VIRTUAL DATABASES FOR DRUG DISCOVERY

Lyubishkin N.R., Kardash O.V., Klenina O.V., Chaban T.I.
Danylo Halytsky Lviv National Medical University, Lviv, Ukraine
nikita_2018@ukr.net

An indispensable condition in performing virtual screening (VS) of drug candidates is the availability of a 3*D* structures of the target protein and ligands to be docked. Some databases were created to store 3*D* structures of molecules. Some of the free databases include Protein Data Bank (PDB), PubChem, ChEMBL, ChemSpider, Zinc, Brazilian Malaria Molecular Targets (BraMMT), Drugbank, and Our Own Molecular Targets (OOMT). In addition, there are some commercially available databases such as the MDL Drug Data Report. Below we are going to present a brief explanation of each of these databases:

• Protein Data Bank (PDB): PDB is the public database where three-dimensional structures of proteins, nucleic acids, and complex molecules have been deposited since 1971. The worldwide PDB organization ensures that PDB files are publicly available to the global community. It is widely used by the academic community and has grown consistently in recent years. In the last 10 years, the number of 3*D* structures of the PDB increased from 48169 at the end of 2008 to 147604 in the end of 2018, an increase of nearly 207 %. This implies that in the last 10 years, almost 9943 new structures have been added to the PDB every year, just over 27 structures per day, on average. The pace of this growth has increased. At the beginning of this decade approximately 25 new entries were added per day on average. In 2018, over 31 new structures were added per day, an average daily growth of 24 % compared to 2010.

• PubChem: PubChem is a public database, aggregating information from smaller, more specific databases. It has more than 97 million compounds available.

• ChEMBL: ChEMBL is a database of bioactive molecules with medicinal properties maintained by the European Institute of Bioinformatics (EBI) of the European Molecular Biology Laboratory (EMBL). Currently, it has almost 2.3 million compounds and 15.2 million known biological activities.

• DrugBank: DrugBank is a collection of 13857 drug entities including 2661 approved drug molecules and 1425 approved biologics (peptides, proteins, and vaccines).

• Zinc: Zinc is a free database of commercially available compounds for VS. Zinc has more than 230 million commercially available compounds in the 3*D* format. Zinc is maintained by Irwin and Shoichet Laboratories of the Department of Pharmaceutical Chemistry at the University of California, San Francisco (UCSF). Several options are available at http://zinc20.docking.org/tranches/home/ for selecting different subsets of ligands for virtual screening. Important criteria such as selecting between 2*D*/3*D*, purchasability, charge, molecular weight and logP are highlighted. To download compounds, different methods such as downloading as an index file or directly downloading with cURL and WGET are shown.

• Our Own Molecular Target (OOMT): OOMT is a special molecular target database because it has the biological assay for all its molecular targets, and includes specific targets for cancer, dengue, and malaria. OOMT was created by a group of researchers from Federal University of Sao Joao del Rei (UFSJ).

• Brazilian Malaria Molecular Targets (BraMMT): The BRAMMT database comprises thirty-five molecular targets for *Plasmodium falciparum* retrieved from the PDB database. This database allows *in silico* virtual high throughput screening (vHTS) experiments against a pool of *P. falciparum* molecular targets.

• Drugbank: DrugBank is a database that contains comprehensive molecular information about drugs, their mechanisms, their interactions, and their targets. The database contains more than 11900 drug entries, including nearly 2538 FDA-approved small molecule drugs, 1670 biotechnology (protein / peptide) drugs approved by the FDA, 129 nutraceuticals and nearly 6000 investigational drugs.

Commercially available Databases:

• MDL Drug Data Report (MDDR): MDDR is a commercial database built from patent databases, publications and congresses. It has more than 260,000 biologically relevant compounds and approximately 10000 compounds are added every year.

• ChemSpider: ChemSpider is a database of chemical substances owned by the Royal Society of Chemistry. It has more than 71 million chemical structures from over 250 data sources. ChemSpider allows downloading up to 1000 structures per day. Previous contact is needed for the download of more structures, and ChemSpider is therefore not a totally free database.

The association-based identification of drug targets is a commonly used approach. Several web servers for target prediction are publically available:

• Open Targets Platform Open Targets Platform is a knowledge-based platform that provides evidence about the association between known drug targets with diseases and enables the identification and prioritization of drug targets. It integrates diverse sources, including omics data, experimental results from animal models, and text-mined data from the literature. The platform then ranks genes according to their association with disease.

• Harmonizome is a collection of comprehensive and processed knowledge gathered from over 70 major online resources оп genes and proteins. It enables the discovery of novel relationships and functional associations between biological entities (proteins/genes).

• Similarity Ensemble Approach (SEA) ranks target proteins based оп the chemical similarity between ligands. 65000 ligands are assigned to groups of human protein targets. Ligand topology is used to calculate a similarity score.

• SwissTargetPrediction performs a similarity search to predict the potential drug targets of queried molecules. The updated version contains 376342 experimentally active compounds and 3068 macromolecular targets.

Thus computer-aided drug discovery CADD techniques are now an essential part of the drug discovery process. Over the past few decades *in silico* drug discovery has accelerated due to rapid advancements in accumulating publicly available biological data.