Coconut Online: Collection of Open Natural Products Database

Maria Sorokina (✉ maria.sorokina@uni-jena.de)
University Friedrich-Schiller, Lessing Strasse 8, 07743, Jena, Germany; maria.sorokina@uni-jena.de; ORCID: 0000-0001-9359-7149 https://orcid.org/0000-0001-9359-7149

Peter Merseburger
Friedrich-Schiller-Universitat Jena https://orcid.org/0000-0002-5154-8696

Kohulan Rajan
Friedrich-Schiller-Universitat Jena https://orcid.org/0000-0003-1066-7792

Mehmet Aziz Yirik
Friedrich-Schiller-Universitat Jena https://orcid.org/0000-0001-7520-7215

Christoph Steinbeck
Friedrich-Schiller-Universitat Jena https://orcid.org/0000-0001-6966-0814

Database

Keywords: natural products, database, NoSQL, MongoDB, molecular similarity search, molecular substructure search

Posted Date: September 16th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-75600/v1

Version of Record: A version of this preprint was published on January 10th, 2021. See the published version at https://doi.org/10.1186/s13321-020-00478-9.

Abstract

Natural products (NPs) are small molecules produced by living organisms with potential applications in pharmacology and other industries for their high bioactivities. Over the years a multiplication of thematic NP databases has been observed. However, there is no online resource regrouping all known NPs in just one place, which would greatly simplify NP research and allow computational screening and other *in silico* applications. Here we present the COlleCtion of Open Natural prodUcTs (COCONUT): an aggregated dataset of NPs available in different open sources and a subsequent web interface to browse, search and easily and quickly download NPs. COCONUT web is freely available at https://coconut.naturalproducts.net.

Introduction

Natural products (NPs) have received constant attention from the scientific community due to their relevance in drug discovery, chemical ecology and molecular biology in general. In a recently published review on natural products databases [1] we inventoried over 120 natural products databases that have been published and used in the last 20 years. However, 16% of these are not online anymore, 40% are commercial and the open ones are in general specialized on a particular type of NPs or lack annotations. Among the NP databases that are still online, Super Natural II [2] is the largest one: however, it seems not to be maintained anymore and is mainly composed of compounds that can be purchased. Another recent database, NPAtlas [3], is constantly growing, however, it is focusing on microbial NPs only. Some databases focusing on NPs produced by plants are also available online, such as KnapSack [4], CMAUP [5] or TCM@Taiwan [6]. In addition to these relatively big databases, there is a plethora of smaller, even more specialized NP collections. There is, therefore, a need for a generalistic NP database, that will efficiently compile NP information from various sources, improve its annotation and offer a pleasant user experience. To achieve this, we first assembled the most complete up-to-date COlleCtion of Open Natural ProdUcTs (COCONUT) that we have been continuously curating and annotating. This data is now available to the scientific community as a full-fledged online natural products database, available at https://coconut.naturalproducts.net.

The COCONUT database is free and open to all users and there is no login required to use it. Its web interface allows for diverse searches (e.g. by molecule name, InChI, InChI key, SMILES, drawn structure, molecular formula), advanced search by molecular features, substructure and similarity searches. It also allows the users to download the whole dataset in different formats, together with separate search results. The database can also be queried via a REST API. It is deployed as a Docker container, making it easily portable for hosting other sets of NPs and to be deployed offline.

Construction And Content

Web interface and technical specificities

All COCONUT data is stored with MongoDB, a cross-platform document-oriented NoSQL database program. The smallest unit in MongoDB is a document, composed of key and value pairs that are similar to JSON objects. Documents of the same nature are organized in collections, which are the equivalent of the SQL-based databases tables. MongoDB is particularly adapted to big and complex data, supports multiple indexing, including text indexing allowing enhanced text search in text-indexed fields and contains a wide range of in-build search and analysis functions.

Two major collections are present in the COCONUT database: SourceNaturalProduct, which contains the original NP data collected from the open sources, and UniqueNaturalProduct, the unified and curated collection of NPs. The full version of COCONUT with all the calculated features can be accessed as a MongoDB dump in the Downloads section of the website. Requests for displaying additional crucial features in the web interface and making them searchable through the advanced search interface are welcome via the COCONUT GitHub tracker (see below).

The COCONUT online front-end is developed entirely with React.js [7], a JavaScript library to build responsive and efficient user interfaces. The OpenChemLib library [8] is used to handle the chemical editor for the search functions. The COCONUT back-end, allowing to process the front-end requests and to communicate with the database is written in Kotlin and Java 11 using the Spring framework. The CDK [9] library is used to process chemical information and formats.

