 doi:10.20944/preprints201912.0332.v1

On the redundancy of natural products public databases and where to find data in 2020 - a review on natural products databases

Dr. Maria Sorokina (ORCID 0000-0001-9359-7149)
Prof. Dr. Christoph Steinbeck (ORCID 0000-0001-6966-0814)
Institute for Inorganic and Analytical Chemistry
Friedrich-Schiller-University
Lessingstr. 8
07743 Jena

Abstract
Natural products (NPs) have been the centre of attention of the scientific community in the last decencies and the interest around them continues to grow incessantly. As a consequence, in the last 20 years, there was a rapid multiplication of various databases and collections as generalistic or thematic resources for NP information. In this review, we establish a complete overview of these resources, and the numbers are overwhelming: over 120 different NP databases and collections were published and re-used since 2000. 98 of them are still somehow accessible and only 50 are open access. The latter include not only databases but also big collections of NPs published as supplementary material in scientific publications and collections that were backed up in the ZINC database for commercially-available compounds. Some databases, even published relatively recently are already not accessible anymore, which leads to a dramatic loss of data on NPs. The data sources are presented in this manuscript, together with the comparison of the content of open ones.
With this review, we also compiled the open-access natural compounds in one single dataset a COlleCtion of Open NatUral producTs (COCONUT), which is available on Zenodo and contains structures and sparse annotations for over 400000 non-redundant NPs, which makes it the biggest open collection of NPs available to this date.

Keywords: natural products, databases

Introduction

Natural products (NPs), are broadly defined as chemicals produced by living organisms. More precise definitions of NPs exist, but they do not always meet a consensus: some of the of natural products include all small molecules that result from metabolic reactions, others classify as “NP” only products of secondary, or

 doi:10.20944/preprints201912.0332.v1

non-essential, metabolism. In this review, we made the choice to exclude molecules that participate in the primary, or essential, metabolism, such as energy or anabolic pathways, and consider only molecules that are smaller than 1500 Da, and are produced by living organisms in order to accomplish a “higher” function, such as signalling or defence. Natural products have evolved over millions of years and acquired a unique chemical diversity, which consequently results in the diversity of their biological activities and drug-like properties. Therefore, even before the rise of the modern chemical pharmacology, NPs have been used for centuries as components of traditional medicines, in particular as active components of herbal remedies. Nowadays, some of the traditional healing practices, such as Indian Ayurveda, traditional Chinese medicine or African herbal medicines, remain the primary treatment option for many people across the world, due to economic reasons, to personal beliefs or to the difficulty in accessing pharmaceutical products. In modern pharmacology too, natural products have become one of the most important resources for developing new lead compounds and scaffolds [1–3]. Every week, scientific articles in peer-reviewed journals are published describing the positive effects of NPs on the healing process of various human and animal diseases. Major classes of antibiotics and antifungals are based on NPs isolated from microorganisms. Drugs used in the treatment of various cancers, cardiovascular diseases, diabetes, and more are often natural products or their derivatives. For instance, between 1981 and 2014 over 50% of newly developed drugs were developed from NPs [1]. NPs and their derivatives are also actively studied in food [4–7], cosmetic industries [8,9] and in agriculture, with natural pesticides development [10].

This growing interest over natural products and their application resolved in uncontrollable growth of the number of published open and commercial databases, industrial catalogues, books of natural products and collections of structures provided in supplementary materials or research articles, compiling NPs from various organisms, geographical locations, targeted diseases and traditional uses. It became, therefore, a real challenge to find a complete and comprehensive open database for natural products. For instance, in the preparation of this review, we listed a total of 122 resources listing NP structures, among them 92 are open and only 50 contain molecular structures that we could retrieve for analyses and compilation.

One other major problem is the publication of structures only in graphical format, such as in the annual reviews of Marine Natural Products [2]: these are not easily retrievable to be computationally analysed and they are not automatically integrated into public molecular databases. Virtual natural products collections are therefore required for virtual screening, which is the first step in all exploratory molecular analyses and to some extent, in the discovery of NP-based drug or other types of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 December 2019 doi:10.20944/preprints201912.0332.v1

active components. For example, the prior virtual screening of known NPs can prevent loss of time with extracting and purifying samples, postponing the wet lab step to the moment of theoretical identification of best candidates. In this way, the usage of modern cheminformatics technologies allows to accelerate research and save time and money for better results.

The previous reviews on natural products databases are either too old and do not reference the actual state of NP resources [11], either focus on one particular type of application for such databases [12,13], in particular databases that can be used for dereplication [14], a particular geographic origin of NPs [15] or simply do not refer a significant part of NP resources [16].

In the present review are collected 117 NP databases and other types of virtual NP libraries, cited in the scientific literature after 2000. NP structures from 50 of them could be retrieved, and their content and overlaps between them were analysed. Finally, the non-redundant collection of natural products from these open resources has been assembled in a MongoDB COlleCtion of Open Natural prodUcTs (COCONUT).

Natural products online resources: availability and characteristics

For now, there is no globally accepted community resource for natural products, where their structures and annotations can be submitted, edited and queried by a large public, like there is UniProt [17] for proteins or NCBI Taxonomy [18] for the classification of living organisms. This leads to an impressive (122) amount of various, open and commercial, with different scope and differently structured resources for NP structures and their annotations. Mentions of NP databases, datasets and collections across publications from 2000 to 2019 and in omicX [19], a catalogue of scientific databases and software, were collected and are listed in Table1.

The databases are sorted by alphabetical order of their names and the table lists their various features such as: if they are open or commercial, if they are maintained and updated, what type of NPs they contain and their origin, the approximative number of molecular structures they contain, most recent publication of the collection, if a registration is required to access the data, if an extensive metadata is available (taxonomy of the organism producing the natural product, tissue, geographical location where it is isolated, it's application in (traditional) medicine, diseases it targets, etc) and if the download of the molecular structures for a local use (such as virtual screening) is easy.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 December 2019 doi:10.20944/preprints201912.0332.v1

For the purpose of this review, the first classification level of the NP databases is their open or commercial access. Next, among the open-access databases, we distinguish databases of metabolites (that contain natural products but also products of primary metabolism), generalistic databases, that do not limit themselves to a particular geographic location or taxonomic classification, databases containing experimental spectra of natural products (NMR, mass spectrometry) and can be used for dereplication applications, thematic databases, that focus on traditional medicine, on drug-like natural products, on the biodiversity of a particular geographic region or on a particular taxonomic group and, finally, open-access industrial catalogues, that are virtual collections of natural products that chemical companies synthesize or isolate and sell. Of course, this segregation is not the only one possible and was made here uniquely for the readability purpose.

Commercial databases

Commercial databases sell the data, access or licence, and in general, it is quite expensive [20], even for academic use (from 6600 US$ per year for the Dictionary of Natural Products [21] to over 40 000 US$ for Reaxys [22] and SciFinder [23]).

The Chemical Abstracts Service (CAS) launched in 1995 SciFinder [23], a curated database of chemical information. Originally available as desktop software, the web version of SciFinder is available since 2008. As it is CAS that assigns a unique registry number to every chemical substance described in the scientific literature since 1957, the SciFinder contains one, if not the biggest collection of curated chemicals, and, subsequently, of natural products. It is estimated that the number of NPs in SciFinder is over 300000.
Reaxys [22] is a database for substances, reactions and documents compiled and maintained by the editor Elsevier. It contains over 10^7 compounds in total, over 200000 of which are natural products.
The Dictionary of Natural products (DNP) [21] and it's autonomous sub-sections, the Dictionary of Marine Natural Products (DNMP) [24] and the Dictionary of Food Compounds [25], are the considered as the most complete and best-curated resources for NP.
NaprAlert [26] was created by researchers at the University of Chicago and contains manually curated information on natural products from literature with rich metadata. Nowadays offers limited free searchers under conditions for academic researchers.
National Institute of Standards and Technology-NIST (version 17) [27] is one of the standard reference databases for mass spectra (MS) data and is developed and maintained at the NIH in the USA. The main library contains over 250000 molecules of natural origin (the separation between primary metabolites and natural products is not clearly marked) and is only purchasable on a compact disk.

 doi:10.20944/preprints201912.0332.v1

MarinLit [28,29] is a database of marine natural products based on literature reviews and contain highly curated data that has been collected since the 1970s at the University of Canterbury, New Zealand, and since several years is maintained by the Royal Society of Chemistry (RSC). AntiMarin [30,31] is a historic database of marine natural products that have a described antibiotic activity. While it is still widely cited in thematic studies, the database itself is not accessible anymore, as was apparently merged with MarinLit.

AntiBase [32] is a comprehensive database of more than 40000 NPs from microorganisms and higher fungi with very rich metadata collected from literature and manually validated. It is not updated since 2014 and is only available for purchase on Wiley's website [33].

eBasis (Bioactive Substances in Food Information Systems) is an online, manually curated collection of 267 foods and 794 active compounds that they contain. The database offers rich and high-quality metadata on food NP activities and structures and limited free access to scientists to try the resource.

The Natural Product Discovery System (NADI) [34] contains over 3000 natural compounds from more than 15000 Malaysian plant species. Despite being developed and maintained by the University Sains Malaysia, it is not open for academic use.

ChemTCM [35] is a database of natural products from plants used in traditional Chinese herbal medicine. The original part of this dataset resides not only in the very rich metadata but also in the predicted activity of NPs against common Western therapeutic targets and their estimated molecular activity according to traditional Chinese herbal medicine categories. The database was developed at King's College London, in the UK, in part with the support of Innovation China-UK.

The Natural Products Library (NPL) [36] was described in a paper by AstraZeneca, a famous pharmaceutical company, but the data, containing at the moment of publication over 800 well-curated and annotated NPs, only remained as an in-house collection.

The Ayurveda dataset [37] was initially a published database of NPs extracted from the Indian traditional medicine plants. The link in the mentioned publication is still working but redirects to a website that provides software solutions for NP and chemistry research in general. Maybe the database is still available together with the software, but the access to it is for subscriptions only.

The Berdy's Bioactive Natural Products Database [38] database is mentioned in publications from the 2000s and early beginning of 2010s but is not accessible anymore not even for the purchase of an older version. Originally, Birdy's company was sending the database as a paper version and with the rise of accessible digital storage, on a digital medium upon order. The company doesn't seem to exist anymore.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 December 2019 doi:10.20944/preprints201912.0332.v1

Open-access databases

We could identify a total of 92 open-access natural products resources across the literature in the last 20 years. The concept of “Open-access” encourages and prioritizes free and open online access to academic information, such as data and scientific publications. For a dataset, whether in a database or attached as additional information to an article, it means that anyone can read, download, copy, distribute, print, search for and within and re-use all or parts of data that are contained in it. For this review, we have endeavoured to compile an exhaustive list of open-access natural products resources that have been cited at least ones in a peer-reviewed scientific publication after the year 2000. As the number of such sources is quite substantial (87), a thematic classification for them has been established. First, we present larger databases of organic molecules that also contain metabolites and natural products. These are followed by the presentation of databases containing molecular spectra (mass spectrometry or NMR) that can be used for the dereplication process for the identification of organic molecules and, in particular, of natural products in experimental data. Next, the scope will be narrowed with databases containing only natural products but without any taxonomic, usage or geographic selection on them. The most diverse data source category is the so-called “thematic” one: it contains databases of NPs that focus on a particular taxonomy (e.g. plants, bacteria, fungi), on a particular usage (e.g. Chinese, Indian or African traditional medicine, NPs found in food or toxic NPs) or on a particular geographic location (e.g. marine NPs, Brazilian and Mexican biodiversity NPs). Finally, are introduced industrial catalogues of natural products. These are made available by chemical companies that synthesize or purify NPs on command.

