NPBS database: a chemical data resource with relational data between natural products and biological sources

Tingjun Xu (✉ xutingjun@sioc.ac.cn)
Weiming Chen
Junhong Zhou
Jingfang Dai
Yingyong Li
Yingli Zhao

Data Note

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

DOI: https://doi.org/10.21203/rs.3.rs-15952/v2

License: © This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License
Abstract

NPBS (Natural Products & Biological Sources) database is a chemical data resource with relational data between natural products and biological sources, manually curated from literatures of natural product research. The relational data links a specific species and all the natural products derived from it, contrarily, links a specific natural product and all the biological sources. The biological sources cover diverse species of plant, bacterial, fungal and marine organism, the natural molecules have proper chemical structure data and computable molecular properties, and all the relational data have corresponding references. NPBS database provides a wider choice of biological sources, and can be used for dereplication to prevent re-isolation and re-characterization of already known natural products.

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 already known 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, which condensed the widespread information from literatures and combined them logically for further applications. There are already several data resources useful for natural product researches (Table 1). The data resources are various from free-access and commercial, from comprehensive and specialized, nevertheless, there are lack of available datasets that have clarity about the relationships between natural products and biological sources, as natural product deriving achievements are increasingly published in journals (13-16).

Table 1. Existing data resources of natural product researches.
| Name                                                   | Availability                                      | References |
|--------------------------------------------------------|---------------------------------------------------|------------|
| Dictionary of Natural Products (DNP)                    | http://dnp.chemnetbase.com                        |            |
| Reaxys                                                 | https://www.reaxys.com                            |            |
| Super Natural II                                       | http://bioinf-applied.charite.de/supernatural_new | 4          |
| Universal Natural Products Database (UNPD)              | http://pkuxxj.pku.edu.cn/UNPD                     |            |
| TCM database@ Taiwan                                   | http://tcm.cmu.edu.tw                            | 5          |
| TCMID                                                   | http://www.megabionet.org/tcmid                   | 6          |
| Chem-TCM                                               | http://www.chemtcn.com                            | 7          |
| NuBBE database                                         | http://nubbe.iq.unesp.br/portal/nubbedb.html      | 8          |
| Dictionary of Marine Natural Products (DMNP)            | http://dmnp.chemnetbase.com                       |            |
| AntiBase                                               | http://wwwuser.gwdg.de/~hlaatsc/antibase.htm      |            |
| MarinLit                                               | http://pubs.rsc.org/marinlit                      |            |
| TIPdb                                                  | http://cwtung.kmu.edu.tw/tipdb                    | 9          |
| NPCARE                                                 | http://silver.sejong.ac.kr/npcare                 | 10         |
| NANPDB                                                 | http://www.african-compounds.org/nanpdb           | 11         |
| StreptomeDB                                            | http://www.pharmaceutical-bioinformatics.de/streptomedb | 12       |

Data collections of chemical information (molecular structures, syntheses and reactivities, chemical and biological properties, etc.) are essential tools in chemical research. These databases mentioned above may be focused on a special group of chemical compounds (natural products, marine compounds, plant chemicals etc.). The main feature of these databases is, however, that the search result is chemical information of natural product when given a query of molecular structure, or documents of reporting natural product when given a query of biological species name. The search results could be direct information of biological sources when search a natural product, or information of natural products when search a biological species name, if the relationships between natural products and biological sources have been established. We herein describe a chemical data resource, NPBS (Natural Products & Biological Sources) database, 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 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 NPBS database. 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.

**Materials And Methods**

The core idea of this work is to link natural products and biological sources, and our priority is obtaining such information of relationships. We then focused on scientific literatures, seeing that literatures of natural product deriving researches invariably report biological species from which they extract the natural molecules. As a strategy of screening literatures for data sources, our data analysts browse the contents of each journal by issues and volumes, select the required articles based on rough judgments of the titles and abstracts, and obtain the full text PDF version of the articles, or abstract texts and bibliographic information when the PDF file is unavailable. The collected literatures would be screened again by inspecting main body of the articles before extract data from them, to exclude literatures which have no descriptions of natural products and biological sources. The journals involved in this project cover the major publications of natural product deriving researches both domestic (Chinese) and international (English), 45% of them are extensive research on natural products, others are on phytochemistry, traditional Chinese medicine, food industry and miscellaneous, as shown in Figure 1. Classification of a selected journal is based on the scope description of the journal. The main list of publications that NPBS database covered is provided as supplemental material.

