Published in final edited form as:
Nat Prod Rep. 2021 January 01; 38(1): 264–278. doi:10.1039/d0np00053a.

Microbial natural product databases: Moving forward in the multi-omics era

Jeffrey A. van Santen[a], **Satria A. Kautsar**[b], **Marnix H. Medema**[b], **Roger G. Linington**[a]

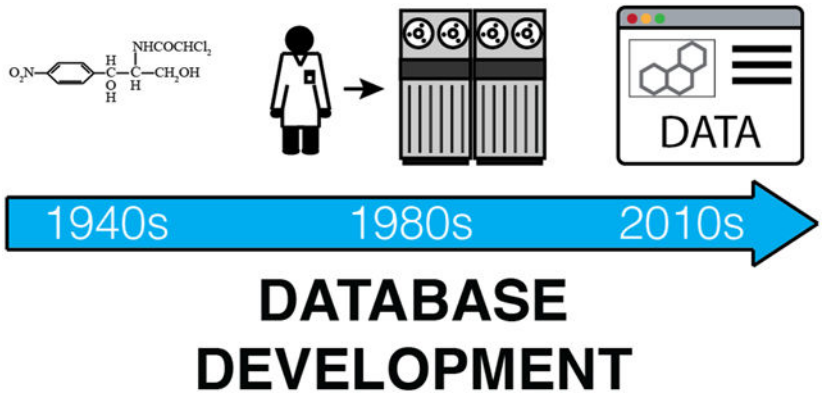
[a]Department of Chemistry, Simon Fraser University, Burnaby, CA, USA

[b]Bioinformatics Group, Wageningen University, Wageningen, NL

Abstract

The digital revolution is driving significant changes in how people store, distribute, and use information. With the advent of new technologies around linked data, machine learning and large-scale network inference, the natural products research field is beginning to embrace real-time sharing and large-scale analysis of digitized experimental data. Databases play a key role in this, as they allow systematic annotation and storage of data for both basic and advanced applications. The quality of the content, structure, and accessibility of these databases all contribute to their usefulness for the scientific community in practice. This review covers the development of databases relevant for microbial natural product discovery during the past decade (2010-2020), including repositories of chemical structures/properties, metabolomics, and genomic data (biosynthetic gene clusters). It provides an overview of the most important databases and their functionalities, highlights some early meta-analyses using such databases, and discusses some basic principles to enable widespread interoperability between databases. Furthermore, it points out conceptual and practical challenges in the curation and usage of natural products databases. Finally, the review closes with a discussion of key action points required for the field moving forward, not only for database developers but for any scientist active in the field.

Graphical Abstract



rliningt@sfu.ca.

6. Conflicts of interest

MHM is a co-founder of Design Pharmaceuticals and a member of the scientific advisory board of Hexagon Bio.

Online databases are becoming key to natural product research, as publication of data is increasingly digitized. Here, we review databases of chemical structures, gene clusters and analytical data, and discuss key challenges and opportunities.

1. Introduction

Information management remains a central limitation in natural products science. Access to comprehensive, structured, freely available repositories containing key data allows researchers to determine what has been found to date, understand how previous discoveries relate to new findings, and identify how new results fit into the broader picture of natural products diversity and biosynthesis. In this review we will present the current landscape of databases for microbial natural products science, and discuss how to address the challenges and limitations facing the field as we move towards the implementation of large, comprehensive, integrated data architectures for natural products data and metadata.

1.1. A brief history of natural products data management

Although we now take for granted the rapid, facile access to electronic data on natural products, this is a relatively recent development (Fig. 1). Prior to the 1990s, there were essentially no online scientific databases containing information on natural products. Instead, most data management strategies involved the laborious transcription of key data from print journals to index cards for use in individual laboratories. It cannot be overstated how much this lack of access to comprehensive, ordered datasets has negatively impacted our field. Asking senior researchers about historical data management approaches yields a litany of stories describing painful days spent chasing information through the print literature. These stories include such historical curiosities as punch cards, 8” floppy discs, photocopier accounts, suitcase sized ‘laptops’, and early mainframe computers.

During this period, numerous print reference books were maintained that collated key data from the scientific literature. Of particular note were the Chemical Abstracts series, the Ring Systems Handbook,[1] Fungal Metabolites volumes 1 and 2,[2,3] the Handbook of Antibiotic Compounds (volumes 1 - 14),[4] the Index on Antibiotics from Actinomycetes volumes 1 and 2,[5,6] and the Encyclopedia of Antibiotics.[7]

Searching through such compendia was inherently slow, and instances of rediscovery were common. To reduce redundant effort by individual researchers, many organizations began to develop their own in-house data collections. A representative example of this type of resource is the system developed by the pharmaceutical company Lederle Laboratories beginning in the early 1960s, as described by Dr. Guy Carter:

“Lederle Laboratories maintained its own ‘database’, dubbed the Antibiotic Properties file, of which we were very proud. The database consisted of a series of 3-ring binders, arranged in alphabetical order, holding a single page of information on each antibiotic including structure (if known, and a surprising number were not), biological spectrum and any other bio data, like cytotoxicity, and chemical properties that were known, like elemental analysis, mw and most importantly a UV spectrum - frequently xeroxed from the original paper and

pasted on the form. The database was maintained by the Lederle library staff, and was compiled by Lederle retirees, who were hired to review the literature for new compounds - quite a system!”

In the 1980s, several important electronic resources began to emerge. CAS and Beilstein began developing the large-scale literature databases that have become Scifinder and Reaxys. Initially these tools had very strict fee-for-search models that often limited the number of searches that researchers could perform in a given month. Gradually, this evolved to the institution subscription model we know today. In the area of natural products, two academic efforts are of particular note. Professor Hartmut Laatsch created AntiBase,[8] a database of microbial natural products, while Professors John Blunt and Murray Munro created MarinLit, a database of articles on marine natural products. Both resources were originally available on CD-ROM by paying an annual subscription to the developers to support development costs.

Commercial publishers were also developing electronic databases. For example, CRC Press began to publish the Dictionary of Natural Products,[9] which also came with a CD-ROM containing a basic search engine. These various electronic resources developed incrementally over the following decades, and remain the reference tools of choice for many natural products research groups around the world today.

1.2. A new age in natural products discovery

The early 2010s were marked by the emergence of new tools that made data-centric methods accessible to the ‘average’ natural products scientist; one without a dedicated training in programming or computer science. Examples of such tools include NaPDoS[10] and eSNaPD,[11] for assessing the biosynthetic diversity of microbial strains, FuSiOn[12] for the de novo prediction of compound modes of action, and iSNAP[13] for the dereplication of non-ribosomal peptides from mass spectrometry data.

One tool that had a significant impact on the adoption of new data technologies was antiSMASH.[14-18] First released in 2011, antiSMASH provided a simple, freely accessible web interface for the identification of biosynthetic gene clusters (BGCs) from genomic sequence data. The natural products community quickly recognized the power that such analyses could bring to many aspects of their research programs, and antiSMASH became a mainstay tool for many natural product programs. Instead of requiring subject experts to scan raw sequence data by hand, antiSMASH offered users a straightforward mechanism to generate initial automated annotations, which could then be prioritized for further investigation. The accessibility and power of this new resource set the tone for natural product tool development, and generated an immediate demand for new tools that would provide the same level of functionality in other areas of natural products.

1.3. Data storage, dissemination and collaboration

The exponential growth in omics research and so called “Big Data” is self-evident. The world's data volume has grown from about 1.5 zettabytes (ZB, 10^{21}) in 2009 to a projected 44 ZB by 2020.[19] Current models suggest that the global data volume will reach 175 ZB by 2025.[20]

In this age of internet and digital information, there is an increasing need to store and share not only raw experimental data but also analysis results, processed data, research protocols, knowledge materials and scientific findings. Gone are the days where scientists spent days scouring the library for answers and waiting for the next delivery of printed journals to keep track of what was happening in their field. Nowadays, people can disseminate, query, and even collaborate on research data with others around the globe in real time and in a large-scale fashion (e.g. crowdsource efforts). In this modern approach to science, databases play an essential role in ensuring that the data being generated are stored, processed, presented and shared in the most effective means.

To enable effective data storage and collaboration, databases should adhere to FAIR (findable, accessible, interoperable and reusable) principles in their implementation.[21,22] This is particularly important for inclusion of researchers from developing nations, where subscription cost for commercial tools can present an insurmountable barrier to access. Many companies provide mechanisms for reduced cost or free journal access to researchers from selected countries, but for low-to-middle income countries that are not included, data access remains a significant barrier to scientific development. This barrier can be significantly reduced by creating high-quality FAIR-compliant resources.

