NPDB: a natural product database with relational data between natural products and biological sources

Tingjun Xu (✉ xutingjun@sioc.ac.cn)

Shanghai Institute of Organic Chemistry <https://orcid.org/0000-0002-5529-1875>

Jingfang Dai

Shanghai Institute of Organic Chemistry <https://orcid.org/0000-0002-8742-5671>

Weiming Chen

Shanghai Institute of Organic Chemistry <https://orcid.org/0000-0002-2660-5492>

Yingyong Li

Shanghai Institute of Organic Chemistry <https://orcid.org/0000-0002-6604-2858>

Yingli Zhao

Shanghai Institute of Organic Chemistry <https://orcid.org/0000-0003-1214-0513>

Junhong Zhou

Shanghai Institute of Organic Chemistry <https://orcid.org/0000-0001-5589-7861>

Database

Keywords: Biological sources, Chemical text-mining, Drug discovery, Natural products, Natural Product Database

Posted Date: March 4th, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-15952/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

We designed a natural product database filled with hand-curated data of natural product information from journal literatures, to establish the relationships between natural products and biological sources. Therefore, we can figure out all the biological sources of a specific natural product reported in literatures, and also can inspect for the holistic view of a specific organism's components, metabolites or extracts.

Introduction

The studies of deriving natural products from biological sources presented in publications provide abundant information of diversity both in biology and chemistry. Although knowledge of natural products has inspired new medicines, agrochemicals and materials, broad research on universal natural products derived from various organisms can be deficient.[1–3] For example, the correlations between the molecules of natural products and their host species are still ambiguous. What are the complete chemical components or metabolites of one specific species? What are the relationships between species sharing identical chemical components or metabolites? Those mysteries may be studied by computational approaches, there are already several data resources useful for natural product researches.[4–14] The data resources are various from free-access and commercial, from comprehensive and specialized, nevertheless, the demand of scientific exploration is still beyond the amount of available datasets, as natural product deriving achievements are increasingly published in journals.[4–16]

We herein describe a natural product database (NPDB), a data resource includes the information of relationships between natural products and biological sources reported in publications. The relational data links a specific species and all the natural products derived from it, and contrarily links a specific natural product and all the biological sources, and each relational data has corresponding bibliographic references. In this database, the natural products are represented by molecular structures of the molecules derived from organisms, and the biological sources are represented by species names of the organisms. Other available information like deriving-parts of the organisms, names of the natural products, and computable molecular properties are also include in NPDB. The volume of the database is extending continuously, as journal literatures of natural product deriving researches are on the increase, and we intend to involve more publications in the data acquisition stage.

Utility And Discussion

At the time of writing, there are 33,377 unique species of biological sources (distinguished by species names), 122,776 unique molecules of natural products (distinguished by InChIKeys),[17] and 898,294 relational data records included in the NPDB. The biological sources cover the diverse species of plant, bacterial, fungal and marine organism, the molecules have proper chemical structure data and computable molecular properties, and all the relational data have corresponding references. The features of the current NPDB are shown in following tables and figures.

The top 10 species of biological sources provide most natural products in NPDB, as shown in Table 1, are all plants as expected, since terrestrial plants are the most abundant and accessible biological source on the earth, and human beings have a long history of studying plant's chemical compositions as food, medicines or materials. On the other hand, literatures of phytochemistry are the majority of the publications we covered at present. Interestingly, the top 2 and the 9th species are fruits, other five species are used as seasoning and spice, each one provide over 700 natural products. Two traditional Chinese herbal medicines *Artemisia annua* and *Hypericum perforatum* have been demonstrated by modern science, their special constituents show significant antimalarial and antidepressant activity. [18,19]

Insert Table 1 here.

The top 10 molecules of natural products derived from most biological sources in NPDB, as shown in Table 2, are 8 terpenoids, 1 steroid, and 1 aliphatic acid, each one is derived from over 4,000 biological sources. Terpenoids are a large group of substances which occur in most organisms, playing vital roles of biofunctionality such as antioxidants and nutritions.[20] The steroid “ β -sitosterol” and the aliphatic acid “palmitic acid” also widely exist in organisms as important classes of bioorganic molecules. [21]

Insert Table 2 here.

