Published online 2 November 2017 *Nucleic Acids Research, 2018, Vol. 46, Database issue* ***D1217–D1222***
doi: 10.1093/nar/gkx1026

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

Xian Zeng[1,2], Peng Zhang[2], Weidong He[2], Chu Qin[2], Shangying Chen[2], Lin Tao[2,3], Yali Wang[2], Ying Tan[1], Dan Gao[1], Bohua Wang[4,5], Zhe Chen[3], Weiping Chen[4,*], Yu Yang Jiang[1,*] and Yu Zong Chen[1,2,*]

[1]Breeding Base-Shenzhen Key Laboratory of Chemical Biology, the Graduate School at Shenzhen, Tsinghua University, Shenzhen Kivita Innovative Drug Discovery Institute, Shenzhen 518055, PR China, [2]Bioinformatics and Drug Design group, Department of Pharmacy, National University of Singapore, Singapore 117543, Singapore, [3]Zhejiang 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, [4]Key Lab of Agricultural Products Processing and Quality Control of Nanchang City, Jiangxi Agricultural University, Nanchang 330045, PR China and [5]College 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-

*To whom correspondence should be addressed. Tel: +65 6516 6877; Fax: +65 6774 6756; Email: phacyz@nus.edu.sg
Correspondence may also be addressed to Y.Y. Jiang. Tel: +86 755 2603 6430; Fax: +86 755 2603 6430; Email: jiangyy@sz.tsinghua.edu.cn
Correspondence may also be addressed to W.P. Chen. Tel: +86 791 8381 3420; Fax: +86 791 8381 3655; Email: iaochen@163.com

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), TCMID (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/clinical trial drugs, species and targets.

DATA COLLECTION AND PROCESSING

The NPs with available experimentally-determined quantitative activity values were searched from PubMed database (https://www.ncbi.nlm.nih.gov/pubmed/) by using the combinations of the keywords 'natural product', 'NP', 'nature', 'marine', 'plant', 'microbe', 'microbial', 'bacterium', 'bacteria', 'bacterial', 'fungus', 'fungi', 'fungal', 'species', 'traditional medicine', 'medicinal', 'indigenous', 'folk', 'herb', 'herbal', 'herbalism', 'Chinese medicine', 'TCM', 'Ayurveda', 'activity', 'active', 'bioactive', 'potent', 'potency', 'IC50', 'IC90','Ki', 'Kd', 'EC50', 'EC90', 'GI50', AC50, 'AC90', 'MIC' and 'IZ'. The searched literatures were evaluated for finding the NPs with both common/taxonomic name of species sources and the quantitative activity data. The quantitative activity is in types of inhibition concentration IC50/IC90/ID50, equilibrium inhibition constant K_i, equilibrium binding constant K_d, activity concentration AC/AC40/AC50, percentage inhibition at fixed concentration, microbial inhibitory or lethal concentration MIC/MFC/MBC/FC, growth inhibitory concentration GI/GI50/TGI, EC50/EC90/ED50/ED90, lethal concentration LC/LC50/LC90/LD50/LD90, inhibition zone IZ, inhibition/activity ratios ratio IC50/ratio EC50/ratio K_i, cytotoxic concentration CC25/CC50/CC90/CC100 and toxic concentration TC50/TD50 values against a macromolecular target, cell or microbe.

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), TCMID (25), TCMSP (26), TM-MC (27), StreptomeDB (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