COCONUT web interface, back-end and database are entirely Dockerised, allowing a quick and easy deployment on local servers and cloud. All the code, for both front-end and back-end, is available on GitHub (https://github.com/mSorok/NaturalProductsOnline).

Data provenance, model and content

Data present in COCONUT has been retrieved from 53 various data sources and some manually collected NPs from literature, shown in Table 1. In the current COCONUT release (August 2020), there are 426,916 unique “flat” (with no stereochemistry) natural products, and a total of 746626 NPs where stereochemistry has been preserved when available.

Every molecule collected from external sources must pass a quality control and registration procedure, where its structure is checked for size (between 5 and 210 heavy atoms), connectivity (only the biggest connected structure is kept), pseudo-atoms, if implicit and explicit hydrogens are correct, and if the bonds are correct and the valences are conserved. The Kekulé representation is also assigned to the aromatic systems of each compound. Then, NPs from different provenance are unified based on the identity of their InChI keys without stereochemistry. This unification step is performed in this way as in different data sources stereochemistry is not uniformly present and can be represented differently. When available, the original molecular structure with stereochemistry is carefully preserved and can be visualized for each NP entry.

The authors are well aware that different stereoisomers of a compound can have very different biological activity. The procedure described above was a necessary step to create a unified resource out of distributed databases of varying quality. Further curation will gradually improve stereochemical assignments and linkage to original source articles.

Each unique NP is then assigned a unique identifier, composed of the “CNP” prefix and 7 digits. An automatic curation for NP metadata is performed, which comprises the retrieval of its official name, synonyms, cross-references to other major chemical databases. Then, a range of molecular properties, descriptors and fingerprints (full list in Table 2) are computed using the in-build CDK libraries. As the number of the computed properties is quite big (73 fields in each document corresponding to one unique NP), only a selected fraction of them is displayed on the COCONUT web interface. Finally, the first round of automatic curation of NP metadata, in particular the molecular name synonyms, cross-references with other major chemical databases, correction of the literature references (PubMed identifiers and DOIs) and taxonomy is performed. All original data, unified NPs and the derived and calculated information are stored in MongoDB. The chemical classification of all NPs in COCONUT is performed with ClassyFire [10] and, when available, is present in the corresponding section of the compound page. Additionally, frameworks facilitating NP analyses for their chemical and therapeutic properties are computed for NPs, such Murcko frameworks [11], Ertl Functional Groups [12] and deep SMILES [13].

Last, the annotation level of each NP in COCONUT is computed. It is a 5-star-based system, where 1 star is the lowest annotation quality (no verified common name, no organism annotation, no literature reference and no trusted data source) and 5 stars is the highest quality, with all the intermediate annotation qualities reflected by 2, 3 and 4 stars. A “trusted” data source here is one that has a high curation level for NPs: ChEBI [14], KNApSAcK [4], ChEMBL [15], CMAUP [5], NP Atlas [3] and, of course, the manually picked data. The annotation level represented with stars is visible for each NP on its page.

Natural product naming

NP common names in COCONUT have been retrieved, when available from their databases of origin. The remaining NPs were searched by InChI in major chemical databases (PubChem, ChEMBL and ChEBI) and common names and synonyms were retrieved when the compound was present there. IUPAC names were systematically computed with ChemAxon, and when no common name was available for the compound, the latter were assigned as one. Therefore, all NPs in COCONUT have an assigned molecular name. An IUPAC name is computed for each NP using ChemAxon’s MolCovert [16], and when any name for the molecule could be found, the IUPAC name is assigned as the main one.

Computed molecular features

Figure 2 demonstrates the distributions and relationships of a small selection of computed molecular features within COCONUT. Sugar moieties are one frequent, but not mandatory, feature of NPs. To track their influence on other features, their absence and presence are colour-mapped (no sugar moiety in the molecular structure in blue, and the presence of at least one sugar moiety in orange). The wide molecular weight range is typical for NPs; it is, however, interesting to notice its correlation with the number of oxygen atoms in the molecule, regardless of the presence and absence of sugar. Another interesting correlation to be noted is between the molecular weight and the nitrogen atom number in sugar-free molecules. The NP-likeness score [17] has a typical distribution for an NP set, where most molecules have a positive score.

Counting rings in a molecule can be a complex task, as the outer perimeter of two fused rings can be counted as one big ring. With more condensed rings, the number of fused ring perimeters (aka as the set of all rings) can grow steeply. In Fig. 2, only the minimal ring count (the minimal cycle base) is represented.