In the early stage of this project, we collected the raw data of NPBS database 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 text-mining system (NPBS$_{sys}$) for computer-aided data acquisition, and attempted to extract the required data information from textual description of the journal literatures automatically (Figure 2). The NPBS$_{sys}$ used the technology of Named Entity Recognition (NER), more specifically, a methodology of NER based on rules and dictionaries was implemented in this system (17). The biological information of species names and parts were matched by dictionaries we built, for example, “Salvia kiaometiensis” and “roots” in the sentence of “study the chemical constituents of the roots of Salvia kiaometiensis”. The chemical names of natural products were matched by rule templates we designed, for example, “euscaphic acid” and “hyptadienic acid” in the sentence of “two compounds were obtained and identified as euscaphic acid and hyptadienic acid”.

The biological information recognized in NPBS$_{sys}$ is species description of the organism, such as “endophytic fungus Alternaria alternate” in Figure 2.B (18), and deriving-parts of the organisms, such as “secondary metabolites”, “leaves”, and “aerial parts”. The chemical informations recognized in NPBSsys 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 NPBS\textsubscript{sys} not only get the chemical entities in literatures but also take an attempt to recognise the connections between the entities. For example, the relationship of natural product “$\beta$-carboline alkaloid” and biological source “\textit{Carthamus tinctorius}” recognised in the sentence “A new $\beta$-carboline alkaloid was isolated from the leaves of \textit{Carthamus tinctorius}” is the connection between these two 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, most available data in NPBS database rely on manual work at present.

In the process of NPBS\textsubscript{sys}, the structures of the molecules are generated by “name to structure module” of ACDLabs (http://www.acdlabs.com), ChemOffice (http://www.cambridgesoft.com) and OPSIN (19), and we use ChemDraw (http://www.cambridgesoft.com) and Reaxys (https://www.reaxys.com) for drawing and searching the structures manually. A machine translation tool of chemical nomenclature has been used for Chinese compound names to English translation (http://www.organchem.csdb.cn/translate). 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 (http://www.rdkit.org) has been applied in Python (https://www.python.org/) for molecular properties computing in NPBS database.

The raw data had been processed properly before added into NPBS database. We first assume the correctness of the primary literatures, unless there are apparent errors like typos, and 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 (20), we label the relational data as “unconfirmed”, once other literatures reported data corroborated the natural products from one of those biological sources, the relational data will be “confirmed”.

\textbf{Results And Discussion}

At the time of writing, there are 33,377 unique biological sources (distinguished by species names), 122,776 unique molecules of natural products (distinguished by InChIKey) (21), and 898,294 relational data records included in NPBS database. 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 entity
The relationship diagram of NPBS database is shown in Figure 3, other features of the current database are shown in following tables and figures.

The top 10 species of biological sources provide most natural products in NPBS database, as shown in Table 2, are all plants as expected. On one hand, literatures of phytochemistry are the majority of the publications we covered at present. On the other hand, terrestrial plants are the most abundant and accessible biological source on the earth, and human beings have a long history of taking plants as food, medicines and materials. 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 (22,23).

### Table 2. The top 10 species of biological sources provide most natural products.

| No. | Species name             | Common name         | Number of natural products |
|-----|--------------------------|---------------------|---------------------------|
| 1   | *Mangifera indica*       | mango               | 1152                      |
| 2   | *Vitis vinifera*         | grape               | 846                       |
| 3   | *Rosmarinus officinalis* | rosemary            | 790                       |
| 4   | *Artemisia annua*        | sweet sagewort      | 797                       |
| 5   | *Capsicum annuum*        | cayenne pepper      | 764                       |
| 6   | *Ocimum basilicum*       | sweet basil         | 761                       |
| 7   | *Hypericum perforatum*  | common St. John's wort | 731                   |
| 8   | *Foeniculum vulgare*     | sweet fennel        | 724                       |
| 9   | *Psidium guajava*        | guava               | 723                       |
| 10  | *Coriandrum sativum*     | coriander           | 723                       |

The top 10 molecules of natural products derived from most biological sources in NPBS database, as shown in Table 3, 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 (24). The steroid “β-sitosterol” and the aliphatic acid “palmitic acid” also widely exist in organisms as important classes of bioorganic molecules (25).