Databases of metabolites and chemicals

The first starting points in the search for structures for organic molecules are these big chemical libraries. They contain a wide range of organic compounds, and metabolites and natural products are well identifiable in them. The reference libraries, widely accepted by the scientific community as sources of reliable molecular information are: ChEBI [39], ChEMBL [40], ChemSpider [41], PubChem [42] and ChemBank [43]. ChEBI is developed and maintained at the European Bioinformatics Institute (EBI) and its main focus is chemical ontologies, i.e. structural relationships between molecules; it contains over 15000 clearly identified NPs. ChEMBL is also the product of EBI but it has a wider focus and is considered as a repository for experimentally elucidated molecular structures and, in particular, drugs and drug-like chemical; it contains over 1800 NPs, but this number is very probably underestimated because of the unclear labelling of molecules as NP in this database. PubChem is an integrated platform of small molecules and biological activities is an initiative of the US National Institute of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 December 2019 doi:10.20944/preprints201912.0332.v1

Health (NIH) and is one of the major sources for biomolecules discovery and submission. It contains over 3500 NPs, although, similarly to ChEMBL, this number is very underestimated due to the unclear labelling of compounds as natural products. ChemSpider is a chemical database offering very rich metadata, cross-references to a lot of other chemical sources and advanced search. It is maintained by the Royal Society of Chemistry and contains over 9700 easily findable NPs. ChemBank was developed by the Broad Institute of Harvard and MIT and was dedicated to the storage of raw screening data of small organic molecules. This resource is unfortunately not available anymore due to maintenance difficulties, although all data remains available for a bulk download, but is not as handy to search.

There are also databases that focus only on metabolites, chemicals that are produced by living organisms (generally, but not only through enzyme-catalyzed reactions) and that are involved in primary and secondary metabolisms. The two major and most comprehensive databases for metabolites covering most of the domains of life are KEGG [44] and MetaCyc [45]. They contain an equivalent amount of chemicals, also involved in secondary metabolism, *i.e.* natural products, but present a different point of view on data organization and have been widely compared in the literature [46]. The BRENDA database [47] focuses on enzyme activities, but also contains the compounds involved in enzyme-catalyzed reactions, and this, covering most of all known domains of life. The particularity of this database is the manually validated compounds, reactions and enzyme activities in its main part, and exhaustive taxonomic origins for enzymes and compounds; however, NPs and primary metabolites are not clearly separated in this resource, so it is difficult to estimate their respective numbers. The Chemical Structure Lookup Service (CSLS) [48] was developed for a very rapid metabolite structure lookup in an aggregated collection of more than 80 databases comprising more than 27 million unique structures in 2007. Not updated anymore, it is still possible to download the datasets, but the lookup service is not available so the extraction of natural products only requires an extensive data curation. The last database presented in this section is BiGG [49]: a platform for highly-curated genome-scale metabolic models. It contains, as parts of the metabolic models metabolites, but the distinction of primary and secondary metabolism is not clear, so it requires a lot of efforts to extract information on natural products only.

Databases for dereplication

Dereplication is one important step in experimental NP discovery as it prevents re-isolation and re-characterization of already known molecules. It consists of a lookup in databases with annotated experimental data (mainly mass spectrometry

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 December 2019 doi:10.20944/preprints201912.0332.v1

(MS) and Nuclear Magnetic Resonance (NMR) spectra) for comparison to newly obtained experimental data, and its annotation in case of found spectral identity. There are two big categories of databases used for dereplication based on the type of spectra they contain, MS and NMR.

Databases for dereplication for MS data

There are three distinct databases called “MassBank”: the MassBank of North America (MoNa) [50], the European MassBank [51] and the Japanese MSSJ MassBank [52]. The three contain reference MS spectra for metabolites and extensive metadata. MoNa tends to be favoured by the scientific community as it integrates data from more sources than the two others, contains rich and community-curated metadata and facilitates the submission of new datasets.

METLIN [53] is a database that allows the characterization of known metabolites and a technology platform for the identification of known and unknown metabolites and other chemical entities. It is a comprehensive resource containing over 1 million molecules including primary metabolites, toxins, small peptides, and natural products. METLIN's high-resolution tandem mass spectrometry (MS/MS) database, which plays a key role in the identification process, has data generated from both reference standards and their labelled stable isotope analogues, facilitated by METLIN-guided analysis of isotope-labelled microorganisms. However, it does not allow an easy download of the data, but the access to the platform is free for academic use.

The Human Metabolome Database (HMDB) [54] is a metabolomic database containing comprehensive information on human metabolites with very extensive metadata and reference spectra. It contains human-produced NPs together with NPs that are essential for the function of the human organism. However, as it is the case in a lot of previously described databases, the separation between NPs and primary metabolites is tricky.

The RIKEN MSn spectral database for phytochemicals (ReSpect) is a collection of in-house and literature MS plant natural products spectra. The website is still maintained and is usable but the last dataset has been added in 2013.

The Global Natural Products Social Molecular Networking (GNPS) [55] is a web-based knowledge base containing MS spectra for natural products only and is intended to be the base for the community-wide organization and sharing of raw, processed or identified data. In addition to providing access to spectra, it is also possible to download solely the structures of the NPs from this database.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 December 2019 doi:10.20944/preprints201912.0332.v1

Databases for dereplication for NMR data

NMRshiftDB [56] an open and peer-reviewed database for organic molecules structures and their NMR spectra. It contains a big number of easily identifiable NP spectra that makes it the reference tool for NP dereplication applications.
NMRdata [57] is a Chinese initiative for the storage and elucidation of NP structures from NMR data. Unfortunately, the main website is in Chinese and the English version is limited. To access the data one needs an account in a university that participates in the NMRdata project. At the moment of the writing of this manuscript, NMRdata contains 1167468 spectra, which theoretically makes it the biggest resource for NMR data in the world but it is under-used due to the language barrier.
NAPROC-13 [58] is a database containing 13C spectral information of over 6000 natural compounds. All data is accessible and searchable online, however, it is not possible to download the subsequent structures.
Spektraris NMR database [59] is a collection of NMR spectra that are focusing on plant natural products. The more than 400 spectra from more than 200 compounds in this database were manually transcribed from the literature. Spectra from this database are also submitted to NMRshiftDB to profit of the advanced technological aspects of the latter.

Generalistic databases of natural products

Generalistic public databases for natural products are not specialized in any particular type of NP nor on NP origins or usages. They are generally intended as catalogues for various purposes, such as in silico screening for activity prediction, molecular docking and so on. Seven generalistic public NP databases that have been active in the last 20 years have been identified from the literature.

SuperNatural II [60] is a database that contains over 300000 NPs together with their 2D structures, computed physicochemical properties and predicted toxicity. It also provides references to the chemical suppliers for the actual purchase of the molecules, but not to other chemical databases. The database is maintained but is probably not updated anymore as some of the companies selling molecules are not active anymore (such as MDPI [61]). Unfortunately, SuperNatural does not provide a bulk download, even if the download of separate MOL files for molecules is possible and erroneously doesn't contain only NPs (e.g it contains dodecahedrane, identified in this database under SN00136231 and it is not a natural product), so this resource needs to be used with caution despite its wide fame in the scientific community.

The Universal Natural Products Database (UNPD) [62] was an effort to compile all know NPs in one collection for in silico drug screening. The last accessible version of the UNPD contains over 200000 NP structures. The database is not accessible anymore through the link provided in the original publication, but a copy of the molecular structures contained in it is still maintained on the ISDB [63] website (a database for in-silico predicted MS/MS spectra for natural products).

ZINC [64] is a public access database and toolset that was initially developed to enable easy access to chemical compounds for virtual screening purposes and that became ever widely used for a big range of cheminformatic applications. It has a very clear separation of molecules in catalogues, in particular on their origin, and contains an easily searchable and retrievable collection of over 85000 natural products.

The Natural Product Activity and Species Source Database (NPASS) [65] contains over 30000 natural products from plants, bacteria, fungi and animals and is developed and maintained at the National University of Singapore. This database was created to provide a reliable source for highly curated NPs with structures, experimental activity values and the organisms that synthesize them.

RIKEN Natural Products Encyclopedia (NPEdia) [66] contains over 25000 secondary metabolites isolated from various species and annotated with rich metadata, such as molecule origin and physicochemical and biological properties. The database is still available online but is not updated since 2014.

3DMET [67] is a database that was created in 2005 in the National Institute of Agrobiological Sciences in Japan and is still maintained and updated until now. The idea of such a database came during the conversion from 2D to 3D NP structures and the errors that were occurring during it that needed manual curation. Currently, the database contains over 18000 entries, cross-referenced to the KEGG database [44], but unfortunately, the download of the structures is not possible.

The Chinese Natural Products Database (CNPD) [68] is a generalistic database created by Chinese researchers in order to facilitate the virtual screening of natural products for drug discovery purposes. This database is mentioned in a lot of papers until 2010 but is impossible to localize, as there is no URL provided in the original publication of the database and the dataset is not added as supplementary information to it. It is therefore probably incorrect to cite this database as a data source for NP, as the only possible sources found (from NeoTrident Technology Ltd) are in Chinese only.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 December 2019 doi:10.20944/preprints201912.0332.v1

One big negative point is that in ZINC, SuperNatural II and UNPD databases, the three biggest ones in terms of the number of NPs, the taxonomic nor geographic origins of the organism that produced the compound cannot be identified and in general they lack metadata and literature references.

For the completeness of this list, it is also necessary to site two major tools for the discovery and prediction of natural products from protein sequence data: antiSMASH [69] and PRISM [70]. Both are trained on, among others, NP data, but the latter is not provided directly to the public.

Thematic databases

Thematic databases for natural products focus on one particular origin or application of these secondary metabolites. Here we list databases that contain NPs produced by a particular domain of life (e.g. plants, fungi, bacteria), produced by organisms living in a particular geographical location (e.g. marine organisms, South American organisms) or by its application (traditional medicines, food or drugs). Apart from some rare exceptions, thematic databases tend to be small (less than 3000 entries) and very specialized.

In order to avoid biological provenance confusion, it needs to be noted that in some cases, NPs isolated from plants and animals can actually be synthesized by microorganisms that live on or in the host [71]. This is particularly the case of endophytes, bacteria living inside plant cells and very difficult to differentiate from the latter during preparation for metabolomics experiments [72]. Although the confusion is rare due to the improvement of identification methods and genetic approaches, it can create a bias in reproducibility of the NP isolation and needs, therefore, to be taken into account.

Natural products by the taxonomy of the synthesizing organism

Plants

KNApSaCK [73] is a comprehensive database for plant natural products that contains over 10000 retrievable 2D and 3D structures, information on the relationships between the NPs and their expressing organism(s). It is pretty difficult to navigate despite the original design choices, and it doesn't offer a bulk download of the dataset.