Natural product annotation

In addition to their structure and computable structural properties, NPs need to be annotated with at least one literature reference, mentioning where, when and from which organism the NP was isolated. As a consequence, an NP entry should be associated with at least one organism, preferentially with an NCBI taxonomy identifier and the geographic location where the organism was collected. Unfortunately, this metadata is often omitted in public databases and datasets from which COCONUT was assembled. Therefore, only 31.7% (135,352) of NPs in COCONUT are annotated with at least one organism taxa, for 15.4% (66,068) of NPs the geographic location (on the continent level) of the organism collection is known and 16.6% (70,730) of NPs have at least one literature reference. These numbers combine both the retrieval of the original NP annotations from their sources and our efforts to retrieve more extensive information from major trusted chemical databases, PubChem [18], ChEMBL [15], ChEBI [14], CMAUP [5] and KnapSacK [4]. Despite our efforts, most of the links between the original publication of the structure elucidation of an NP and its reference, source organism and its geographical location are still missing. A possible solution to fill these gaps is manual curation, but the amount of data in COCONUT is redhibitory for even considering this approach. Another solution is to use unsupervised machine learning for optical recognition approaches, to parse modern peer-reviewed literature and books to re-establish links between NP structures and their provenance.

We analysed the taxonomic classification of known NP producers together with overlaps in NP production between superkingdom for the 31% of the NPs in COCONUT for which the provenance organism is known (Fig. 3). Here are distinguished five taxonomic categories: plants, bacteria, fungi, animals and marine. The last one is not a proper monoclade classification, but rather reflects a group of organisms that are found only in marine and oceanic environments, and therefore can overlap in terms of its species and NP content with other categories, which are more stringent taxonomically. A large part (65%) of these annotated NPs are produced only by plants, and only very few (0.5%) are from animal origin. Main overlaps in terms of NP production between the taxonomic kingdoms are between plants and marine organisms (which is unsurprising, as there can be real plants among the marine entities) and surprisingly between plants and fungi. The other overlaps between taxonomic kingdoms are not as significant. It needs to be pointed out here that multicellular organisms, such as plants, animals and some of the fungi are most of the time in symbiosis with microorganisms, in particular bacteria. Therefore, NPs isolated from such a multicellular organism can be synthesized and secreted by their symbionts or microbiomes, and therefore mistakenly assigned to an incorrect organism.

The geographic location of the collection or the natural presence of the NP-producing organism is a piece of information that is even more difficult to obtain. Nowadays, a range of organisms, and in particular plants, can be found in different parts of the planet due to globalisation and their success in human consumption (e.g. garlic, tomatoes, curcuma or ginger). It is, therefore, difficult, if not impossible, to determine their original provenance. Also, the geographical information is often omitted in literature and most NP databases. When available, the geographical provenance is stored in the MongoDB dump of COCONUT, but not displayed on the website.

For NPs where geographical information is available, it appears that most of them are produced by organisms that have been isolated in Asia (Fig. 4). This bias is introduced by the intensive study by scientists of the traditional Chinese and Indian medicines and by the big efforts in isolation and elucidation of NPs from medicinal plants. NPs from the African continent are also well represented in COCONUT (Fig. 4), mainly due to the scientific interest in African traditional medicines and African biodiversity. There is, for now, no data from the biodiversity of the Australian continent, and only very little data for NPs isolated from endemic European organisms. NPs from the Americas are mainly extracted and solved while Brazilian and Mexican biodiversity exploration. Only a few NPs are present in more than one continent, mainly in Asia and Africa, and the overlap values are biased by the very different NP set sizes between the different continents.

Searching the database

COCONUT online is intended to be a full-fledged chemical database, with all the subsequent functions, in particular the chemical search. At the moment, the chemical search is uncommon with MongoDB, therefore several approaches have been implemented to run molecular substructure and similarity searches.

Simple search

First, a simple search can be performed using the header search bar. The query can be performed on molecule names (e.g. “curcumin”), SMILES, InChI, InChi key, COCONUT id or molecular formula. Name search uses native MongoDB text indexing, allowing fuzzy flexible search in the “name” and “synonyms” fields. First, the input string type is identified using regular expressions, then the DB is queried against the appropriate fields, and the result, when exists, is returned to the front-end.