### Table 3. The top 10 molecules of natural products derived from most biological sources.
The molecular features of the natural products in NPBS database, as shown in Figure 4, are perceived as chemically different from the molecules in other chemical databases (http://www.organchem.csdb.cn/). For the structural complexity, more than 86% of the natural products have ring system, over one third have more than 3 rings (Figure 4.A), and 56% of them are heterocycles (Figure 4.B). Approximately half of the natural products are aliphatic, 58% of them have chiral centers (Figure 4.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 4.C).

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, and have insight into “drug-likeness” of the molecules in NPBS database (26). Over half of the natural products are within the bounds of Lipinski’s primary five parameters (Figure 5): 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.

For evaluation of NPBS database, we carried out an experiment of comparison with other accessible natural product databases by searching for several common natural products and biological sources. The result as listed in Table 4, shows that NPBS database may be more applicable for searching wide biological sources of a specific natural product, and molecular properties of natural products derived from a specific biological source. For example, NPBS database shows 17 biological sources that contain natural product of “aconitine”, and 49 natural products derived from biological source of “Artemisia apiacea”.

| No. | Molecule | Name            | Number of biological sources |
|-----|----------|-----------------|------------------------------|
| 1   | α-pinene | 6609            |
| 2   | limonene | 6064            |
| 3   | β-pinene | 5899            |
| 4   | (-)β-linalool | 5730         |
| 5   | p-cymene | 5354            |
| 6   | myrcene  | 5304            |
| 7   | β-sitosterol | 5112         |
| 8   | palmitic acid | 4871         |
| 9   | (-)-terpinen-4-ol | 4749     |
| 10  | camphene | 4636            |
Table 4. Comparison with NPBS and other accessible natural product databases.

| Data resource | Result of searching natural products | Result of searching biological sources. |
|---------------|--------------------------------------|----------------------------------------|
| Super Natural II | Molecular properties | Not available |
| TCM Database | Molecular properties | Only some of Chinese Medicinal Herbs available |
| Reaxys | Substances and documents | Documents |
| TCMID | Not available | Only some of Traditional Chinese Medicines available |
| NuBBE database | Number of results | Number of results |
| TIPdb | No result | No result |
| NPBS | Molecular properties and biological sources | Natural products and molecular properties |

Conclusions

Compare to the top databases of natural product (Table 1.), NPBS database has not exceed their volume or coverage, but has greater clarity about the relationships between natural products and biological sources. With continuous upgrading and optimizing of the text-mining system, we might have an efficient and low-cost tool to expand the volume of NPBS database in the future. However, we haven’t concentrated on the classification and analysis of the biological source data. Nevertheless, current datasets in NPBS database 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.

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 may 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. Adopters can also use the natural product structures from a specific biological source for dereplication with experimental data, to prevent re-isolation and re-characterization of already known molecules. 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

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 National Natural Science Foundation of China [21805303], CSDB [XXH135] and SGST [18DZ2294000].

Acknowledgements

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

References

1. Butler, M.S. (2004) The Role of Natural Product Chemistry in Drug Discovery. J. Nat. Prod. 67, 2141–2153.
2. Dayan, F.E., Cantrell, C.L. and Duke, S.O. (2009) Natural products in crop protection. Bioorg. Med. Chem., 17, 4022–4034.
3. Zhang, X., Jiang, M., Niu, N. et al. (2017) Natural-Product-Derived Carbon Dots: From Natural Products to Functional Materials. ChemSusChem, 11, 11–24.
4. Banerjee, P., Erehman, J., Gohlke, B.O. et al. (2014) Super Natural II—a database of natural products. Nucleic Acids Res., 43, D935–D939.
5. He, M., Yan, X., Zhou, J. et al. (2001) Traditional Chinese Medicine Database and Application on the Web. J. Chem. Inf. Model., 41, 273–277.
6. Xue, R., Fang, Z., Zhang, M. et al. (2012) TCMID: traditional Chinese medicine integrative database for herb molecular mechanism analysis. Nucleic Acids Res., 41, D1089–D1095.
7. Ehrman, T.M., Barlow, D.J. and Hylands, P.J. (2007) Phytochemical Informatics of Traditional Chinese Medicine and Therapeutic Relevance. J. Chem. Inf. Model., 47, 2316–2334.
8. Valli, M., dos Santos, R.N., Figueira, L.D. et al. (2013) Development of a Natural Products Database from the Biodiversity of Brazil. J. Nat. Prod., 76, 439–444.
9. Lin, Y.C., Wang, C.C., Chen, I.S. et al. (2013) A Database of Anticancer, Antiplatelet, and Antituberculosis Phytochemicals from Indigenous Plants in Taiwan. Sci. World J., 2013, 736386.
10. Choi, H., Cho, S.Y., Pak, H.J. et al. (2017) NPCARE: Database of Natural Products and Fractional Extracts for Cancer Regulation. J. Cheminf., 9, 2.
11. Ntie Kang, F., Telukunta, K.K., Döring, K. et al. (2017) NANPDB: A Resource for Natural Products from Northern African Sources. J. Nat. Prod., 80, 2067–2076.
12. Klementz, D., Döring, K., Lucas, X. et al. (2016) StreptomeDB 2.0 - an Extended Resource of Natural Products Produced by Streptomycetes. *Nucleic Acids Res.*, **44**, D509–D514.