2. Databases for microbial natural products research

2.1. Chemical structure and properties databases

The current landscape for natural product structural databases is highly fragmented. A recent comprehensive review by Sorokina and Steinbeck[23] lists an astonishing 122 resources for natural product structures developed since the year 2000. This list includes both commercial and non-commercial repositories, covering a wide range of source organisms and geographic locations. However, despite the breadth of natural product databases available, the options for microbial natural product scientists are surprisingly limited. From the 122 resources, 50 permit access to the full set of structures. Of these, 11 contain entries for bacterial natural products, and only three (NPASS, StreptomeDB and the Natural Products Atlas) permit filtering by taxonomic origin to extract only the microbially-derived compounds. These three resources therefore currently represent the best freely available sources of information on microbial natural products structures (Fig. 2).

NPASS [http://bidd2.nus.edu.sg/NPASS/].[25]—NPASS is a recently developed natural products database (2018) designed to provide both source organisms and biological activities for natural products. It contains partial coverage of the chemical space of natural products from several taxonomic sources, including plants, invertebrates and microorganisms. In total it contains 35,032 compounds, of which approximately 9,000 are microbial in origin.

StreptomeDB [www.pharmbioinf.uni-freiburg.de/streptomedb3/].[26]—StreptomeDB is a targeted database that focuses exclusively on the bacterial genus *Streptomyces*. Recently updated in 2020, it contains 7,125 compounds with source organism information, as well as some bioactivity and spectral data.

The Natural Products Atlas [https://www.npatlas.org/].[27]**—**The Natural Products Atlas is a new resource (2019) designed to provide comprehensive coverage of all microbially-derived natural product structures. It currently contains 25,523 compounds (*v2019_12*) and is under active development. It features bi-directional links to two other natural products resources; the MIBiG database of biosynthetic gene clusters and the GNPS database of natural products mass spectra.

In addition to open source databases, a number of high-quality commercial platforms are available. Of these, the Dictionary of Natural Products (DNP), MarinLit and AntiBase are the most well established, although AntiBase was last updated in 2014. All three of these databases are large (>30,000 compounds) and contain rich metadata. They have broad coverage of the published literature and are generally very accurate. However, they have high annual subscription costs and do not permit bulk export of structural data or other information to external applications. This limits their utility to individual searches and precludes their integration with other natural products-based data resources.

DNP [http://dnp.chemnetbase.com/].—DNP contains over 290,000 entries (accessed Feb. 2020) and includes natural products from all major source organism groups, as well as physicochemical and biological data. The database is continually updated through an extensive process of manual curation by subject experts, ensuring high data quality standards. However, spot checks on the dataset based on compound names suggest that coverage is not universal, even for some well-known compound classes (e.g., abyssomicins).

MarinLit [http://pubs.rsc.org/marinlit/].—MarinLit is a literature database of marine natural products, including structures, taxonomy, and reports on total synthesis for 35,015 compounds (accessed Feb. 2020). It includes compounds from invertebrates and algae, as well as 8,082 compounds from marine-derived microorganisms. Impressively, this database is updated almost daily, making it the most contemporary resource in this area.

Dictionary of Antibiotics and Related Substances.[28]**—**The Dictionary of Antibiotics and Related Substances is a reference text of over 2,000 pages listing all known naturally occurring antibiotic substances (>10,000). It was recently updated (2013) from the original edition from the 1980s, and now includes many entries from the BMIC database, which was maintained for many years by Dr. Janos Berdy and was the foundational database for the Handbook of Antibiotic Compounds. It is accompanied by a searchable CD-ROM. There also exist numerous natural products databases from biotech and pharmaceutical companies, as discussed in section 1.1. Unfortunately, many of these are difficult, if not impossible, to obtain. Most are not under active development, and are archived in only physical formats, or in legacy database structures. Despite willingness from some companies to release these data to the wider community, access can be precluded by practical challenges such as completing liability release documentation; a task of typically low priority for legal departments.

Finally, it is worth mentioning the natural products coverage of the two largest chemical literature databases; Scifinder and Reaxys. Both of these platforms include the majority of compounds from the natural products literature. However, neither is particularly well suited

to natural products-based queries beyond simple structure searches. Scifinder does not include any flags identifying compounds as natural products, making it impossible to separate natural products from synthetic compounds. Reaxys does include the term ‘Isolated from Natural Source’ but many known natural products are not annotated with this flag, meaning that searches performed using this filter are not comprehensive.

2.2. Biosynthetic gene cluster databases

As the rate of BGC discovery began to accelerate in the early 2000s, the biosynthesis community faced many of the same challenges that had been encountered by the natural products structure elucidation community thirty years earlier. In particular, information about BGC discovery was becoming scattered across the scientific literature, or stored in a less structured manner in genomic databases such as NCBI GenBank. As with structure-based discovery, this limited the possibilities for cross-linking between resources and prevented programmable access to exploit the knowledge within. To address this issue, several databases of BGC data have been developed.

ClusterMine360 [http://clustermine360.ca].[29]**—**Made available in 2013, ClusterMine360 was one of the first platforms to venture in to the task of cataloguing the information on experimentally validated BGCs with known products. Focusing on the Nonribosomal Peptide (NRP) and Polyketide (PK) classes, it contains 300 BGCs linked to their chemical products. While initially prepared for continuous expansion via user-submitted annotations, it seems that the total number of BGCs covered by the database has not increased significantly since its initial release.

DoBISCUIT [https://www.nite.go.jp/en/nbrc/genome/dobiscuit.html].[30]**—**Released around the same time as ClusterMine360, DoBISCUIT published an initial collection of 72 known PK BGCs. Unfortunately, the database is no longer accessible, although its main page is still active and shows a final log of 108 BGCs recorded on December 27, 2016.

MIBiG Repository [https://mibig.secondarymetabolites.org].[31,32]**—**In 2015, a coordinated effort of more than 150 natural product scientists resulted in the publication of the Minimum Information about a Biosynthetic Gene Cluster (MIBiG) data standard and repository for known and experimentally characterized BGCs. Holding information on more than a thousand of characterized BGCs, MIBiG was quickly adopted by the community as a central reference database for BGC data. Notably, antiSMASH[16] automatically compares each detected BGC to all reference gene clusters from MIBiG. Four years after the initial release, in 2019, a second iteration of both the database and schema was announced, highlighting an accumulated total of 2,021 BGC entries and a major overhaul of its online repository infrastructure. MIBiG contains only BGCs which have been experimentally verified to be responsible for the production of one or more known natural products. MIBiG entries are also subject to extensive manual curation and annotation by both the developers and the scientific community, further increasing the information content and data quality in this repository.

IMG-ABC [https://img.igi.doe.gov/cgi-bin/abc/main.cgi].[33]—Taking advantage of the Joint Genome Institute (JGI)'s extensive bacterial genomic platform, IMG/M, the IMG-ABC sets out to be the most comprehensive and feature-rich database of known (indirectly sourced from MIBiG) and computationally predicted bacterial BGCs. Prior to IMG-ABC v5, the database comprised a total of more than one million BGCs predicted using both antiSMASH and the ClusterFinder algorithm.[34] The latter approach has since been dropped in favour of the more stringent but more 'high-confidence' BGC class detection of antiSMASH 5. This has resulted in a drop of total BGCs provided by IMG-ABC, with 410,558 BGCs available as of 29 June 2020.

An important detail to note is that, due to the JGI's Data Usage policy (https://jgi.doe.gov/user-programs/pmo-overview/policies/), it is not advisable to do bulk-analysis and publication of IMG-ABC's data as some of the genomes may still be under embargo. In the future, we recommend that IMG/M (and IMG-ABC) should follow the footsteps of their fungal genome database counterpart, MycoCosm (https://mycocosm.jgi.doe.gov/),[35] to provide a simple filtering of embargoed genomes, thus enabling a 'safe' bulk-download and analysis of their data.

antiSMASH Database [https://antismash-db.secondarymetabolites.org].[36,37]—The antiSMASH database (antiSMASH-DB) was initially released in 2016 by the same team who developed antiSMASH to act as a central repository for precomputed antiSMASH runs. In contrast to the IMG-ABC, antiSMASH-DB aims to provide a limited, dereplicated list of putative BGCs sourced from the highest quality bacterial genomes. For sets of highly similar genomes (e.g., thousands of *Escherichia coli* genomes with only a few single nucleotide polymorphisms), representatives have been picked instead of providing results for all strains individually. One key reason to do this is to provide a seamless integration with antiSMASH via its 'ClusterBlast' module, which performs a sequence comparison of each detected BGC with those in the database. Following its second release in 2018, antiSMASH-DB harbours a total of 152,106 BGCs pre-calculated from 24,776 bacterial genomes (of which 32,548 BGCs were derived from 6,200 complete genomes) from the NCBI RefSeq database.[38] The upcoming third release will include BGCs from high-quality fungal genomes as well.

2.3. Databases for metabolomics and analytical chemistry

A number of resources for the sharing and analysis of metabolomics data have arisen in the last decade. Many of these resources focus around the FAIR sharing of data to enable more productive natural products discovery, and are not limited to the scope of microbial natural products science.

