Chapter 1

ONLINE DRUG DATA BANKS AND MOLECULAR LIBRARIES

G. Dicky John Davis[1,*], R. Archana Preetha[2] and M. Gowri Neelima[2]

[1]Department of Bioinformatics,
Sri Ramachandra Institute of Higher Education & Research,
Chennai, Tamil Nadu, India
[2]Department of Biotechnology,
Maharani Lakshmi Ammanni College for Women, Bengaluru,
Karnataka, India

ABSTRACT

Small molecule structure databases have seen an exponential growth and have become a key knowledge base that fosters cheminformatics and bioinformatics research. Most of the oral/injections/external drugs present in the market today are small molecules and they regulate or modify biological targets like receptors, enzymes and channels in the disease process. Natural compounds are gaining more attention as they contain bioactive compounds which are very effective and less toxic. Advancement

* Corresponding Author's Email: dicky@sriramachandra.edu.in.

in omics and use of bio-based technologies make them lucrative and acquiescent target drug candidates or for designing by screening. Further, a complete overview of these resources will maximize their expediency, allowing researchers in applicable fields to utilize these resources and exploit them to their potential. This chapter focuses on various online drug databanks and molecular libraries which have been branded for virtual screening. Databases vary in size and scope such as conventional or specific to a topic of source information, wherein have been classified as commercial, public aggregators, scientific and natural products. Data resources presently accessible are described with their salient characteristics to accelerate the drug development process.

1. INTRODUCTION

For a long time, small molecules are the basis for drug discovery. Small and active substances have been the cornerstone for classic drug development. These molecules can easily enter into the bloodstream and reach the desired target for the treatment. Small molecules are molecules with low molecular weight and have very simple structure and drugs developed using them were very effective as they penetrate through cell membranes effectively. These small molecules were well known for their therapeutic value as they provide insight into cellular mechanisms which imbibes scope in computer aided drug design (Schreiber, 2019).

Even before the time of technological advancement with high-throughput screening and combinatorial chemistry these small molecules were extracted from natural substances. Typically, a small molecule weighs <1000 Daltons and this size accredits for its easy mass-production into tablets for orally administered drugs that are readily absorbed by the gastrointestinal lumen. Most of the orally active small molecule compounds available in the market do follow the Rule of Five: ≤ 500 Daltons; log *P* ≤ 5; H-bond donors ≤ 5; H-bond acceptors ≤ 10 (Lipinski 2004), which is commonly used while designing a drug. The rotatable bonds>10 and polar surface area>120 $Å^2$ correlates with reduced bioavailability of the candidate drug (Veberet al. 2002; Kelderet al. 1999).

Drugs include small molecules and proteins or biologics, wherein the mode of action for small molecules is quite unique (Eggert 2013). For instance, small molecules developed for anticancer treatment, infiltrate the cell membranes, reach the nucleus and interact with DNA (Adams et al. 2015). Hence, modulating transcription factors, initiating cell death and stops cancerous growth (Vogler et al. 2009; Lambert et al.,2018). This capacity of small molecules to enter cells has empowered oncology therapies from broad spectrum of cytotoxic chemotherapy, to targeted genome specific treatment in personalized medicine (Hoelder, Clarke, and Workman, 2012).

Currently, studies on small molecules are into pharmaceutical research in varied fields of studies like Apoptosis, Cell Signaling, Heat Shock Proteins, Oxidative Stress, Ion Channels and Transporters. Compounds that can target critical regulatory proteins are needed toidentify a challenging way to inhibit a function, restore or enhance a function and confer a new function (Schreiber 2019).Central nervous system therapeutics involves small molecules that enter the brain predominantly through transcellular diffusion, while others use transport systems such as L-dopa (Banks and Greig 2019).

2. EXPANSE OF DRUG DATABANKS

The driving force in drug discovery is the availability of chemical structures and linked bioactivity that are publicly accessible. These increasingly exceptionally large collections of data include both experimental and in-silico compounds, as well as visionary compounds from patents (Lipinski et al. 2015). The data resources include both public domains such as PubChem, ChEMBL, Chemspider and BindingDB, and commercial databases such as SciFinder from Chemical Abstracts Service (CAS) and Reaxys from Elsevier (Nicola, Liu, and Gilson 2012).

