3.3: Public Chemical Databases

PubChem: chemical information repository at the U.S. NIH

PubChem (https://pubchem.ncbi.nlm.nih.gov)[1,2,3] is a public repository of information on small molecules and their biological activities, developed and maintained by the National Library of Medicine (NLM), an institute within the U.S. National Institutes of Health (NIH). Since its launch in 2004 as a component of the NIH's Molecular Libraries Roadmap Initiatives, it has been rapidly growing, and now serves as a key chemical information resource for researchers in many biomedical science areas, including cheminformatics, chemical biology, and medicinal chemistry. Detailed information on PubChem can be found in these three papers:

- PubChem Substance and Compound databases

  S. Kim *et al.*, Nucleic Acids Research **2016**, *44*, D1202-D1213

  (https://doi.org/10.1093/nar/gkv951)

- PubChem BioAssay: 2017 update

  Wang Y. *et al.* Nucleic Acids Research **2017**, *45*, D955-D963

  (https://doi.org/10.1093/nar/gkw1118)

- Getting the most out of PubChem for virtual screening

  S. Kim, Expert Opin. Drug Discov. **2016**, *11*, 843-855

  (http://dx.doi.org/10.1080/17460441.2016.1216967)

As of February 2017, PubChem contains more than 235 million depositor-provided substances, 94 million unique chemical structures, and one million biological assays, which cover about 10 thousand protein target sequences. For efficient use of this vast amount of data, PubChem provides various search and analysis tools. Some of these search tools will be used later in this course for demonstration purposes.

ChemSpider: a chemical database integrated with RSC's publishing process

ChemSpider (http://www.chemspider.com/)[4,5] is a free chemical structure database, containing information on 34 million structures collected from ~500 data sources. It also provides information on chemical reactions through ChemSpider SyntheticPages (CSSP)[6]. ChemSpider uses a crowdsourcing approach that allows registered users for manual comment and correction of ChemSpider records. Owned by the Royal Society of Chemistry (RSC), which publishes ~40 peer-reviewed chemistry journals, ChemSpider is integrated with the RSC publishing process, whereby new chemicals identified in newly published RSC articles also become available in ChemSpider.

ChEMBL: literature-extracted biological activity information

ChEMBL (https://www.ebi.ac.uk/chembl/)[7,8] is a large bioactivity database, developed and maintained by the European Bioinformatics Institute (EBI), which is part of the European Molecular Biology Laboratory (EMBL). The core activity data in ChEMBL are "manually" extracted from the full text of peer-reviewed scientific publications in select chemistry journals, such as *Journal of Medicinal Chemistry*, *Bioorganic Medicinal Chemistry Letters*, and *Journal of Natural products*. From each publication, details of the compounds tested, the assays performed and any target information for these assays are abstracted. ChEMBL also integrates screening results and bioactivity data from other public databases (such as PubChem BioAssay) and information on approved drugs from the U.S. FDA Orange Book[9] and the NLM's DailyMed[10].

ChEBI: a dictionary of small molecular entity

ChEBI (https://www.ebi.ac.uk/chebi/)[11,12] stands for "Chemical Entities of Biological Interest". It is a freely available database of "small" molecular entities, developed at the European Bioinformatics Institute (EBI). The molecular entities in ChEBI are either natural or synthetic products used to intervene the processes of living organisms. As a rule, however, ChEBI does not contain macromolecules directly encoded by genome (e.g., nucleic acids, proteins, and peptides derived from protein by cleavage). ChEBI provides "standardized" descriptions of molecular entities that enable other databases to annotate their entries in a consistent fashion. ChEBI focuses on high-quality manual annotation, non-redundancy, and provision of a chemical ontology rather than full

coverage of the vast chemical space. Note that both ChEMBL and ChEBI are developed and maintained by the EMBL-EBI. While ChEMBL focuses on "bioactivity" of a large number of bioactive molecules (currently ~2.0 millions), ChEBI is a "dictionary" that provides high-quality standardized descriptions for a relatively small number of molecules (currently ~50 thousands).