The Global Natural Products Social molecular network (GNPS) [https://gnps.ucsd.edu/].[39]—The GNPS system is an ecosystem for sharing and analyzing tandem mass spectrometry data. It is built on the MassIVE platform, and features an impressive suite of internally connected tools. It also provides functionality for complete data lifecycle management, from data acquisition through to publication. One of the most popular features is molecular networking, which enables the visualization relationships

between spectra from MS/MS experiments. Data submitted for analysis in GNPS are organized into datasets, which can either be kept private or made public. To date there are 1,413 public datasets available online (accessed Feb. 24, 2020). In addition, GNPS houses a number of public MS/MS spectral libraries, containing 74,130 annotated spectra.

MetaboLights [https://www.ebi.ac.uk/metabolights/].[40]—MetaboLights is a database run by EMBL-EBI that was originally created in 2012, and overhauled in 2019. It is a database for metabolomics data with capabilities for storing and reporting on a large variety of data types, including NMR, GC/MS, LC/MS, as well as metabolite structures, their reference spectra, and biological roles. MetaboLights is the recommended repository for metabolomics data for a number of journals based on the FAIRsharing initiative [https://fairsharing.org/biodbcore-000168/].

NMR Metabolomics—A recent comprehensive review by McAlpine *et al.*[41] established the state of NMR dereplication with respect to the field of natural products. The review demonstrates that there remains an urgent need for a comprehensive and open data exchange of NMR data for natural products. Following publication of this review, the National Center for Complementary and Integrative Health and the Office of Dietary Supplements at the NIH in the US initiated a call for proposals to develop such a resource.[42] This call resulted in the establishment in 2020 of the **Natural Products Magnetic Resonance Database** (NP-MRD; www.np-mrd.org) which aims to create an open access repository of experimental and calculated spectra for natural products structures.

In addition to this new initiative there are a number of current databases and tools which have addressed this problem with both experimental and predicted NMR spectra.

NAPROC-13 [http://c13.usal.es/].[43]—NAPROC-13 is a database which contains ^{13}C NMR spectra for over 6,000 natural product compounds. The database has a web interface allowing for rapid identification of compounds present in complex mixtures, as well as providing structural information useful for novel structure elucidation.

NMRshiftDB [https://nmrshiftdb.nmr.uni-koeln.de/].[44]—NMRshiftDB contains many similar features to NAPROC-13 as well as NMR from other nuclei. However, it is not exclusive to natural products chemistry.

Biological Magnetic Resonance Data Bank (BMRB) [http://www.bmrb.wisc.edu/].[45]—BMRB contains a wide variety of experimental and simulated NMR data from proteins, peptides, nucleic acids, and other biomolecules. BMRB is not exclusive to microbial natural products, and also contains data from all realms of natural products and metabolomics. BMRB also maintains a library of NMR pulse sequences and computational software for biomolecular NMR.

Human Metabolome Database (HMDB) [https://hmdb.ca/].[46]—HMDB is an open-access database which provides detailed information about metabolites found in the human body, thus including those essential to the human microbiome. Many metabolites also contain experimental 1D and 2D NMR spectra, freely available for download.

CH-NMR-NP [https://www.j-resonance.com/en/nmrdb/].[47]—CH-NMR-NP is a database hosted by JEOL of NMR data compiled from a list of journals from 2000-2014. It contains ^{1}H and ^{13}C NMR data from approximately 35,500 natural products and is not exclusive to microbial natural products. CH-NMR-NP is searchable online and permits download of the NMR data in the JEOL Delta data format on a compound-by-compound basis.

3. Database curation and usage

3.1. Practical challenges for database users

Surprisingly, it remains very difficult to compare data between resources in this area. Chemical structure and compound name are the common terms connecting many of these databases. In principle it should be possible to associate data from one resource (e.g. biosynthetic gene cluster) with data from another (e.g. NMR or MS data) via the chemical structure. In practice however, there is no agreed upon standardization method for chemical structures which provides a unique, machine readable structural representation without information loss. For example, several SMILES strings are possible for a single structure, standard InChI representations do not retain information on preferred tautomers, and MOL files are large blocks of text that are unwieldy to store in most database formats. These issues mean that databases typically align poorly by structure without significant additional manual curation.

Compound names are similarly challenging. Small changes in punctuation, the inclusion and encoding of special characters, or the absence of trivial names for many compounds in the literature all contribute to poor overlap between resources. This is further complicated by the assignment of new synonyms for existing compounds and, occasionally, the erroneous assignment of the same name to multiple structures. To add further complication, some compound classes receive several different parent names, often in an attempt to increase the visibility of new discoveries. Conversely, some researchers use the same parent name for all compounds isolated from a given organism, regardless of structural relatedness. Both of these issues complicate the grouping of related structures based on trivial names.

Some resources have invested substantial effort in improving interoperability. For example, the Natural Products Atlas and MIBiG teams have manually reviewed every entry in the MIBiG database and identified the appropriate Natural Products Atlas entry in each case. These two resources now include bi-directional links between data pages, and offer exportable tables that list links between primary keys in each platform. Similar links have been set up with the GNPS platform.

Investing similar effort to align other key resources by structure could have a significant impact on the development of new crossdiscipline discovery tools. An example of an effective cross-referencing system is provided by UniChem,[48] a system set up by the EMBL-EBI to connect chemical structures across multiple databases by assigning a UniChem identifier to each unique chemical structure, and linking this identifier to all the databases affiliated with the UniChem system.

3.2. Practical challenges for database creation and management

The current publishing model is not well suited to large-scale database creation and maintenance. Each journal has its own format and data requirements, and no journals produce standardized, machine readable files containing key primary data (Fig. 3). Rather, these data are often provided as supplementary materials in a wide variety of formats. Deposition of data to public resources (e.g. depositing biosynthetic gene clusters with NCBI) is valuable, but accession numbers must still be extracted manually from the methods or data availability sections of the papers, slowing the rate of data curation.

For chemical structures, the situation is even more difficult. Most authors do not deposit new structures to public databases (e.g. PubChem[49] or ChEBI[50]), meaning that structures start as computerized representations, (e.g. ChemDraw files) are reproduced by journals as flat images in PDFs, and must then be manually re-entered in machine readable formats. This medieval approach to information dissemination is a significant barrier to data integration efforts, and one that the community must urgently address. The American Chemical Society style guide includes a clear summary of many of the challenges surrounding machine interpretation of printed structures.[51]

We propose that editors require a SMILES string in the manuscript for every new compound, as an additional component of the experimental data section. Although this is not a substitute for a separate structured data file (e.g. MDL SDF or structured JSON), it is easy to implement and would improve the digitalization of natural products research results by increasing structure availability and reducing error rates caused by manual re-entry of compound structures. Initiatives of some journals, such as *Nature Chemical Biology,*[52] to collect such data and automatically submit all published structures to the PubChem database in a computer-readable format show that this is feasible.

For BGCs the problem is sometimes even worse as, unlike chemical structures, digital representations of BGC sequences cannot be reconstructed from images in a paper. Hence, deposition of the data to a public repository is absolutely required in order to assess a scientific paper on its merits, and to reproduce and leverage these results. The fact that many journals, even highly regarded ones such as the *Journal of the American Chemical Society,* regularly publish papers on BGCs without the sequence being made available anywhere is highly problematic. As is the case for proteins,[53] we feel that it is imperative that accession numbers to GenBank entries containing the BGC are explicitly mentioned in the paper. When a BGC is characterized from a genome sequence previously published by another research group, authors should refer to the accession number of that genome and the coordinates of the BGC within it, or at least provide locus tags of the genes or accession numbers of the encoded proteins, to allow readers and database developers to find the underlying data.

Ideally, every database should relate each data point to the appropriate reference from which these data were derived. This would allow users to evaluate data more carefully than aggregated datasets where data provenance is unknown. Fortunately, the digital object identifier (DOI) system provides a unique identifier for journal articles that is easily converted to a hyperlink to each article and provides a simple method for storing article

information. Frustratingly however, some publishers have not assigned DOIs to their legacy article collections. Because DOIs are not universally assigned, database systems must therefore handle both DOIs and full reference data (journal, volume, issue, pages). With the advent of e-journals that use non-standard citation formats, this has quickly become a complicated and error prone process. We therefore present a second recommendation that publishers review their legacy holdings and, where appropriate, assign DOIs to these back catalogues. This simple action would have a significant impact on the information content and interoperability of separate natural product-based data resources.