Substructure search implementation

Searching for an exact substructure in a MongoDB database of molecules appeared to be surprisingly easy. Each molecule in the database needs to have their fingerprints of choice (in COCONUT are used the PubChem fingerprints) to be precomputed and stored as a list of bytes (BinData type in MongoDB). The query molecule (substructure) then needs to have its fingerprint to be also computed and to be matched against the database using the $allBitsSet function [19]. This native to MongoDB function allows to select documents in a collection where a BinData field has all the query bits set to “on” (but can have bits set to “on” that are not present in the query). Then, to confirm the substructure match, the Ullmann pattern matching [20] is performed using CDK methods.

Similarity search implementation

Similarity search with MongoDB was implemented following the excellent ChEBML blog post tutorial on LSH-based similarity search in MongoDB [21] and adapting it to Java, Kotlin and Spring data. In this approach, the MongoDB aggregation framework is used to perform inverted indexing search against PubChem fingerprints stored in a separate table and referencing COCONUT identifiers that contain the molecular features encoded by each bit.

Advanced search

The advanced search allows searching for NPs in COCONUT according to a range of parameters, such as molecular formula, molecular descriptor values, number of rings, type of sugar moieties present in them, etc.

Querying COCONUT through the API

An API is also available to programmatically query COCONUT. It relies on Kotlin API functionalities and it’s usage, together with some examples, is described in detail in the documentation section of the website (https://coconut.naturalproducts.net/documentation).

Documentation

Complete documentation describing COCONUT, its data and functionalities are available at the documentation section of the website https://coconut.naturalproducts.net/documentation.

Utility And Discussion

The online COCONUT database is an open tool for researchers in the natural products community. COCONUT is the biggest collection of NPs in 2020 and the data it contains already benefits researchers in NPs with various aims, such as biodiversity research and drug discovery. The web interface allows querying and parsing the data collection in various, chemically relevant ways with adequate performance. It is also the first big chemical database using MongoDB as a storage management system.

A wide range of molecular descriptors are pre-computed and literature, producer taxonomy and their geography are as much annotated as currently possible without extensive manual curation. The web database can be searched in multiple ways, by molecular structure, by compound name and by molecular features, making this repository a complete chemical database. The user interface is modern and easy to use. Besides, the whole content of COCONUT is available for download in multiple formats.

In the close future, COCONUT will support user registration to enable user-driven NP curation and submission and will undergo a better data annotation, in particular regarding the organisms that are producing the NPs, their geography and the corresponding literature, using deep learning approaches.

Feedback

Bugs, annotation issues and requests of new COCONUT entries or re-annotation of existing ones can be reported at the project issues tracker (https://github.com/mSorok/NaturalProductsOnline/issues). Suggestions for new features are also welcome.

Availability

All COCONUT data, code to process raw NP data, data quality control and annotation, and the code for the font- and the back-end of the COCONUT online website are freely available without any restriction. The latest COCONUT data, as MongoDB full dump can be downloaded at https://coconut.naturalproducts.net/download. Code for data assembly, processing and quality control process codes is available on GitHub at https://github.com/mSorok/COCONUT. The code for the front-end and back-end is also available on GitHub at https://github.com/mSorok/NaturalProductsOnline.

Conclusions

COCONUT is the largest open collection of natural products at this time. It may be of particular importance for the NP community as it gathers in one single place most of open NP knowledge, and makes it easily accessible and queryable.

The final aim of COCONUT is to provide to the scientific community NP structures and their provenance, *i.e*. organisms that synthesize them and geographic location of the latter. However, a lot of data curation, in particular using new generation deep learning-based methods of extracting information from publications and books, together with website functionalities developments are still need to be done for COCONUT, but the database as it is now is already an important tool to facilitate NP and medicinal chemistry research.

Abbreviations

- NP: natural products
- COCONUT: COlleCtion of Open Natural ProdUcTs
- CDK: Chemistry Development Kit
- IUPAC: International Union of Pure and Applied Chemistry

- API: Application Programming Interface
- REST: REpresentational State Transfer
- DOI: Digital Object Identifier
- JSON: JavaScript Object Notation
- (No)SQL: (non-/not only) Structured Query Language

Declarations

- Availability of data and materials: the source code of the web interface and the back-end is available on GitHub at https://github.com/mSorok/NaturalProductsOnline. The data was curated and processed using the COCONUT code suite available on GitHub at https://github.com/mSorok/COCONUT. All COCONUT data can be accessed on the website at https://coconut.naturalproducts.net/ and downloaded entirely or partially in several formats (MongoDB dump, SDF and CSV).
- Competing interests: The authors declare no competing interests.
- Funding: This work was supported by the German Research Foundation within the framework CRC1127 ChemBioSys.
- Authors' contributions: MS coordinated the study, participated in the implementation of the front-end and the backend, collected; processed and curated the COCONUT data and drafted the manuscript. PM participated in the implementation of the front-end and the back-end of COCONUT online. KR and MAY helped with COCONUT data curation. CS designed and supervised the study. All authors read the manuscript and agree on its content.
- Acknowledgements: not applicable.