Collective Molecular Activities of Useful Plants (CMAUP) [74], a relatively new database, contains very extensive information on plants that are linked to human activities together with their chemical constituents, i.e. natural products. The database offers very rich metadata for NPs, such as the plants that produce them and their geographical distributions.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 December 2019 doi:10.20944/preprints201912.0332.v1

TriForC [75] is a European Union-funded project that aims for the “discovery and production of known and novel bioactive triterpenes for pharmaceutical and agrochemical development”. The database contains a pipeline for triterpenes discovery and 266 NPs together with the enzymes and pathways leading to their production. It contains metadata for the compounds, but no structures in computer-readable format nor the possibility of downloading them.
Alkamid database [76] references over 300 N-alkylamides from plants, a promising group of bioactive compounds in drug and crops research. The database is fully open and offers rich metadata, in particular, the taxonomical classification of plants that produces the NPs, but doesn’t allow a bulk download of any information from it.
The Tea Metabolome Database (TMDB) [5] is a curated and literature-based database for tea components. Not accessible anymore, it contained over 1300 constituents found in tea.

Microorganisms

StreptomeDB [77] is a collection of natural products from bacteria from the Streptomyces genus, which is very important for the production of natural bioactive compounds such as antibiotics, antitumour and immunosuppressant drugs. These bacteria are of particular importance in pharmacological research as around two-thirds of all known natural antibiotics are produced by them. While collecting data for this review, we encountered some difficulties to access the website, but the data was downloadable. In addition, an old dataset is available on ZINC.
The Natural Products Atlas (NP Atlas) [78] is maintained at the Simon Fraser University in Canada and is curated by a consortium of data curators around the world. It is designed to cover natural products from microbes (bacteria, fungi, lichens and cyanobacteria) published in the peer-reviewed literature. The resource is actively updated, allows a bulk download of all data and metadata and since September 2019 is completely open.
ProCarDB [79] is a database for carotenoids produced by bacteria. It contains over 300 compounds with rich metadata and structures but doesn’t offer any download option.
PAMDB [80] is a comprehensive Pseudomonas aeruginosa metabolome database, well-curated, with rich metadata and offering bulk download. However, it does not contain only natural products but also results of the primary metabolism, so it was not included in the COCONUT collection.
The Lichen Database [81] is a collection of over 200 metabolites that have been isolated and identified experimentally in lichens. The database is not available yet, but the data has been already published in the MetaboLights [82] repository for metabolomics experimental data.

 doi:10.20944/preprints201912.0332.v1

Natural products by use

Traditional medicines

The World Health Organization listed between 1999 and 2009 a list of over 21 000 plants used for medicinal purposes all over the world [83,84]. This effort was made for proper identification of safe plants, as it is estimated that plant-based traditional medicines are used by 60 % of the world's population [85]. In addition to efforts to establish formal, DNA-based identification of such plants for wider use [86], collections of medicinal plant species, and in particular of phytochemicals, natural products produced by plants, associated to their therapeutic activities and physicochemical properties are being established around the world. This is particularly the case in Asia and Africa, where traditional medicines remain an important part of everyday life for cultural, traditional and economic reasons.

Traditional Chinese Medicine (TCM) is naturally part of the Chinese public health system [87,88]. It is therefore coherent that in this country the scientific study of natural compounds from plants used in TCM is very advanced and is receiving strong governmental support, and they have developed a plethora of databases containing NPs, their sources and effects.
The biggest database containing NPs used in TCM is TCM@Taiwan [89]. It contains over 58000 entries and is directly feeding iSMART [90], an integrated cloud computing web server for online virtual screening, evolution studies and drug design. In addition to this, there are several other, smaller, databases for NPs TCM that can be cited, such as the Chinese Ethnic Minority Traditional Drug Database (CEMTDD) [91], that is maintained, but not updated and contains 4000 NPs, the Chinese Traditional Medicinal Herbs Database (CHDD) [92] , not maintained anymore, but according to the publication contained over 30000 entries, now not accessible and probably lost for the scientific community. Some other databases containing phytochemicals and other active compounds used in TCM can be cited, such as the Comprehensive Herbal Medicine Information System for Cancer (CHMIS-C) [93] that is not maintained anymore, the Encyclopaedia of Traditional Chinese Medicine (ETCM) [94], that is maintained but the chemical structures it contains are not easily retrievable, the database of medicinal materials and chemical compounds in Northeast Asian TM (TM-MC) [95], which is maintained, updated, but no structures but contains precise plant species for all compounds, the Traditional Chinese Medicine Integrative Database (TCMID) [96], maintained, but not updated anymore, The Traditional Chinese Medicine Systems Pharmacology database and analysis platform (TCMSP) [97], that is also not maintained anymore but used to contain over 29000 NPs. One can quickly realize that there is a lot of databases that focus on chemical compounds used in TCM, and creators of the latter recognize it: there is

even a database called “Yet Another Traditional Chinese Medicine Database” (YaTCM) [98] that was published in 2018. Mainly, all these databases differ in the number of compounds they cover, in the richness of their metadata and on the availability of the datasets they contain.

Another extremely important traditional medicine in Asia is the Indian Ayurveda, that also got a wide popularization worldwide over the past decade. There are, however, very few databases listing natural compounds from plants, insects and animals used in Ayurveda, and they do not contain as many entries as the Chinese ones. Only two are currently online and open. The first one, IMPPAT [99] is the manually curated database of over 10000 phytochemicals extracted from 1700 Indian medicinal plants, their phytochemistry and their therapeutic effects. The other, MedPServer [100] contains natural products from plants from North-East India used in traditional medicine. It aims towards the understanding of the therapeutic mechanisms of action of the 1124 NPs from these plants by integrating ligand-based and structure-based approaches. NeMedPlant [101] is a small (over 100 NPs) database of active compounds from plants used in North-East Indian traditional medicine, with rich metadata focused on the plants that produce the compound but without possibilities of downloading any information and is not updated anymore. Because it was cited in several peer-reviewed papers, we also need to mention TIM [85], the database created in 2011 for the Prediction of Biologically Active Natural Products from Ayurveda Traditional Medicine but never linked to an actual database not listing the NPs in the supplementary material of the publication.

Phytochemica [102] is a small database of plant-derived chemicals that contains plants from Himalaya used in both Chinese and Indian traditional medicines. There are also some databases of natural products that specialize in traditional medicines of other parts of Asia, such as the Database of Indonesian Medicinal Plants [103] and TIPdb [104] for plants from Taiwan, but most of them are relatively small and contain in general only few hundreds of compounds.

African Traditional Medicine (ATM) is the other extremely rich and developed traditional medicine with a lot of modern efforts to study, rationalize and put its teachings to the benefit of modern medicine. As for the CTM and the Ayurveda, it requires inventorying plants used by African traditional doctors, identifying the parts that are used to efficiently cure and then identify the active components that they contain. It exists also a certain number of databases focusing on natural products from plants used in traditional medicines on the African continent. Among those, the most famous and the most generalistic is AfroDB [105], although it is only accessible through the ZINC catalogues. The pan-African natural products library (p-ANAPL) also needs to be cited here, as it focuses on plants used in ATM and is available as the supplementary information if its publication [106]. Three datasets, AfroCancer [107], AfroMalariaDB [108] and Afrotryp [109], available as supplementary information of their respective publications link NPs from plants used in traditional

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 December 2019 doi:10.20944/preprints201912.0332.v1

medicines to their potential targets involved in the treatment of cancer, malaria and Trypanosoma. There are then country-specific and relatively small databases for NPs extracted from ATM plants, such as the Cameroon Medicinal Natural Products database (CamMedNP) [110], Central African Medicinal Plants database (ConMedNP) [111] and the Ethiopian Traditional Medicine Database (ETM-DB) [112].

Databases of drug-like natural compounds

Not linked, at least directly, to the traditional medicines, there is a lot of pharmacological research around the therapeutic properties of natural products, and these are compiled in the databases for drugs and drug candidates. In these databases, natural compounds are generally associated with a type of disease or molecular targets or receptors they interact with, and a rich description of their molecular and overall effects on the state of a patient or of a healthy person. The reference database in this category is DrugBank [113]. It latest version, which was greatly modified and curated compared to previous ones, contains over 10000 drugs, among which 3732 are approved drugs and 200 approved drugs that have been produced by a living organism. In order to select only the latter, one needs to search for “nutraceuticals” in the search bar of the DrugBank website [114]. The previous version of Drugbank, 4.0 [115], contained over 8000 nutraceuticals, and they were added to COCONUT.

BindingBD [116] is an interesting database for pharmaceutical research as it contains measured binding affinities of proteins that are supposedly targets of drugs, with small drug-like molecules. Although it does contain natural products and their protein targets, they are not clearly distinguishable from synthetic drugs in this database.

The Novel Antibiotics Database [117], that is still surprisingly online, is not updated since 2003 and contains 5430 compounds of natural origin with an antibiotic activity that have been published in the Journal of Antibiotics between 1947 and 2003. However, no structure is available for download, only compound names, their activity and the organisms they were isolated from.

ChemIDplus [118] is a database part of the TOXicology DataNETwork and chemicals that have a relationship with diseases, environment, environmental health and poisoning. It contains rich metadata for each chemical, including its physicochemical properties but also its impact on health and environment. A simple search for "natural product" returns more than 9000 entries, it is however not possible to bulk download the results of the query.

The Herbal Ingredient Targets (HIT) [119] and the Herbal Ingredients in-vivo Metabolism (HIM) [120] databases are two inter-connected collections of natural

 doi:10.20944/preprints201912.0332.v1

products from mainly (but not only) Chinese plants. Both are not accessible online anymore, but the structures of the NPs they contained are available on ZINC. They contained very extensive metadata on the molecular targets of the herbal active ingredients, their toxicity, a wide range of pharmacologically relevant molecular descriptors and their therapeutic effects. Unfortunately, this metadata is not available on ZINC and is probably lost.

There are several databases that focus on collecting information on NPs with anticancer properties and their mechanisms of action. The first one, NPCARE [121] contains over 6000 NPs from plants, marine organisms, fungi and bacteria with validated anticancer activities and contains extensive metadata. The website is available and seems updated but cannot be accessed sometimes, probably due to server failures on the maintenance side. The Indian Plant Anticancer Compounds Database (InPACdb) [122] is not available anymore but used to contain very broad information covering pharmaceutical and physicochemical properties of 144 natural products, cancer types and molecular targets. Fortunately, the data is still available on GitHub [123]. Another database, containing phytochemicals with anti-cancer properties is the Naturally Occurring Plant-based Anti-cancer Compound-Activity-Target (NPACT) database [124] is still maintained and accessible It contains 1574 manually curated entries with rich metadata on NPs and their therapeutical mechanisms on different types of cancer. The US National Cancer Institute also maintains and makes freely available a number of small (390 on average) natural compound datasets [125] that are selected as of interest in anticancer research and are currently undergoing tests in various research groups from the US NIH.

InflamNat [126] is a small (200 NPs) but well-curated dataset of natural products with anti-inflammatory activity. The dataset consists of NP structures, their type and origin and literature references, and is available as supplementary information for its publication.

BioPhytMol [127] is a manually curated database of natural compounds from plants that have an antibacterial effect. The database has over 2500 entries with very rich metadata, in particular regarding the plant species from which the compounds were extracted. The database is open and maintained but does not offer a bulk download option to be used to further analyses.

