NPASS: natural product activity and species source database for natural product research, discovery and tool development

Xian Zeng1,2, Peng Zhang2, Weidong He2, Chu Qin2, Shangying Chen2, Lin Tao2,3, Yali Wang2, Ying Tan1, Dan Gao1, Bohua Wang4,5, Zhe Chen3, Weiping Chen4,*, Yu Yang Jiang1,* and Yu Zong Chen1,2,*

1Breeding Base-Shenzhen Key Laboratory of Chemical Biology, the Graduate School at Shenzhen, Tsinghua University, Shenzhen Kivita Innovative Drug Discovery Institute, Shenzhen 518055, PR China, 2Bioinformatics and Drug Design group, Department of Pharmacy, National University of Singapore, Singapore 117543, Singapore, 3Zhejiang Key Laboratory of Gastro-intestinal Pathophysiology, Zhejiang Hospital of Traditional Chinese Medicine, Zhejiang Chinese Medical University, School of Medicine, Hangzhou Normal University, Hangzhou 310006, RP China, 4Key Lab of Agricultural Products Processing and Quality Control of Nanchang City, Jiangxi Agricultural University, Nanchang 330045, PR China and 5College of Life and Environmental Sciences, Collaborative Innovation Center for Efficient and Health Production of Fisheries in Hunan Province, Hunan University of Arts and Science, Changde, Hunan 415000, PR China

Received August 15, 2017; Revised October 13, 2017; Editorial Decision October 16, 2017; Accepted October 18, 2017

ABSTRACT

There has been renewed interests in the exploration of natural products (NPs) for drug discovery, and continuous investigations of the therapeutic claims and mechanisms of traditional and herbal medicines. In-silico methods have been employed for facilitating these studies. These studies and the optimization of in-silico algorithms for NP applications can be facilitated by the quantitative activity and species source data of the NPs. A number of databases collectively provide the structural and other information of ∼470 000 NPs, including qualitative activity information for many NPs, but only ∼4000 NPs are with the experimental activity values. There is a need for the activity and species source data of more NPs. We therefore developed a new database, NPASS (Natural Product Activity and Species Source) to complement other databases by providing the experimental activity values and species sources of 35 032 NPs from 25 041 species targeting 5863 targets (2946 proteins, 1352 microbial species and 1227 cell-lines). NPASS contains 446 552 quantitative activity records (e.g. IC50, Ki, EC50, GI50 or MIC mainly in units of nM) of 222 092 NP-target pairs and 288 002 NP-species pairs. NPASS, http://bidd2.nus.edu.sg/NPASS/, is freely accessible with its contents searchable by keywords, physicochemical property range, structural similarity, species and target search facilities.

INTRODUCTION

Modern drug discovery has been benefited from nature (1,2), with >50% of approved drugs provided by or derived from nature (3–5). There have been revitalized interests in the exploration of natural products (NPs) for drug discovery (1,6). Continuous efforts have been directed at the studies of the therapeutic claims and mechanisms of the traditional and herbal medicines (7) used by large populations in the world (8). These efforts can be facilitated by the availability of the structural, functional and species source data of NPs, particularly the quantitative activity data. The discovery and functional investigation of bioactive NPs have been facilitated by the in-silico chemoinformatics (9), molecular modeling and docking (10,11), quantitative structure–activity relationship (12), and machine learning (13,14) methods. The usefulness of the in-silico algorithms in the study of NPs can be further improved if they can be optimized by using the structural and quantitative activity data of the NPs.

It has been reported that about a million NPs are known (15), many of which have been experimentally studied for determining their activities. The structure, activity and species source data of these NPs is highly useful for NP dis-
covery, investigation, and in-silico tool development. Established chemical databases such as ChEMBL (16), PubChem Bioassay (17) and BindingDB (18) provide experimentally-determined quantitative activity data (e.g. IC50, ki, GI50, MIC values etc.) for a large number of chemical molecules. But, a small portion of these molecules (e.g. ∼1200 compounds in ChEMBL) are explicitly labelled as NPs and without annotation of their species sources.