13. Gu, J., Gui, Y., Chen, L. et al. (2013) Use of Natural Products as Chemical Library for Drug Discovery and Network Pharmacology. *PLoS ONE*, **8**, e62839, 10.1371/journal.pone.0062839.

14. Chen, Y., de Bruyn Kops and C., Kirchmair, J. (2017) Data Resources for the Computer-Guided Discovery of Bioactive Natural Products. *J. Chem. Inf. Model.*, **57**, 2099–2111.

15. Medema, M.H. and Fischbach, M.A. Computational approaches to natural product discovery. *Nat. Chem. Biol.*, **11**, 639–648.

16. Rodrigues, T., Reker, D., Schneider, P. et al. (2016) Counting on natural products for drug design. *Nat. Chem.*, **8**, 531–541.

17. Xu, T., Chen, W., Zhao, Y., Dai, J., Li Y. and Zhou J. (2020) Patent CN201911006389.X.

18. Wang, J., Ma, Z., Wang, G., Liu, J. et al. (2019) Study on secondary metabolites of endophytic fungus Alternaria alternate from Paeonia lactiflora. *Zhongcaoyao*, **50**, 1061-1065. 10.7501/j.issn.0253-2670.2019.05.006.

19. Lowe, D.M., Corbett, P.T., Murray-Rust, P. et al. (2011) Chemical Name to Structure: OPSIN, an Open Source Solution. *J. Chem. Inf. Model.*, **51**, 739–753.

20. Li, W., Gu, Z., Yang, Y. et al. (2014) Non-volatile taste components of several cultivated mushrooms. *Food Chem.*, **143**, 427–431.

21. Heller, S.R., McNaught, A., Pletnev, I., et al. (2015) InChI, the IUPAC International Chemical Identifier. *J. Cheminf.*, **7**, 23.

22. Tu, Y. (2011) The discovery of artemisinin (qinghaosu) and gifts from Chinese medicine. *Nat. Med.*, **17**, 1217–1220.

23. Butterweck, V., Jürgenliemk, G., Nahrstedt, A. et al. (2000) Flavonoids from Hypericum perforatum Show Antidepressant Activity in the Forced Swimming Test. *Planta. Med.*, **66**, 3–6.

24. Wagner, K.H. and Elmadfa, I. (2003) Biological relevance of terpenoids. *Ann. Nutr. Metab.*, **47**, 95-106.

25. Saeidnia, S., Manayi, A., Gohari, A.R. et al. (2014) The story of beta-sitosterol-a review. *Eur. J. Med. Plants*, **4**,590-609.

26. Lipinski, C.A., Lombardo, F., Dominy, B.W. et al. (2012) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.*, **64**, 4–17.

**Figures**
Figure 1

The coverage of the journals involved in NPBS database
Figure 2

The design of the text-mining system. (A) Flow process of the data text-mining system. (B) Demonstration of the procedure for data acquisition.
Figure 3

The entity relationship diagram of NPBS database
Figure 4

The molecular features of the natural products in NPBS database. (A) Number of rings. (B) Structural features. (C) Number of Atoms.
Figure 5

The “drug-likeness” features of the natural products in NPBS database.