The last database in this section is the Open Source Malaria [128], which is a very nice project as it is a totally open-source collaborative project for anti-malarial drugs discovery that already encountered certain success [129]. Drug candidates tested in this project are often of natural origin, but as the focus of this database is to collect their effects, it is not always specified, so the content of OSM was not integrated into COCONUT.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 December 2019 doi:10.20944/preprints201912.0332.v1

Food

FooDB [7] is the reference database on chemical food constituents associated with extremely rich and diverse metadata. It is developed by the Wishart research group and supported by the Canadian Institutes of Health Research. In total it contains over 22000 NPs and offers a convenient bulk download their structures.
BitterDB [6] collects bitter-tasting natural compounds associated with rich metadata on their receptors. However, it also contains synthetic molecules with a bitter taste, and in this database, it is difficult to separate them from the natural ones.
Phenol-Explorer [130] is a comprehensive database on polyphenol content in food. It currently contains over 800 phenol structures from over 400 foods. Data is derived from the scientific literature, and all data is associated with rich metadata and is available for download.
PhytoHub [131] is a database of dietary phytochemicals and the human and animal metabolites that derive from them. Over 1200 NPs from more than 350 foods are available in this resource, together with rich metadata and references to other chemical and spectral databases. it, unfortunately, doesn't offer a bulk download for the moment.
The SuperSweet database [4] is a collection of various molecules, mainly from plant origin, but also synthetics that have a sweet taste. Their structures together with information on their number of calories, therapeutic uses and sweetness index are available. The database is still maintained but is not updated since 2011 and does not provide a bulk download of its content.

Toxins

A toxin is a substance that is toxic for one or more living organisms and that has a plant or animal origin. Despite this original definition, more and more resources on toxins also integrate molecules from non-organic origin massively present in the environment as they also have a harmful effect on the living organisms. For instance, Exposome-explorer [132] is a manually curated database of biomarkers of exposure to environmental and dietary factors, and it also contains these factors and their structures. A lot of the toxic environmental and dietary factors in it are from natural origin, but also, approximately half of the compounds in this database are not NPs, which is reasonable, as, for example, environmental pollution is anthropogenic. In the same way can be mentioned the T3DB [133], the toxin and toxin-target database, as it contains a number of toxins produced by the living organism but its focus is on synthetic toxins and how human metabolism reacts to them.

The biggest (over a 1000) database of animal toxins was the Animal Toxin Database (ATDB) [134], designed originally to collect toxin structures, origins and effects, but it is not available anymore at the URL provided in the publication. More specialized databases were also published, such as the International Venom and Toxin

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 December 2019 doi:10.20944/preprints201912.0332.v1

Database [135], the Snake Neurotoxin Database [136], the Mollusk Toxin Database [137] or the Scorpion Toxin Database [138]. Unfortunately, most of these databases were based on unformatted text and were lacking effective systems for data query, and none of them is not accessible anymore. It is also unknown if the data contained in these databases is lost or is still available in some generalistic resources.
The last in this section, the Toxic Plants - Phytotoxins Database (TPPT) [139], is accessible and is maintained and updated by the Agroscope in Switzerland. It contains over 1500 phytotoxins from Central Europe and offers high-quality metadata and a convenient bulk download.

Other

The two databases described next could not be fitted in any of the previous categories. The Carotenoids database [140] is a collection of NPs produced by a wide range of organisms and that share common substructures (polyene with possibly terminating rings) and properties as they are all yellow, orange or red pigments. Carotenoids produced by plants have particular importance for the nutritional value of the consumed food [141], but plants are not the only producers of this molecular type which is demonstrated in the Carotenoids database. This database is developed and maintained at the RIKEN institute. SuperScent [9] is a database of volatile compounds essential from an organic origin that can be scented by humans and animals. It contains over 2000 compounds with their structures and properties but does not offer any download and most of the compound pages are now working. This database is maintained at Charité Belin but is not updated since 2010.

Natural products by the geographic origin of producing organisms

There is a number of country-level efforts to catalogue the biodiversity of natural products in particular geographical zones, generally defined by country political borders. These databases are mainly plant-focused, but can also integrate NP produced by insects, by microorganisms and animal toxins. In this part, the databases are cited in the geographical order from West to East. The last part is describing collections of NPs from organisms in marine and ocean environments.
BIOFAQUIM [142] is a database published in 2019 and offers for full download over 400 unique natural products from plants, fungi and propolis from Mexican flora and fauna, the species from which the compounds were extracted and their geographical location. The Nuclei of Bioassays, Ecophysiology and Biosynthesis of Natural Products Database (NUBBEDB) [143] is the first natural product library from Brazilian biodiversity. It currently contains over 2000 NPs, highly curated and good quality metadata and easy download of all or partial data. The UEFS dataset [144] is

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 December 2019 doi:10.20944/preprints201912.0332.v1

a collection of NPs isolated from Brazilian plants and maintained by the State University of Ferriera de Santana in Bahia, Brazil. The NPs in this collection have been published separately but there is no common publication nor public database for it, it is however accessible via ZINC.

Three databases contain natural products from the African flora and fauna. The Northern African Natural Products Database (NANPDB) [145] contains over 4500 NPs from plants, endophytes, fungi and bacteria. The database provides rich metadata, literature references, cross-references to major chemical databases and an easy bulk download. The South African natural compound database (SANCDB) [146] is very similar to NANPDB in its quality and contains over 600 NPs isolated from South African biodiversity. It is also possible to submit new molecules and to participate in the curation of the database. The Mitishamba database [147] contains 1100 NPs isolated from Kenyan plants. The database is still maintained but doesn't seem to be updated and it is possible to download data from it only by requesting an account.

ChemDB [3] and MAPS database [148] are two databases for natural compounds from Pakistani plants. Unfortunately, none of them is accessible anymore. VIETHERB [149] is a database published in 2018 with the aim of providing high-quality and literature-based data on herbs and active compounds from them. Despite the novelty of the database, it is not accessible anymore.
The oceans cover 71% of the surface of the Earth, therefore databases that collect natural products from marine organisms are expected to be broad, complex and cover a wide range of organisms. Unfortunately, the biggest repositories for marine NP structures are commercial (e.g. MarineLit [29] and DMNP [24] presented above). In the marine NP community, the major trend is to publish newly discovered molecules in specialised journals (such as the Journal of Natural Products [150] or Marine Drugs [151]) as images and rich textual description that are not, for now, easily machine-retrievable.
In the last 20 years, four databases containing structures of marine NPs and their metadata were published. Two of them are not accessible anymore: the Marine Compound Database (MCDB) [152] and the Marine Natural Product Database (MNPD) [153]. Both contained only a few hundreds of entries according to their respective publications but these were comprising rich metadata which is now lost. The Dragon Exploration System on Marine Sponge Compounds Interactions (DESMCI) [154] is still accessible but seems not to be maintained as the actual data, such as molecular structures and the corresponding metadata is not visible when one tries to access it. The Seaweed Metabolite Database (SWMD) [155] is the only one really maintained and it contains 1110 entries, with only 423 unique structures. Molecular structures in this database are annotated with the species of the algae that

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 December 2019 doi:10.20944/preprints201912.0332.v1

produce them, together with the geographical origin of the latter, biological activity of the compound and its physicochemical properties.

Industrial catalogues

A lot of companies that are synthesizing and isolating chemical compounds offer a catalogue of their products, and in some cases, these catalogues also contain the structures and annotations. These catalogues are often cited in the scientific literature as sources of NP structures, therefore it was important to mention the most used catalogues in this review. Surprisingly, a non-negligible number of cited catalogues of NP structures are accessible only to clients, on-demand or to registered users. This is the case of the NP catalogues from Ambinter-Greenpharma natural compound library [156], ChemBridge diversity datasets [157] (their NP catalogue seems to be not available anymore), LOPAC1280 by Merk [158], Prestwick [159] and TargetMol [160]. Open NP catalogues are provided by the following: AnalytiCon Discovery [161], InterBioScreen [162], Indofine Chemical Company [163], Pi Chemicals Systems [164] and Specs [165]. The website of the latter is not offering the download of their natural products catalogue anymore, but a dataset is available on ZINC [166]. Note that only the most famous and cited in academic research are listed and more industrial catalogues for NPs exist.

Problems

The biggest problem nowadays is that there are too many sources for natural products. A non-experienced researcher (and even a more experienced one) in natural products will just get lost in this variety and diversity of data sources. The next major problem is access to data and its maintenance. Indeed, a lot of publication point to a website that is not maintained anymore. This is the case of the majority of animal toxins databases, but also of a number of small regional or traditional medicine databases. In the list of NP sources presented in Table 1, over 20% are not maintained anymore or the access is intermittent. In some rare cases, the information on the NP structures is still recoverable via the ZINC database, but it is not the case of more modern databases and ZINC doesn't store any metadata from these collections. Also, the description and origins of the natural products, beyond their structure are generally lacking, and it is especially the case in data aggregators that are nevertheless the most commonly used. This leads to cases where *in silico* screening reveals potentially interesting compounds but requires way more efforts and investigations to identify its origins and the way of obtaining it experimentally. Only 40% of NP databases offer an easy bulk download of molecular structures that they contain for further analyses with local tools.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 December 2019 doi:10.20944/preprints201912.0332.v1

This multiplicity of databases comes also from the publishing pressure on scientists, the infamous “publish or perish”. Nowadays, publishing a dataset or a database is a relatively easy publication and have the potential to generate a high number of citations. However, this trend generates a plethora of databases that are unmaintained beyond the publication time (like it is the case of VIETHERB [149] for example, published only one year prior to the writing of the present review and already not accessible anymore), despite the journals requirements to provide accessibility to the published datasets and databases for a number of years ahead.

Comparison and analysis of the content of open NP databases

The 50 NP collections from which NP structures could be downloaded were analysed in order to evaluate their overlap in terms of molecular structures and coherence of their content. 19 physicochemical properties, such as molecular weight, NP-likeness [167,168], logP, TPSA Efficiency, and Zagreb Index, were computed and their distributions are shown in an interactive graphic at https://npreview.naturalproducts.net. Due to the high number of databases to compare, a non-interactive would not be visible. Globally, the physicochemical properties of all datasets are comparable. The NP subset of Drugbank contains molecules that are less likely to be NPs, which can be explained by its high content in NP-derived drugs and the difficulty in dissociating the latter from synthetic ones. The average mass of all NPs in the assembled collection is of 454 Da, and the Spektraris and TCM@Taiwan databases contain the heaviest molecules: both contain molecules with an average of 612Da. The logP is a lipophilicity measure commonly used in analytical chemistry; the more it is positive, the more lipophilic is the compound and the more negative, the more hydrophilic. Here, the logP was computed with two algorithms, AlogP and XlogP available in the CDK [169]. In general, NPs tend to be lipophilic, which allows them to have higher membrane penetration, but all datasets also contain in lesser amounts, hydrophilic molecules. CarotenoidsDB and the SeaWeed Metabolites Database outstand from others with their very lipophilic content. On the other side, ReSpect contains more hydrophilic molecules than other datasets.

The overlap in terms of molecular structures between the databases was also calculated and is presented in figure 1 and in Table 2. In figure 1, which represents a network of overlap between databases, there is a directed edge between database A and database B if more than 50% of the unique molecules from database A are

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 December 2019 doi:10.20944/preprints201912.0332.v1

present in database B. An interactive version of this network, where the user can change the percentage of similarity between databases to display is available at https://npreview.naturalproducts.net. It should be noted that 40 of the 50 open NP databases have an overlap of at least 50% with at least one other open database. Except for the Lichen Database, all datasets share at least 10% of their compounds with at least one other open dataset.