One final and often overlooked point is the cost of running and maintaining a database. Servers, IT staff, and continued software development are often forgotten in planning the longevity of data tools. Furthermore, a database may reach the end of its life due to funding or being superseded by another platform. Currently when this happens, data are often simply lost. One simple and effective solution is to store versioned releases of data dumps on a free scientific data storage solutions such as Zenodo (run by CERN and OpenAIRE,https://zenodo.org/) or GigaDB (run by the GigaScience journal,http://gigadb.org/). Otherwise, standard steps can be followed to archive a database.[54] Doing so can prevent the relegation of data to the annals of lost and forgotten databases and is best practice for FAIR data.

3.3. Curating microbial natural products data in 2020

Curating natural products data from the primary literature remains a predominantly manual process. It requires three main steps; identification of articles pertaining to microbial natural products discovery, extraction of structures, gene clusters and other data from each article, and organization of these data into a structured format. The most challenging of these is the identification of relevant articles. Traditionally, more than 50% of all microbial natural products discoveries were published in either the *Journal of Antibiotics* or the *Journal of Natural Products.* However, as natural products research has broadened in scope, the number of venues for reporting natural products discovery has increased. This creates challenges for data curation. Manual inspection of titles and abstracts for all published articles is now an impossibly large task. Instead, curation efforts must rely on either targeted curation of key journals, or text mining strategies using keywords to find relevant articles from public data sources such as PubMed. Both of these approaches have limitations that impact the coverage of curation efforts. Focus on a targeted list of journals can exclude reports in peripherally related areas (e.g. marine chemical ecology or microbiome studies) while text mining approaches are likely to miss core articles and are susceptible to bias depending on the algorithm(s) used for filtering. Authors can assist with this effort by ensuring that the discovery of new natural products or BGCs is prominently described in the abstract. In most cases, curators do not have bulk access to the full text versions of articles, meaning that the title and abstract are the only information available for article prioritization. A clear statement describing new compound or BGC discovery in the abstract is therefore the most effective method to ensure that new data are included in curation efforts.

3.4. Community contributions

A second route to data curation is through investigator-initiated submissions directly to databases. This approach has many clear advantages. It makes curation a distributed effort,

rather than relying on a small number of volunteers. This in turn improves both coverage and accuracy, because the original authors are providing the key data directly. It reduces effort because these data (e.g. structures) are already in an appropriate electronic format, and reduces error rates by eliminating instances where curators incorrectly interpret data from original articles.

There are however a number of disadvantages to the community contribution model. Databases without control over data insertion can quickly become corrupted through either accidental or malicious behavior. This may often be unintentional, as it is easy to misinterpret a step in a submission form and input the wrong data. In addition, submissions from external users may not conform to the defined scope of the database. Without appropriate care, the contents of the database can quickly become heterogeneous, making it difficult or impossible to perform meaningful analyses on the entire dataset.

To address these challenges, most platforms include a secondary curation step, where external submissions are reviewed by subject experts for appropriateness and completeness. This approach is much faster than de novo literature searching, as the core data have already been submitted in an appropriate format. To make sure that submitted data are as unambiguous as possible, a clear ontology detailing the options for each data field is required, as well as clear instructions and tutorials for submission.[55] From our experience with the Natural Products Atlas and MIBiG, approximately 50% of community submissions are accepted ‘as is’, with a further 35% requiring format or content corrections, and 15% being rejected as outside the scope of the database.

Currently, the Natural Products Atlas, MIBiG, MetaboLights and GNPS are four of the only natural products resources that accept external submissions. This is likely in part due to low demand, because of ‘submission fatigue’ from the ever-increasing list of requirements placed on corresponding authors. Initial submissions now require extensive information about authors and grants, and accepted articles must often be separately deposited in open repositories to satisfy funding agencies. To add to this, sequence data must typically be deposited in an open repository (e.g. NCBI) and crystal structures deposited with the Protein Data Bank or the Cambridge Structural Database. Understandably, uptake for voluntary submission of additional data is low. However, the power provided to the scientific community offered by the accumulation of data in these repositories cannot be overstated. It is up to the natural products field to lead the way in data deposition, and to develop new strategies that improve data coverage in these areas without increasing the burden on lead investigators. There are clear incentives for researchers to do so, including increased visibility and citation rates for their science, as well as the ability to see and use these data when navigating publicly available data resources.

4. Integration and Interoperability between database

4.1. Multi-omics and meta-analysis driven microbial natural products discovery

This area of natural products science is still in its infancy, but a number of important discoveries have already been enabled by the availability of comprehensive, well-structured datasets.

Global analyses performed with natural product databases—Several groups have performed recent meta-analyses on natural products science using natural product databases. Pye *et al.*[56] investigated the rate of novel compound discovery as a function of time and source organism type using a combination of commercial and in-house databases. They showed that, while the absolute number of novel scaffolds being discovered each year remains roughly constant, the number of derivative compounds being reported has increased dramatically over the past 30 years; currently, less than 10% of new marine and microbial compounds can be considered 'novel' scaffolds.

Pascolutti *et al.*[57] used the Dictionary of Natural Products (DNP) to identify small, 'fragment-like' natural products, and evaluate their physicochemical properties. They demonstrated that a subset of structures was representative of a large percentage of the total motif diversity in this sample set, and suggested that these molecules could form the foundation for future fragment-based screening libraries.

O'Hagan and Kell[58] took this premise one step further to ask which combination of 96, 384, 1152 or 1920 compounds would best represent the chemical space in Nature. Using a combination of the now-defunct Universal Natural Products Database[59] and DNP they were able to identify libraries that covered up to 30% of overall chemical space, and to propose a high coverage library made up entirely of commercially available natural products.

Global analyses have also been performed for BGCs, such as the study by Cimermancic et al.[34] in 2014, which surveyed the biosynthetic landscape across 1,154 sequenced bacterial and archaeal genomes, revealing widely distributed BGC classes of unknown function. Since then, the size of genomic databases has grown by orders of magnitude, however. As an example, NCBI RefSeq now holds more than 190,000 bacterial genomes compared to ±29,000 in late 2014, not to mention the rising availability of metagenome-assembled genome (MAG) sequences.[60-63] These newly available genomic data provide exciting opportunities to assess, for example, which taxonomic groups encode the richest natural product biosynthetic diversity and should therefore be targeted for discovery efforts, or how biosynthetic diversity is governed by species phylogeny versus ecology.[64]

New uses for structure databases—The availability of curated structure databases has enabled the development of a number of exciting extensions to existing analytical platforms. Reher *et al.*[65] recently published a new version of the Small Molecule Accurate Recognition Technology platform, termed SMART 2.0. This tool uses neural networks to match HSQC NMR spectra of unknown compounds against a database of known compounds. Using this approach, the SMART 2.0 algorithm predicts the identities of compound classes for unknown molecules directly from a single NMR spectrum. In this new release, the authors included calculated HSQC spectra based on structures from several natural products databases. This dramatically increased the number of reference spectra, from 2,054 in the original report to >53,000 in this new version.

In the area of mass spectrometry, a number of tools have been developed for the prediction of MS/MS fragmentation patterns.[66-69] These approaches provide a powerful new discovery modality for natural products researchers by providing an alternative to the need for

validated synthetic standards for all compounds. For example, the latest version of the CFM-ID platform, CFM-ID 3.0,[68] includes a large reference library of pre-calculated spectra, as well as online and local options for calculating spectra for bespoke compound libraries. Similarly, the new release of the SIRIUS platform (SIRIUS 4)[69] incorporates the CSI:FingerID platform[70] and predicts the most likely structure for signals from mass spectrometry data, based on comparison with a database of known structures. These complement additional tools, such as MS2LDA[71] and the associated MotifDB,[72] which provide annotation of metabolite substructures based on motifs found across databases of tandem mass spectra. The availability of both compound databases and tools like CFM-ID and SIRIUS therefore enables the creation of targeted annotation libraries based on specific parameters relevant to a given study (taxonomic origin, compound class, etc).

New uses for BGC databases—One of the most obvious uses of BGC databases is in the process of dereplication: identifying whether BGCs detected in a set of (meta)genome sequences are likely to encode known biosynthetic pathways or not. For example, Crits-Cristoph et al.[73] used the MIBiG database to show that >90% of BGCs they identified in metagenome-assembled genomes from uncultivated Acidobacteria, Verrucomicobia, Gemmatimonadetes, and Rokubacteria were likely to encode novel pathways. This process of dereplication can now also be automated for large genomic datasets using the BiG-SCAPE algorithm.[74] BiG-SCAPE computes sequence similarity networks from user-specified antiSMASH results together with all MIBiG database BGCs and reconstructs gene cluster families (GCFs), from which one can assess which BGCs are similar to a known BGC from MIBiG and which are not.

Another clear use case of BGC databases is to annotate functions in, for example, microbiome studies and using these annotations to infer ecological interactions. For example, Bahram et al.[75] used a set of MIBiG entries linked to products with proven antimicrobial functions to assess whether fungal antibiotic production potential is associated with the frequency of bacterial antibiotic resistance genes across topsoil metagenomes.