On the other hand, a number of NP databases have been developed for providing comprehensive information about NPs. These include general databases such as SuperNatural (∼325 000 NPs with 2D structure, physicochemical properties, and the predicted toxicity and targets) (19), UNPD (∼229 000 NPs with 3D structures, 31 000 NPs with species source annotation) (20), ZINC (∼80 000 NPs with 2D structures) (21) and specialized databases of NPs of specific functional classes. The first group of specialized databases are of specific indigenous medicines such as TCM-ID (22), TCM@Taiwan/iSMART (23,24), TC MID (25), TCMSP (26) for Traditional Chinese Medicine, TM-MC for Asian traditional medicines (27), NuBBE for NPs from Brazil medicinal plants (28), SANCDB (29) and AfroCancer (30) for NPs from South African regions. The second group of specialize databases are for specific NP classes such as HIT for herbal ingredients and their targets (31), NPACT for anticancer NPs (32) and BioPhyMol for anti-mycobacterial NPs (33). However, a small portion (∼4000) of the NPs in these general and specialized databases are provided with the experimentally determined quantitative activity values.

There is a need for the experimentally-determined quantitative activity data and species source information for more NPs. To complement the existing databases in catering this need, we developed a new database, NPASS Natural Product Activity and Species Source database, for providing the literature-reported experimentally-determined activity values and species sources of significantly higher number of NPs than those provided in the existing databases. The activity and species source of the NPs were obtained from the comprehensive literature search and manual reviews, with particular focus on the experimentally-determined activity (e.g. IC50, Ki, EC50, GI50 and MIC) values of the NPs against macromolecule or cell targets, and the taxonomy of the species sources of these NPs together with collection location and date while available. The contents of NPASS can be conveniently accessed by multiple search modes including keywords, physicochemical property range, structural similarity to the NPASS NP entries and additionally approved/c Clinical trial drugs, species and targets.

DATA COLLECTION AND PROCESSING


The 2D or 3D structures of the identified NPs were searched from the PubChem (17), ZINC (21), ChEMBL (16) and BindingDB (18), using the title, author name, and/or NP name/structure, or PubChem compound links provided by PubMed of the respective literatures. For those NPs not found in these databases but with structures provided in the respective literatures, their structures were drawn by using Marvin Sketch software (Marvin Sketch, Chem Axon). The NP structures in multiple formats (sdf, mol, inchi, smiles etc.) were uniformly converted into canonical SMILES using OpenBabel (version 2.4.1) (34) with the configuration of ‘Remove all but the largest contiguous fragment’. The derived SMILES strings were then used to generate the InChI, InChIKey, MOL formats using OpenBabel. Duplicates were removed based on the InChIKey strings. It was reported that GRAPE/GARLIC algorithm and circular fingerprint (especially the FCFP6) show better performance than conventional fingerprint algorithms in natural product similarity search (35,36). Given that there is no open-accessed software for calculating GRAPE/GARLIC fingerprints, the FCFP6 fingerprint was used to code compounds for similarity calculation. Besides, the widely used PubChem 881-bit substructure fingerprint was also adopted as an option in the database. Chemical structure similarity between compounds was defined by using Tanimoto coefficient ($T_c$) with the compounds represented by the fingerprints. These fingerprints were computed by using chemfp toolkit (37).