Large databases were pumped in with an overload of data generated through high-throughput screening and novel platforms for data integration and mining. “Illuminating the Druggable Genome” (IDG) program from the National Institutes of Health (NIH) quantifies data from a wide range of chemical, biological and clinical resources to

prioritize protein targets for precision medicine (Oprea 2019). The databases having million chemical structures linked to bioactivity and biological data is listed in Table 1.

Table 1. Summary of Statistics for Chemical databases which are above a million structures, mostly linked to bioactivity and biological data

Database Name	Database Link	No. of entries	Description
Generated database (GDB-17)	http://www.gdb.unibe.ch/	166 billion	Downloadable database of in-silico compounds having no biological activity
SciFinder	http://www.cas.org/products/scifinder	166 million	Unique organic and inorganic chemical substances with CAS Registry Number
UniChem	https://www.ebi.ac.uk/unichem/	175 million	EMBL-EBI database of non-redundant chemical structures and its resources
PubChem Compounds	https://pubchem.ncbi.nlm.nih.gov/	111 million	Records are systematically arrangedas Substance, Compound, and BioAssay databases which are explicitly cross-linked.
ChemSpider	http://www.chemspider.com/	90 million	Integrates and links compounds from hundreds of data sources with spectra, or syntheses
Reaxys	https://www.elsevier.com/solutions/reaxys	118 million	Organic, inorganic and organometallic substances, including natural products with their species derivation
ZINC	http://zinc.docking.org/	230 million	Biologically relevant, ready-to-dock formats of purchasable “drug-like” compounds
BindingDB	http://www.bindingdb.org/bind/index.jsp	0.9 million	Database concentrates on the interactions of protein with small, drug-like molecules which are potential drug targets
Cambridge Structural Database	https://www.ccdc.cam.ac.uk/solutions/csd-system/components/csd/	1 million	small-molecule organic and metal-organic crystal structures from x-ray and neutron diffraction analyses

Natural products (NPs) are distinct chemicals synthesized by biotic organisms having molecular weight <1500 Daltons and possess specific biological activity such as molecular signaling or defense (Sorokina and Steinbeck, 2020). NPs have unique chemical diversity,

which eventually have varied bioactivity and drug-like properties. NPs are the vital components of folk medicine and are now the most potential resources for the development of new drug entities and scaffolds (Newman and Cragg 2016).

Databases vary in size and scope such as conventional or specific to a topic of source information, public or commercial data, free or patent data, real or virtual compounds, natural products or synthetic compounds, structures with biological data and bioactivity. Further, a complete overview of these existing resources will maximize their expediency by allowing the researchers in relevant fields to utilize these resources and exploit them to their potential.

This chapter focuses on various online drug databanks and molecular libraries which have been branded for virtual screening. Here, the major databases are classified as commercial, public aggregators, scientific and natural products.

3. COMMERCIAL MOLECULAR LIBRARIES

Reaxys and SciFinder are web-based major chemical information retrieval products that are linked to data from published literature, including journals and patents.

Reaxys

Reaxys is an independent database, owned and updated by Elsevier monthly. Reaxys has a comprehensive collection since pre-1960 organic literature. The literature citations from Beilstein handbook of organic chemistry which are indexed in Reaxys and contain over 9.6 million organic compounds and over 10 million reactions. Similar data is also indexed in Gmelin handbook of inorganic and organometallic chemistry which have references since 1700s which has over 2.4 million inorganic and organometallic compounds, 1.8 million reactions and 1.2 million references. Reaxys database has a choice of algorithmic natural language queries (NLQs) interpretation and also

has customized searches through graphical user interfaces called query lets(Mutton and Ridley 2019).

SciFinder

SciFinderis a comprehensive database produced by Chemical Abstracts Service (CAS) which has search engine that can be queried based on author, topic, substances by name or CAS Registry Number, OR use of editor to draw chemical structures, substructures, or reactions. It is updated daily and exceptionally has a greater coverage of literature as well as patents since 1967.It also has a comprehensive registration for submission of any type of chemical compounds, mixtures, polymers and organic reactions from 1985 onwards. SciFinder text search queries is also based on algorithmic based interpretation of NLQs, but the interpretations are not available. Where as in Reaxys, interpretations of NLQs will be displayed as function Edit present in Query Builder.SciFinder coverage deepens to allied fields encompassing chemical or civil engineering, materials science and also physics (Ridley 2001).