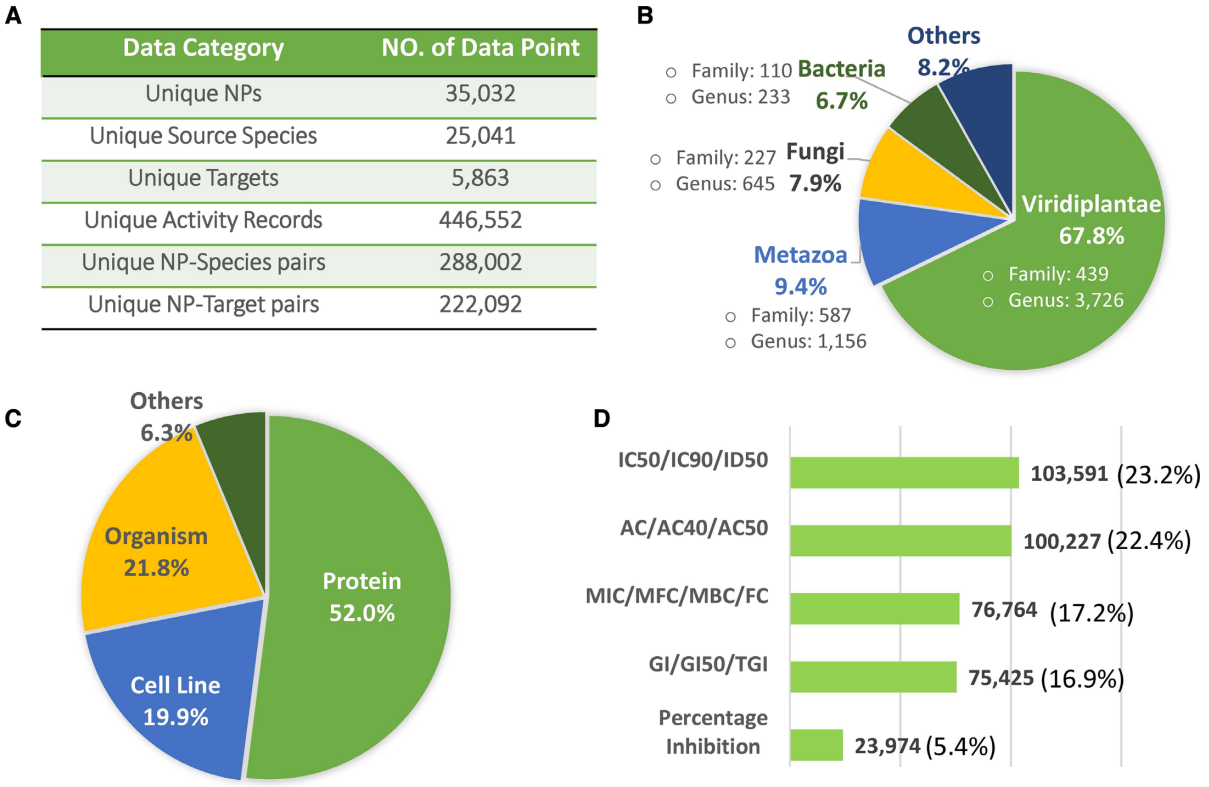

Figure 1. Statistics of NPASS contents. (**A**) Statistics of data entries of each data category. Distributions of source species kingdom/superkingdom (**B**), target types (**C**) and top 5 activity categories (**D**) were summarized above.

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, InChIKey), highest clinical development stage (Approved/Phase1–3/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://bidd2.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, and 87 leukemia cancer cell lines. There are 446 552 quantitative activity records in terms of inhibition concentration IC50/IC90/ID50 (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 K_i (2.7%),

Figure 2. Search modes provided in NPASS. (**A**) Users can browse the database by natural product, species sources and targets and click each chart column to access specific data. Besides, a hierarchical tree-like mode is provided to easily locate specific category which classified based on taxonomic information of proteins, species, and compounds. (**B**) NPASS can be searched by range of properties, structure, species sources and targets.

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

Singapore Academic Research Fund (in part) [R-148-000-208-112, R-148-000-230-114, R-148-000-239-114]; National Basic Research Program [2013CB967204]; National Natural Science Foundation of China [81325021]; Shenzhen Municipal Government [JSGG20141016150327538, 20150113A0410006]; Shenzhen Reform Commission (Disciplinary Development Program for Chemical Biology) and China Scholarship Council. Funding for open access charge: National Basic Research Program [2013CB967204]; National Natural Science Foundation of China [81325021].
Conflict of interest statement. None declared.

