistry.** *Adv Food Nutr Res.* 2009; **58**: 33–56.
PubMed Abstract | Publisher Full Text

Martinez-Mayorga K, Medina-Franco JL: **Foodinformatics: Applications of chemical information to food chemistry.** Springer. 2014;
Publisher Full Text

Martinez-Mayorga K, Peppard TL, López-Vallejo F, *et al.*: **Systematic mining of Generally Recognized as Safe (GRAS) flavor chemicals for bioactive compounds.** *J Agric Food Chem.* 2013; **61**(31): 7507–14.
PubMed Abstract | Publisher Full Text

Medina-Franco JL, Martínez-Mayorga K, Bender A, *et al.*: **Scaffold diversity analysis of compound data sets using an entropy-based measure.** *QSAR Comb Sci.* 2009; **28**(11–12): 1551–1560.
Publisher Full Text

Medina-Franco JL, Martínez-Mayorga K, Peppard TL, *et al.*: **Chemoinformatic analysis of GRAS (Generally Recognized as Safe) flavor chemicals and natural products.** *PLoS One.* 2012; **7**(11): e50798.
PubMed Abstract | Publisher Full Text | Free Full Text

Méndez-Lucio O, Medina-Franco JL: **The many roles of molecular complexity in drug discovery.** *Drug Discov Today.* 2017; **22**(1): 120–126.
PubMed Abstract | Publisher Full Text

Minkiewicz P, Darewicz M, Iwaniak A, *et al.*: **Internet databases of the properties, enzymatic reactions, and metabolism of small molecules-search options and applications in food science.** *Int J Mol Sci.* 2016; **17**(12): pii: E2039.
PubMed Abstract | Publisher Full Text | Free Full Text

Naveja JJ, Medina-Franco JL: **ChemMaps: Towards an approach for visualizing the chemical space based on adaptive satellite compounds [version 2; referees: 3 approved with reservations].** *F1000Res.* 2017; **6**: pii: Chem Inf Sci-1134.
PubMed Abstract | Publisher Full Text | Free Full Text

Naveja JJ, Medina-Franco JL: **Insights from pharmacological similarity of epigenetic targets in epipolypharmacology.** *Drug Discov Today.* 2018; **23**(1): 141–150.
PubMed Abstract | Publisher Full Text

Naveja JJ, Rico-Hidalgo MP, Medina-Franco JL: **Dataset 1 in: Analysis of a large food chemical database: chemical space, diversity, and complexity.** *F1000Research.* 2018a.
http://www.doi.org/10.5256/f1000research.15440.d209071

Naveja JJ, Rico-Hidalgo MP, Medina-Franco JL: **Dataset 2 in : Analysis of a large food chemical database: chemical space, diversity, and complexity.** *F1000Research.* 2018b.
http://www.doi.org/10.5256/f1000research.15440.d209072

Neveu V, Perez-Jiménez J, Vos F, *et al.*: **Phenol-Explorer: an online comprehensive database on polyphenol contents in foods.** *Database (Oxford).* 2010; **2010**: bap024.
PubMed Abstract | Publisher Full Text | Free Full Text

Oprea TI, Gottfries J: **Chemography: the art of navigating in chemical space.** *J Comb Chem.* 2001; **3**(2): 157–166.
PubMed Abstract | Publisher Full Text

Peña-Castillo A, Méndez-Lucio O, Owen JR, *et al.*: **Chemoinformatics in Food Science**. In J. Gasteiger & T. Engel (Eds.), *Chemoinformatics - Volume 2: From Methods to Applications.* Weinheim, Germany: Wiley-VCH. 2018.
Publisher Full Text

Rasouli H, Farzei MH, Khodarahmi R: **Polyphenols and their benefits: A review.** *Int J Food Prop.* 2017; **20**(sup2): 1700–1741.
Publisher Full Text

Ruddigkeit L, Reymond JL: **The chemical space of flavours**. In K. Martinez-Mayorga & J. L. Medina-Franco (Eds.), *Foodinformatics.* Cham: Springer International Publishing. 2014; 83–96.
Publisher Full Text

Scalbert A, Johnson IT, Saltmarsh M: **Polyphenols: antioxidants and beyond.** *Am J Clin Nutr.* 2005; **81**(1 Suppl): 215S–217S.
PubMed Abstract | Publisher Full Text

Schneider P, Schneider G: **Privileged Structures Revisited.** *Angew Chem Int Ed Engl.* 2017; **56**(27): 7971–7974.
PubMed Abstract | Publisher Full Text | Free Full Text

Sheridan RP, Kearsley SK: **Why do we need so many chemical similarity search methods?** *Drug Discov Today.* 2002; **7**(17): 903–911.
PubMed Abstract | Publisher Full Text