The molecular features of the natural products in NPDB, as shown in Figure 1, are perceived as chemically differing from the molecules in our general chemical databases.[22] For the structural complexity, more than 86% of the natural products have ring system, over one third have more than 3 rings (Figure 1. A), 56% of them are heterocycles (Figure 1. B). Approximately half of the natural products are aliphatic, 58% of them have chiral centers (Figure 1. B). The natural products also present extremely higher oxygen content on average, over 93% of them have oxygen atom, and the percentage of the natural products having more than 10 oxygen atoms reach 16%, it seems odd when compared with nitrogen content (Figure 1. C).

Insert Figure 1 here.

For the interest of taking natural products as starting points for medicinal chemistry and drug discovery, the Lipinski's rule of five parameters may have significant referential value to insight into “drug-likeness” of the molecules in NPDB.[23] Over half of the natural products are within the bounds of Lipinski's primary five parameters (Figure 2. F): molecular weight less than 500, number of hydrogen bond donors less than 5, number of hydrogen bond acceptors less than 10, number of rotatable bonds less than 10, and LogP less than 5. There is no doubt that the natural products are the treasure of potential drug candidates.

Insert Figure 2 here.

Construction And Content

As a strategy of screening literatures for data sources, our data analysts browse the contents of each journals by issues and volumes, select the required articles based on judgments of the titles and columns, obtain the PDF version of the full text, or abstracts and bibliographic information when the PDF file is unavailable. The journals involved cover the major publications of natural product deriving researches both domestic (Chinese) and international (English), 45% of them are extensive research on natural product, others are on phytochemistry, traditional chinese medicine, food industry and miscellaneous, as shown in Fig. 3. The main list of publications that NPDB covered is included in the supporting information file provided as supplemental materials .

Insert Figure 3 here.

In the early stage of this work, we collected the raw data of NPDB from publications manually. Our data analysts reviewed the literatures, indexed the information of biological sources and the molecules of natural products. With the practical experience and plenty of hand-curated data, we developed a rule-based text mining system for natural product data acquisition (NPDsys), and attempted to extracted the required data information from textual description of the journal literatures automatically (Fig. 3).

Insert Figure 4 here.

The biological informations recognized in NPDsys are species names of the organisms, such as “*Alternaria alternate*” in Fig. 3. B,[24] and derive-parts of the organisms, such as “secondary metabolites”, “leaves”, and “aerial parts”. The chemical informations recognized in NPDsys are trivial, systematic or semi-systematic names of the natural product molecules, and the author-numbers of the molecules. The author-numbers such as “Compound A”, “Compound (1)”, are used for associating different representations of the same natural product appear in abstract, introduction, results and experimental section, or English and Chinese names of the same natural product in Chinese literatures. When the data meet the definition of “large-scale”, that information will be complementary in natural product name translation and molecular structure converting. As an exploration of chemical text-mining, the NPDsys not only get the chemical entities in literatures but also take an attempt to recognise the connections between the entities. Nevertheless, a great quantity of literatures are not available as text-mining materials, and some literatures do not provide appropriate systematic names for the “new-found” natural product molecules, lead to extra procedure of hand-drawing structures or web-searching. Therefore, the available data set in NPDB is hand-curated data at present.

The raw data had been processed properly before added into NPDB. We first assume the correctness of the primary literatures, unless there are apparent errors like typos, then backtrack on the original document when encounter abnormal data in the subsequent processing. For the same biological source from multiple literatures, we merge the data and remain the distinct natural products, and list all the references. We have similar approach for the same natural product derived from different biological sources reported by multiple literatures. When encounter multiple biological sources in one literature (For

example, components of mixed species researches),[25] we label the relational data as “optional”, once other literatures reported data corroborated the natural products from one of those biological sources, the relational data will be “confirmed”.