REFERENCES

1. Li,J.W. and Vederas,J.C. (2009) Drug discovery and natural products: end of an era or an endless frontier? *Science*, **325**, 161–165.
2. Tao,L., Zhu,F., Qin,C., Zhang,C., Xu,F., Tan,C.Y., Jiang,Y.Y. and Chen,Y.Z. (2014) Nature's contribution to today's pharmacopeia. *Nat. Biotechnol.*, **32**, 979–980.
3. Newman,D.J. and Cragg,G.M. (2016) Natural products as sources of new drugs from 1981 to 2014. *J. Nat. Prod.*, **79**, 629–661.
4. Zhu,F., Qin,C., Tao,L., Liu,X., Shi,Z., Ma,X., Jia,J., Tan,Y., Cui,C., Lin,J. *et al.* (2011) Clustered patterns of species origins of nature-derived drugs and clues for future bioprospecting. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 12943–12948.
5. Tao,L., Zhu,F., Qin,C., Zhang,C., Chen,S., Zhang,P., Zhang,C., Tan,C., Gao,C., Chen,Z. *et al.* (2015) Clustered distribution of natural product leads of drugs in the chemical space as influenced by the privileged target-sites. *Sci. Rep.*, **5**, 9325.
6. Harvey,A.L., Edrada-Ebel,R. and Quinn,R.J. (2015) The re-emergence of natural products for drug discovery in the genomics era. *Nat. Rev. Drug Discov.*, **14**, 111–129.
7. Leonti,M. and Casu,L. (2013) Traditional medicines and globalization: current and future perspectives in ethnopharmacology. *Front. Pharmacol.*, **4**, 92.
8. Kayne,S.B. (2010) *Traditional Medicine: A Global Perspective*. Pharmaceutical Press, London.
9. Rodrigues,T., Reker,D., Schneider,P. and Schneider,G. (2016) Counting on natural products for drug design. *Nat. Chem.*, **8**, 531–541.
10. Chen,X., Ung,C.Y. and Chen,Y. (2003) Can an in silico drug-target search method be used to probe potential mechanisms of medicinal plant ingredients? *Nat. Prod. Rep.*, **20**, 432–444.
11. Chen,Y.Z. and Ung,C.Y. (2002) Computer automated prediction of potential therapeutic and toxicity protein targets of bioactive compounds from Chinese medicinal plants. *Am. J. Chin. Med.*, **30**, 139–154.
12. Stahura,F.L., Godden,J.W., Xue,L. and Bajorath,J. (2000) Distinguishing between natural products and synthetic molecules by descriptor Shannon entropy analysis and binary QSAR calculations. *J. Chem. Inf. Comput. Sci.*, **40**, 1245–1252.
13. Rupp,M., Bauer,M.R., Wilcken,R., Lange,A., Reutlinger,M., Boeckler,F.M. and Schneider,G. (2014) Machine learning estimates of natural product conformational energies. *PLoS Comput. Biol.*, **10**, e1003400.
14. Ung,C.Y., Li,H., Cao,Z.W., Li,Y.X. and Chen,Y.Z. (2007) Are herb-pairs of traditional Chinese medicine distinguishable from others? Pattern analysis and artificial intelligence classification study of traditionally defined herbal properties. *J. Ethnopharmacol.*, **111**, 371–377.
15. Demain,A.L. and Vaishnav,P. (2011) Natural products for cancer chemotherapy. *Microb. Biotechnol.*, **4**, 687–699.
16. Gaulton,A., Hersey,A., Nowotka,M., Bento,A.P., Chambers,J., Mendez,D., Mutowo,P., Atkinson,F., Bellis,L.J., Cibrian-Uhalte,E. *et al.* (2017) The ChEMBL database in 2017. *Nucleic Acids Res.*, **45**, D945–D954.
17. Wang,Y., Bryant,S.H., Cheng,T., Wang,J., Gindulyte,A., Shoemaker,B.A., Thiessen,P.A., He,S. and Zhang,J. (2017) PubChem BioAssay: 2017 update. *Nucleic Acids Res*, **45**, D955–D963.
18. Gilson,M.K., Liu,T., Baitaluk,M., Nicola,G., Hwang,L. and Chong,J. (2016) BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res.*, **44**, D1045–D1053.