Singh N, Guha R, Giulianotti MA, *et al.*: **Chemoinformatic analysis of combinatorial libraries, drugs, natural products, and molecular libraries small molecule repository.** *J Chem Inf Model.* 2009; **49**(4): 1010–1024.
PubMed Abstract | Publisher Full Text | Free Full Text

Stumpfe D, de la Vega De León A, Dimova D, *et al.*: **Advancing the activity cliff concept, part II.** *F1000Res.* 2014; 3: 75.
PubMed Abstract | Publisher Full Text | Free Full Text

Tang GY: **Why Polyphenols have Promiscuous Actions? An Investigation by Chemical Bioinformatics.** *Nat Prod Commun.* 2016; **11**(5): 655–656.
PubMed Abstract

The Metabolomics Innovation Centre: **FooDB (Version 1).** Computer software, Canada: The Metabolomics Innovation Centre. 2017.
Reference Source

Yongye AB, Waddell J, Medina-Franco JL: **Molecular scaffold analysis of natural products databases in the public domain.** *Chem Biol Drug Des.* 2012; **80**(5): 717–724.
PubMed Abstract | Publisher Full Text

Open Peer Review

Current Referee Status: ✔ ✔ ✔

Version 2

Referee Report 14 August 2018

doi:10.5256/f1000research.17367.r37062

✔ **Piotr Minkiewicz**
Department of Food Biochemistry, Faculty of Food Science, University of Warmia and Mazury in Olsztyn, Olsztyn-Kortowo, Poland

I fully approve the recent version of the artcle. I can recommend it as very valuable for readers representing the areas of food science and pharmaceutical science.

Competing Interests: No competing interests were disclosed.

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

Version 1

Referee Report 07 August 2018

doi:10.5256/f1000research.16825.r36226

✔ **Rachelle J. Bienstock**
RJB Computational Modeling LLC, Chapel Hill, NC, USA

The paper on chemical diversity of FooDB compared to several other databases, including GRAS and DrugBank and drug-like natural products fromZINC12, by Naveja, Rico-Hidalgo, and Medina-Franco was an interesting, informative and nicely presented analysis. The figures and graphical presentation of ChemMaps results in particular is very clear. One thing which I think would be interesting for a further study and analysis, (since epigenetics and some other diseases and health implications are mentioned in regards to polyphenols) is an analysis regarding vitamins and other compounds and dietary supplements which have had specific health claims made. ChemMaps analysis of these compounds according to properties in these databases and correlation with biological pathways would be interesting for future 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.

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

Author Response 07 Aug 2018

José L. Medina-Franco, Universidad Nacional Autónoma de México, Mexico

We are grateful for the positive comments and thank the reviewer for the excellent suggestions to expand this work in future studies.

Competing Interests: I have no competing interests.

Referee Report 30 July 2018

doi:10.5256/f1000research.16825.r36288

Khushbu Shah [1,2]
[1] Division of Medicinal Chemistry, Graduate School of Pharmaceutical Sciences, Duquesne University, Pittsburgh, PA, USA
[2] Kramer Levin Naftalis Frankel LLP, New York, NY, USA

This manuscript purports to analyze and disclose the chemical diversity of the FooDB database. It is an interesting study with a logical flow based on appropriate methods.

There a few *optional suggestions* that the authors could adapt in the manuscript:

- It would be advisable for the authors to add the rationale behind selecting the three versions – GRAS, DrugBank and ZINC for data curation.
- Since acyclic compounds represented the most common scaffold in FooDB, the authors could expand upon the types of functional groups commonly observed in these acyclic compounds in FooDB.

- Further, the authors point out that there are more polyphenols in FooDB vs. Phenol-explorer. The authors could include the dataset from Phenol-explorer in a consensus diversity plot (like Figure 4) to clearly represent their reults.

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.

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

Author Response 07 Aug 2018

José L. Medina-Franco, Universidad Nacional Autónoma de México, Mexico

We really appreciate the reviewer´s feedback and value the optional suggestions. In the revised manuscript we added the rationale for selecting the ´specific version of the three data sets. We also included a comment that a systematic analysis of the functional groups present in the acyclic structures is highly relevant. This excellent suggestion, as well as the comparison of the polyphenols in FooDB with those in Phenol-explorer, will be reported in a follow-up study.

Competing Interests: I have no competing interests.