Furthermore, people have been using BGC databases like antiSMASH-DB to identify BGCs that contain specific combinations of genes of interest. For example, Krause et al.[76] performed pattern matching to chart the occurrence and diversity of PapR2-like regulators (SARP-type DNA-binding proteins with potential as generic activators for silent BGCs) within antiSMASH-DB, which revealed its widespread distribution across Actinobacterial genomes.

Another straightforward use of a BGC database is to chart the biosynthetic diversity of organisms within a larger taxonomic group.[77] Databases such as antiSMASH-DB make these analyses straightforward, by providing ready to use, pre-calculated BGC data and metadata (e.g., on their taxonomic origins) that can be accessed via an Application Programming Interface (API).

Finally, BGC databases also have potential to function as a ‘parts catalogue’ for pathway engineering using synthetic biology. For example, the ClusterCAD software[78] allows users to design new modular polyketide synthase assembly lines by sourcing polyketide BGCs and

polyketide synthase modules from MIBiG, and providing a graphical interface to mix and match these to build novel polyketide structures of interest. In principle, this type of computer-aided design could be expanded in various ways, e.g. by sourcing and searching any BGC from publicly available data in IMG/ABC or the antiSMASH database, or by, for example, including searches for genes encoding tailoring enzymes.

Examples of data integration between databases—There are very few examples of natural products discoveries made directly through the integration of multiple databases. This is no doubt due to the poor interoperability between most current resources, and the weak standardization of core data (structure representation, taxonomy, etc.). Some innovative research has been powered by combining chemical structure data with BGC data. For example, the GRAPE-GARLIC software pipeline[79] used retrobiosynthesis on an in-house database of chemical structures to reconstruct their monomer composition, which was the matched to monomers computationally predicted from BGC sequences found in public sequence databases. Similarly, integrating BGC data with metabolomics data has led to a range of approaches to (semi-)automatically link molecules to the genes involved in their biosynthesis based on pattern matching strategies.[80-82] There is clearly a vast opportunity for the development of new tools in these areas, and we look forward to seeing what the next decade will bring.

4.2. Enabling interoperability between databases

Natural products databases span a wide range of subject areas (structures, biosynthetic gene clusters, geographic origin, taxonomic origin etc). However, because the field is very large and data curation is slow, most databases are designed with narrow scope. This has led to a proliferation of small databases with partial overlap in terms of content, and no standardization of included fields.

A number of technologies exist which could facilitate the exchange of data between databases. In particular, the advent of the specifications for the Semantic Web (or Web 3.0) by the World Wide Web Consortium (W3C, https://www.w3.org/standards/semanticweb/) would greatly facilitate data interchange. These technologies include Resource Description Framework, Web Ontology Language, and JSON-LD, amongst many others. Implementing tools like this affords structured and linked datasets and is currently driving a change in how data is handled on the internet. Practically, these technologies make data machine-readable and are currently leveraged heavily by the web's largest driving forces, including Google and Amazon. Unfortunately, we have yet to see these technologies realized in the field of natural products. This is due in large part to the depth of technical knowledge required to implement these requirements.

A simpler approach is the development of web APIs with well-defined schemas for existing online tools. APIs can deliver data in JSON or XML format, permitting real-time extraction of information from different resources, and eliminating the need for the duplicate storage of key data. Replication of the same data in different repositories is a basic 'no-no' in database science, because of the challenges associated with ensuring that both copies are always correctly synchronized.

Creating APIs not only enables the faster development of front-end tools such as data summary dashboards or detailed data pages, but it also provides informaticians with methods to more easily access and interrogate data. This in turn reduces the barrier to access to ask new questions in the field, and catalyses the exploration and development of new ideas.

To be interoperable, databases require at least one unique field that is the same in each dataset (the 'primary key'). Realistically, chemical structures are the only practical option as the primary key between natural products databases. To be useful, structures must therefore be entered consistently in all cases. Database creators must decide how to handle a large number of complicated situations including: entering racemates as one compound or two, including or excluding salt forms, handling atropisomers and metal complexes, managing partial and missing configurations, identifying and updating structures that have been corrected in subsequent studies, etc.

An ideal scenario would be to have a central, comprehensive database of all natural product structures to which other resources could refer. This would vastly increase the speed of database creation (by eliminating the need to curate the structure component) and would automatically align all of these resources (via the central structure ID). Sadly, no such database currently exists. In the absence of such a resource, database managers are encouraged to cooperatively define compound standardization strategies, and to manually review and align structural data between resources. This unglamorous task receives little recognition in the community, meaning that it is a low priority for most academic research groups. Until the natural products community develops guidelines and standards for data curation, this situation will likely persist, which presents a considerable threat that the value and opportunity offered by comparing datasets from different subject areas will be lost.

5. Future perspective

Data-centric approaches have fundamentally altered the landscape in many areas of natural science. For example, from the laborious early determination of protein crystal structures in the 1960s, protein biochemistry has evolved to a sophisticated field where even non-experts can perform large-scale, automated docking studies of virtual libraries against almost any biological target. Similarly, the longstanding effort to create KEGG as an encyclopaedia of gene function[83] is enabling the development of tools for the automated annotation of gene function across genomes and metagenomes (e.g. BlastKOALA and GhostKOALA[84]).

Natural products science has yet to take full advantage of this changing landscape of scientific discovery. Many discovery programs remain focused on manual methods, without effectively leveraging prior knowledge in the field. This is evidenced by high rates of compound rediscovery and the heterologous expression of 'unusual' BGCs that turn out to produce well-known compound classes. While this cannot always be avoided, better data integration of chemical structure data, genomic data and metabolomic data has a clear potential to improve prioritization of research efforts.

The opportunities offered by developing new data-driven discovery methods are clear. However, it is unreasonable to expect that researchers involved in tool development will also create the basal datasets required to power these tools. Instead, we must commit resources to the creation of large, well-structured repositories of key information, and must develop a culture where data deposition of new results is a standard and expected part of the discovery workflow. If we can accomplish these goals, the return on this investment will be felt powerfully in every corner of natural products science.

Acknowledgements

We thank Drs. G. Carter, C. Pearce, J. Gloer, M. Balunas, S. Singh, J. Blunt and D. Newman for helpful discussions, and Dr. H. Potter for providing statistics on MarinLit content.

Funding for this work was provided by NIH grants U41-AT008718 and U24-AT010811 (RGL) and the Graduate School for Experimental Plant Sciences, The Netherlands (SAK).

8. References

1. Schulz H, Georgy U, Schulz H and Georgy U, in From CA to CAS online, Springer Berlin Heidelberg, 1994, pp. 118–123.
2. Turner WB, Fungal Metabolites *(Volume* 1), Academic Press Inc, 1971.
3. Turner WB, Fungal Metabolites *(Volume* 2), Academic Press Inc, 1983.
4. Bérdy J, CRC Handbook of Antibiotic Compounds, CRC Press, Boca Raton, Fla, 1980.
5. Umezawa H, Index of Antibiotics From Actinomycetes *(Volume* 1), University Park Press, 1967.
6. Umezawa H, Index of Antibiotics From Actinomycetes *(Volume* 2), University Park Press, 1979.
7. Glasby JS, Encyclopedia of Antibiotics, Wiley-Blackwell, 3rd edn., 1993.
8. Laatsch H, AntiBase: The Natural Compound Identifier, Wiley-VCH, 2017.
9. Buckingham J, Dictionary of Natural Products, CRC Press, 1993.
10. Ziemert N, Podell S, Penn K, Badger JH, Allen E and Jensen PR, PLoS One, 2012, 7, e34064. [PubMed: 22479523]
11. Reddy BVB, Milshteyn A, Charlop-Powers Z and Brady SF, Chem. Biol, 2014, 21, 1023–1033. [PubMed: 25065533]
12. Potts MB, Kim HS, Fisher KW, Hu Y, Carrasco YP, Bulut GB, Ou Y-H, Herrera-Herrera ML, Cubillos F, Mendiratta S, Xiao G, Hofree M, Ideker T, Xie Y, Huang L. J. -s., Lewis RE, MacMillan JB and White MA, Sci. Signal, 2013, 6, ra90. [PubMed: 24129700]
13. Ibrahim A, Yang L, Johnston C, Liu X, Ma B and Magarvey NA, Proc. Natl. Acad. Sci, 2012, 109, 19196–19201. [PubMed: 23132949]
14. Medema MH, Blin K, Cimermancic P, de Jager V, Zakrzewski P, Fischbach MA, Weber T, Takano E and Breitling R, Nucleic Acids Res., 2011, 39, W339–W346. [PubMed: 21672958]
15. Blin K, Medema MH, Kazempour D, Fischbach MA, Breitling R, Takano E and Weber T, Nucleic Acids Res., 2013, 41, W204–W212. [PubMed: 23737449]
16. Weber T, Blin K, Duddela S, Krug D, Kim HU, Bruccoleri R, Lee SY, Fischbach MA, Müller R, Wohlleben W, Breitling R, Takano E and Medema MH, Nucleic Acids Res., 2015, 43, W237–W243. [PubMed: 25948579]
17. Blin K, Wolf T, Chevrette MG, Lu X, Schwalen CJ, Kautsar SA, Suarez Duran HG, de los Santos ELC, Kim HU, Nave M, Dickschat JS, Mitchell DA, Shelest E, Breitling R, Takano E, Lee SY, Weber T and Medema MH, Nucleic Acids Res., 2017, 45, W36–W41. [PubMed: 28460038]
18. Blin K, Shaw S, Steinke K, Villebro R, Ziemert N, Lee SY, Medema MH and Weber T, Nucleic Acids Res., 2019, 47, W81–W87. [PubMed: 31032519]
19. Wang W and Krishnan E, JMIR Med. Informatics, 2014, 2, e1.
20. Reinsel D, Gantz J and Rydning J, The Digitization of the World - From Edge to Core. IDC White Paper, 2018.