4. PUBLIC DATA AGGREGATORS

Generic chemical resources comprise of catalogs and tools to mine bioactivity screening techniques and also has repository of varied chemical information. The chemical data are explicitly linked to relevant records in associated databases such as genes, proteins, compounds and molecular pathways, providing wider knowledge and scope.

PubChem

PubChem is a free and publically available resource at the U.S. National Center for Biotechnology Information (NCBI). It provides descriptions of small molecule and its bioactivity data achieved from

primary literatures. PubChem encompasses three primary databases which are explicitly inter-linked namely Substance, Compound, and BioAssay. PubChem Substance archives depositor-contributed descriptions of chemical substances (Kim et al. 2016). PubChem Compound has the data of varied Compound extracts and stores the unique structures submitted in Substance through an automated process called chemical structure standardization (Hähnke, Kim, and Bolton 2018). PubChem BioAssay contains 273 million bioactivity data which was deposited by over 80 organizations across the world (Wang et al. 2017).

PubChem also contains spectral information of compounds and the scope has been extended to include chemicals used in agricultural and food industry (Kim et al. 2019). It has several services where gene or protein targets are depicted in Target View page. Biological activity information is represented in Bioactivity dyad page, Patented chemicals are summarized in Patent View page, Sources page presents where data comes fromWidgets feature in PubChem allows data to be displayed on other webpages, and Power User Gateway View allows programmers an open access to annotations present in the Summary.

ChEMBL

ChEMBL is an open large-scale database that conceals a wide range of annotated and curated data, physically mined from the medicinal chemistry primary literatures. This database also includes the information of drug metabolism and disposition data and bioactivity data from patents; crop protection data and even neglected disease screening. ChEMBLalso has updated information of approved drugs and clinical development applicants on small molecules along with their biological activity, mechanism of action and therapeutic value. Biological activity information is periodically swapped with Binding DB and PubChem BioAssay (Mendez et al. 2019). REST – a Representational State Transfer allows the retrieval of ChEMBL data, an additional web services capability that employs an extensive mark-up language or JavaScript Object Notation (JSON) for programmatic retrieval (Gaulton et al. 2017).

UniChem

UniChem is a non-redundant database that cross references between EBI ChEMBL and ChEBI which are primary chemistry resources as well as other resources like Gene Expression Atlas, PDBe. It provides a wider cross-referencing service by mapping the data from PubChem, ZINC, etc., which are external resources. It utilizes the Standard InChI to normalize the connectivity of molecules in different protonation or charge states.A source-specific downloading mechanism and parser is made available as configuration and resources of various databases differ widely (Chambers et al. 2013).

ChemSpider

ChemSpider was created by Antony Williams as a free chemical structure database which was later acquired by the Royal Society of Chemistry in May 2009. It is developed by curation from over 278 data resources. ChemSpider authorize it's registered users to curate and annotate the records and forms a community for chemists based on crowd sourcing (Pence and Williams 2010). It provides NMR ^{1}H or ^{13}C spectroscopic data of small molecules which is a versatile data to exploit for learning, teaching and research. The information is also accessible as a mobile app for Android and IOS through Chemspider APIs.

5. Scientific Drug Databanks

KEGG COMPOUND

KEGG COMPOUND(Kyoto Encyclopedia of Genes and Genomes) is an assembly of biochemical compounds as well as biopolymer

structures of ligand and small molecules which are designated by C number. It also interconnects the compound's activity against a specific pathway involved in a disease through specific links of KEGG pathway maps/modules/drug/network.

ZINC

ZINC is a highly accessed database for virtual screening that has data of purchasable compounds. All the 35 million compounds present are in3D formats, which can be readily used for docking and are also commercially-available. ZINCis hosted by Irwin and Shoichet Laboratories, Department of Pharmaceutical Chemistry, University of California, San Francisco (UCSF). (Irwin et al., 2012).

6. NATURAL PRODUCTSDATABASE

Natural products are biological active metabolites synthesized in nature by living organism that serve for their own survival and function. These metabolites have acquired unique chemical mélange and drug-like properties through evolution over millions of years. Nearly 100 different natural product databases are available online wherein only 50 are open access.

