From big data to complex network: a navigation through the maze of drug–target interaction

18

Ze Wang[1], *Min Li*[2], *Muyun Tang*[2] *and Guang Hu*[2]
[1]Department of Pharmaceutical Sciences, Zunyi Medical University at Zhuhai Campus, Zhuhai, P.R. China, [2]Department of Bioinformatics, Center for Systems Biology, School of Biology and Basic Medical Sciences, Soochow University, Suzhou, P.R. China

18.1 Introduction

At an average cost of $985 million per drug and at least a decade to reach the market, drug discovery and development are highly expensive, time-consuming, and complex processes (Wouters et al., 2020; Mullard, 2020; Mohs and Greig, 2017). In fact, the attrition rate of drug discovery and the number of clinical trial failures has increased in the last decades (Bolognesi and Cavalli, 2016; Chaudhari, et al., 2017). As pointed out by Hopkins, the fundamental problem may be the core philosophy in drug discovery, which traditionally assumes that the primary goal as designing exquisitely selective "magic bullets" to bind with a single disease target (Hopkins, 2008). With the development of systems biology, scientists realized the one-bullet-one-target assumption is oversimplified and accepted the concept of network pharmacology as a paradigm shift in drug discovery (Hopkins, 2008; Loscalzo and Barabasi, 2011; Yildirim et al., 2007; Liang and Hu, 2016; Yan et al., 2018). Compared to the traditional one-bullet-one-target paradigm, network pharmacology attempts to uncover drug action by considering the interaction between drug molecules and their potential targets through a holistic network, which has great potential to facilitate disease mechanism understanding and drug discovery (Wang et al., 2021).

Identification and discovery of potential therapeutic targets for drugs have largely benefited from high-throughput experimental techniques, which generate numerous biological data (Russell et al., 2013). On the other hand, clarification and characterization of active ingredients from herbal plants also deposited a huge amount of chemical data (French et al., 2018). With the continuous collection and deposition of big data from high-throughput experiments, modern drug discovery and development are moving into the big data era (Zhu, 2020). It is now realized that big data in drug discovery is proposing four challenges to traditional data management and analysis methodologies, including the scale of data, the growth speed of data, the diversity of data source, and the uncertainty of data (Ciallella and Zhu, 2019; Lee and Yoon, 2017).

Big Data Analytics in Chemoinformatics and Bioinformatics. DOI: https://doi.org/10.1016/B978-0-323-85713-0.00016-5

For example, several million compounds were typically investigated in high-throughput experiments in drug development (Santos et al., 2017). More importantly, data uncertainty, especially when considering complex biological mechanisms (e.g., drug responses, side effects), has brought further obstacles to using this data. Therefore, the development of new data analysis tools and computational algorithms to manage and utilize these data is necessary for drug discovery and development.

Modeling the action of drugs through the big data has given birth to the complex network view of drug−target interaction (Hopkins, 2008), which is composed of nodes and lines representing molecular entities (for both drug and target molecules) and their relations, respectively. Network science, which originates from the great mathematician Euler in the Königsberg problem, is growing as a systematic tool for the analysis of complex networks emerging from a wide range of disciplines (Borgatti and Halgin, 2011; Newman, 2003; Parkhe et al., 2006). As shown by Yildirim et al., the application of network science to drug screening data has demonstrated a network map rather than isolated, bipartite nodes for drugs and targets, which revolutionized our understanding of drug−target interactions (Fig. 18.1) (Yildirim et al., 2007). Currently, computational identification and analysis of drug−target interaction are becoming a cutting-edge research areas in drug discovery and development.

Figure 18.1 Drug−target interaction network constructed from FDA-approved drugs (Yildirim et al., 2007). In the network, drugs and targets are represented as circular and rectangular nodes, respectively. The area of node is proportional to the number of interactions, which are shown as lines. Different colors are used to classify drugs and targets, according to Anatomical Therapeutic Chemical Classification and the Gene Ontology database, respectively.
Source: With permission from Yildirim, M.A., Goh, K.I., Cusick, M.E., Barabasi, A.L., Vidal, M., 2007. Drug-target network. Nat. Biotechnol. 25, 1119−1126.

In this chapter, we reviewed the important data sources in drug discovery and development, including drug screening, active ingredient profiling, and target fishing. These databases are building blocks for the construction and prediction of drug−target interactions. Then, we introduced the algorithms and methodologies in the construction, prediction, and analysis of drug−target interaction. The prediction methods can be roughly divided into structure-based, similarity-based, and machine learning-based. Although structure-based methods showed high accuracy, the application of these methods is often limited by the lack of three-dimensional structures. Therefore, we only focused on the other two methods. In the second part, we also reviewed important computational tools and methods in network construction and analysis. We hope the content of this chapter will highlight the critical role of the network view of drug−target interaction, which is driven by the continuously expanding databases.

18.2 Databases

The construction of drug−target interaction networks relies on databases, which are generally composed of a hierarchical collection of alphabetical, numerical, graphical, and structural data. This section will introduce the most commonly used databases covering small molecules (Table 18.1), biological macromolecules (Table 18.2), and traditional Chinese medicine (TCM) (Table 18.3), as well as their interactions.

18.2.1 Chemical databases

18.2.1.1 DrugBank

Released in 2006, DrugBank (https://go.drugbank.com/) is one of the most used drug-related resources for bioinformatics, chemoinformatics, and medicinal chemistry (Wishart et al., 2017). It is a freely available internet-based database that aims to comprehensively include detailed information on targets, mechanisms, and interactions of both FDA-approved and investigational drugs. The current version contains a total number of 14,460 drug entries, including 2683 FDA-approved small molecule drugs, 2585 biotech drugs such as proteins and peptides, 6643 phase I/II/III drugs, and 131 nutraceuticals (Table 18.1, Fig. 18.2, data collected at the end of April 2021) (Wishart et al., 2017). Besides, 5236 nonredundant protein sequences and annotations related to the drugs were included. Each drug entry is composed of over 200 distinct data fields covering chemical identification, pharmacology, pharmaceutics, clinical trial, target sequence, pathway, and spectra information (Wishart et al., 2017). Data in DrugBank can be accessed and retrieved from a field search engine. Additionally, the database provides alternative format and datasets for data mining and analysis. For example, DrugBank contains a portal for machine-learning algorithms, which require labeled datasets including drug, target, side-effect, and toxicity (Wishart et al., 2017).