Referee Report 11 July 2018

doi:10.5256/f1000research.16825.r35684

? **Piotr Minkiewicz**
Department of Food Biochemistry, Faculty of Food Science, University of Warmia and Mazury in Olsztyn, Olsztyn-Kortowo, Poland

I have no critical remarks concerning methods, correctness of work. Discussion is also appropriate from the point of view of scientists working in the areas of cheminformatics and/or pharmacology.
I would like to ask some questions concerning relevance of the article for food science.
The analysis performed reveals similarity in structural and physico-chemical features between compounds from FooDB and DrugBank. Does it mean that more detailed studies may reveal similar biological activity (i.e. interactions with the same target) of drugs and bioactive food components.
Are Authors' results consistent with these published in the following articles concerning similarity of effects of drugs and food components?
Jensen K. et al. PLoS Comput Biol, 10, (2014)[1]
Jensen K. et al. PLoS Comput Biol, 11, (2015)[2]
Proteins interacting with polyphenols and described in the following article: Lacroix S. et al. Sci Rep, 8, (2018)[3] are also annotated in DrugBank as drug targets. Is the above finding consistent with the Authors' conclusions?
I would like to ask Authors to add few sentences concerning limitations of the proposed methodology (for instance limitations occurring due to presence of activity cliffs).

References

1. Jensen K, Panagiotou G, Kouskoumvekaki I: Integrated text mining and chemoinformatics analysis associates diet to health benefit at molecular level.*PLoS Comput Biol*. 2014; **10** (1): e1003432 PubMed Abstract | Publisher Full Text
2. Jensen K, Ni Y, Panagiotou G, Kouskoumvekaki I: Developing a molecular roadmap of drug-food interactions.*PLoS Comput Biol*. 2015; **11** (2): e1004048 PubMed Abstract | Publisher Full Text
3. Lacroix S, Klicic Badoux J, Scott-Boyer MP, Parolo S, Matone A, Priami C, Morine MJ, Kaput J, Moco S: A computationally driven analysis of the polyphenol-protein interactome.*Sci Rep*. 2018; **8** (1): 2232 PubMed Abstract | Publisher Full Text

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

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?
Partly

Competing Interests: No competing interests were disclosed.

I have read this submission. I 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 16 Jul 2018

José L. Medina-Franco, Universidad Nacional Autónoma de México, Mexico

Thank the reviewer for critically reading our manuscript and the valuable feedback. Hereunder we provide a point-by-point response to each comment.

Comment: "I have no critical remarks concerning methods, correctness of work. Discussion is also appropriate from the point of view of scientists working in the areas of cheminformatics and/or pharmacology.
I would like to ask some questions concerning relevance of the article for food science.
The analysis performed reveals similarity in structural and physico-chemical features between compounds from FooDB and DrugBank. Does it mean that more detailed studies may reveal similar biological activity (i.e. interactions with the same target) of drugs and bioactive food components."
Response: We agree with the valuable input. Indeed, as the reviewers points out, similar physico-chemical properties between compounds from FoodDB and DrugBank encourages additional systematics investigations for bioactivity of food components. In the revised version of the manuscript, that is under editing and will be uploaded in due course, we will expand the discussion of the manuscript elaborating more on the significance of the work.

Comment: "Are Authors' results consistent with these published in the following articles concerning similarity of effects of drugs and food components?
Jensen K. et al. PLoS Comput Biol, 10, (2014)1
Jensen K. et al. PLoS Comput Biol, 11, (2015)2'
Response: We are grateful to the reviewer for pointing out the two papers of Jensen K. et al. As stated in the manuscript, the goal of this study was to characterize the chemical content, diversity and complexity of the chemical structures of a large and public database of food chemicals. The studies of Jensen et al. are focused on finding food-disease associations and food-drug interactions. Following the reviewers advice, we addressed this comment in the revised manuscript stating that as a Perspective of our current work, the FooDB can be used to further augment the current knowledge of food-disease associations and food-drug interactions. The two suggested references are being added to the revised manuscript.

Comment: "Proteins interacting with polyphenols and described in the following article: Lacroix S. et al. Sci Rep, 8, (2018)3 are also annotated in DrugBank as drug targets. Is the above finding consistent with the Authors' conclusions?"
Response: Thank the reviewer for bring to our attention the work of Lacroix S. et al. Our results are consistent with this study. In particular, the number of polyphenol compounds found in the FooDB is larger than the amount of compounds found in the Phenol-Explorer database. This point is being addressed in section "*3.4.1. Polyphenols"* of the revised manuscript. In the revised manuscript we added the suggested reference. In addition, in the Conclusions section, we are also stating that the set of polyhpenols from FooDB identified in this work can further enrich the on-going efforts of the polyphenol-protein interactome studies such as the one published by Lacroix S. et al.

Comment: "I would like to ask Authors to add few sentences concerning limitations of the proposed methodology (for instance limitations occurring due to presence of activity cliffs)."
Response: Following the reviewers' advice, we added a discussion of the limitations of the methodology addressing the caution that needs to be taken while dealing with activity cliffs. Relevant references to activity cliffs are being added.

Competing Interests: We have no competing interests.