NIST Webbook: thermodynamic and spectroscopic data of chemicals

The U.S. National Institutes of Standards and Technology (NIST) compiles chemical and physical property data for chemical species and distributes them through the web site called the NIST Chemistry WebBook (http://webbook.nist.gov)[13,14]. These data include thermochemical data (e.g., enthalpy of formation, heat capacity, and vapor pressure), reaction thermochemistry data (e.g., enthalpy of reaction and free energy of reaction), spectroscopic data (e.g., IR and UV/Vis spectra), gas chromatographic data, ion energetics data, and so on.

DrugBank: comprehensive information on drug molecules

DrugBank[15,16,17] (http://www.drugbank.ca/) is a comprehensive online database containing biochemical and pharmacological information about ~8,000 drug molecules, including U.S. Food and Drug Administration (FDA)-approved small-molecule drugs and biotech drugs (e.g., protein/peptide drugs) as well as experimental drugs. DrugBank provides a wide range of drug information, including drug targets, mechanism of action, adverse drug reactions, food-drug and drug-drug interactions, experimental and theoretical ADMET properties (*i.e.*, Absorption, Distribution, Metabolism, Excretion, and Toxicity), and many others. Most of these data are curated from primary literature sources, by domain-specific experts and skilled biocurators.

HMDB: the Human Metabolome Database

The Human Metabolome Database (HMDB) (http://www.hmdb.ca)[18,19,20] is comprehensive information on human metabolites and human metabolism data. This database contains curated information derived from scientific literature, as well as experimentally determined metabolite concentrations in human tissue or biofluid (e.g., urine, blood, cerebrospinal fluid and so on). Reference Mass spectra (MS) and nuclear magnetic resonance (NMR) spectra for metabolites are also provided when available. In addition to data for "detected" metabolites (those with measured concentrations or experimental confirmation of their existence), the HMDB also provides information on "expected" metabolites (those for which biochemical pathways are known or human intake/exposure is frequent but the compound has yet to be detected in the body).

TOXNET: a collection of toxicological information

TOXNET (http://toxnet.nlm.nih.gov/)[21,22,23,24] maintained by the National Library of Medicine (NLM) at NIH, is a group of databases covering toxicology, hazardous chemicals, toxic releases, environmental and occupational health, risk assessment. Currently, 16 databases are integrated into the TOXNET system, and users can search all these databases either at once or individually. While all the 16 databases provide valuable information, three of them may be worth mentioning in the context of this course.

- ChemIDPlus[25,26] is a dictionary of over 400,000 chemical records (names, synonyms, and structures) and provides access to the structure and nomenclature files used for the identification of chemical substances in the TOXNET system and other NLM databases.
- The Hazardous Substances Data Bank (HSDB)[27,28] focuses on the toxicology of potentially hazardous chemicals, providing information on human exposure, industrial hygiene, emergency handling procedures, environmental fate, regulatory requirements, nanomaterials, and related areas. All HSDB data are referenced and derived from a core set of books, government documents, technical reports and selected primary journal literature. Importantly, HSDB is peer-reviewed by the Scientific Review Panel (SRP), a committee of experts in the major subject areas within the data bank's scope.
- The Comparative Toxicogenomics Database (CTD)[29,30] contains manually curated data describing interactions of chemicals with genes/proteins and diseases. This database provides insight into the molecular mechanisms underlying variable susceptibility for environmentally influenced diseases.

A brief overview of TOXNET and its databases can be found in the TOXNET Fact Sheet[22] and a recent paper by Fowler and Schnall[24].