Table 18.1 Chemical databases for drugs and small molecules.

Database	Description	Database statistics	Website	Reference
DrugBank	FDA-approved and investigational drug	2683 FDA-approved small molecule drugs, 2585 biotech drugs, 6643 investigational drugs, 131 nutraceuticals, 5236 nonredundant protein sequences, and annotations	https://go.drugbank.com/	Wishart et al. (2017)
PubChem	Resources for chemical compounds	270,998,024 chemical entities (109,891,884 unique chemical structures) and 1,366,263 bioassays	https://pubchem.ncbi.nlm.nih.gov/	Kim et al. (2018)
ChEMBL	Structure, bioassays, affinity data for drug	More than 2 million compounds,17 million activity data, >1600 distinct cell lines, 500 tissues/organs, 3600 organisms, >14,300 targets	https://www.ebi.ac.uk/chembl/	Mendez et al. (2018)
ChemSpider	Pure chemical structure and property	103 million chemical structures and links to original data sources	http://www.chemspider.com/	Pence and Williams (2010)

Table 18.2 Biological databases for targets.

Database	Description	Database statistics	Website	Reference
UniProt	Comprehensive database for protein sequence and annotations	564,638 reviewed protein sequences for over 84 thousand species	https://www.uniprot.org/	Consortium (2018)
PDB	Structural data for biomacromolecules	177,009 structural entities for biological macromolecules	https://www.rcsb.org/	Burley et al. (2018)
STRING	Interaction database	24,584,628 proteins, 3,123,056,667 total interactions from 5090 organisms	https://string-db.org/	Szklarczyk et al. (2018)
BindingDB	Affinity database	2.2 million protein−ligand affinity data, involving 977,487 small molecules and 8516 targets	https://www.bindingdb.org/	Gilson et al. (2015)

Table 18.3 Databases for traditional Chinese medicine.

Database	Description	Database statistics	Website	Reference
TCM database@Taiwan	Currently the most comprehensive and largest noncommercial TCM database available for download	37,170 (32,364 nonduplicate) TCM compounds from 352 TCM ingredients.	http://tcm.cmu.edu.tw/	Chen (2011)
TCMSP	Highlight the role that the systems pharmacology plays across the TCM discipline.	All the 499 herbs registered in Chinese pharmacopoeia (2010), with a total of 12144 chemicals.	https://www.tcmspw.com/tcmsp.php	Ru et al. (2014)
TCMID	Information on all respects of TCM including formulae, herbs, and herbal ingredients, and information for drugs and diseases	8159 herbs, 46,914 TCM formulae, and more than 25,210 herb ingredients.	http://119.3.41.228:8000/tcmid/	Huang et al. (2017)

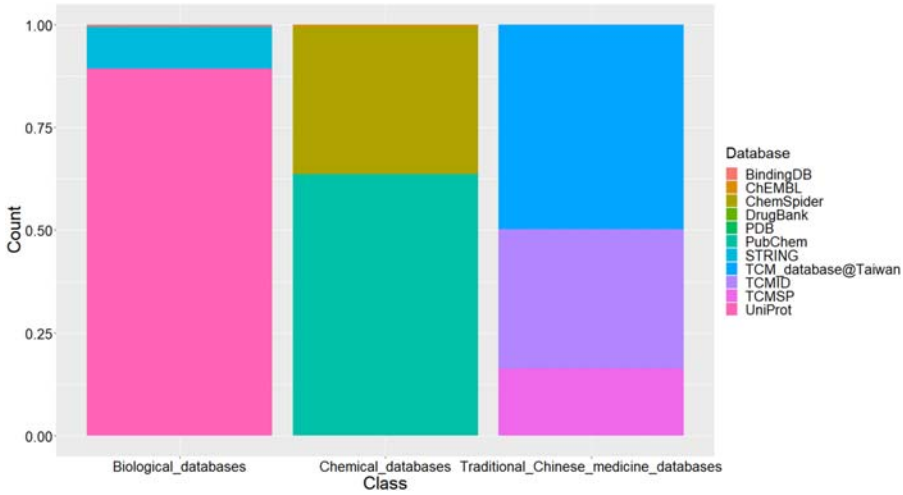

Figure 18.2 Statistics for data size of each database in three different categories, that is, small molecules, biological targets, and traditional Chinese medicine.

18.2.1.2 PubChem

Initiated and maintained by the US National Institutes of Health (NIH), PubChem (https://pubchem.ncbi.nlm.nih.gov/) is an open database that collects chemical information and resources (Kim et al., 2018). PubChem supports bidirectional data transfer between users and the database, allowing contributors to create, upload, and edit data freely. Since its first version in 2004, PubChem has continually become a huge chemical database that contains 270,998,024 chemical entities (109,891,884 unique chemical structures) and 1,366,263 bioassays (carbohydrates, nucleotides, peptides, etc.) contributed by PubChem users (Table 18.1, Fig. 18.2) (Kim et al., 2018). It provides spectral information including ^{1}H NMR, ^{13}C NMR, 2D NMR, FT-IR, Ms, UV-Vis, and Raman data for more than 590,000 compounds. Spectral data in PubChem are linked with external spectral databases such as SpectraBase (http://spectrabase.com) and the MassBank of North America (https://mona.fiehnlab.ucdavis.edu/). By the end of April 2021, the database archived 296,907,771 biological activity data, 90,426 gene data, 96,561 protein data, 4849 taxonomy, and 237,925 pathways involved with chemical entities (Kim et al., 2018). Data in PubChem is organized as three dependent databases, including substance which collects descriptions of substances contributed by users, Compound which enumerates chemical compounds according to unique chemical structure, and Bioassay containing biological assays and experiments related to the compounds.