In the process of the natural product data acquisition, the structures of the molecules are generated by “name to structure module” of ACDLabs, ChemOffice and OPSIN for computer-aided data analysing, and we use ChemDraw and Reaxys for drawing and searching the structures manually.[26–29] A machine translation tool of chemical nomenclature has been used for Chinese compound names to/from English translation.[30] We have an evaluation of the structures generated by different toolkits, scores have been made by molecular formula comparison in order to evaluate the different structures from the same compound name, and the eventual structures are standardized to MDL Molfile format. RDKit has been applied in Python for molecular properties computing in NPDB.[31, 32]

Conclusions

With continuous upgrading and optimizing of NPDSys, we might have an efficient and low-cost tool to expand the volume of NPDB in the future. Owing to the limitation of professional, we haven't concentrated on the classification and analysis of the biological source data. Nevertheless, current datasets in NPDB has shown considerable diversity both in biology and chemistry, and the relational data provide significant clues in the correlations between some special natural products and their host species. Compare to the top databases of natural product, the NPDB has not exceed their volume or coverage, but has greater clarity about the relationships between natural products and biological sources. [4–9, 14] Besides mining information from literatures, to make published data quickly accessible and to combine them logically for further applications, the main purpose or destination of this database is clarifying about the relationships between natural products and biological sources, since the relationships are intersecting. For example, the same natural product and its homologues can originate from various biological sources, adopters can have a wider choice on biological sources from their own concerns. On the other hand, integrated constituent information from a specific biological source, can inspire novel applications of the biological source, for example, adopters can analyse the potential toxicology or pharmacology of a Chinese medicinal herb at molecular level. We are looking forward to discovering more secrets of natural products using new approach of cheminformatics, and providing sophisticated data-support for pharmaceutical research by web interface and retrieval functionality.

Declarations

Availability

NPDB is available at <http://www.organchem.csdb.cn/scdb/NaturalProduct>

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

Weiming Chen and Tingjun Xu: Original idea, database design. Jingfang Dai: Web site design and development. Yingyong Li, Yingli Zhao and Junhong Zhou: Data processing and analysis.

Funding

This work was supported by NSFC (21805303), CSDB (XXH135) and SGST (18DZ2294000).

Acknowledgements

We are grateful to our data analysis group for the high quality work.

References

1. Butler MS (2004) The Role of Natural Product Chemistry in Drug Discovery. *J Nat Prod* 67:2141–2153. doi:10.1021/np040106y.
2. Dayan FE, Cantrell CL, Duke SO (2009) Natural products in crop protection. *Bioorg Med Chem* 17:4022–4034. doi:10.1016/j.bmc.2009.01.046.
3. Zhang X, Jiang M, Niu N, Chen Z, Li S, Liu S, Li J (2017) Natural-Product-Derived Carbon Dots: From Natural Products to Functional Materials. *ChemSusChem* 11:11–24. doi:10.1002/cssc.201701847.
4. Dictionary of Natural Products (DNP). <http://dnp.chemnetbase.com>. Accessed Feb 14 2020.
5. Banerjee P, Erehman J, Gohlke BO, Wilhelm T, Preissner R, Dunkel M (2014) Super Natural II—a database of natural products. *Nucleic Acids Res* 43:D935–D939. doi:10.1093/nar/gku886.
6. Gu J, Gui Y, Chen L, Yuan G, Lu HZ, Xu X (2013) Use of Natural Products as Chemical Library for Drug Discovery and Network Pharmacology. *PLoS ONE* 8:e62839. doi:10.1371/journal.pone.0062839.
7. Universal Natural Products Database (UNPD). <http://pkuxxj.pku.edu.cn/UNPD>. Accessed Oct 17, 2016.
8. He M, Yan X, Zhou J, Xie G (2001) Traditional Chinese Medicine Database and Application on the Web. *J Chem Inf Model* 41:273–277. doi:10.1021/ci0003101.
9. Xue R, Fang Z, Zhang M, Yi Z, Wen C, Shi T (2012) TCMID: traditional Chinese medicine integrative database for herb molecular mechanism analysis. *Nucleic Acids Res* 41: D1089–D1095. doi:10.1093/nar/gks1100.
10. Chem-TCM. <http://www.chemtcm.com>. Accessed Feb 14 2020.
11. Núcleo de Bioensaios, Biossíntese e Ecofisiologia de Produtos Naturais (NuBBE). <http://nubbe.iq.unesp.br/portal/nubbedb.html>. Accessed Feb 14 2020.
12. Dictionary of Marine Natural Products. <http://dmnp.chemnetbase.com>. Accessed Feb 14 2020.
13. Valli M, dos Santos RN, Figueira LD, Nakajima CH, Castro-Gamboa I, Andricopulo AD, Bolzani VS (2013) Development of a Natural Products Database from the Biodiversity of Brazil. *J Nat Prod* 76:439–444. doi:10.1021/np3006875.