References

1. Sorokina M, Steinbeck C. Review on natural products databases: where to find data in 2020. J Cheminformatics. 2020;12:20. doi:10.1186/s13321-020-00424-9.
2. Banerjee P, Erehman J, Gohlke B-O, Wilhelm T, Preissner R, Dunkel M. Super Natural II—a database of natural products. Nucleic Acids Res. 2015;43:D935–9. doi:10.1093/nar/gku886.
3. van Santen JA, Jacob G, Singh AL, Aniebok V, Balunas MJ, Bunsko D, et al. The Natural Products Atlas: An Open Access Knowledge Base for Microbial Natural Products Discovery. ACS Cent Sci. 2019;5:1824–33. doi:10.1021/acscentsci.9b00806.
4. Nakamura K, Shimura N, Otabe Y, Hirai-Morita A, Nakamura Y, Ono N, et al. KNApSAcK-3D: A Three-Dimensional Structure Database of Plant Metabolites. Plant Cell Physiol. 2013;54:e4–e4. doi:10.1093/pcp/pcs186.
5. Zeng X, Zhang P, Wang Y, Qin C, Chen S, He W, et al. CMAUP: a database of collective molecular activities of useful plants. Nucleic Acids Res. 2019;47:D1118–27.

6. Chen CY-C. TCM Database@Taiwan: The World's Largest Traditional Chinese Medicine Database for Drug Screening In Silico. PLOS ONE. 2011;6:e15939. doi:10.1371/journal.pone.0015939.
7. React – A JavaScript library for building user interfaces. https://reactjs.org/. Accessed 21 Aug 2020.
8. OpenChemLib (https://github.com/cheminfo/openchemlib-js). https://github.com/cheminfo/openchemlib-js.
9. Willighagen EL, Mayfield JW, Alvarsson J, Berg A, Carlsson L, Jeliazkova N, et al. The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. J Cheminformatics. 2017;9:33. doi:10.1186/s13321-017-0220-4.
10. Djoumbou Feunang Y, Eisner R, Knox C, Chepelev L, Hastings J, Owen G, et al. ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. J Cheminformatics. 2016;8:61. doi:10.1186/s13321-016-0174-y.
11. Bemis GW, Murcko MA. The Properties of Known Drugs. 1. Molecular Frameworks. J Med Chem. 1996;39:2887–93. doi:10.1021/jm9602928.
12. Fritsch S, Neumann S, Schaub J, Steinbeck C, Zielesny A. ErtlFunctionalGroupsFinder: automated rule-based functional group detection with the Chemistry Development Kit (CDK). J Cheminformatics. 2019;11:37. doi:10.1186/s13321-019-0361-8.
13. O'Boyle N, Dalke A. DeepSMILES: An Adaptation of SMILES for Use in Machine-Learning of Chemical Structures. 2018. doi:10.26434/chemrxiv.7097960.v1.
14. Hastings J, de Matos P, Dekker A, Ennis M, Harsha B, Kale N, et al. The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. Nucleic Acids Res. 2013;41 Database issue:D456–63. doi:10.1093/nar/gks1146.
15. Gaulton A, Hersey A, Nowotka M, Bento AP, Chambers J, Mendez D, et al. The ChEMBL database in 2017. Nucleic Acids Res. 2017;45:D945–54. doi:10.1093/nar/gkw1074.
16. Marvin | ChemAxon. https://chemaxon.com/products/marvin. Accessed 20 Aug 2020.
17. Ertl P, Roggo S, Schuffenhauer A. Natural Product-likeness Score and Its Application for Prioritization of Compound Libraries. J Chem Inf Model. 2008;48:68–74. doi:10.1021/ci700286x.
18. Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, et al. PubChem Substance and Compound databases. Nucleic Acids Res. 2016;44:D1202–13. doi:10.1093/nar/gkv951.
19. $bitsAllSet — MongoDB Manual. https://github.com/mongodb/docs/blob/master/source/reference/operator/query/bitsAllSet.txt. https://docs.mongodb.com/manual/reference/operator/query/bitsAllSet. Accessed 21 Aug 2020.
20. Ullmann (CDK 2.3 API). http://cdk.github.io/cdk/latest/docs/api/index.html. Accessed 21 Aug 2020.
21. Michał. LSH-based similarity search in MongoDB is faster than Postgres cartridge. THE CHEMBL-OG The Organization of Drug Discovery Data. http://chembl.blogspot.com/2015/08/lsh-based-similarity-search-in-mongodb.html. Accessed 21 Aug 2020.