18.2.1.3 ChEMBL

ChEMBL (https://www.ebi.ac.uk/chembl/) is a manually maintained drug discovery database that deposits medicinal chemistry data from clinical development

candidates and academic journals including *Bioorganic & Medicinal Chemistry Letter, Journal of Medicinal Chemistry, Bioorganic & Medicinal Chemistry, Journal of Natural Products, European Journal of Medicinal Chemistry, MedChemComm, ACS Medicinal Chemistry Letters*, etc. (Mendez et al., 2018). Structures of compounds, assays, and activity information were manually extracted from the literature by ChEMBL curators. Since information such as structure connectivity, stereochemistry, and quantitative values are prone to error, it is encouraged to contribute to ChEMBL data by depositing chemical and biological information during scientific publication (Mendez et al., 2018). The current released version ChEMBL 28 (at the end of April 2021) contains over 2 million compounds from over 80,000 publications and patents. It includes over 17 million activity data annotating from over 1600 distinct cell lines, 500 tissues/organs, and 3600 organisms (Table 18.1, Fig. 18.2) (Mendez et al., 2018). The number of targets in ChEMBL has exceeded 14,300, with 6311 human proteins (Mendez et al., 2018). Except for human, mouse, and rat targets, the database also contains plenty of experimental data from other model organisms such as *Staphylococcus aureus*. ChEBML is embracing new data sources from bacteria, viruses, and pathogens, making it an ideal platform for multipurpose drug development (e.g., antimicrobial). Clinical data in ChEMBL is continuing to be incorporated with other public databases such as the ClinicalTrials.gov database (https://clinicaltrials.gov/), FDA Orange Book (https://www.accessdata.fda.gov/scripts/cder/ob/), FDA New Drug Approvals (https://www.fda.gov/Drugs/DevelopmentApprovalProcess/DrugInnovation/default.htm), the British National Formulary (https://bnf.nice.org.uk/), Medicinal Subject Headings (MeSH, https://www.nlm.nih.gov/mesh/). Bioactivity data are also timely exchanged with external databases like PubChem (https://pubchem.ncbi.nlm.nih.gov/) and BindingDB (http://www.bindingdb.org/). Other properties of deposited compounds were calculated by RDKit (https://www.rdkit.org/). For data accessibility, ChEMBL supports text search through its webpage and download from FTP site (ftp://ftp.ebi.ac.uk/pub/databases/chembl/ChEMBLdb/latest/) with a variety of data formats including SD file and FASTA file (Mendez et al., 2018).

18.2.1.4 ChemSpider

From the perspective of pure chemical structure and property, researchers hope to obtain a variety of information about a compound, including molecular structure, systematic nomenclature, physical properties, spectral data, reactions and synthetic methods, and safety information. The information is typically distributed in different literatures, libraries, and databases. ChemSpider (http://www.chemspider.com/) was born to collectively integrate chemical structure-related information from different data sources (Table 18.1, Fig. 18.2) (Pence and Williams, 2010). In 2009, ChemSpider was purchased by the Royal Society of Chemistry (RSC), allowing the accessibility of a wealth of information from RSC, that is, scientific publications and databases. ChemSpider has also been connected with other databases such as Wikipedia (https://en.jinzhao.wiki/wiki/Main_Page), PubChem, and Kyoto Encyclopedia of Genes and Genomes (KEGG) (https://www.kegg.jp/). To avoid

errors in the data input process, ChemSpider is curated by only registered users. The data in ChemSpider can be accessed from text search, structure searches as well as substructure search. With over 103 million chemical structures and links to original data sources, ChemSpider is becoming a portal to the property, annotation, synthesis, spectral information of the expanding chemical universe (Table 18.1) (Pence and Williams, 2010).

18.2.2 Databases for targets

18.2.2.1 UniProt

The Universal Protein Resource (UniProt, https://www.uniprot.org/) is aimed to provide a comprehensive and high-quality data source of protein sequences and annotations (Table 18.2) (Consortium, 2018). The behavior and physiology of cells are defined by proteins that respond to environmental signals. Understanding the time-dependent protein expression at a whole proteome level is crucial to interpret life in a quantitative way. With the improvements of experimental techniques, the information on protein sequence, structure and function is increasing broadly and deeply. It is therefore challenging to manage the information and make it conveniently accessible to users. UniProt data are managed by more than 100 experts hosted by the collaboration of the European Bioinformatics Institute (EMBL-EBI), the Swiss Institute of Bioinformatics (SIB), and the Protein Information Resource (PIR). UniProt (release 2020_05) now provides 564,638 reviewed entries for over 84 thousand species including humans, rice, *Arabidopsis thaliana*, mouse, zebrafish, etc. UniProt entry is composed of the core data field (protein sequence, protein name, description, taxonomy, citation) and peripheral field including as much annotation information (Consortium, 2018). Although the database can provide rich information by simple text query and search, it actively supports in-depth data mining through various online training such as webinars (https://www.ebi.ac.uk/training/online/), YouTube videos (https://www.youtube.com/user/uniprotvideos/), Facebook (https://www.facebook.com/uniprot.org/), and Twitter (@uniprot).

18.2.2.2 Protein Data Bank

Structural biology has witnessed frequent advances in the structural determination of proteins, RNA, DNA, and their complexes with small molecules. Since 1971, the Protein Data Bank (PDB, https://www.rcsb.org/) established an open-access database in structural biology by depositing only seven protein structures at the beginning (Table 18.2) (Burley et al., 2018). With continuing development, PDB has grown up to a comprehensive database consisting of 177,009 structural entities for biological macromolecules (Fig. 18.2) (Burley et al., 2018). PDB data entry is originated from experimental sources including X-ray diffraction, nuclear magnetic spectroscopy (NMR), and three-dimensional electron microscopy (Table 18.2). Structural data are validated and biocurated by a global expert team to ensure the accurate representation of the structural data and the underlying annotation

information. Data exploration service in PDB allows convenient accessibility to every structural entry via any popular web browser (e.g., Chrome, Firefox, Microsoft Edge). The website rcsb.org supports the keywords and unstructured text search, whilst the obtained data are sorted and tabulated to include atomic coordinates, experimental methods, sequence, description, citation, specific chemical components, taxonomy, and enzyme classification. Additionally, PDB data can be explored by multiple online tools for data manipulation and visualization. For example, the PDB website enables metabolic pathway mapping for user-interested structures, drug, and ligand discovery through external links such as DrugBank and BindingDB, as well as the fast and interactive three-dimensional display through NGL Viewer (Burley et al., 2018).