21. Wilkinson MD, Dumontier M, Aalbersberg Ij. J., Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJG, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PAC, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S-A, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J and Mons B, Sci. Data, 2016, 3, 160018. [PubMed: 26978244]
22. Boeckhout M, Zielhuis GA and Bredenoord AL, Eur. J. Hum. Genet, 2018, 26, 931–936. [PubMed: 29777206]
23. Sorokina M and Steinbeck C, J. Cheminform, 2020, 12, 20. [PubMed: 33431011]
24. Micallef L and Rodgers P, PLoS One, 2014, 9, e101717. [PubMed: 25032825]
25. Zeng X, Zhang P, He W, Qin C, Chen S, Tao L, Wang Y, Tan Y, Gao D, Wang B, Chen Z, Chen W, Jiang YY and Chen YZ, Nucleic Acids Res., 2018, 46, D1217–D1222. [PubMed: 29106619]
26. Klementz D, Döring K, Lucas X, Telukunta KK, Erxleben A, Deubel D, Erber A, Santillana I, Thomas OS, Bechthold A and Gunther S, Nucleic Acids Res., 2016, 44, D509–D514. [PubMed: 26615197]
27. van Santen JA, Jacob G, Singh AL, Aniebok V, Balunas MJ, Bunsko D, Carnevale Neto F, Castaño-Espriu L, Chang C, Clark TN, Cleary Little JL, Delgadillo DA, Dorrestein PC, Duncan KR, Egan JM, Galey MM, Haeckl FPJ, Hua A, Hughes AH, Iskakova D, Khadilkar A, Lee J-H, Lee S, LeGrow N, Liu DY, Macho JM, McCaughey CS, Medema MH, Neupane RP, O'Donnell TJ, Paula JS, Sanchez LM, Shaikh AF, Soldatou S, Terlouw BR, Tran TA, Valentine M, van der Hooft JJJ, Vo DA, Wang M, Wilson D, Zink KE and Linington RG, ACS Cent. Sci, 2019, 5, 1824–1833. [PubMed: 31807684]
28. Bycroft BW and Payne DJ, Dictionary of Antibiotics and Related Substances, CRC Press, 2nd edn., 2013.
29. Conway KR and Boddy CN, Nucleic Acids Res., 2012, 41, D402–D407. [PubMed: 23104377]
30. Ichikawa N, Sasagawa M, Yamamoto M, Komaki H, Yoshida Y, Yamazaki S and Fujita N, Nucleic Acids Res., 2012, 41, D408–D414. [PubMed: 23185043]
31. Medema MH, Kottmann R, Yilmaz P, Cummings M, Biggins JB, Blin K, de Bruijn I, Chooi YH, Claesen J, Coates RC, Cruz-Morales P, Duddela S, Dusterhus S, Edwards DJ, Fewer DP, Garg N, Geiger C, Gomez-Escribano JP, Greule A, Hadjithomas M, Haines AS, Helfrich EJN, Hillwig ML, Ishida K, Jones AC, Jones CS, Jungmann K, Kegler C, Kim HU, Kötter P, Krug D, Masschelein J, Melnik AV, Mantovani SM, Monroe EA, Moore M, Moss N, Nützmann H-W, Pan G, Pati A, Petras D, Reen FJ, Rosconi F, Rui Z, Tian Z, Tobias NJ, Tsunematsu Y, Wiemann P, Wyckoff E, Yan X, Yim G, Yu F, Xie Y, Aigle B, Apel AK, Balibar CJ, Balskus EP, Barona-Gómez F, Bechthold A, Bode HB, Borriss R, Brady SF, Brakhage AA, Caffrey P, Cheng Y-Q, Clardy J, Cox RJ, De Mot R, Donadio S, Donia MS, van der Donk WA, Dorrestein PC, Doyle S, Driessen AJM, Ehling-Schulz M, Entian K-D, Fischbach MA, Gerwick L, Gerwick WH, Gross H, Gust B, Hertweck C, Höfte M, Jensen SE, Ju J, Katz L, Kaysser L, Klassen JL, Keller NP, Kormanec J, Kuipers OP, Kuzuyama T, Kyrpides NC, Kwon H-J, Lautru S, Lavigne R, Lee CY, Linquan B, Liu X, Liu W, Luzhetskyy A, Mahmud T, Mast Y, Méndez C, Metsä-Ketelä M, Micklefield J, Mitchell DA, Moore BS, Moreira LM, Muller R, Neilan BA, Nett M, Nielsen J, O'Gara F, Oikawa H, Osbourn A, Osburne MS, Ostash B, Payne SM, Pernodet J-L, Petricek M, Piel J, Ploux O, Raaijmakers JM, Salas JA, Schmitt EK, Scott B, Seipke RF, Shen B, Sherman DH, Sivonen K, Smanski MJ, Sosio M, Stegmann E, Süssmuth RD, Tahlan K, Thomas CM, Tang Y, Truman AW, Viaud M, Walton JD, Walsh CT, Weber T, van Wezel GP, Wilkinson B, Willey JM, Wohlleben W, Wright GD, Ziemert N, Zhang C, Zotchev SB, Breitling R, Takano E and Glöckner FO, Nat. Chem. Biol, 2015, 11, 625–631. [PubMed: 26284661]
32. Kautsar SA, Blin K, Shaw S, Navarro-Muñoz JC, Terlouw BR, van der Hooft JJJ, van Santen JA, Tracanna V, Suarez Duran HG, Pascal Andreu V, Selem-Mojica N, Alanjary M, Robinson SL, Lund G, Epstein SC, Sisto AC, Charkoudian LK, Collemare J, Linington RG, Weber T and Medema MH, Nucleic Acids Res., 2019, 48, D454–D458.