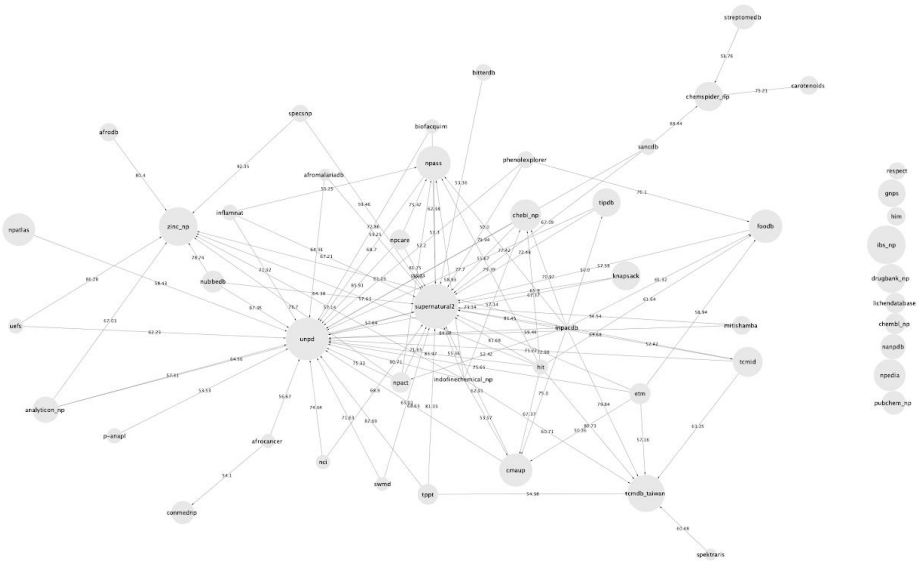

Figure 1: Network of content similarity between the 50 open natural products databases. The network is directed, and there is an arrow from database A to database B if more than 50% of molecules in database A are also present in database B. The interactive version of this network is available at https://npreview.naturalproducts.net

Five natural products are found in 34 of these 50 databases: apigenin, quercetin, kaemferol, catechin and naringenin. Interestingly, belong all to the flavanol group, part of the flavonoids family and share a common skeleton (Figure 2a) with only differences in hydroxy groups. In the top ten most frequent molecules in open databases, in addition to more flavonoids, there is also coumaric acid (Figure 2b), gallic acid (Figure 2c), scopoletin (Figure 2d) and ellagic acid (Figure 2e). According to the literature, all these compounds are well-known plant products, however, most of the flavanols, coumaric acid and scopoletin are also present in the bacterial NP database, StreptomeDB.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 December 2019 doi:10.20944/preprints201912.0332.v1

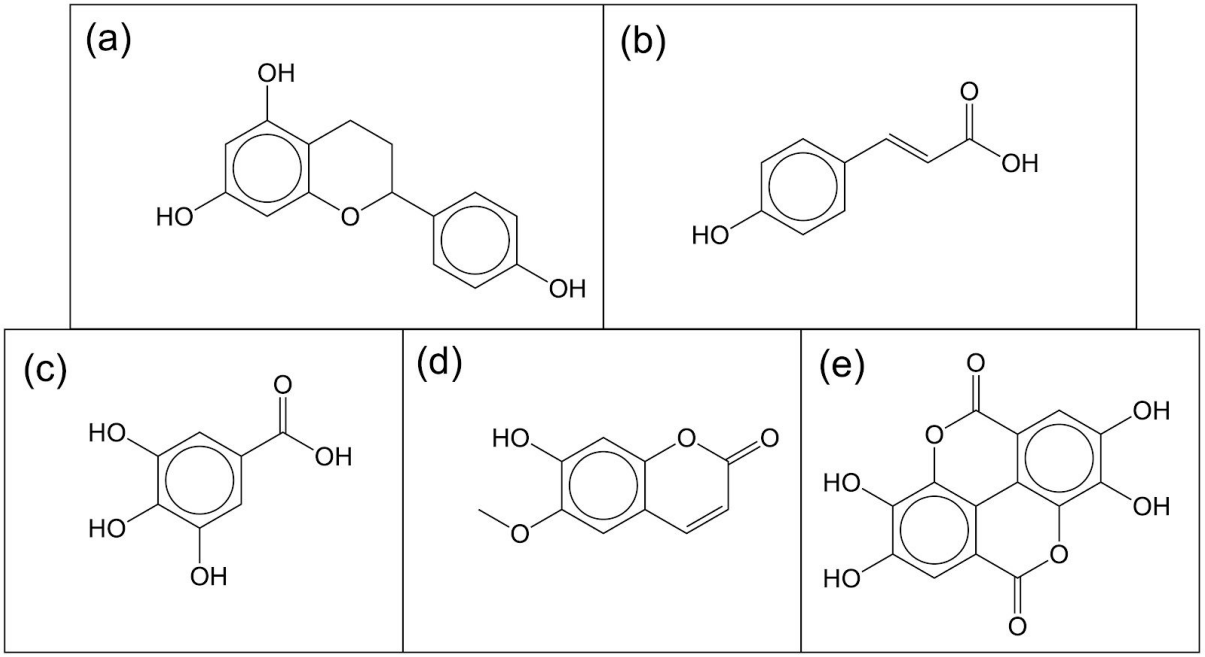

Figure 2. Most frequent molecules in open databases. (a) Common biggest substructure in the top 5 most frequent molecules, found in 34 out of 50 open databases. (b) Coumaric acid; (c) Gallic acid; (d) Scopoletin; (e) Ellagic acid.

COlleCtion of Open NatUral producTs (COCONUT)

In its current version, COCONUT contains 411621 unique molecules, depleted for now from their stereochemistry for unification purposes, that were collected from 50 open and accessible NP databases, listed in table 1. This number is big, as this dataset still needs to undergo a curation process, as, despite their claims, some of the NP collections do not contain only natural compounds. The collection is available as a MongoDB dump and a CSV file on Zenodo (https://doi.org/10.5281/zenodo.3547718) and a user-friendly web interface to browse it is under development.

Discussion

There are currently 122 data collections of natural products (NPs) that have been published and cited in the scientific literature between 2000 and 2019. Only 50 of them are open access or have their content accessible (in ZINC for example) and among them, the overlap of their content is significant, as 40 of these datasets share at least 50% of the compounds they contain with at least one other dataset.

There are several aggregators, such as the ZINC catalogue for natural products, SuperNatural II and UNPD (not maintained anymore), but they do not cover the entire space of known NPs and do not allow submissions of newly discovered compounds.

There is a need for an aggregator database for natural products, that will be commonly recognized, well organized and allowing an easy submission of newly found molecules, like it is the case for UniProt for proteins.

Conclusions

Natural products are important molecules for medical, chemical and social research. There is no, for now, any universal, community-accepted database for NP discovery, screening and dereplication. Instead, there is an extremely high number of very diverse databases and datasets, not all maintained or open access in 2020, which represents a serious loss of knowledge. There is a need for a unified universal repository for NPs, to avoid the unnecessary duplication of online resources and facilitate NP research. For the purpose of this review, a COlleCtion of Open Natural prodUcTs (COCONUT) has been assembled, analyzed and made available in Zenodo (https://doi.org/10.5281/zenodo.3547718). A web interface is currently under development for user-friendly querying, exploration and download of the known open natural products space. In the future, the annotations of the molecules contained in COCONUT will be improved, in particular, systematically linking the compound to the first publication where it was described and to the organisms that synthesize it.

Materials and methods

All databases in Table 1 were downloaded in July and September 2019. Molecular structures were processed with CDK 2.2 and, when available, annotations were parsed with Java (code available on GitHub https://github.com/mSorok/COCONUT). Resulting original and non-redundant collections of natural products are stored in a MongoDB database, available as a dump on Zenodo (https://doi.org/10.5281/zenodo.3547718). Redundancy was eliminated based on InChi Keys. All network representations of overlaps between databases are made with Cytoscape [170]. Plots and comparative analyses made with Python and the Plotly and Dash libraries. The code for the interactive plots is available on GitHub at https://github.com/mSorok/NPDBReviewDash.

Bibliography:

1. Newman DJ, Cragg GM. Natural Products as Sources of New Drugs from 1981 to 2014. J Nat Prod. 2016;79: 629–661. doi:10.1021/acs.jnatprod.5b01055
2. Khalifa SA, Elias N, Farag MA, Chen L, Saeed A, Hegazy M-EF, et al. Marine Natural Products: A Source of Novel Anticancer Drugs. Mar Drugs. 2019;17: 491.

3. Bano Mirza S, Bokhari H, Qaiser Fatmi M. Exploring Natural Products from the Biodiversity of Pakistan for Computational Drug Discovery Studies: Collection, Optimization, Design and Development of A Chemical Database (ChemDP). 2015 [cited 9 Sep 2019]. Available: https://www.ingentaconnect.com/content/ben/cad/2015/00000011/00000002/art00003
4. Ahmed J, Preissner S, Dunkel M, Worth CL, Eckert A, Preissner R. SuperSweet—a resource on natural and artificial sweetening agents. Nucleic Acids Res. 2011;39: D377–D382. doi:10.1093/nar/gkq917
5. Yue Y, Chu G-X, Liu X-S, Tang X, Wang W, Liu G-J, et al. TMDB: A literature-curated database for small molecular compounds found from tea. BMC Plant Biol. 2014;14: 243. doi:10.1186/s12870-014-0243-1
6. Dagan-Wiener A, Di Pizio A, Nissim I, Bahia MS, Dubovski N, Margulis E, et al. BitterDB: taste ligands and receptors database in 2019. Nucleic Acids Res. 2019;47: D1179–D1185. doi:10.1093/nar/gky974
7. FooDB. [cited 3 Oct 2019]. Available: http://foodb.ca/
8. Mahesh SK, Fathima J, Veena VG. Cosmetic Potential of Natural Products: Industrial Applications. In: Swamy MK, Akhtar MS, editors. Natural Bio-active Compounds: Volume 2: Chemistry, Pharmacology and Health Care Practices. Singapore: Springer Singapore; 2019. pp. 215–250. doi:10.1007/978-981-13-7205-6_10
9. Dunkel M, Schmidt U, Struck S, Berger L, Gruening B, Hossbach J, et al. SuperScent—a database of flavors and scents. Nucleic Acids Res. 2009;37: D291–D294. doi:10.1093/nar/gkn695
10. Sparks TC, Wessels FJ, Lorsbach BA, Nugent BM, Watson GB. The new age of insecticide discovery-the crop protection industry and the impact of natural products. Pestic Biochem Physiol. 2019 [cited 24 Oct 2019]. doi:10.1016/j.pestbp.2019.09.002
11. Füllbeck M, Michalsky E, Dunkel M, Preissner R. Natural products: sources and databases. Nat Prod Rep. 2006;23: 347–356. doi:10.1039/B513504B
12. Johnson SR, Lange BM. Open-Access Metabolomics Databases for Natural Product Research: Present Capabilities and Future Potential. Front Bioeng Biotechnol. 2015;3. doi:10.3389/fbioe.2015.00022
13. Tawfike AF, Viegelmann C, Edrada-Ebel R. Metabolomics and Dereplication Strategies in Natural Products. In: Roessner U, Dias DA, editors. Metabolomics Tools for Natural Product Discovery: Methods and Protocols. Totowa, NJ: Humana Press; 2013. pp. 227–244. doi:10.1007/978-1-62703-577-4_17
14. Chen Y, de Bruyn Kops C, Kirchmair J. Data Resources for the Computer-Guided Discovery of Bioactive Natural Products. J Chem Inf Model. 2017;57: 2099–2111. doi:10.1021/acs.jcim.7b00341
15. Pereira F, Aires-de-Sousa J. Computational Methodologies in the Exploration of Marine Natural Product Leads. Mar Drugs. 2018;16.
16. Harvey AL, Edrada-Ebel R, Quinn RJ. The re-emergence of natural products for drug discovery in the genomics era. Nat Rev Drug Discov. 2015;14: 111–129. doi:10.1038/nrd4510
17. UniProt Consortium T. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 2018;46: 2699. doi:10.1093/nar/gky092
18. Federhen S. The NCBI Taxonomy database. Nucleic Acids Res. 2012;40: D136–D143. doi:10.1093/nar/gkr1178
19. omicX. In: omicX [Internet]. [cited 9 Oct 2019]. Available: https://omictools.com/
20. Williams AJ, Martin GE, Rovnyak D. Modern NMR Approaches to the Structure Elucidation of Natural Products: Volume 1: Instrumentation and Software. Royal Society of Chemistry; 2016.
21. Dictionary of Natural Products 28.1. [cited 9 Oct 2019]. Available:

http://dnp.chemnetbase.com/faces/chemical/ChemicalSearch.xhtml;jsessionid=DB01289ACAA79C222859E1CD8A98A894
22. Reaxys. [cited 9 Oct 2019]. Available: https://www.reaxys.com/#/search/quick
23. Gabrielson SW. SciFinder. J Med Libr Assoc. 2018;106: 588–590. doi:10.5195/jmla.2018.515
24. Dictionary of Marine Natural Products 2018. [cited 9 Oct 2019]. Available: http://dmnp.chemnetbase.com/faces/chemical/ChemicalSearch.xhtml;jsessionid=824F3121F9A123D4684A7A8289F618E2
25. Dictionary of Food Compounds 2018. [cited 18 Oct 2019]. Available: http://dfc.chemnetbase.com/faces/chemical/ChemicalSearch.xhtml;jsessionid=60BDE6E1AE536A1C52AFB65A680DC289
26. Loub WD, Farnsworth NR, Soejarto DD, Quinn ML. NAPRALERT: computer handling of natural product research data. J Chem Inf Model. 1985;25: 99–103. doi:10.1021/ci00046a009
27. Johnson SG. NIST Standard Reference Database 1A v17. In: NIST [Internet]. 19 Jun 2014 [cited 9 Oct 2019]. Available: https://www.nist.gov/srd/nist-standard-reference-database-1a-v17
28. Blunt JW, Carroll AR, Copp BR, Davis RA, Keyzers RA, Prinsep MR. Marine natural products. Nat Prod Rep. 2018;35: 8–53. doi:10.1039/C7NP00052A
29. MarinLit. Available: http://pubs.rsc.org/marinlit/
30. Blunt JW, Munro MHG, Laatsch H. Antimarin database. Univ Canterb. 2006;432.
31. Blunt J, Munro M, Upjohn M. The role of databases in marine natural products research. Handb Mar Nat Prod. 2012; 389–421.
32. AntiBase. [cited 9 Oct 2019]. Available: https://application.wiley-vch.de/stmdata/antibase.php
33. Wiley-VCH - AntiBase. [cited 21 Oct 2019]. Available: https://application.wiley-vch.de/stmdata/antibase.php
34. Ikram NKK, Durrant JD, Muchtaridi M, Zalaludin AS, Purwitasari N, Mohamed N, et al. A Virtual Screening Approach For Identifying Plants with Anti H5N1 Neuraminidase Activity. J Chem Inf Model. 2015;55: 308–316. doi:10.1021/ci500405g
35. Ehrman TM, Barlow DJ, Hylands PJ. In silico search for multi-target anti-inflammatories in Chinese herbs and formulas. Bioorg Med Chem. 2010;18: 2204–2218. doi:10.1016/j.bmc.2010.01.070
36. Quinn RJ, Carroll AR, Pham NB, Baron P, Palframan ME, Suraweera L, et al. Developing a Drug-like Natural Product Library. J Nat Prod. 2008;71: 464–468. doi:10.1021/np070526y
37. Lagunin AA, Druzhilovsky DS, Rudik AV, Filimonov DA, Gawande D, Suresh K, et al. [Computer evaluation of hidden potential of phytochemicals of medicinal plants of the traditional Indian ayurvedic medicine]. Biomeditsinskaia Khimiia. 2015;61: 286–297. doi:10.18097/PBMC20156102286
38. Berdy J, Kertesz M. Bioactive natural products database: an aid for natural products identification. In: Collier HR, editor. Chemical Information. Springer Berlin Heidelberg; 1989. pp. 237–251.
39. Hastings J, de Matos P, Dekker A, Ennis M, Harsha B, Kale N, et al. The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. Nucleic Acids Res. 2013;41: D456--63. doi:10.1093/nar/gks1146
40. Gaulton A, Hersey A, Nowotka M, Bento AP, Chambers J, Mendez D, et al. The ChEMBL database in 2017. Nucleic Acids Res. 2017;45: D945–D954. doi:10.1093/nar/gkw1074
41. Pence HE, Williams A. ChemSpider: An Online Chemical Information Resource. J Chem Educ. 2010;87: 1123–1124. doi:10.1021/ed100697w

42. Hähnke VD, Kim S, Bolton EE. PubChem chemical structure standardization. J Cheminformatics. 2018;10: 36. doi:10.1186/s13321-018-0293-8
43. Seiler KP, Kuehn H, Happ MP, DeCaprio D, Clemons PA. Using ChemBank to Probe Chemical Biology. Curr Protoc Bioinforma. 2008;22: 14.5.1-14.5.26. doi:10.1002/0471250953.bi1405s22
44. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2016; gkw1092. doi:10.1093/nar/gkw1092
45. Caspi R, Billington R, Fulcher CA, Keseler IM, Kothari A, Krummenacker M, et al. The MetaCyc database of metabolic pathways and enzymes. Nucleic Acids Res. 2018;46: D633–D639. doi:10.1093/nar/gkx935
46. Altman T, Travers M, Kothari A, Caspi R, Karp PD. A systematic comparison of the MetaCyc and KEGG pathway databases. BMC Bioinformatics. 2013;14: 112. Available: http://www.biomedcentral.com/1471-2105/14/112/abstract
47. Jeske L, Placzek S, Schomburg I, Chang A, Schomburg D. BRENDA in 2019: a European ELIXIR core data resource. Nucleic Acids Res. 2019;47: D542–D549. doi:10.1093/nar/gky1048
48. Sitzmann M, Filippov IV, Nicklaus MC. Internet resources integrating many small-molecule databases1. SAR QSAR Environ Res. 2008;19: 1–9. doi:10.1080/10629360701843540
49. King ZA, Lu J, Dräger A, Miller P, Federowicz S, Lerman JA, et al. BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. Nucleic Acids Res. 2016;44: D515–D522. doi:10.1093/nar/gkv1049
50. MassBank of North America (MoNa). [cited 16 Oct 2019]. Available: http://mona.fiehnlab.ucdavis.edu/
51. MassBank | European MassBank (NORMAN MassBank) Mass Spectral DataBase. [cited 16 Oct 2019]. Available: http://massbank.normandata.eu/MassBank/
52. MassBank | MSSJ MassBank Mass Spectral DataBase. [cited 16 Oct 2019]. Available: http://www.massbank.jp/
53. Guijas C, Montenegro-Burke JR, Domingo-Almenara X, Palermo A, Warth B, Hermann G, et al. METLIN: A Technology Platform for Identifying Knowns and Unknowns. Anal Chem. 2018;90: 3156–3164. doi:10.1021/acs.analchem.7b04424
54. Wishart DS, Feunang YD, Marcu A, Guo AC, Liang K, Vázquez-Fresno R, et al. HMDB 4.0: the human metabolome database for 2018. Nucleic Acids Res. 2018;46: D608–D617. doi:10.1093/nar/gkx1089
55. Wang M, Carver JJ, Phelan VV, Sanchez LM, Garg N, Peng Y, et al. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. Nat Biotechnol. 2016;34: 828. Available: https://doi.org/10.1038/nbt.3597
56. Kuhn S, Schlörer NE. Facilitating quality control for spectra assignments of small organic molecules: nmrshiftdb2– a free in-house NMR database with integrated LIMS for academic service laboratories. Magn Reson Chem. 2015;53: 582–589. doi:10.1002/mrc.4263
57. NMRdata. [cited 15 Oct 2019]. Available: http://www.nmrdata.com/
58. López-Pérez JL, Therón R, del Olmo E, Díaz D. NAPROC-13: a database for the dereplication of natural product mixtures in bioassay-guided protocols. Bioinformatics. 2007;23: 3256–3257. doi:10.1093/bioinformatics/btm516
59. Fischedick JT, Johnson SR, Ketchum REB, Croteau RB, Lange BM. NMR spectroscopic search module for Spektraris, an online resource for plant natural product identification – Taxane diterpenoids from Taxus×media cell suspension cultures as a case study. Phytochemistry. 2015;113: 87–95.

doi:10.1016/j.phytochem.2014.11.020
60. Banerjee P, Erehman J, Gohlke B-O, Wilhelm T, Preissner R, Dunkel M. Super Natural II—a database of natural products. Nucleic Acids Res. 2015;43: D935–D939. doi:10.1093/nar/gku886
61. Molecular Diversity Preservation International (MDPI). [cited 15 Oct 2019]. Available: https://www.mdpi.org/
62. Gu J, Gui Y, Chen L, Yuan G, Lu H-Z, Xu X. Use of Natural Products as Chemical Library for Drug Discovery and Network Pharmacology. PLOS ONE. 2013;8: e62839. doi:10.1371/journal.pone.0062839
63. ISDB by oolonek. [cited 15 Oct 2019]. Available: http://oolonek.github.io/ISDB/
64. Sterling T, Irwin JJ. ZINC 15 – Ligand Discovery for Everyone. J Chem Inf Model. 2015;55: 2324–2337. doi:10.1021/acs.jcim.5b00559
65. Zeng X, Zhang P, He W, Qin C, Chen S, Tao L, et al. NPASS: natural product activity and species source database for natural product research, discovery and tool development. Nucleic Acids Res. 2018;46: D1217–D1222. doi:10.1093/nar/gkx1026
66. Tomiki T, Saito T, Ueki M, Konno H, Asaoka T, Suzuki R, et al. RIKEN natural products encyclopedia (RIKEN NPEdia), a chemical database of RIKEN natural products depository (RIKEN NPDepo). J Comput Aid Chem. 2006;7: 157–162.
67. Maeda MH, Kondo K. Three-Dimensional Structure Database of Natural Metabolites (3DMET): A Novel Database of Curated 3D Structures. J Chem Inf Model. 2013;53: 527–533. doi:10.1021/ci300309k
68. Shen J, Xu X, Cheng F, Liu H, Luo X, Shen J, et al. Virtual Screening on Natural Products for Discovering Active Compounds and Target Information. 2003 [cited 20 May 2019]. doi:info:doi/10.2174/0929867033456729
69. Blin K, Wolf T, Chevrette MG, Lu X, Schwalen CJ, Kautsar SA, et al. antiSMASH 4.0—improvements in chemistry prediction and gene cluster boundary identification. Nucleic Acids Res. 2017;45: W36–W41. doi:10.1093/nar/gkx319
70. Skinnider MA, Dejong CA, Rees PN, Johnston CW, Li H, Webster ALH, et al. Genomes to natural products PRediction Informatics for Secondary Metabolomes (PRISM). Nucleic Acids Res. 2015;43: 9645–9662. doi:10.1093/nar/gkv1012
71. M. Crawford J, Clardy J. Bacterial symbionts and natural products. Chem Commun. 2011;47: 7559–7566. doi:10.1039/C1CC11574J
72. Sarethy IP, Srivastava N, Pan S. Endophytes: The Unmapped Repository for Natural Products. In: Akhtar MS, Swamy MK, Sinniah UR, editors. Natural Bio-active Compounds: Volume 1: Production and Applications. Singapore: Springer; 2019. pp. 41–70. doi:10.1007/978-981-13-7154-7_2
73. Nakamura K, Shimura N, Otabe Y, Hirai-Morita A, Nakamura Y, Ono N, et al. KNApSAcK-3D: A Three-Dimensional Structure Database of Plant Metabolites. Plant Cell Physiol. 2013;54: e4–e4. doi:10.1093/pcp/pcs186
74. Zeng X, Zhang P, Wang Y, Qin C, Chen S, He W, et al. CMAUP: a database of collective molecular activities of useful plants. Nucleic Acids Res. 2019;47: D1118–D1127. doi:10.1093/nar/gky965
75. Miettinen K, Iñigo S, Kreft L, Pollier J, De Bo C, Botzki A, et al. The TriForC database: a comprehensive up-to-date resource of plant triterpene biosynthesis. Nucleic Acids Res. 2018;46: D586–D594. doi:10.1093/nar/gkx925
76. Boonen J, Bronselaer A, Nielandt J, Veryser L, De Tré G, De Spiegeleer B. Alkamid database: Chemistry, occurrence and functionality of plant N-alkylamides. J Ethnopharmacol. 2012;142: 563–590. doi:10.1016/j.jep.2012.05.038
77. Klementz D, Döring K, Lucas X, Telukunta KK, Erxleben A, Deubel D, et al. StreptomeDB 2.0—an extended resource of natural products produced by streptomycetes. Nucleic Acids Res. 2016;44: D509–D514. doi:10.1093/nar/gkv1319