18.2.2.3 String

With impressive advances in elucidating the interaction between individual proteins, it is realized cellular machinery depends on the global network of physical (direct) and functional (indirect) protein−protein interactions. The information space of protein−protein interactions is far more complicated than the intrinsic properties and annotations of individual proteins. STRING (https://string-db.org/) is a knowledgebase of known and computationally predicted protein−protein interactions (Szklarczyk et al., 2018). It collects and stores protein−protein interaction data from a variety of publicly available data sources: genomic predictions, high-throughput experiments, co-expression, automated text-mining, and online databases such as Database of Interacting Proteins (DIP, http://dip.doe-mbi.ucla.edu/), Biomolecular Interaction Network Database (BIND, http://bind.ca/), Molecular Interaction Database (MINT, http://mint.bio.uniroma2.it/mint/), KEGG (http://www.kegg.jp/), and Reactome (http://www.reactome.org/). STRING v11.0 contains 24,584,628 proteins and 3,123,056,667 total interactions from 5090 organisms including Homo sapiens, Mus musculus, *A. thaliana*, and so on (Table 18.2, Fig. 18.2) (Szklarczyk et al., 2018). STRING defines a functional association unit as the basic building blocks, which is an edge between two proteins both having functional contributions to a specific biological process. By the definition, protein−protein interaction does not necessarily require physical contact between proteins. STRING website provides user-friendly access to the interaction network for single protein and multiple proteins, which can be enquired either by name or sequence. Also, through the STRING online server, users can compute functional enrichment for a set of proteins involving the interaction network (Szklarczyk et al., 2018).

18.2.2.4 BindingDB

BindingDB (https://www.bindingdb.org/) is an open database of experimental affinity data of protein−ligand interaction (Gilson et al., 2015). With steady growth since 2000, BindingDB now contains about 2.2 million protein−ligand affinity data, involving 977,487 small molecules and 8516 protein targets (Table 18.2,

Fig. 18.2) (Gilson et al., 2015). The data source for BindingDB includes scientific publications and patents. Affinity data of at least one protein−ligand complex is supplied in the database along with information on publication source and experimental conditions (e.g., temperature, pH, buffer composition). BindingDB supports interactive connection to several public databases including PDB, UniProt, DrugBank, ChEMBL, PubChem, Reactome, MarinLit (http://pubs.rsc.org/marinlit), and ZINC (http://zinc.docking.org/). Data in BindingDB is organized as hyperlinks listed in a table format and can be accessed through flexible web tools for query, browsing, download, visualization, and analysis (Gilson et al., 2015).

18.2.3 Databases for traditional Chinese medicine

TCM often comprises over thousands of chemical compounds from different botanical species, hitting multiple biological targets (Cheung, 2011). The herbal compounds and corresponding targets form a complex network that involves various nodes and edges (Li et al., 2011; Tao et al., 2013). To comprehensively characterize and analyze the network, the wet experiment is time-consuming and expensive due to the dozens of chemical entities and biological targets involved. Systems pharmacology is a big data-driven strategy that deals with prior experimental data of herbal compounds as well as biological assays (Li et al., 2011; Ru et al., 2014). With increasing attention towards discovering novel lead compounds from TCM, a database for TCM is necessary. Besides, the prediction power of systems pharmacology is enhanced by online target prediction algorithms. This section briefly reviews some typical TCM databases (Table 18.3).

18.2.3.1 Traditional Chinese medicine Database@Taiwan

TCM Database@Taiwan includes more than 20000 chemical compounds from 453 herbs, animals, and minerals in TCM (Chen, 2011). The database is evolving to cover more compound data from folk herbs. In TCM Database@Taiwan, drug molecules were classified into 22 different categories according to clinical applications (Chen, 2011). The classification model is based on the theories of TCM involving the Yin-yang and the Five Elements theory. TCM ingredients were collected from publications on Medline and ISI Web of Knowledge. Through simple and advanced search, TCM Database@Taiwan provides both two-dimensional and three-dimensional structures of each TCM constituent, as well as physical properties such as ALogP, polar surface area, rotatable bonds, and so on (Table 18.3, Fig. 18.2) (Chen, 2011).

18.2.3.2 Traditional Chinese medicine systems pharmacology

The TCM systems pharmacology (TCMSP) database and analysis platform is built for this purpose (Ru et al., 2014). TCMSP contains 499 Chinese herbs collected in Chinese Pharmacopeia (Ru et al., 2014). Through deep data mining and analysis, 29,384 chemical compounds, 3311 targets, and 837 associated diseases were

manually curated in the database (Table 18.3, Fig. 18.2) (Ru et al., 2014). ADME-related properties were computed in TCMSP, including oral bioavailability, half-life, drug-likeness, Caco-2 permeability, blood−brain barrier, and Lipinski's rule of five (Ru et al., 2014). For drug targets, TCMSP includes all experimentally validated targets and SysDT model predicted targets. The strengths of the TCMSP platform allow the analytical decomposition of TCM through data and network methodology (Ru et al., 2014).

18.2.3.3 Traditional Chinese medicine integrated database

TCM integrated database (TCMID) is aiming to provide convenient online information on TCM for pharmacologists and scholars (Huang et al., 2017). Established in 2013, TCMID integrated online databases including TCM Database@Taiwan (Chen, 2011), HIT (Ye et al., 2010) to collect over 49,000 prescriptions, 8159 herbs, 25,210 ingredients, 3791 diseases, 6828 drugs and 17,521 targets (Table 18.3, Fig. 18.2). Since most publications on TCM, especially separation and pharmacological research, were written in Chinese, TCMID manually collects original data from the Chinese national knowledge infrastructure (CNKI) and translated the related information into English. Users can easily retrieve detailed descriptions and information from external databases such as Drugbank, OMIM, and STITCH. Additionally, TCMID has documented mass spectra (Ms) of Chinese herbs through CNKI. TCMID collects the original place of the herb, Ms spectrum, chromatography spectrum, as well as compound information (Huang et al., 2017). With new features added, TCMID is growing as an important hub for the modernization of TCM.

18.3 Prediction, construction, and analysis of drug−target network

The drug−target network is mathematically described as a bipartite network graph $G(D, T, P)$. In the network, the drug set D and target set T is defined as

$$D = D(d_1, \quad d_2, \ldots, d_n)$$

$$T = T(t_1, \quad t_2, \ldots, t_m)$$

And the interaction set P is defined as a Kronecker matrix:

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ & \cdots & & \\ p_{m1} & p_{m2} & \cdots & p_{nm} \end{bmatrix}$$

where $p_{kl} = 1$ when drug d_k binds with target t_l, otherwise $p_{kl} = 0$. Practically, a binding affinity threshold is used to obtain the interaction p_{kl}. The purpose of the prediction,

construction, and analysis of the drug−target network is to identify drug targets from the whole target pool, formulate the interactome configuration, and characterize the property and module both globally and locally. A landscape on the drug−target interaction network is crucial to the understanding of therapeutic mechanisms and side effects. In this section, we briefly review advances in algorithms, computational tools, and network analysis methods in drug−target interaction.