19. Banerjee,P., Erehman,J., Gohlke,B.O., Wilhelm,T., Preissner,R. and Dunkel,M. (2015) Super Natural II–a database of natural products. *Nucleic Acids Res.*, **43**, D935–D939.
20. Gu,J., Gui,Y., Chen,L., Yuan,G., Lu,H.Z. and Xu,X. (2013) Use of natural products as chemical library for drug discovery and network pharmacology. *PLoS One*, **8**, e62839.
21. Sterling,T. and Irwin,J.J. (2015) ZINC 15—ligand discovery for everyone. *J. Chem. Inf. Model.*, **55**, 2324–2337.
22. Wang,J.F., Zhou,H., Han,L.Y., Chen,X., Chen,Y.Z. and Cao,Z.W. (2005) Traditional Chinese medicine information database. *Clin. Pharmacol. Ther.*, **78**, 92–93.
23. Chen,C.Y. (2011) TCM Database@Taiwan: the world's largest traditional Chinese medicine database for drug screening in silico. *PLoS One*, **6**, e15939.
24. Chang,K.W., Tsai,T.Y., Chen,K.C., Yang,S.C., Huang,H.J., Chang,T.T., Sun,M.F., Chen,H.Y., Tsai,F.J. and Chen,C.Y. (2011) iSMART: an integrated cloud computing web server for traditional Chinese medicine for online virtual screening, de novo evolution and drug design. *J. Biomol. Struct. Dyn.*, **29**, 243–250.
25. Xue,R., Fang,Z., Zhang,M., Yi,Z., Wen,C. and Shi,T. (2013) TCMID: Traditional Chinese Medicine integrative database for herb molecular mechanism analysis. *Nucleic Acids Res.*, **41**, D1089–D1095.
26. Ru,J., Li,P., Wang,J., Zhou,W., Li,B., Huang,C., Li,P., Guo,Z., Tao,W., Yang,Y. *et al.* (2014) TCMSP: a database of systems pharmacology for drug discovery from herbal medicines. *J. Cheminform.*, **6**, 13.
27. Kim,S.K., Nam,S., Jang,H., Kim,A. and Lee,J.J. (2015) TM-MC: a database of medicinal materials and chemical compounds in Northeast Asian traditional medicine. *BMC Complement Altern. Med.*, **15**, 218.
28. Valli,M., dos Santos,R.N., Figueira,L.D., Nakajima,C.H., Castro-Gamboa,I., Andricopulo,A.D. and Bolzani,V.S. (2013) Development of a natural products database from the biodiversity of Brazil. *J. Nat. Prod.*, **76**, 439–444.
29. Hatherley,R., Brown,D.K., Musyoka,T.M., Penkler,D.L., Faya,N., Lobb,K.A. and Tastan Bishop,O. (2015) SANCDB: a South African natural compound database. *J. Cheminform.*, **7**, 29.
30. Ntie-Kang,F., Simoben,C.V., Karaman,B., Ngwa,V.F., Judson,P.N., Sippl,W. and Mbaze,L.M. (2016) Pharmacophore modeling and in silico toxicity assessment of potential anticancer agents from African medicinal plants. *Drug Des Devel Ther*, **10**, 2137–2154.
31. Ye,H., Ye,L., Kang,H., Zhang,D., Tao,L., Tang,K., Liu,X., Zhu,R., Liu,Q., Chen,Y.Z. *et al.* (2011) HIT: linking herbal active ingredients to targets. *Nucleic Acids Res.*, **39**, D1055–D1059.
32. Mangal,M., Sagar,P., Singh,H., Raghava,G.P. and Agarwal,S.M. (2013) NPACT: Naturally Occurring Plant-based Anti-cancer Compound-Activity-Target database. *Nucleic Acids Res.*, **41**, D1124–D1129.
33. Sharma,A., Dutta,P., Sharma,M., Rajput,N.K., Dodiya,B., Georrge,J.J., Kholia,T., Consortium,O. and Bhardwaj,A. (2014) BioPhytMol: a drug discovery community resource on anti-mycobacterial phytomolecules and plant extracts. *J. Cheminform.*, **6**, 46.
34. O'Boyle,N.M., Banck,M., James,C.A., Morley,C., Vandermeersch,T. and Hutchison,G.R. (2011) Open Babel: an open chemical toolbox. *J. Cheminform.*, **3**, 33.
35. Johnston,C.W., Skinnider,M.A., Dejong,C.A., Rees,P.N., Chen,G.M., Walker,C.G., French,S., Brown,E.D., Bérdy,J., Liu,D.Y. *et al.* (2016) Assembly and clustering of natural antibiotics guides target identification. *Nat. Chem. Biol.*, **12**, 233–239.