33. Chen I-MA, Chu K, Palaniappan K, Pillay M, Ratner A, Huang J, Huntemann M, Varghese N, White JR, Seshadri R, Smirnova T, Kirton E, Jungbluth SP, Woyke T, Eloe-Fadrosh EA, Ivanova NN and Kyrpides NC, Nucleic Acids Res., 2019, 47, D666–D677. [PubMed: 30289528]
34. Cimermancic P, Medema MH, Claesen J, Kurita K, Wieland Brown LC, Mavrommatis K, Pati A, Godfrey PA, Koehrsen M, Clardy J, Birren BW, Takano E, Sali A, Linington RG and Fischbach MA, Cell, 2014, 158, 412–421. [PubMed: 25036635]
35. Grigoriev IV, Nikitin R, Haridas S, Kuo A, Ohm R, Otillar R, Riley R, Salamov A, Zhao X, Korzeniewski F, Smirnova T, Nordberg H, Dubchak I and Shabalov I, Nucleic Acids Res., 2014, 42, D699–D704. [PubMed: 24297253]
36. Blin K, Medema MH, Kottmann R, Lee SY and Weber T, Nucleic Acids Res., 2017, 45, D555–D559. [PubMed: 27924032]
37. Blin K, Pascal Andreu V, de los Santos ELC, Del Carratore F, Lee SY, Medema MH and Weber T, Nucleic Acids Res., 2019, 47, D625–D630. [PubMed: 30395294]
38. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, Tatusova T, DiCuccio M, Kitts P, Murphy TD and Pruitt KD, Nucleic Acids Res., 2016, 44, D733–D745. [PubMed: 26553804]
39. Wang M, Carver JJ, Phelan VV, Sanchez LM, Garg N, Peng Y, Nguyen DD, Watrous J, Kapono CA, Luzzatto-Knaan T, Porto C, Bouslimani A, Melnik AV, Meehan MJ, Liu W-T, Crüsemann M, Boudreau PD, Esquenazi E, Sandoval-Calderón M, Kersten RD, Pace LA, Quinn RA, Duncan KR, Hsu C-C, Floros DJ, Gavilan RG, Kleigrewe K, Northen T, Dutton RJ, Parrot D, Carlson EE, Aigle B, Michelsen CF, Jelsbak L, Sohlenkamp C, Pevzner P, Edlund A, McLean J, Piel J, Murphy BT, Gerwick L, Liaw C-C, Yang Y-L, Humpf H-U, Maansson M, Keyzers RA, Sims AC, Johnson AR, Sidebottom AM, Sedio BD, Klitgaard A, Larson CB, Boya P CA, Torres-Mendoza D, Gonzalez DJ, Silva DB, Marques LM, Demarque DP, Pociute E, O'Neill EC, Briand E, Helfrich EJN, Granatosky EA, Glukhov E, Ryffel F, Houson H, Mohimani H, Kharbush JJ, Zeng Y, Vorholt JA, Kurita KL, Charusanti P, McPhail KL, Nielsen KF, Vuong L, Elfeki M, Traxler MF, Engene N, Koyama N, Vining OB, Baric R, Silva RR, Mascuch SJ, Tomasi S, Jenkins S, Macherla V, Hoffman T, Agarwal V, Williams PG, Dai J, Neupane R, Gurr J, Rodríguez AMC, Lamsa A, Zhang C, Dorrestein K, Duggan BM, Almaliti J, Allard P-M, Phapale P, Nothias L-F, Alexandrov T, Litaudon M, Wolfender J-L, Kyle JE, Metz TO, Peryea T, Nguyen D-T, VanLeer D, Shinn P, Jadhav A, Müller R, Waters KM, Shi W, Liu X, Zhang L, Knight R, Jensen PR, Palsson BØ, Pogliano K, Linington RG, Gutiérrez M, Lopes NP, Gerwick WH, Moore BS, Dorrestein PC and Bandeira N, Nat. Biotechnol, 2016, 34, 828–837. [PubMed: 27504778]
40. Haug K, Cochrane K, Nainala VC, Williams M, Chang J, Jayaseelan KV and O'Donovan C, Nucleic Acids Res., 2019, 48, D440–D444.
41. McAlpine JB, Chen S-N, Kutateladze A, MacMillan JB, Appendino G, Barison A, Beniddir MA, Biavatti MW, Bluml S, Boufridi A, Butler MS, Capon RJ, Choi YH, Coppage D, Crews P, Crimmins MT, Csete M, Dewapriya P, Egan JM, Garson MJ, Genta-Jouve G, Gerwick WH, Gross H, Harper MK, Hermanto P, Hook JM, Hunter L, Jeannerat D, Ji N-Y, Johnson TA, Kingston DGI, Koshino H, Lee H-W, Lewin G, Li J, Linington RG, Liu M, McPhail KL, Molinski TF, Moore BS, Nam J-W, Neupane RP, Niemitz M, Nuzillard J-M, Oberlies NH, Ocampos FMM, Pan G, Quinn RJ, Reddy DS, Renault J-H, Rivera-Chávez J, Robien W, Saunders CM, Schmidt TJ, Seger C, Shen B, Steinbeck C, Stuppner H, Sturm S, Taglialatela-Scafati O, Tantillo DJ, Verpoorte R, Wang B-G, Williams CM, Williams PG, Wist J, Yue J-M, Zhang C, Xu Z, Simmler C, Lankin DC, Bisson J and Pauli GF, Nat. Prod. Rep, 2019, 36, 35–107. [PubMed: 30003207]
42. Sorkin BC, Betz JM and Hopp DC, Org. Lett, 2020, 22, 2867–2867. [PubMed: 32243184]
43. Lopez-Perez JL, Theron R, del Olmo E and Diaz D, Bioinformatics, 2007, 23, 3256–3257. [PubMed: 17956876]
44. Steinbeck C and Kuhn S, Phytochemistry, 2004, 65, 2711–2717. [PubMed: 15464159]

45. Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J, Livny M, Mading S, Maziuk D, Miller Z, Nakatani E, Schulte CF, Tolmie DE, Kent Wenger R, Yao H and Markley JL, Nucleic Acids Res., 2007, 36, D402–D408. [PubMed: 17984079]
46. Wishart DS, Feunang YD, Marcu A, Guo AC, Liang K, Vázquez-Fresno R, Sajed T, Johnson D, Li C, Karu N, Sayeeda Z, Lo E, Assempour N, Berjanskii M, Singhal S, Arndt D, Liang Y, Badran H, Grant J, Serra-Cayuela A, Liu Y, Mandal R, Neveu V, Pon A, Knox C, Wilson M, Manach C and Scalbert A, Nucleic Acids Res., 2018, 46, D608–D617. [PubMed: 29140435]
47. Asakura K, J. Synth. Org. Chem. Japan, 2015, 73, 1247–1252.
48. Chambers J, Davies M, Gaulton A, Hersey A, Velankar S, Petryszak R, Hastings J, Bellis L, McGlinchey S and Overington JP, J. Cheminform, 2013, 5, 3. [PubMed: 23317286]
49. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoemaker BA, Thiessen PA, Yu B, Zaslavsky L, Zhang J and Bolton EE, Nucleic Acids Res., 2019, 47, D1102–D1109. [PubMed: 30371825]
50. Hastings J, Owen G, Dekker A, Ennis M, Kale N, Muthukrishnan V, Turner S, Swainston N, Mendes P and Steinbeck C, Nucleic Acids Res., 2016, 44, D1214–D1219. [PubMed: 26467479]
51. Banik GM, in The ACS Guide to Scholarly Communication, eds. Banik GM, Baysinger G, Kamat PV and Pienta NJ, American Chemical Society, Washington, DC, 2020.
52. Nat. Chem. Biol, 2007, 3, 297–297. [PubMed: 17510641]
53. Gerlt JA, Ed., Biochemistry, 2018, 57, 4239–4240. [PubMed: 30037234]
54. Olson JE, Database Archiving, Elsevier, 2009.
55. Epstein SC, Charkoudian LK and Medema MH, Stand. Genomic Sci, 2018, 13, 16. [PubMed: 30008988]
56. Pye CR, Bertin MJ, Lokey RS, Gerwick WH and Linington RG, Proc. Natl. Acad. Sci, 2017, 114, 5601–5606. [PubMed: 28461474]
57. Pascolutti M, Campitelli M, Nguyen B, Pham N, Gorse A-D and Quinn RJ, PLoS One, 2015, 10, e0120942. [PubMed: 25902039]
58. O'Hagan S and Kell DB, Biotechnol. J, 2018, 13, 1700503.
59. Gu J, Gui Y, Chen L, Yuan G, Lu H-Z and Xu X, PLoS One, 2013, 8, e62839. [PubMed: 23638153]
60. Tully BJ, Graham ED and Heidelberg JF, Sci. Data, 2018, 5, 170203. [PubMed: 29337314]
61. Parks DH, Rinke C, Chuvochina M, Chaumeil P-A, Woodcroft BJ, Evans PN, Hugenholtz P and Tyson GW, Nat. Microbiol, 2017, 2, 1533–1542. [PubMed: 28894102]
62. Stewart RD, Auffret MD, Warr A, Walker AW, Roehe R and Watson M, Nat. Biotechnol, 2019, 37, 953–961. [PubMed: 31375809]
63. Almeida A, Nayfach S, Boland M, Strozzi F, Beracochea M, Shi ZJ, Pollard KS, Parks DH, Hugenholtz P, Segata N, Kyrpides NC and Finn RD, bioRxiv, , DOI:10.1101/762682.
64. Hoffmann T, Krug D, Bozkurt N, Duddela S, Jansen R, Garcia R, Gerth K, Steinmetz H and Müller R, Nat. Commun, 2018, 9, 803. [PubMed: 29476047]
65. Reher R, Kim HW, Zhang C, Mao HH, Wang M, Nothias L-F, Caraballo-Rodriguez AM, Glukhov E, Teke B, Leao T, Alexander KL, Duggan BM, Van Everbroeck EL, Dorrestein PC, Cottrell GW and Gerwick WH, J. Am. Chem. Soc, 2020, 142, 4114–4120. [PubMed: 32045230]
66. Ruttkies C, Schymanski EL, Wolf S, Hollender J and Neumann S, J. Cheminform, 2016, 8, 3. [PubMed: 26834843]
67. Mohimani H, Gurevich A, Shlemov A, Mikheenko A, Korobeynikov A, Cao L, Shcherbin E, Nothias L-F, Dorrestein PC and Pevzner PA, Nat. Commun, 2018, 9, 4035. [PubMed: 30279420]
68. Djoumbou-Feunang Y, Pon A, Karu N, Zheng J, Li C, Arndt D, Gautam M, Allen F and Wishart DS, Metabolites, 2019, 9, 72.
69. Dührkop K, Fleischauer M, Ludwig M, Aksenov AA, Melnik AV, Meusel M, Dorrestein PC, Rousu J and Böcker S, Nat. Methods, 2019, 16, 299–302. [PubMed: 30886413]
70. Dührkop K, Shen H, Meusel M, Rousu J and Böcker S, Proc. Natl. Acad. Sci, 2015, 112, 12580–12585. [PubMed: 26392543]
71. van der Hooft JJJ, Wandy J, Barrett MP, Burgess KEV and Rogers S, Proc. Natl. Acad. Sci, 2016, 113, 13738–13743. [PubMed: 27856765]