The species sources of the identified NPs were collected from the respective literatures or by using the NP name/structure to search the TCM-ID (22), TCM@Taiwan (23), TC MID (25), TCMSP (26), TM-MC (27), Streptomyces (38) and HerDing (39) databases. While available, we also collected the collection location, earliest collection time (format: Year-Month, such as 2015-MAR), and organism part of the species for extracting the NP. The earliest isolation year of NPs was tentatively determined based on the publication time of the literature that either claimed the NP as a novel structure or was the first paper reporting the NP. Synthetic gene clusters associated to NPs were collected from MIBiG database (40). The phylogenetic information of the species sources of the NPs was obtained from NCBI Taxonomy database by querying the database with the respective species name. The approved and clinical trial drugs were used in NPASS for facilitating the users to
find the drugs that are similar in structure to a user-input compound. These drugs were obtained from the TTD version 2016 (41), ChEMBL version 23 (16), DrugBank version 5.0 (42), and IUPHAR/BPS version 2017.4 (43) after removing inorganic drugs and duplicates. Drug structure (SMILES, InChi, InChI-Key), highest clinical development stage (Approved/Phase 1–3/Withdrawn), and crosslinks to other databases (ChEMBL, ChEBI (44), PubChem, DrugBank, TTD, IUPHAR/BPS, KEGGdrug (45), PharmGKB (46) and CAS-Number) were curated.

DATABASE CONTENTS, STRUCTURE AND ACCESS

NPASS, freely accessible at http://bidd.nus.edu.sg/NPASS/, currently contains 35 032 unique NPs from 25 041 species targeting 5,863 targets, with a total of 288 002 NP-species pairs and 222 092 NP-target pairs (Figure 1A). NPs are classified into 18 different chemical superclasses using ClassyFire webserver (47), which includes ‘Alkaloids and derivatives’, ‘Benzenoids’, ‘Lipids and lipid-like molecules’, ‘Lignans, neolignans and related compounds’, and so on. The distribution of NPs, active NPs, and potent NPs in each superclasses can be found in Supplementary Table S1. The species sources of these NPs are from 6,814 genus in the kingdom or super-kingdom of viridiplantae (67.8%), metazoan (9.4%), fungi (7.9%) and bacteria (6.7%) (Figure 1B). The distribution of NPs, active NPs, and potent NPs in each super-kingdom or kingdom can be found in Supplementary Table S2. 59.6% and 93.1% species sources are with identifiable taxonomic information at species and genus level, respectively, which were obtained through matching species name against NCBI Taxonomy database. For the remaining 6.9% species without taxonomic information at genus level, querying the NCBI Taxonomy database with their species name were without returns, and these entries were subject to further manual evaluation to resolve this problem. Around 71% of NPs are annotated with source species directly from publications, and source species of the remaining NPs are curated from existing databases (Supplementary Table S3). The 5863 targets include 2946 proteins, 121 unspecified members of protein families, 143 protein complexes, 8 nucleic acids, 5 unspecified subcellular targets, 1352 microbial and pathogenic organisms, 1227 cell-lines and 51 tissue targets (Figure 1C). Protein targets are classified into categories, such as enzyme, membrane receptor, ion channel, and transporter. While the enzymes are further classified according to EC classification system. Cell line targets includes 900 cancer cell lines which are further categorized based on disease relevance, such as 161 lung cancer cell lines, 98 brain cancer cell lines, 98 brain cancer cell lines, and 87 leukemia cancer cell lines. There are 446 552 quantitative activity records in terms of inhibition concentration IC50/IC90/ICD50 (23.2%), activity concentration AC/AC40/AC50 (22.4%), microbial inhibitory or lethal concentration MIC/MFC/MBC/FC (17.2%), growth inhibitory concentration GI/GI50/TGI (16.9%), percentage inhibition at fixed concentration (5.4%), EC50/EC90/ED50/ED90 (5.0%), equilibrium inhibition constant Ki (2.7%)

Figure 1. Statistics of NPASS contents. (A) Statistics of data entries of each data category. Distributions of source species kingdom/superkingdom (B), target types (C) and top 5 activity categories (D) were summarized above.
lethal concentration LC/LC50/LC90/LD50/LD90 (1.7%), inhibition zone IZ (1.7%), equilibrium binding constant $K_d$ (1.5%), inhibition/activity ratios/ratio IC50/ratio EC50/ratio $K_i$ (1.5%), cytotoxic concentration CC25/CC50/CC90/CC100 (0.4%), and toxic concentration TC50/TD50 (0.1%) (Figure 1D).