14. Chen Y, de Bruyn Kops C, Kirchmair J (2017) Data Resources for the Computer-Guided Discovery of Bioactive Natural Products. *J Chem Inf Model* 57:2099–2111. doi:10.1021/acs.jcim.7b00341.
15. Medema MH, Fischbach MA (2015) Computational approaches to natural product discovery. *Nat Chem Biol*, 11:639–648. doi:10.1038/nchembio..
16. Rodrigues T, Reker D, Schneider P, Schneider G (2016) Counting on natural products for drug design. *Nat Chem* 8:531–541. doi:10.1038/nchem.2479.
17. Heller SR, McNaught A, Pletnev I, Stein S, Tchekhovskoi D (2015) InChI, the IUPAC International Chemical Identifier. *J Cheminf.* doi:10.1186/s13321-015-0068-4.
18. Tu Y (2011) The discovery of artemisinin (qinghaosu) and gifts from Chinese medicine. *Nat Med* 17:1217–1220. doi:10.1038/nm.2471.
19. Butterweck V, Jürgenliemk G, Nahrstedt A, Winterhoff H (2000) Flavonoids from *Hypericum perforatum* Show Antidepressant Activity in the Forced Swimming Test. *Planta Med* 66:3–6. doi:10.1055/s-2000-11119.
20. Wagner KH, Elmadfa I (2003) Biological relevance of terpenoids. *Ann Nutr Metab* 47:95-106. doi: 1159/000070030.
21. Saeidnia S, Manayi A, Gohari AR, Abdollahi M (2014) The story of beta-sitosterol-a review. *Eur J Med Plants* 4:590-609. doi: 10.9734/EJMP/2014/7764.
22. Shanghai Institute of Organic Chemistry of CAS (2019) Chemistry Database. <http://www.organchem.csdb.cn>. Accessed Feb 14 2020.
23. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (2012) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Delivery Rev* 64:4–17. doi:10.1016/j.addr.2012.09.019.
24. Wang J, Ma Z, Wang G, Liu J, Xu F, Peng D, Wang G (2019) Study on secondary metabolites of endophytic fungus *Alternaria alternate* from *Paeonia lactiflora*. *Zhongcaoyao* 50:1061-1065. doi: 7501/j.issn.0253-2670.2019.05.006.
25. Li W, Gu Z, Yang Y, Zhou S, Liu Y, Zhang J (2014) Non-volatile taste components of several cultivated mushrooms. *Food Chem* 143:427–431. doi:10.1016/j.foodchem.2013.08.006
26. Advanced Chemistry Development (2019) ACDlabs. <http://www.acdlabs.com>. Accessed Feb 14 2020.
27. PerkinElmer Informatics (2019) ChemOffice. <http://www.cambridgesoft.com/software/overview.aspx>. Accessed Feb 14 2020.
28. Lowe DM, Corbett PT, Murray-Rust P, Glen RC (2011) Chemical Name to Structure: OPSIN, an Open Source Solution. *J Chem Inf Model* 51:739–753. doi:10.1021/ci100384d.
29. Elsevier (2019) Reaxys. <https://www.reaxys.com>. Accessed Feb 14 2020.
30. Shanghai Institute of Organic Chemistry of CAS (2019) Machine translation tool. <http://www.organchem.csdb.cn/translate>. Accessed Feb 14 2020.
31. Landrum G (2019) RDKit. <http://www.rdkit.org>. Accessed Feb 14 2020.

Tables

Due to technical limitations, all tables are only available for download from the Supplementary Files section.

Figures

Figure 1

The molecular features of the natural products in NPdb. (A) Number of rings. (B) Structural features. (C) Number of Atoms.

Figure 2

The “drug-likeness” features of the natural products in NPDB. (A) Calculated LogP value. (B) Number of hydrogen bond donors. (C) Number of hydrogen bond acceptors. (D) Number of rotatable bonds. (E) Molecular weight. (F) Lipinski’s five parameters.

Figure 3

The coverage of the journals involved in the database.

Figure 4

The design of the natural product data acquisition system. (A) Illustration of natural product data acquisition system. (B) Demonstration of the procedure for natural product data acquisition.

Supplementary Files

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

- [Supportinginformation.pdf](#)
- [tables.docx](#)
- [graphicalabstract.jpg](#)