72. Rogers S, Ong CW, Wandy J, Ernst M, Ridder L and van der Hooft JJJ, Faraday Discuss., 2019, 218, 284–302. [PubMed: 31120050]

73. Crits-Christoph A, Diamond S, Butterfield CN, Thomas BC and Banfield JF, Nature, 2018, 558, 440–444. [PubMed: 29899444]

74. Navarro-Muñoz JC, Selem-Mojica N, Mullowney MW, Kautsar SA, Tryon JH, Parkinson EI, De Los Santos ELC, Yeong M, Cruz-Morales P, Abubucker S, Roeters A, Lokhorst W, Fernandez-Guerra A, Cappelini LTD, Goering AW, Thomson RJ, Metcalf WW, Kelleher NL, Barona-Gómez F and Medema MH, Nat. Chem. Biol, 2020, 16, 60–68. [PubMed: 31768033]

75. Bahram M, Hildebrand F, Forslund SK, Anderson JL, Soudzilovskaia NA, Bodegom PM, Bengtsson-Palme J, Anslan S, Coelho LP, Harend H, Huerta-Cepas J, Medema MH, Maltz MR, Mundra S, Olsson PA, Pent M, Põlme S, Sunagawa S, Ryberg M, Tedersoo L and Bork P, Nature, 2018, 560, 233–237. [PubMed: 30069051]

76. Krause J, Handayani I, Blin K, Kulik A and Mast Y, Front. Microbiol, 2020, 11, 225. [PubMed: 32132989]

77. Gregory K, Salvador LA, Akbar S, Adaikpoh BI and Stevens DC, Microorganisms, 2019, 7, 181.

78. Eng CH, Backman TWH, Bailey CB, Magnan C, García Martín H, Katz L, Baldi P and Keasling JD, Nucleic Acids Res., 2018, 46, D509–D515. [PubMed: 29040649]

79. Dejong CA, Chen GM, Li H, Johnston CW, Edwards MR, Rees PN, Skinnider MA, Webster ALH and Magarvey NA, Nat. Chem. Biol, 2016, 12, 1007–1014. [PubMed: 27694801]

80. Doroghazi JR, Albright JC, Goering AW, Ju K-S, Haines RR, Tchalukov KA, Labeda DP, Kelleher NL and Metcalf WW, Nat. Chem. Biol, 2014, 10, 963–968. [PubMed: 25262415]

81. Goering AW, McClure RA, Doroghazi JR, Albright JC, Haverland NA, Zhang Y, Ju K-S, Thomson RJ, Metcalf WW and Kelleher NL, ACS Cent. Sci, 2016, 2, 99–108. [PubMed: 27163034]

82. Eldjárn GH, Ramsay A, van der Hooft JJJ, Duncan KR, Soldatou S, Rousu J and Rogers S, bioRxiv, , D0I:10.1101/2020.06.12.148205.

83. Kanehisa M, Furumichi M, Tanabe M, Sato Y and Morishima K, Nucleic Acids Res., 2017, 45, D353–D361. [PubMed: 27899662]

84. Kanehisa M, Sato Y and Morishima K, J. Mol. Biol, 2016, 428, 726–731. [PubMed: 26585406]

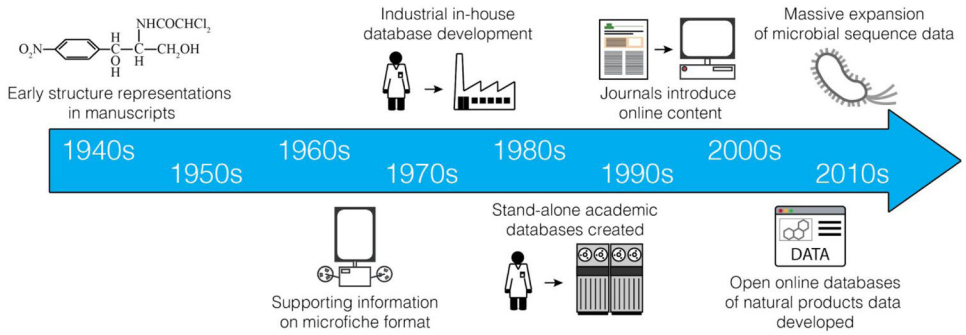


Fig. 1:
Timeline of data distribution methods for natural products.

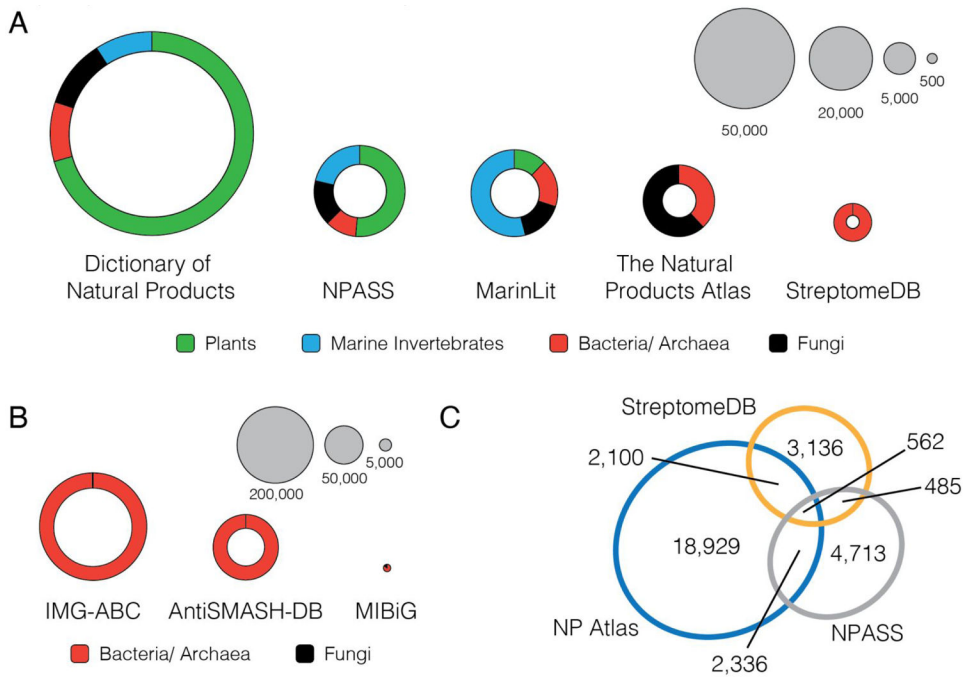


Fig. 2:

A) Distribution of compound source types in selected natural products databases. B) Distribution of biosynthetic gene cluster source types in selected biosynthetic gene cluster databases C) Overlap of microbial natural product InChIKey structure representations between open access databases. Microbial database overlap was calculated using the unique sets of the InChIKey connectivity hashes from each database. This decreases the compound count in each database because sets of configurational isomers are reduced to single flat structures: NP Atlas 25,523 to 23,927, NPASS 8,729 to 8,096, and StreptomeDB 7,125 to 6,283. The Proportional Venn Diagram was created using eulerAPE v3.[24]

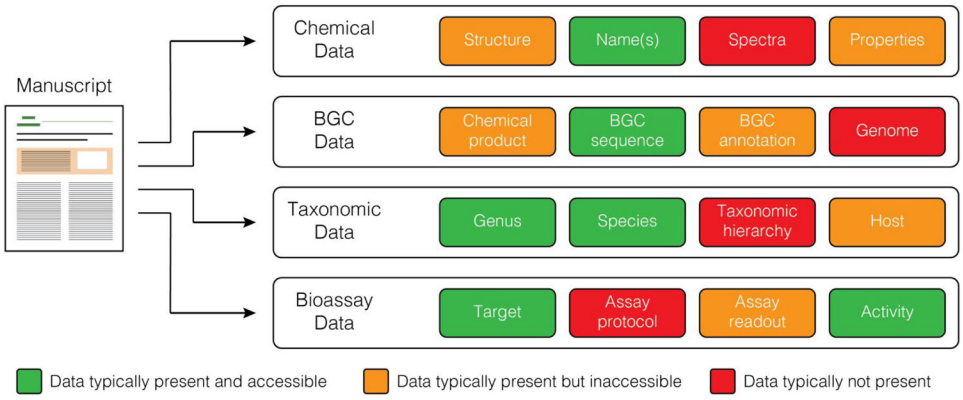


Fig. 3:

Data types and their relative accessibility from published articles in the primary scientific literature