Tables

Due to technical limitations, table 1-2 is only available as a download in the Supplemental Files section.

Figures

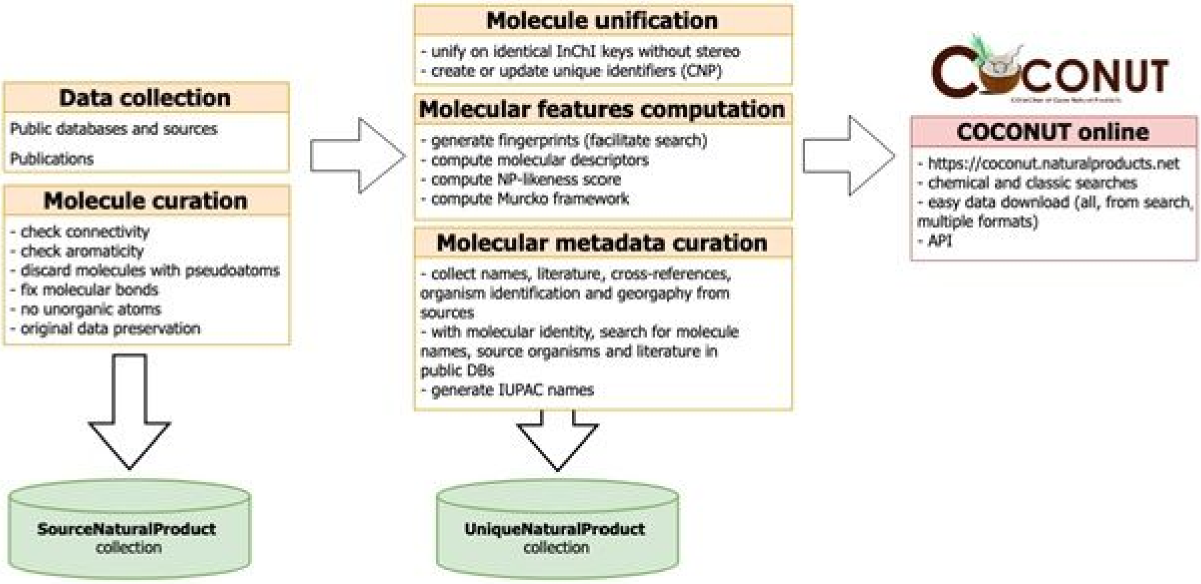
Data collection
Public databases and sources
Publications
Molecule curation
- check connectivity
- check aromaticity
- discard molecules with pseudoatoms
- fix molecular bonds
- no unorganic atoms
- original data preservation
SourceNaturalProduct collection
Molecule unification
- unify on identical InChI keys without stereo
- create or update unique identifiers (CNP)
Molecular features computation
- generate fingerprints (facilitate search)
- compute molecular descriptors
- compute NP-likeness score
- compute Murcko framework
Molecular metadata curation
- collect names, literature, cross-references, organism identification and georgaphy from sources
- with molecular identity, search for molecule names, source organisms and literature in public DBs
- generate IUPAC names
UniqueNaturalProduct collection
COCONUT
COCONUT online
- https://coconut.naturalproducts.net
- chemical and classic searches
- easy data download (all, from search, multiple formats)
- API

Figure 1

Construction and curation of COCONUT online.