78. Natural Products Atlas. [cited 16 Oct 2019]. Available: https://www.npatlas.org/joomla/
79. Nupur LNU, Vats A, Dhanda SK, Raghava GPS, Pinnaka AK, Kumar A. ProCarDB: a database of bacterial carotenoids. BMC Microbiol. 2016;16: 96. doi:10.1186/s12866-016-0715-6
80. Huang W, Brewer LK, Jones JW, Nguyen AT, Marcu A, Wishart DS, et al. PAMDB: a comprehensive Pseudomonas aeruginosa metabolome database. Nucleic Acids Res. 2018;46: D575–D580. doi:10.1093/nar/gkx1061
81. Lichen Database. In: MTBLS999:A database of high-resolution MS/MS spectra for lichen metabolites [Internet]. [cited 16 Oct 2019]. Available: https://www.ebi.ac.uk/metabolights/MTBLS999
82. Haug K, Salek RM, Conesa P, Hastings J, de Matos P, Rijnbeek M, et al. MetaboLights—an open-access general-purpose repository for metabolomics studies and associated meta-data. Nucleic Acids Res. 2013;41: D781–D786. doi:10.1093/nar/gks1004
83. Organization WH. WHO Monographs on Selected Medicinal Plants - Volume 2. World Health Organization; 1999.
84. Organization WH. WHO Monographs on Selected Medicinal Plants - Volume 4. World Health Organization; 2009.
85. Polur H, Joshi T, Workman CT, Lavekar G, Kouskoumvekaki I. Back to the Roots: Prediction of Biologically Active Natural Products from Ayurveda Traditional Medicine. Mol Inform. 2011;30: 181–187. doi:10.1002/minf.201000163
86. Palhares RM, Gonçalves Drummond M, dos Santos Alves Figueiredo Brasil B, Pereira Cosenza G, das Graças Lins Brandão M, Oliveira G. Medicinal Plants Recommended by the World Health Organization: DNA Barcode Identification Associated with Chemical Analyses Guarantees Their Quality. PLoS ONE. 2015;10. doi:10.1371/journal.pone.0127866
87. Xu J, Yang Y. Traditional Chinese medicine in the Chinese health care system. Health Policy. 2009;90: 133–139. doi:10.1016/j.healthpol.2008.09.003
88. Yuan H, Ma Q, Ye L, Piao G. The Traditional Medicine and Modern Medicine from Natural Products. Molecules. 2016;21: 559. doi:10.3390/molecules21050559
89. Chen CY-C. TCM Database@Taiwan: The World's Largest Traditional Chinese Medicine Database for Drug Screening In Silico. PLOS ONE. 2011;6: e15939. doi:10.1371/journal.pone.0015939
90. Chang K-W, Tsai T-Y, Chen K-C, Yang S-C, Huang H-J, Chang T-T, et al. 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. 2011;29: 243–250. doi:10.1080/073911011010524988
91. Huang J, Zheng Y, Wu W, Xie T, Yao H, Pang X, et al. CEMTDD: The database for elucidating the relationships among herbs, compounds, targets and related diseases for Chinese ethnic minority traditional drugs. Oncotarget. 2015;6: 17675–17684. doi:10.18632/oncotarget.3789
92. Qiao X, Hou T, Zhang W, Guo S, Xu X. A 3D Structure Database of Components from Chinese Traditional Medicinal Herbs. J Chem Inf Comput Sci. 2002;42: 481–489. doi:10.1021/ci010113h
93. Fang X, Shao L, Zhang H, Wang S. CHMIS-C: A Comprehensive Herbal Medicine Information System for Cancer. J Med Chem. 2005;48: 1481–1488. doi:10.1021/jm049838d
94. Xu H-Y, Zhang Y-Q, Liu Z-M, Chen T, Lv C-Y, Tang S-H, et al. ETCM: an encyclopaedia of traditional Chinese medicine. Nucleic Acids Res. 2019;47: D976–D982. doi:10.1093/nar/gky987
95. Kim S-K, Nam S, Jang H, Kim A, Lee J-J. TM-MC: a database of medicinal materials

and chemical compounds in Northeast Asian traditional medicine. BMC Complement Altern Med. 2015;15: 218. doi:10.1186/s12906-015-0758-5

96. TCMID: traditional Chinese medicine integrative database for herb molecular mechanism analysis. [cited 29 Apr 2019]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3531123/
97. Ru J, Li P, Wang J, Zhou W, Li B, Huang C, et al. TCMSP: a database of systems pharmacology for drug discovery from herbal medicines. J Cheminformatics. 2014;6: 13. doi:10.1186/1758-2946-6-13
98. Li B, Ma C, Zhao X, Hu Z, Du T, Xu X, et al. YaTCM: Yet another Traditional Chinese Medicine Database for Drug Discovery. Comput Struct Biotechnol J. 2018;16: 600–610. doi:10.1016/j.csbj.2018.11.002
99. Mohanraj K, Karthikeyan BS, Vivek-Ananth RP, Chand RPB, Aparna SR, Mangalapandi P, et al. IMPPAT: A curated database of Indian Medicinal Plants, Phytochemistry And Therapeutics. Sci Rep. 2018;8. doi:10.1038/s41598-018-22631-z
100. Potshangbam AM, Polavarapu R, Rathore RS, Naresh D, Prabhu NP, Potshangbam N, et al. MedPServer: A database for identification of therapeutic targets and novel leads pertaining to natural products. Chem Biol Drug Des. 2019;93: 438–446. doi:10.1111/cbdd.13430
101. Meetei PA, Singh P, Nongdam P, Prabhu NP, Rathore R, Vindal V. NeMedPlant: a database of therapeutic applications and chemical constituents of medicinal plants from north-east region of India. Bioinformation. 2012;8: 209–211. doi:10.6026/97320630008209
102. Pathania S, Ramakrishnan SM, Bagler G. Phytochemica: a platform to explore phytochemicals of medicinal plants. Database. 2015;2015. doi:10.1093/database/bav075
103. Yanuar A, Mun'im A, Lagho ABA, Syahdi RR, Rahmat M, Suhartanto H. Medicinal Plants Database and Three Dimensional Structure of the Chemical Compounds from Medicinal Plants in Indonesia. ArXiv11117183 Q-Bio. 2011 [cited 22 Oct 2019]. Available: http://arxiv.org/abs/1111.7183
104. Tung C-W, Lin Y-C, Chang H-S, Wang C-C, Chen I-S, Jheng J-L, et al. TIPdb-3D: the three-dimensional structure database of phytochemicals from Taiwan indigenous plants. Database. 2014;2014. doi:10.1093/database/bau055
105. Ntie-Kang F, Zofou D, Babiaka SB, Meudom R, Scharfe M, Lifongo LL, et al. AfroDb: A Select Highly Potent and Diverse Natural Product Library from African Medicinal Plants. PLOS ONE. 2013;8: e78085. doi:10.1371/journal.pone.0078085
106. Ntie-Kang F, Onguéné PA, Fotso GW, Andrae-Marobela K, Bezabih M, Ndom JC, et al. Virtualizing the p-ANAPL Library: A Step towards Drug Discovery from African Medicinal Plants. PLOS ONE. 2014;9: e90655. doi:10.1371/journal.pone.0090655
107. Ntie-Kang F, Nwodo JN, Ibezim A, Simoben CV, Karaman B, Ngwa VF, et al. Molecular Modeling of Potential Anticancer Agents from African Medicinal Plants. J Chem Inf Model. 2014;54: 2433–2450. doi:10.1021/ci5003697
108. Onguéné PA, Ntie-Kang F, Mbah JA, Lifongo LL, Ndom JC, Sippl W, et al. The potential of anti-malarial compounds derived from African medicinal plants, part III: an in silico evaluation of drug metabolism and pharmacokinetics profiling. Org Med Chem Lett. 2014;4: 6. doi:10.1186/s13588-014-0006-x
109. Ibezim A, Debnath B, Ntie-Kang F, Mbah CJ, Nwodo NJ. Binding of anti-Trypanosoma natural products from African flora against selected drug targets: a docking study. Med Chem Res. 2017;26: 562–579. doi:10.1007/s00044-016-1764-y
110. Ntie-Kang F, Mbah JA, Mbaze LM, Lifongo LL, Scharfe M, Hanna JN, et al. CamMedNP: Building the Cameroonian 3D structural natural products database for virtual screening. BMC Complement Altern Med. 2013;13: 88.