18.3.1 Algorithms to predict drug−target interaction network

Prediction of biological networks containing thousands of compounds and targets is still challenging to the traditional experimental approach, such as high-throughput screening and biological assays (Haggarty et al., 2003; Kuruvilla et al., 2002; Wang et al., 2015; Whitebread et al., 2005). Therefore, the computational prediction method is important for biological network analysis. Although the virtual screening method for three-dimensional compounds and targets is well developed, the lack of three-dimensional structural data and time-consuming algorithms for most biological molecules still makes this approach limited in real application (Cheng et al., 2007; Morris et al., 2009). Alternatively, several knowledge-based computational methods have been developed to efficiently address the drug−target prediction problem (Table 18.4). In this section, we will briefly review the typical algorithms and methods for drug−target prediction.

Table 18.4 Algorithms to predict drug−target interaction.

Algorithms	Classification	Description	Reference
Bipartite graph algorithm	Supervised machine learning	A supervised machine learning algorithm for a BG model, mapping drugs in chemical space and targets in genomic space	Yamanishi et al. (2008)
Advanced BG algorithm	Supervised machine learning	Advanced version of BG algorithm with pharmacological data involved	Yamanishi et al. (2010)
BLM	Supervised machine learning	A BLM incorporating the concepts of local models to predict drug−target interaction	Bleakley and Yamanishi (2009)
BLM-NII	Supervised machine learning	An updated version of BLM by introducing neighbor-based interaction-profile inferring	Mei et al. (2013)
RBM	Supervised machine learning	A RBM with a two-layer graphic model effectively capture the features of drug−target interaction and predict different types of drug−target interaction	Wang and Zeng (2013)

(*Continued*)

Table 18.4 (Continued)

Algorithms	Classification	Description	Reference
Random forest algorithm	Supervised machine learning	Combines the information from chemical, biological, and network features to predict drug−target interaction with high accuracy	Cao et al. (2014)
Negative dataset selection method	Supervised machine learning	Two methods to assist the selection of negative dataset in the machine learning-based algorithms	Wang et al. (2014)
NetLapRLS	Semisupervised machine learning	Adopts both labeled and unlabeled data in machine learning	Xia et al. (2010)
Chemical similarity	Chemical similarity	Chemical similarity method based on the assumption that similar drug structures are more likely to interact with similar targets	Keiser et al. (2009)
Two-step similarity	Chemical similarity	A similarity score was obtained by graph representation and chemical functional group representation in two steps	Chen and Zeng (2013)
Phenotypic side-effect similarity	Network similarity	An algorithm to determine if two drugs will interact with the same target	Campillos et al. (2008)
NRWRH	Network similarity	Based on the framework of random walk and the assumption that similar drugs often corresponding to similar targets	Xia et al. (2010)
DBSI	Network similarity	Drug-based similarity inference based on complex network theory	Cheng et al. (2012)
TBSI	Network similarity	Target-based similarity inference based on complex network theory	Cheng et al. (2012)
NBI	Network similarity	Network-based inference based on complex network theory	Cheng et al. (2012)
Within scores and between scores	Network similarity	Considering features from target similarity and drug similarity	Shi et al. (2015)
MTOI	Network similarity	Multiple target optimal intervention finding algorithm to identify potential drug targets and their optimal combinations restoring to a normal state	Yang et al. (2008)

18.3.1.1 Machine learning-based methods

Yamanishi et al. have developed a bipartite graph (BG) algorithm to probe drug−target interaction for four target classes including enzymes, ion channels, GPCRs, and nuclear receptors (Table 18.4) (Yamanishi et al., 2008). By introducing in-prior knowledge of chemical structures and genomic sequence information, they have built a supervised machine learning algorithm for a BG model, mapping drugs in chemical space and targets in genomic space (Fig. 18.3). The machine learning models f_c and f_g were defined based on a modified kernel regression function:

$$f: \quad X \times X \rightarrow \mathbf{R}^q$$

$$f(x, x_i) = \sum_{i=1}^{n} s(x, x_i)\mathbf{w}_i + \epsilon$$

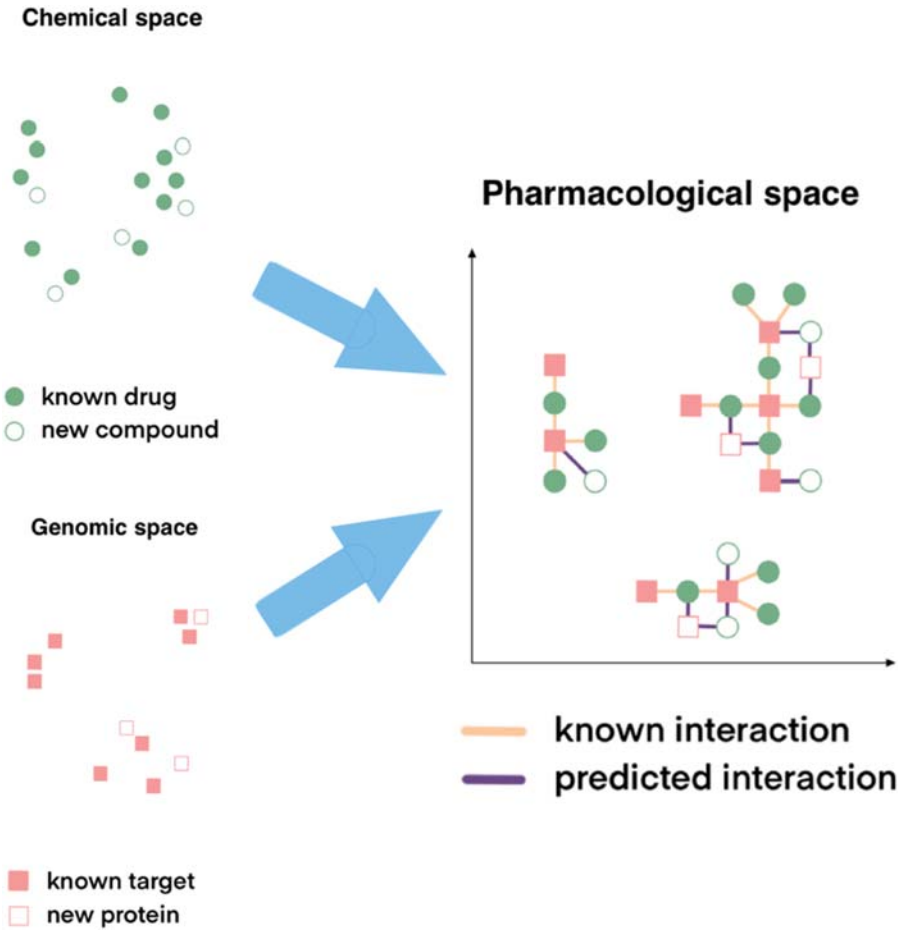

Figure 18.3 An illustration of bipartite graph algorithm.

where **w** represents a weighing vector, and *s* stands for similarity score for chemical structures or sequence (Yamanishi et al., 2008). In this BG pharmacological space algorithm, the structural and sequence similarity were considered, and the drug−target interactions were predicted by the closeness between drugs and targets.

Regarding the side effect of a drug, it is assumed that drugs with similar side effects are more likely to interact with similar targets. Taking pharmacological data into consideration may further improve the performance of a machine learning-based algorithm. Yamanishi et al. have further improved the BG method by involving pharmacological knowledge (Yamanishi et al., 2010). The pharmacological effect similarity, computed from the chemical structures of drugs, was introduced into the BG model to identify drug−target interactions (Table 18.4) (Yamanishi et al., 2010).

Based on the BG method, Bleakley et al. proposed a bipartite local model (BLM) incorporating the concepts of local models to predict drug−target interaction (Table 18.4) (Bleakley and Yamanishi, 2009). By involving local models, the edge-prediction problem was transformed into the binary classification of labeled points (Bleakley and Yamanishi, 2009). Targets are predicted by comparing sequence similarities, and drug−target interactions are predicted based on structural similarities. Finally, independent drug−target interactions were obtained putatively. Since it combines the strengths of the BG model and the local model, the BLM algorithm showed an excellent computational performance to predict drug−target interaction (Bleakley and Yamanishi, 2009).

Despite the computational speed, BLM is not able to predict drug−target interaction without training data. Therefore, the prediction of drug−target interaction for new drug molecules is not possible by using BLM. Mei et al. have proposed an updated version of BLM by introducing neighbor-based interaction-profile inferring (BLM-NII, Table 18.4) (Mei et al., 2013). The BLM-NII method derived the initial weighted interactions for the new drug from its neighbor interaction profile and then labeled this interaction to train the BLM model (Mei et al., 2013). For nuclear receptors, BLM-NII enhances the BLM method, especially for the dataset that contains drug−target with no prior interaction information.

Zeng et al. have developed a restricted Boltzmann machine (RBM) method that can predict drug−target interactions and the types of interaction (Table 18.4) (Wang and Zeng, 2013). In the RBM method, a contrastive divergence algorithm was applied to a two-layer graphic model which represents drug−target interaction. Zeng et al. has tested the RBM method on MATADOR and STITCH database (Günther et al., 2008; Szklarczyk et al., 2015; Wang and Zeng, 2013). It has shown the RBM method can effectively capture the features of drug−target interaction and predict different types of drug−target interaction.

Cao et al. proposed a random forest algorithm to predict drug−target interaction (Table 18.4) (Cao et al., 2014). The novelty of the algorithm was the combination with the information from chemical, biological, and network features. The accuracy of the algorithm was evaluated as 93.52%, 94.84%, 89.68%, and 84.72% for enzymes, ion channels, GPCRs, and nuclear receptors, respectively (Cao et al., 2014). The performance of the algorithm showed the importance of network topology as training information for the prediction of drug−target interaction.

In the prediction of drug−target interaction, a common problem for the machine learning-based method is the lack of a negative dataset. Wang et al. have proposed two methods to assist the selection of negative datasets in the machine learning-based algorithms (Table 18.4) (Wang et al., 2014). In the first method, a drug−protein deviation function is defined as:

$$\xi(\mathbf{X}_i) = \sum_j \left| \frac{\langle x_j \rangle (x_{ij} - \langle x_j \rangle)}{\mathrm{var}(x_j) \sum (\langle x_j \rangle^2 / \mathrm{var}(x_j))} \right|$$

vector $\mathbf{X}_i$ ($i = 1,2,\ldots, m$) is an m-dimension vector representing m properties of the ith target, x stands for the jth value for the property of the ith targets. Wang et al. used $\xi > 0.42$ as a threshold value to select a negative dataset (Wang et al., 2014). In the second method, a probability function for the ith unknown target to be a negative sample was defined as

$$P(\mathbf{X}_i) = \frac{(\xi(\mathbf{X}_i) - \langle \xi_{\mathrm{positive}} \rangle)^2}{\sum (\xi(\mathbf{X}_i) - \langle \xi_{\mathrm{positive}} \rangle)^2}$$

$P = 0.5$ was used to consider the negative dataset. Wang et al. have improved the prediction accuracy and identified 1797 and 227 drug−target interactions by using these two methods, respectively (Wang et al., 2014).