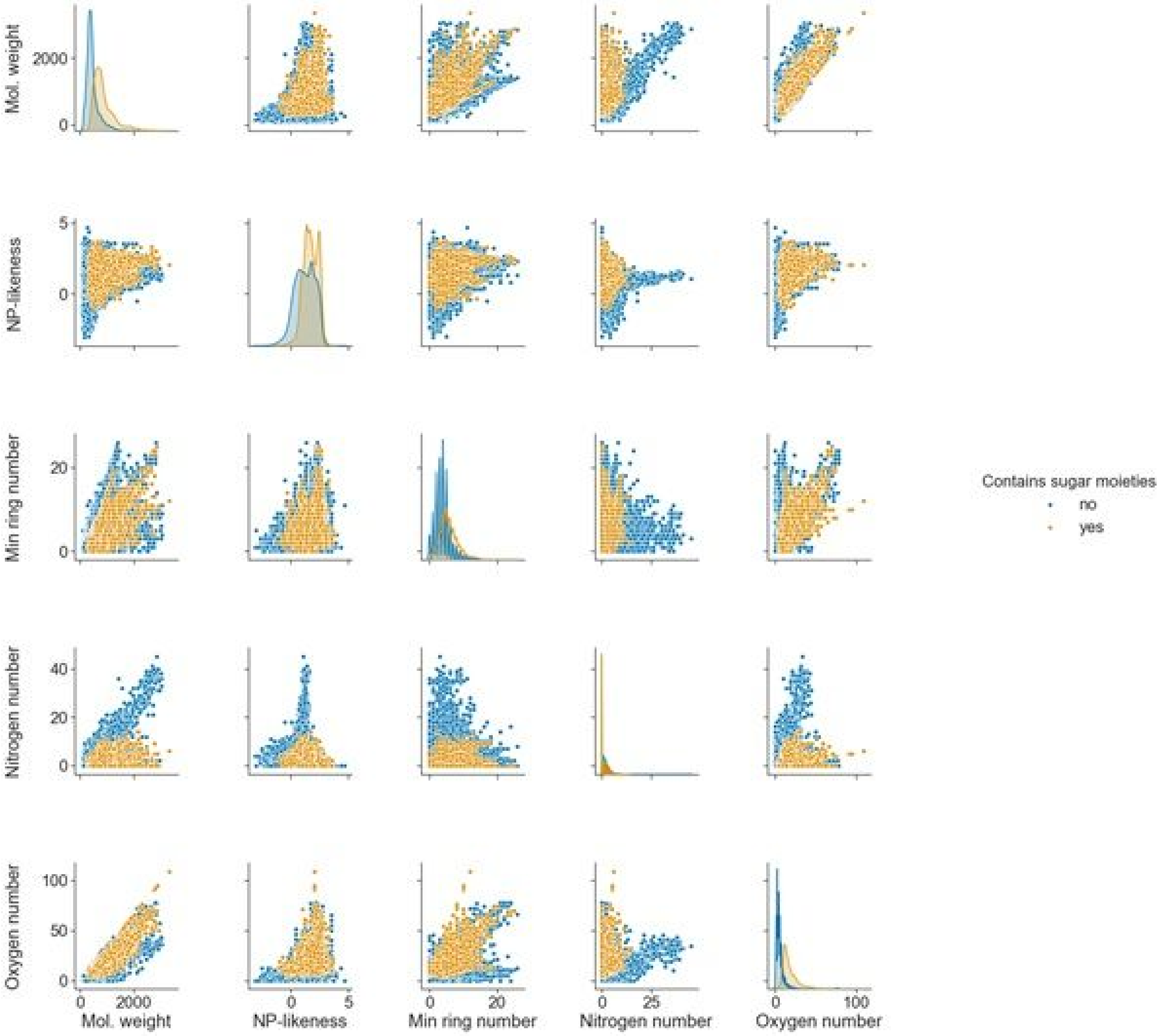

Figure 2

Pair plot of a selection of five of the molecular features available in COCONUT. Colour mapping corresponds to the presence (yellow) and absence (blue) of glycosidic moieties in the molecular structures of the NPs in COCONUT.