MarinLit

MarinLit database is committed to natural products of prime importance in marine research. This database was developed by Professors John Blunt and Murray Munro while they were working at the University of Canterbury, New Zealand in 1970s. It was developed to attain the requisites of the University of Canterbury Marine Group, which in turn has evolved into a unique database with versatile tools for searching and dereplication. MarinLit hovers an extensive range of

marine organism and their metabolites making it database opted by all marine biologist in natural products researcher.

Seaweed Metabolite Database (SWMD)

Seaweed Metabolite Database (SWMD)records data of secondary metabolites from marine algae and is a free open access database hosted in the public domain (http://swmd.co.in/). The compounds can be accessed by text as well as structure search forvaried chemicals extracted from seaweeds.Seaweedsare one of the richest renewable biological resource of secondary metabolites with diverse chemical structures and are commercially important in pharmaceutical industries. SWMD is comprised of around 1110entries having423 unique structures,which are classified based on their Phyla as Phaeophyta (brown algae, n = 266), Chlorophyta (green algae, n = 33),and Rhodophyta (red algae, n =811). It has comprehensive metadata on extraction method, bioactivity and geographical origin of the seaweed. The data can be accessed for virtual screening and is linked to other public databases such as ChemSpider and PubChem(Davis and Vasanthi 2011).

7. CONCLUSION

The databases analyzed are widely used drug discovery resources and provides inspiration for novel drug discoveries. All these resources strive to deliver superior quality content and enable the users an option to easily access the data and facilitate researchers in virtual screening. Artificial Intelligence approaches and applications have revolutionized the molecular libraries to provide high-quality and curated resources. In this chapter, we have described in detail the online drug databanks which will offer both skilled and amateur users a comprehensive purview of the resources accessible in the broad area of drug discovery and data informatics for novel inventions.

REFERENCES

Adams, Jerry L., James Smothers, Roopa Srinivasan. & Axel Hoos. (2015). 'Big Opportunities for Small Molecules in Immuno-Oncology'. *Nature Reviews. Drug Discovery*,*14* (9), 603–22.

Banks, William A. & Nigel H. Greig. (2019). 'Small Molecules as Central Nervous System Therapeutics: Old Challenges, New Directions, and a Philosophic Divide'. *Future Medicinal Chemistry*,*11* (6), 489–93.

Chambers, Jon, Mark Davies, Anna Gaulton, Anne Hersey, Sameer Velankar, Robert Petryszak, Janna Hastings, Louisa Bellis, Shaun McGlinchey. & John P. Overington. (2013). 'UniChem: A Unified Chemical Structure Cross-Referencing and Identifier Tracking System'. *Journal of Cheminformatics*,*5* (1), 3.

Davis, G Dicky John. & a Hannah Rachel Vasanthi. (2011). "Seaweed Metabolite Database (SWMD): A Database of Natural Compounds from Marine Algae." *Bioinformation*,*5* (8), 361–64.

Eggert, Ulrike S. (2013). 'The Why and How of Phenotypic Small-Molecule Screens'. *Nature Chemical Biology*,*9* (4), 206–9.

Gaulton, Anna, Anne Hersey, MichałNowotka, A. Patrícia Bento, Jon Chambers, David Mendez, Prudence Mutowo., et al. (2017). 'The ChEMBL Database in 2017'. *Nucleic Acids Research*,*45* (D1), D945–54.

Hähnke, Volker D., Sunghwan Kim. & Evan E. Bolton. (2018). 'PubChem Chemical Structure Standardization'. *Journal of Cheminformatics*,*10* (1), 36.

Hoelder, Swen, Paul A. Clarke. & Paul Workman. (2012). 'Discovery of Small Molecule Cancer Drugs: Successes, Challenges and Opportunities'. *Molecular Oncology*. https://doi.org/10.1016/j.molonc.2012.02.004.

Irwin, John J., Teague Sterling, Michael M. Mysinger, Erin S. Bolstad. & Ryan G. Coleman. (2012). 'ZINC: a free tool to discover chemistry for biology.' *Journal of chemical information and modeling*, *52*(7), 1757-1768.DOI: 10.1021/ci3001277.

Kelder, Jan, Peter D. J. Grootenhuis, Denis M. Bayada, Leon P. C. Delbressine, & Jan-peter Ploemen. (1999).

'10.1023/A:1015040217741'. *Pharmaceutical Research*. https://doi.org/10.1023/A:1015040217741.