doi:10.1186/1472-6882-13-88
111. Ntie-Kang F, Amoa Onguéné P, Scharfe M, Owono LCO, Megnassan E, Meva'a Mbaze L, et al. ConMedNP: a natural product library from Central African medicinal plants for drug discovery. RSC Adv. 2014;4: 409–419. doi:10.1039/C3RA43754J
112. Bultum LE, Woyessa AM, Lee D. ETM-DB: integrated Ethiopian traditional herbal medicine and phytochemicals database. BMC Complement Altern Med. 2019;19: 212. doi:10.1186/s12906-019-2634-1
113. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. 2018;46: D1074–D1082. doi:10.1093/nar/gkx1037
114. DrugBank | nutraceutical search. [cited 17 Oct 2019]. Available: https://www.drugbank.ca/drugs?utf8=%E2%9C%93&nutraceutical=1&filter=true
115. Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, Liu Y, et al. DrugBank 4.0: shedding new light on drug metabolism. Nucleic Acids Res. 2014;42: D1091–D1097. doi:10.1093/nar/gkt1068
116. Gilson MK, Liu T, Baitaluk M, Nicola G, Hwang L, Chong J. BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. Nucleic Acids Res. 2016;44: D1045–D1053. doi:10.1093/nar/gkv1072
117. Novel Antibiotics Database. [cited 18 Oct 2019]. Available: http://www.antibiotics.or.jp/journal/database/database-top.htm
118. Tomasulo P. ChemIDplus-Super Source for Chemical and Drug Information. Med Ref Serv Q. 2002;21: 53–59. doi:10.1300/J115v21n01_04
119. Ye H, Ye L, Kang H, Zhang D, Tao L, Tang K, et al. HIT: linking herbal active ingredients to targets. Nucleic Acids Res. 2011;39: D1055–D1059. doi:10.1093/nar/gkq1165
120. Kang H, Tang K, Liu Q, Sun Y, Huang Q, Zhu R, et al. HIM-herbal ingredients in-vivo metabolism database. J Cheminformatics. 2013;5: 28. doi:10.1186/1758-2946-5-28
121. Choi H, Cho SY, Pak HJ, Kim Y, Choi J, Lee YJ, et al. NPCARE: database of natural products and fractional extracts for cancer regulation. J Cheminformatics. 2017;9: 2. doi:10.1186/s13321-016-0188-5
122. Vetrivel U, Subramanian N, Pilla K. InPACdb Indian plant anticancer compounds database. Bioinformation. 2009;4: 71–74. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2823384/
123. Dr.V,Umashankar. InPACdb | Indian-Plant-Anticancer-Compound-DB. 29 Mar 2018 [cited 17 Oct 2019]. Available: https://github.com/inpacdb/Indian-Plant-Anticancer-Compound-DB-inpacdb-
124. Mangal M, Sagar P, Singh H, Raghava GPS, Agarwal SM. NPACT: Naturally Occurring Plant-based Anti-cancer Compound-Activity-Target database. Nucleic Acids Res. 2013;41: D1124–D1129. doi:10.1093/nar/gks1047
125. Compound Sets - NCI DTP Data - National Cancer Institute - Confluence Wiki. [cited 18 Oct 2019]. Available: https://wiki.nci.nih.gov/display/NCIDTPdata/Compound+Sets
126. Zhang R, Lin J, Zou Y, Zhang X-J, Xiao W-L. Chemical Space and Biological Target Network of Anti-Inflammatory Natural Products. J Chem Inf Model. 2019;59: 66–73. doi:10.1021/acs.jcim.8b00560
127. Sharma A, Dutta P, Sharma M, Rajput NK, Dodiya B, Georrge JJ, et al. BioPhytMol: a drug discovery community resource on anti-mycobacterial phytomolecules and plant extracts. J Cheminformatics. 2014;6: 46. doi:10.1186/s13321-014-0046-2
128. OSM - Open Source Malaria. [cited 18 Oct 2019]. Available: http://opensourcemalaria.org/
129. Williamson AE, Ylioja PM, Robertson MN, Antonova-Koch Y, Avery V, Baell JB, et al.

Open Source Drug Discovery: Highly Potent Antimalarial Compounds Derived from the Tres Cantos Arylpyrroles. ACS Cent Sci. 2016;2: 687–701. doi:10.1021/acscentsci.6b00086

130. Rothwell JA, Perez-Jimenez J, Neveu V, Medina-Remón A, M'Hiri N, García-Lobato P, et al. Phenol-Explorer 3.0: a major update of the Phenol-Explorer database to incorporate data on the effects of food processing on polyphenol content. Database. 2013;2013. doi:10.1093/database/bat070
131. PhytoHub. [cited 16 Oct 2019]. Available: http://phytohub.eu/
132. Neveu V, Moussy A, Rouaix H, Wedekind R, Pon A, Knox C, et al. Exposome-Explorer: a manually-curated database on biomarkers of exposure to dietary and environmental factors. Nucleic Acids Res. 2017;45: D979–D984. doi:10.1093/nar/gkw980
133. Lim E, Pon A, Djoumbou Y, Knox C, Shrivastava S, Guo AC, et al. T3DB: a comprehensively annotated database of common toxins and their targets. Nucleic Acids Res. 2010;38: D781–D786. doi:10.1093/nar/gkp934
134. He Q-Y, He Q-Z, Deng X-C, Yao L, Meng E, Liu Z-H, et al. ATDB: a uni-database platform for animal toxins. Nucleic Acids Res. 2008;36: D293–D297. doi:10.1093/nar/gkm832
135. International Venom and Toxin Database. Available: http://www.kingsnake.com/toxinology/
136. Snake Neurotoxin Database. Available: http://sdmc.i2r.a-star.edu.sg/Templar/DB/snake_neurotoxin/
137. MOLLUSK toxin database. Available: http://research.i2r.a-star.edu.sg/MOLLUSK/
138. Srinivasan KN, Gopalakrishnakone P, Tan PT, Chew KC, Cheng B, Kini RM, et al. SCORPION, a molecular database of scorpion toxins. Toxicon. 2002;40: 23–31. doi:10.1016/S0041-0101(01)00182-9
139. Günthardt BF, Hollender J, Hungerbühler K, Scheringer M, Bucheli TD. Comprehensive Toxic Plants–Phytotoxins Database and Its Application in Assessing Aquatic Micropollution Potential. J Agric Food Chem. 2018;66: 7577–7588. doi:10.1021/acs.jafc.8b01639
140. Yabuzaki J. Carotenoids Database: structures, chemical fingerprints and distribution among organisms. Database J Biol Databases Curation. 2017;2017. doi:10.1093/database/bax004
141. Rodriguez-Amaya DB, Kimura M, Godoy HT, Amaya-Farfan J. Updated Brazilian database on food carotenoids: Factors affecting carotenoid composition. J Food Compos Anal. 2008;21: 445–463. doi:10.1016/j.jfca.2008.04.001
142. Pilón-Jiménez BA, Saldívar-González FI, Díaz-Eufracio BI, Medina-Franco JL. BIOFACQUIM: A Mexican Compound Database of Natural Products. Biomolecules. 2019;9: 31. doi:10.3390/biom9010031
143. Pilon AC, Valli M, Dametto AC, Pinto MEF, Freire RT, Castro-Gamboa I, et al. NuBBEDB: an updated database to uncover chemical and biological information from Brazilian biodiversity. Sci Rep. 2017;7: 7215. doi:10.1038/s41598-017-07451-x
144. UEFS Natural Products. [cited 6 Nov 2019]. Available: http://zinc12.docking.org/catalogs/uefsnp
145. Ntie-Kang F, Telukunta KK, Döring K, Simoben CV, A. Moumbock AF, Malange YI, et al. NANPDB: A Resource for Natural Products from Northern African Sources. J Nat Prod. 2017;80: 2067–2076. doi:10.1021/acs.jnatprod.7b00283
146. Hatherley R, Brown DK, Musyoka TM, Penkler DL, Faya N, Lobb KA, et al. SANCDB: a South African natural compound database. J Cheminformatics. 2015;7: 29. doi:10.1186/s13321-015-0080-8
147. Derese S, Oyim J, Rogo M, Ndakala A. Mitishamba database: a web based in silico

database of natural products from Kenya plants. 2015.
148. Ashfaq UA, Mumtaz A, Qamar T ul, Fatima T. MAPS Database: Medicinal plant Activities, Phytochemical and Structural Database. Bioinformation. 2013;9: 993–995. doi:10.6026/97320630009993
149. Nguyen-Vo T-H, Le T, Pham D, Nguyen T, Le P, Nguyen A, et al. VIETHERB: A Database for Vietnamese Herbal Species. J Chem Inf Model. 2019;59: 1–9. doi:10.1021/acs.jcim.8b00399
150. Journal of Natural Products. Available: https://pubs.acs.org/journal/jnprdf
151. Marine Drugs. Available: https://www.mdpi.com/journal/marinedrugs
152. A database of natural products and chemical entities from marine habitat. [cited 6 Nov 2019]. Available: http://www.bioinformation.net/003/003000032008.htm
153. Lei J, Zhou J. A Marine Natural Product Database. J Chem Inf Comput Sci. 2002;42: 742–748. doi:10.1021/ci010111x
154. Sagar S, Kaur M, Radovanovic A, Bajic VB. Dragon exploration system on marine sponge compounds interactions. J Cheminformatics. 2013;5: 11. doi:10.1186/1758-2946-5-11
155. Davis GDJ, Vasanthi AHR. Seaweed metabolite database (SWMD): A database of natural compounds from marine algae. Bioinformation. 2011;5: 361–364. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3053594/
156. Ambinter-Greenpharma natural compound library (GPNCL). In: Greenpharma [Internet]. [cited 9 Oct 2019]. Available: https://www.greenpharma.com/products/compound-librairies/
157. ChemBridge | Screening Library | Diversity Libraries. [cited 16 Oct 2019]. Available: https://www.chembridge.com/screening_libraries/diversity_libraries/
158. LOPAC1280 – Library of Pharmacologically Active Compounds. In: Sigma-Aldrich [Internet]. [cited 16 Oct 2019]. Available: https://www.sigmaaldrich.com/life-science/cell-biology/bioactive-small-molecules/lopac1280-navigator.html
159. Prestwick Chemical – The Prestwick Phytochemical Library, a collection of natural products. [cited 16 Oct 2019]. Available: http://www.prestwickchemical.com/libraries-screening-lib-phyto.html
160. Targetmol | Natural Compound Library. [cited 16 Oct 2019]. Available: https://www.targetmol.com/compound-library/Natural-Compounds-Library
161. AnalytiCon Discovery, Screening Libraries. In: AnalytiCon Discovery [Internet]. [cited 16 Oct 2019]. Available: https://ac-discovery.com/screening-libraries/
162. InterBioScreen | Natural Compounds. Available: https://www.ibscreen.com/natural-compounds
163. INDOFINE Chemical Company. [cited 16 Oct 2019]. Available: http://www.indofinechemical.com/Media/sdf/sdf_files.aspx
164. Pi Chemicals System. [cited 16 Oct 2019]. Available: http://www.pipharm.com/catalog_products/list?category=28
165. Specs - Compound Management services and Research Compounds for the Life Science industry. [cited 16 Oct 2019]. Available: https://www.specs.net/index.php
166. ZINC Specs Natural Products. [cited 16 Oct 2019]. Available: http://zinc.docking.org/catalogs/specsnp/
167. Ertl P, Roggo S, Schuffenhauer A. Natural Product-likeness Score and Its Application for Prioritization of Compound Libraries. J Chem Inf Model. 2008;48: 68–74. doi:10.1021/ci700286x
168. Sorokina M, Steinbeck C. NaPLeS: a natural products likeness scorer—web application and database. J Cheminformatics. 2019;11: 55. doi:10.1186/s13321-019-0378-z

169. Willighagen EL, Mayfield JW, Alvarsson J, Berg A, Carlsson L, Jeliazkova N, et al. The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. J Cheminformatics. 2017;9: 33. doi:10.1186/s13321-017-0220-4
170. Otasek D, Morris JH, Bouças J, Pico AR, Demchak B. Cytoscape Automation: empowering workflow-based network analysis. Genome Biol. 2019;20: 185. doi:10.1186/s13059-019-1758-4