As discussed above, labeling the positive or negative dataset is often a challenging problem in the development of algorithms based on supervised machine learning methods. The problem can be addressed by introducing a semisupervised method, which adopts both labeled and unlabeled data in machine learning. Wong et al. developed a manifold regularization semisupervised machine learning method (Table 18.4) (NetLapRLS) to predict drug−target interaction, which generates a biological space by combining information of chemical space, sequence space, and drug−target interaction network (Xia et al., 2010). In the drug domain, classification functions are defined as:

$$\mathbf{F}_d^* = \min_{\mathbf{F}_d} J(\mathbf{F}_d) = \mathbf{Y} - \mathbf{F}_{df}^2 + \beta_d \mathrm{Trace}(\mathbf{F}_d^T \mathbf{L}_d \mathbf{F}_d)$$

$$\mathbf{F}_d^* = \mathbf{W}_d \alpha_d^*$$

$$\alpha_d^* = \arg \min_{\alpha_d \in R^{n_d \times n_p}} \left\{ \mathbf{Y} - \mathbf{W}_d \alpha_d{}_f^2 + \beta_d \mathrm{Trace}(\alpha_d^T \mathbf{W}_d \mathbf{L}_d \mathbf{W}_d \alpha_d) \right\}$$

where $\mathbf{F}_d$ is the prediction function on drug domain, α_d is a cost function, and is Frobenius norm, β_d is the trade-off in the drug domain, Trace is the matrix trace, $\mathbf{Y}$ is the adjacent matrix of the known drug−target interaction network (Xia et al., 2010). The similar function $\mathbf{F}_t$ was defined in the target domain. Applying representer theorem and optimization, the prediction function in drug and protein domain

were derived as:

$$\mathbf{F}_d^* = \mathbf{W}_d(\mathbf{W}_d + \beta_d \mathbf{L}_d \mathbf{W}_d)^{-1}\mathbf{Y}$$

$$\mathbf{F}_t^* = \mathbf{W}_t(\mathbf{W}_t + \beta_t \mathbf{L}_t \mathbf{W}_t)^{-1}\mathbf{Y}$$

The predictions are then obtained by combining drug and target domain as (Xia et al., 2010)

$$\mathbf{F}^* = \frac{\mathbf{F}_d^* + (\mathbf{F}_t^*)^T}{2}$$

18.3.1.2 Similarity-based methods

Side effects and efficacy of a drug could be explained by the multiple physiological targets of a drug. It is reasonable to assume that similar drug structures are more likely to interact with similar targets. But one should keep in mind that molecular similarity should be well defined first, since the concept of similarity is subjective and the similarity space is complex (Basak et al., 2002; Basak et al., 2006). By using two-dimensional chemical similarity method, Keiser et al. predicted thousands of drug−target interactions (Table 18.4) (Keiser et al., 2009). Among them, 23 associations were experimentally confirmed, including the inhibition of the 5-hydroxytryptamine transporter by the ion channel drug Vadilex, and antagonism of the histamine H4 receptor by the enzyme inhibitor Rescriptor (Keiser et al., 2009).

Chen et al. have developed a two-step similarity-based method to predict the target group of drugs (Table 18.4) (Chen and Zeng, 2013). In this method, drugs were encoded as their graph representations. Then the target group $T(d)$ for drug d was defined as a vector containing five elements, which are Boolean values representing whether a drug target belongs to the five target groups, that is, G-protein-coupled receptors (GPCRs), cytokine receptors, nuclear receptors, ion channels and enzymes (Chen and Zeng, 2013). A similarity score was obtained by graph representation and chemical functional group representation in two steps, respectively. The method provided more than one target group for each drug, and the prediction accuracy was 79.01% and 76.43% for the training and test set, respectively (Chen and Zeng, 2013).

Using phenotypic side-effect similarity, which describes the similarity of in vitro target binding profiles of drugs, Kuhn et al. proposed an algorithm to determine if two drugs will interact with the same target (Table 18.4) (Campillos et al., 2008). Two-dimensional Tanimoto similarity coefficient of a chemical structure and a linear function P_{SE} describing the probability of sharing the same considering the side-effect similarity were used in the algorithm. Combining these functions, Kuhn et al. defined a sigmoid function P_{2D} to characterize the probability of sharing the same target from chemical structures (Campillos et al., 2008):

$$P_{2D} = \left(1 + e^{\frac{B-y}{A}}\right)^{-1}$$

where A and B are function parameters. Kuhn et al. used the method to analyze 746 approved drugs, and build a side-effect network with 1018 drug−drug relations, which contains 261 with no chemical similarity (Campillos et al., 2008).

Based on the framework of random walk and the assumption that similar drugs often correspond to similar targets, Yan et al. have developed a network-based random walk with restart on the heterogeneous network (NRWRH) algorithm to predict drug−target interaction (Table 18.4) (Xia et al., 2010). Different from machine learning approaches, the NRWRH algorithm utilized network analysis techniques by introducing random walk on the heterogeneous network. With information on known drug−target interactions, Yan et al. have integrated three different networks into a heterogeneous network, including target−target similarity network, drug−drug similarity network, and drug−target interaction network (Xia et al., 2010). To implement a random walk, a transition matrix $\mathbf{M}$ was calculated as:

$$\mathbf{M} = \begin{bmatrix} \mathbf{M}_{TT} & \mathbf{M}_{TD} \\ \mathbf{M}_{DT} & \mathbf{M}_{TT} \end{bmatrix}$$

where $\mathbf{M}_{TT}$ and $\mathbf{M}_{DD}$ are the probability for target-to-target and drug-to-drug transition in the random walk, respectively; $\mathbf{M}_{TD}$ and $\mathbf{M}_{DT}$ are the transition probability for target-to-drug and drug-to-target, respectively (Xia et al., 2010). Then the random walk was implemented by the following iteration equation:

$$p_{t+1} = (1 - r)M^T p_t + rp_0$$

where the probability p is iteratively calculated with the restart probability r and transition matrix $\mathbf{M}$. Yan et al. has shown that NRWRH has improved prediction performance in four classes of drug−target interactions, that is, enzymes, ion channels, GPCRs, and nuclear receptors (Xia et al., 2010).

Based on complex network theory, Tang et al. proposed three supervised inference methods to predict drug−target interaction (Table 18.4), namely drug-based similarity inference (DBSI), target-based similarity inference (TBSI), and network-based inference (NBI) (Cheng et al., 2012). For the three inference methods, different similarity score functions were defined based on chemical structure similarity, sequence similarity, or network similarity. For example, the final score $f(i)$ of drug d_i in the NBI method is obtained from:

$$f(i) = \sum_{l=1}^{m} \frac{a_{il}}{k(t_l)} \sum_{o=1}^{n} \frac{a_{ol} f_o(o)}{k(d_o)}$$

where $k(d_o)$ represents the number of targets interacting with drug d_o, and $k(t_l)$ denotes the number of drugs interacting with target t_l. NBI method showed the best performance despite it neglects chemical structure similarity (Cheng et al., 2012).

Shi et al. introduced the drug target pair as a vector of within-scores and between-scores, which utilizes features from target similarity and drug similarity

(Table 18.4) (Shi et al., 2015). By doing this, Shi et al. has created a global classifier and a uniform vector of all different types of drug−target pair (Shi et al., 2015). Besides, the unknown drug−target pair can be analyzed in the same visualization space.