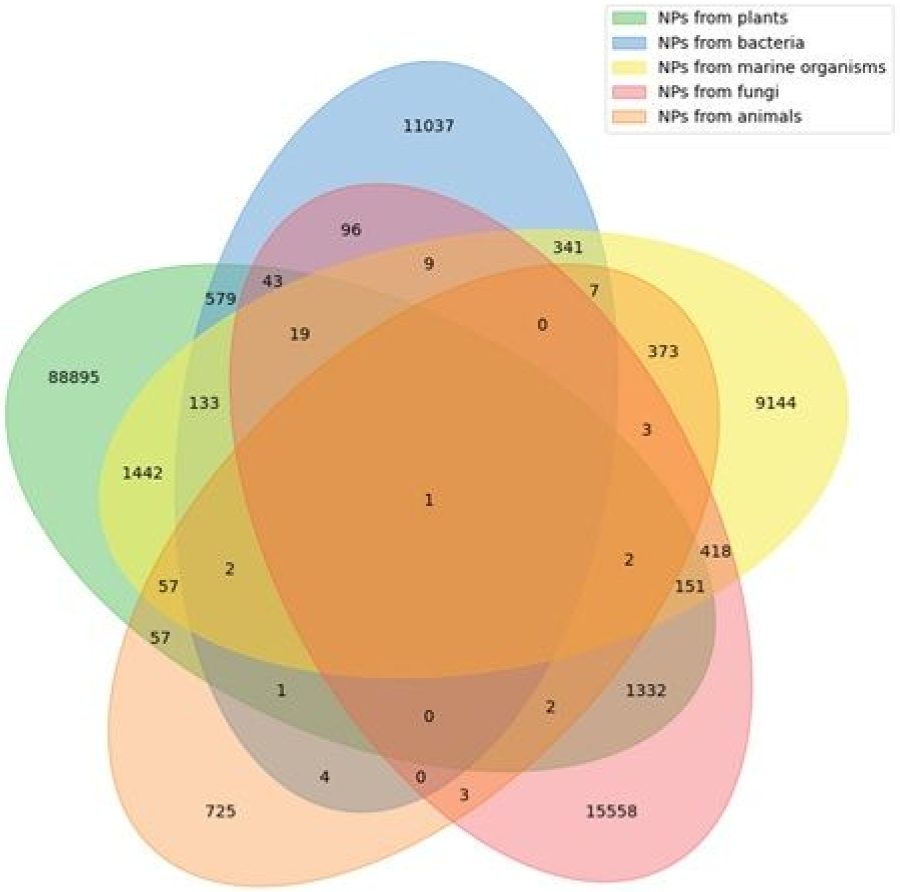

Figure 3

Overlap of NP taxonomic provenance in COCONUT.

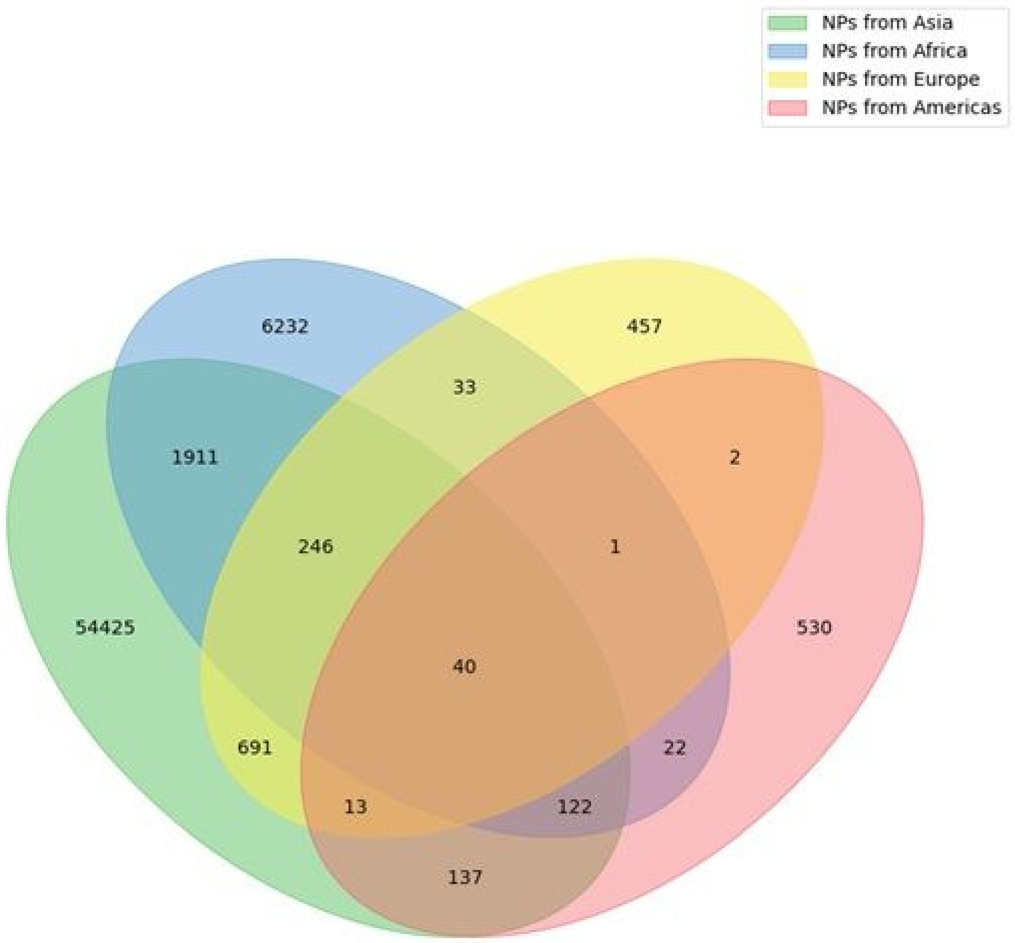

Figure 4

Overlap of NP geographic provenance in COCONUT.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Table1coconutdatasources.xlsx
- Table2fieldsforunps.xlsx