Protein Data Bank (PDB): a key source for protein-bound ligand structures

The Protein Data Bank (PDB) is an archive of the experimentally determined 3-D structures of large biological molecules such as proteins and nucleic acids. These structures were determined primarily by using X-ray crystallography and nuclear magnetic

resonance (NMR) spectroscopy. While PDB is not a small molecule database, it contains the 3-D structures of many proteins with small-molecule ligands bound to them. PDB allows users to search for proteins that an input small molecule binds to. Considering that it is not possible to experimentally determine how small molecules (such as drug or toxic chemicals) actually bind to their target proteins in a living organism, PDB is the most widely used resource for experimentally determined protein-bound structures of small molecules. The PDB are maintained by the Worldwide PDB (wwPDB)[31], and freely accessible via the websites of its member organizations: PDBe (PDB in Europe)[32,33], PDBj (PDB Japan)[34,35], RCSB PDB (Research Collaboratory for Structural Bioinformatics PDB)[36,37].

References

1. Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L. Y.; He, J. E.; He, S. Q.; Shoemaker, B. A.; Wang, J. Y.; Yu, B.; Zhang, J.; Bryant, S. H. *Nucleic Acids Res.* **2016**, *44*, D1202.
2. Wang, Y.; Bryant, S. H.; Cheng, T.; Wang, J.; Gindulyte, A.; Shoemaker, B. A.; Thiessen, P. A.; He, S.; Zhang, J. *Nucleic Acids Res.* **2017**, *45*, D955.
3. Kim, S. *Expert Opinion on Drug Discovery* **2016**, *11*, 843.
4. ChemSpider (http://www.chemspider.com) (Accessed on 2/17/2017).
5. Pence, H. E.; Williams, A. *J. Chem. Educ.* **2010**, *87*, 1123.
6. ChemSpider SyntheticPages (CSSP) (http://cssp.chemspider.com/) (Accessed on 2/17/2017).
7. ChEMBL (https://www.ebi.ac.uk/chembl/) (Accessed on 2/17/2017).
8. Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magariños, M. P.; Overington, J. P.; Papadatos, G.; Smit, I.; Leach, A. R. *Nucleic Acids Res.* **2017**, *45*, D945.
9. Orange Book: Approved Drug Products with Therapeutic Equivalence Evaluations (http://www.accessdata.fda.gov/scripts/cder/ob/default.cfm) (Accessed on 2/17/2017).
10. DailyMed (http://dailymed.nlm.nih.gov/) (Accessed on 2/17/2017).
11. ChEBI (https://www.ebi.ac.uk/chebi/) (Accessed on 2/17/2017).
12. Hastings, J.; de Matos, P.; Dekker, A.; Ennis, M.; Harsha, B.; Kale, N.; Muthukrishnan, V.; Owen, G.; Turner, S.; Williams, M.; Steinbeck, C. *Nucleic Acids Res.* **2013**, *41*, D456.
13. NIST Chemistry Webbook (http://webbook.nist.gov/chemistry/) (Accessed on 2/19/2017).
14. Linstrom, P. J.; Mallard, W. G. *J. Chem. Eng. Data* **2001**, *46*, 1059.
15. DrugBank (http://www.drugbank.ca/) (Accessed on 2/19/2017).
16. About DrugBank (http://www.drugbank.ca/about) (Accessed on 2/19/2017).
17. Law, V.; Knox, C.; Djoumbou, Y.; Jewison, T.; Guo, A. C.; Liu, Y. F.; Maciejewski, A.; Arndt, D.; Wilson, M.; Neveu, V.; Tang, A.; Gabriel, G.; Ly, C.; Adamjee, S.; Dame, Z. T.; Han, B. S.; Zhou, Y.; Wishart, D. S. *Nucleic Acids Res.* **2014**, *42*, D1091.
18. The Human Metabolome Database (HMDB) (http://www.hmdb.ca/) (Accessed on 2/19/2017).
19. About the Human Metabolome Database (HMDB) (http://www.hmdb.ca/about) (Accessed on 2/19/2017).
20. Wishart, D. S.; Jewison, T.; Guo, A. C.; Wilson, M.; Knox, C.; Liu, Y. F.; Djoumbou, Y.; Mandal, R.; Aziat, F.; Dong, E.; Bouatra, S.; Sinelnikov, I.; Arndt, D.; Xia, J. G.; Liu, P.; Yallou, F.; Bjorndahl, T.; Perez-Pineiro, R.; Eisner, R.; Allen, F.; Neveu, V.; Greiner, R.; Scalbert, A. *Nucleic Acids Res.* **2013**, *41*, D801.
21. ToxNet (http://toxnet.nlm.nih.gov/) (Accessed on 2/19/2017).
22. Factsheet - Toxicology Data Network (TOXNET) (http://www.nlm.nih.gov/pubs/factsheets/toxnetfs.html) (Accessed on 2/19/2017).
23. Wexler, P. *Toxicology* **2001**, *157*, 3.
24. Fowler, S.; Schnall, J. G. *Am. J. Nurs.* **2014**, *114*, 61.
25. ChemIDplus (http://chem.sis.nlm.nih.gov/chemidplus/chemidlite.jsp) (Accessed on 2/19/2017).
26. Fact Sheet - ChemIDplus (http://www.nlm.nih.gov/pubs/factsheets/chemidplusfs.html) (Accessed on 2/19/2017).
27. Hazardous Substances Data Bank (HSDB) (http://toxnet.nlm.nih.gov/newtoxnet/hsdb.htm) (Accessed on 2/19/2017).
28. Fact Sheet - Hazardous Substances Data Bank (HSDB) (http://www.nlm.nih.gov/pubs/factsheets/hsdbfs.html) (Accessed on 2/19/2017).
29. Comparative Toxicogenomics Database (CTD) (http://toxnet.nlm.nih.gov/newtoxnet/ctd.htm) (Accessed on 2/19/2017).
30. Fact Sheet - Comparative Toxicogenomics Database (CTD) (http://www.nlm.nih.gov/pubs/factsheets/ctdfs.html) (Accessed on 2/19/2017).
31. Worldwide Protein Data Bank (wwPDB) (http://www.wwpdb.org/) (Accessed on 2/19/2017).