36. Skinnider,M.A., Dejong,C.A., Franczak,B.C., McNicholas,P.D. and Magarvey,N.A. (2017) Comparative analysis of chemical similarity methods for modular natural products with a hypothetical structure enumeration algorithm. *J. Cheminform.*, **9**, 46.
37. Dalke,A. (2013) The FPS fingerprint format and chemfp toolkit. *J. Cheminformatics*, **5**, P36.
38. Klementz,D., Doring,K., Lucas,X., Telukunta,K.K., Erxleben,A., Deubel,D., Erber,A., Santillana,I., Thomas,O.S., Bechthold,A. *et al.* (2016) StreptomeDB 2.0–an extended resource of natural products produced by streptomycetes. *Nucleic Acids Res.*, **44**, D509–D514.
39. Choi,W., Choi,C.H., Kim,Y.R., Kim,S.J., Na,C.S. and Lee,H. (2016) HerDing: herb recommendation system to treat diseases using genes and chemicals. *Database (Oxford)*, **2016**, baw011.
40. Medema,M.H., Kottmann,R., Yilmaz,P., Cummings,M., Biggins,J.B., Blin,K., de Bruijn,I., Chooi,Y.H., Claesen,J., Coates,R.C. *et al.* (2015) Minimum Information about a Biosynthetic Gene cluster. *Nat. Chem. Biol.*, **11**, 625–631.
41. Yang,H., Qin,C., Li,Y.H., Tao,L., Zhou,J., Yu,C.Y., Xu,F., Chen,Z., Zhu,F. and Chen,Y.Z. (2016) Therapeutic target database update 2016: enriched resource for bench to clinical drug target and targeted pathway information. *Nucleic Acids Res.*, **44**, D1069–D1074.
42. Law,V., Knox,C., Djoumbou,Y., Jewison,T., Guo,A.C., Liu,Y., Maciejewski,A., Arndt,D., Wilson,M., Neveu,V. *et al.* (2014) DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res.*, **42**, D1091–D1097.
43. Southan,C., Sharman,J.L., Benson,H.E., Faccenda,E., Pawson,A.J., Alexander,S.P., Buneman,O.P., Davenport,A.P., McGrath,J.C., Peters,J.A. *et al.* (2016) The IUPHAR/BPS Guide to PHARMACOLOGY in 2016: towards curated quantitative interactions between 1300 protein targets and 6000 ligands. *Nucleic Acids Res.*, **44**, D1054–D1068.
44. Hastings,J., Owen,G., Dekker,A., Ennis,M., Kale,N., Muthukrishnan,V., Turner,S., Swainston,N., Mendes,P. and Steinbeck,C. (2016) ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Res.*, **44**, D1214–D1219.
45. Kanehisa,M., Sato,Y., Kawashima,M., Furumichi,M. and Tanabe,M. (2016) KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.*, **44**, D457–D462.
46. Thorn,C.F., Klein,T.E. and Altman,R.B. (2013) PharmGKB: the Pharmacogenomics Knowledge Base. *Methods Mol. Biol.*, **1015**, 311–320.
47. Djoumbou Feunang,Y., Eisner,R., Knox,C., Chepelev,L., Hastings,J., Owen,G., Fahy,E., Steinbeck,C., Subramanian,S., Bolton,E. *et al.* (2016) ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. *J. Cheminform.*, **8**, 61.
48. Bienfait,B. and Ertl,P. (2013) JSME: a free molecule editor in JavaScript. *J. Cheminform.*, **5**, 24.
49. Zhang,C., Shao,Y.M., Ma,X., Cheong,S.L., Qin,C., Tao,L., Zhang,P., Chen,S., Zeng,X., Liu,H. *et al.* (2017) Pharmacological relationships and ligand discovery of G protein-coupled receptors revealed by simultaneous ligand and receptor clustering. *J. Mol. Graph. Model.*, **76**, 136–142.
50. Salvador-Reyes,L.A. and Luesch,H. (2015) Biological targets and mechanisms of action of natural products from marine cyanobacteria. *Nat. Prod. Rep.*, **32**, 478–503.
51. Rao,H., Huangfu,C., Wang,Y., Wang,X., Tang,T., Zeng,X., Li,Z. and Chen,Y. (2015) Physicochemical profiles of the marketed agrochemicals and clues for agrochemical lead discovery and screening library development. *Mol. Inform.*, **34**, 331–338.
52. Saslis-Lagoudakis,C.H., Savolainen,V., Williamson,E.M., Forest,F., Wagstaff,S.J., Baral,S.R., Watson,M.F., Pendry,C.A. and Hawkins,J.A. (2012) Phylogenies reveal predictive power of traditional medicine in bioprospecting. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 15835–15840.