Kim, Sunghwan, Jie Chen, Tiejun Cheng, AstaGindulyte, Jia He, Siqian He, Qingliang Li., et al. (2019). 'PubChem 2019 Update: Improved Access to Chemical Data'. *Nucleic Acids Research*,*47* (D1), D1102–9.

Kim, Sunghwan, Paul A. Thiessen, Evan E. Bolton, Jie Chen, Gang Fu, AstaGindulyte, Lianyi Han., et al. (2016). 'PubChem Substance and Compound Databases'. *Nucleic Acids Research*,*44* (D1), D1202–13.

Lambert, Mélanie, Samy Jambon, Sabine Depauw. & Marie-Hélène David-Cordonnier. (2018). 'Targeting Transcription Factors for Cancer Treatment'. *Molecules*, *23* (6). https://doi.org/10.3390/molecules23061479.

Lipinski, Christopher A. (2004). 'Lead- and Drug-like Compounds: The Rule-of-Five Revolution'. *Drug Discovery Today: Technologies*. https://doi.org/10.1016/j.ddtec.2004.11.007.

Lipinski, Christopher A., Nadia K. Litterman, Christopher Southan, Antony J. Williams, Alex M. Clark. & Sean Ekins. (2015). 'Parallel Worlds of Public and Commercial Bioactive Chemistry Data'. *Journal of Medicinal Chemistry*,*58* (5), 2068–76.

Mendez, David, Anna Gaulton, A. Patrícia Bento, Jon Chambers, Marleen De Veij, Eloy Félix, María Paula Magariños., et al. (2019). 'ChEMBL: Towards Direct Deposition of Bioassay Data'. *Nucleic Acids Research*,*47* (D1), D930–40.

Mutton, Troy. & Damon D. Ridley. (2019). 'Understanding Similarities and Differences between Two Prominent Web-Based Chemical Information and Data Retrieval Tools: Comments on Searches for Research Topics, Substances, and Reactions'. *Journal of Chemical Education*,*96* (10), 2167–79.

Newman, David J. & Gordon M. Cragg. (2016). 'Natural Products as Sources of New Drugs from 1981 to 2014'. *Journal of Natural Products*,*79* (3), 629–61.

Nicola, George, Tiqing Liu. & Michael K. Gilson. (2012). 'Public Domain Databases for Medicinal Chemistry'. *Journal of Medicinal Chemistry*. https://doi.org/10.1021/jm300501t.

Oprea, Tudor I. (2019). 'Exploring the Dark Genome: Implications for Precision Medicine'. *Mammalian Genome: Official Journal of the International Mammalian Genome Society*,*30* (7-8), 192–200.

Pence, Harry E. & Antony Williams. (2010). 'ChemSpider: An Online Chemical Information Resource'. *Journal of Chemical Education*,*87* (11), 1123–24.

Ridley, Damon D. (2001). 'Introduction to Structure Searching with SciFinder Scholar'. *Journal of Chemical Education*,*78* (4), 559.

Schreiber, Stuart L. (2019). 'A Chemical Biology View of Bioactive Small Molecules and a Binder-Based Approach to Connect Biology to Precision Medicines'. *Israel Journal of Chemistry*,*59* (1-2), 52–59.

Sorokina, Maria. & Christoph Steinbeck. (2020). 'Review on Natural Products Databases: Where to Find Data in 2020'. *Journal of Cheminformatics*,*12* (1), 20.

Veber, Daniel F., Stephen R. Johnson, Hung-Yuan Cheng, Brian R. Smith, Keith W. Ward. & Kenneth D. Kopple. (2002). 'Molecular Properties That Influence the Oral Bioavailability of Drug Candidates'. *Journal of Medicinal Chemistry*. https://doi.org/10.1021/jm020017n.

Vogler, M., Dinsdale, D., Dyer, M. J. S. & Cohen, G. M. (2009). 'Bcl-2 Inhibitors: Small Molecules with a Big Impact on Cancer Therapy'. *Cell Death and Differentiation*,*16* (3), 360–67.

Wang, Yanli, Stephen H. Bryant, Tiejun Cheng, Jiyao Wang, AstaGindulyte, Benjamin A. Shoemaker, Paul A. Thiessen, Siqian He. & Jian Zhang. (2017). 'PubChem BioAssay: 2017 Update'. *Nucleic Acids Research*,*45* (D1), D955–63.

Page layout by Anvi Composers