32. Protein Data Bank in Europe (PDBe) (http://www.ebi.ac.uk/pdbe/) (Accessed on 2/19/2017).
33. Gutmanas, A.; Alhroub, Y.; Battle, G. M.; Berrisford, J. M.; Bochet, E.; Conroy, M. J.; Dana, J. M.; Montecelo, M. A. F.; van Ginkel, G.; Gore, S. P.; Haslam, P.; Hendrickx, P. M. S.; Hirshberg, M.; Lagerstedt, I.; Mir, S.; Mukhopadhyay, A.; Oldfield, T. J.; Patwardhan, A.; Rinaldi, L.; Sahni, G.; Sanz-Garcia, E.; Sen, S.; Slowley, R. A.; Velankar, S.; Wainwright, M. E.; Kleywegt, G. J. *Nucleic Acids Res.* **2014**, *42*, D285.
34. Protein Data Bank Japan (PDBj) (http://pdbj.org/) (Accessed on 2/19/2017).
35. Kinjo, A. R.; Suzuki, H.; Yamashita, R.; Ikegawa, Y.; Kudou, T.; Igarashi, R.; Kengaku, Y.; Cho, H.; Standley, D. M.; Nakagawa, A.; Nakamura, H. *Nucleic Acids Res.* **2012**, *40*, D453.
36. RCSB Protein Data Bank (RCSB PDB) (http://www.rcsb.org/pdb/) (Accessed on 2/19/2017).
37. Rose, P. W.; Prlic, A.; Bi, C. X.; Bluhm, W. F.; Christie, C. H.; Dutta, S.; Green, R. K.; Goodsell, D. S.; Westbrook, J. D.; Woo, J.; Young, J.; Zardecki, C.; Berman, H. M.; Bourne, P. E.; Burley, S. K. *Nucleic Acids Res.***2015**, *43*, D345.