The NPASS database was developed on MySQL database and PHP server software. Its web-interfaces were built by using HTML, PHP and JavaScript, and were designed to enable the access of its entries by NP, target or species source using multiple browse and search facilities. While applicable, the NP entries are cross-linked to PubChem, ChEMBL, ZINC, MIBiG and SuperNatural databases. Their species sources are cross-linked to NCBI Taxonomy database. Their targets are cross-linked to Uniprot, ChEMBL, TTD, DrugBank and IUPHAR/BPS. The relevant literatures of the quantitative activity data and species source information are provided by the PubMed identifiers and cross-linked to PubMed. In the NPASS main page, users can click the browse button or search button to visit the respective page (Figure 2A). In the browse page, users can browse the database by NP names, initial of NP names, target names, initial of target names, target types (proteins, unspecified members of protein families, protein complexes, nucleic acids, subcellular targets, microbial and pathogenic organisms, cell lines, and tissue targets), species source names, initial of species sources, range of molecular weights of the NPs. All data can be freely and conveniently downloaded from the respective entry page. A download summary table allows users to download data of interest by selecting specific data section.

In the NPASS search page, users can search the database contents by NP, target or species source respectively (Figure 2B). NP search can be conducted by inputting an NP name or by selecting the range of one or more of the physicochemical properties AlogP, molecular weight, number of hydrogen bond acceptors HBA, number of hydrogen bond donors HBD, and number of rotatable bonds. NP search
may also be conducted by inputting its structure via the SMILES string input field, or by drawing its structure using the JSME Molecular Editor (48) provided in the ‘Search by Structure’ section of the search page. Users can perform chemical similarity search by selecting a user-defined similarity cut-off in terms of the Tanimoto coefficient $T_c$. Calculation of $T_c$ between query molecules and NPs in NPASS is achieved based on chemfp toolkit (37). The default cutoff is high structural similarity $T_c = 0.85$ (49). Other frequently-used cutoffs include intermediate structural similarity $T_c = 0.7$ and remote structural similarity $T_c = 0.56$ (49). The identified structurally similar NPs are given in the order of high, intermediate and remote similarity. Batch search of multiple NPASS entries can be performed by inputting a list of NP names, compound IDs (can be NPASS NP IDs, ChEMBL IDs, PubChem CIDs) or SMILES strings in the ‘Batch Search’ box provided in the ‘Search by Property’ section of the search page. Species search can be made by inputting the species name or taxonomic ID (NCBI Taxonomic IDs). Target search can be performed by inputting a target name or target ID (i.e. Uniprot ID, NPASS Target ID). A target name can be (i) the name of a protein, (ii) the name of a protein family for the targets that are unspecified members of that family, (iii) the name of a protein complex, (iv) the name of a nucleic acid, (v) subcellular component name (i.e. ribosome, proteasome), (vi) the name of a microbial or pathogenic organism (e.g. Human immunodeficiency virus 1, Plasmodium falciparum), (vii) the name of a cell line (e.g. MDA-MB-231, MCF7) and (viii) the name of a tissue (e.g. lung, plasma, brain).

**PERSPECTIVES**

NPASS complements other chemical and NP databases (19–23.25–33.38) in providing the experimentally-determined quantitative activity data and species sources information for more diverse sets of NPs. The expanded coverage of the quantitative activity and other important NP data (19–23.25–33.38) together with the knowledge of the mechanisms (2.50), chemical properties (5.51) and taxonomic profiles (4.52) of the NPs can better facilitate NP-based drug discovery (1.2.6), mechanism study (2.7.50) and in-silico tool development (9.10,12–14). There is a need for more comprehensive mining of the literatures for the experimentally-determined activity data and species source information of NPs, and the exploration of these data for optimizing in-silico algorithms to make them more useful tools for discovering, modeling, predicting and analyzing NPs. Collective efforts are also needed to expand the current databases to more fully cover the structural, functional, phylogenetic and mechanistic data of the NPs.

**SUPPLEMENTARY DATA**

Supplementary Data are available at NAR Online.

**FUNDING**


**Conflict of interest statement.** None declared.

**REFERENCES**