Tang et al. have developed a multiple target optimal intervention (MTOI) finding algorithm which aims to identify potential drug targets and their optimal combinations restoring to a normal state (Table 18.4) (Yang et al., 2008). To implement the algorithm, ODEs and parameters for the network were obtained from the Michaelis−Menten equation and experimental data. Monte Carlo simulation was performed to achieve the desired state by optimizing an objective function F_{obj} (Yang et al., 2008). Tang et al. applied the MTOI method to understand the side-effects of traditional nonsteroidal antiinflammatory drugs in an inflammation-related network (Yang et al., 2008).

18.3.2 Tools for network construction

18.3.2.1 Cytoscape

Modeling complex biological network from a set of experimental data is crucial to understand various layers in systems biology, including biochemical reactions, gene transcription kinetics, cellular physiology, and metabolic control. Researchers have developed different computer-aided software to facilitate the management and visualization of big data from lab experiments and mathematical predictions. Cytoscape is an important tool to build a unified biological framework from high-throughput expression data and bio-molecular states (Table 18.5) (Shannon et al., 2003). The network graph is the core concept in Cytoscape, which represents molecular species and their interactions as nodes and edges, respectively. The basic functionality of Cytoscape generates a graph representation of imported biological data. By defining attributes, nodes are paired according to their names and values. Hierarchical

Table 18.5 Computational tools for network construction.

Tools	Description	Website	Reference
Cytoscape	Build biological framework from high-throughput expression data and bio-molecular states	https://go.drugbank.com/	Shannon et al. (2003)
Pajek	Efficiently analyze large network structures by storing sparse networks	https://pubchem.ncbi.nlm.nih.gov/	Batagelj et al. (2003)
Gephi	Utilizes 3D graphics engine to explore and manipulate large networks in-time	https://www.ebi.ac.uk/chembl/	Bastian et al. (2009)
NetworkX	Python package aiming to create, explore and analyze network structures	http://stitch.embl.de/	Hagberg et al. (2008)

classification is allowed by using graph annotation. Users can customize graph layout, attribute-to-visual mapping, and complete graph selection and graph-filtering with plugin functions (Shannon et al., 2003). With the help of external databases of drug−protein interaction, protein−protein interaction, protein-nucleic acid interaction, and genetic interaction, Cytoscape is powerful in modeling, analyzing, and visualizing biological networks for humans and other organisms.

18.3.2.2 Pajek

Pajek is a program to analyze large network structures efficiently (Table 18.5) (Batagelj et al., 2003). Networks in biological systems are usually large, and contains thousands of nodes and edges. The common network analysis tool is mathematically based on a matrix, which is inefficient when dealing with large graphs. Since modern computers have enough memory for storing sparse networks, Pajek proposed an alternative approach to efficiently analyze large graphs by compensating for space complexity (Batagelj et al., 2003). Data structures in Pajek are implemented as six layers, namely network, permutation, vector, cluster, partition, and hierarchy. Also, different transition methods were defined to allow data structure transformation. Theoretically, most of the algorithms in Pajek have subquadratic time complexities (Batagelj et al., 2003).

18.3.2.3 Gephi

To obtain high-quality visualization and data processing experience, a network exploration tool should develop to incorporate high flexible and scalable interactive functions. Gephi is a freely available program that uses a three-dimensional graphics engine to explore and manipulate large networks in time (Table 18.5) (Bastian et al., 2009). The three-dimensional rendering technique is based on a computer graphic card. Due to its multitask nature, Gephi can deal with large graphs with over 2000 nodes (Bastian et al., 2009). Gephi loaded network data into the workspace where each network can be managed separately. And, the function can be extended with external plugin programming. The manipulated networks can be exported as SVG or PDF files.

18.3.2.4 NetworkX

NetworkX is a Python package aiming to create, explore and analyze network structures (Table 18.5) (Hagberg et al., 2008). NetworkX can deal with arbitrary graph objects including simple graphs, directed graphs, graphs with self-loops, and parallel edges based on its basic data structure. The standard data structure in NetworkX contains edge lists, adjacency matrices, and adjacency lists (Hagberg et al., 2008). Since the computation storage and speed depends on the choice of data structure, NetworkX uses adjacent lists for real-world networks with sparse nature. Search and update algorithms for adjacent lists can be achieved through dictionary data structure in Python (Hagberg et al., 2008). Once a graph object is created in NetworkX, users can analyze the network through standard algorithms, such as

degree distribution, clustering coefficient, shortest path computing, and spectral measures. NetworkX allows graph visualization through its hooks into Matplotlib. In application, NetworkX has been used to perform spectral analysis of network dynamics and to investigate the synchronization of oscillators. The installation of NetworkX is easy, which requires NumPy, SciPy and Matplotlib installed in prior (Hagberg et al., 2008).

18.3.3 Network topological analysis

18.3.3.1 Degree distribution

The degree of a node is the number of edges linking to the node. It has shown in many biological networks are scale-free, which means the degree distribution of a network follows a power-law $k^{-\lambda}$, where λ is the degree exponent. In a scale-free network, the distribution of degrees is not evenly distributed. Cohen et al. showed a scale-free network is very robust to random attacks (Cohen et al., 2000). Therefore, the proteins with a high degree, also named hubs, evolve slowly and are crucial for the cell's survival (Cheng et al., 2014; Eisenberg and Levanon, 2003; Hahn and Kern, 2004; He and Zhang, 2006; Jeong et al., 2001).

18.3.3.2 Path and distance

The shortest path for a pair of nodes is defined as the shortest length linking the two nodes out of all possible path lengths. Analysis of the shortest path is important to investigate regulatory pathways in protein−protein interaction networks through direction assignment (Blokh et al., 2013; Silverbush and Sharan, 2014). By using the concept of path and distance, it is possible to evaluate the proximity in the drug−target network. Guney et al. have proposed different distance measurement methods (including the closest, shortest, kernel, center, and separation distances) to analyze the therapeutic effect of drugs (Guney et al., 2016). They have investigated 238 drugs used in 78 diseases and found that the therapeutic effect is localized in the neighborhood of a small network (Guney et al., 2016). Guney et al. have shown the network-based distance analysis are useful in drug repurposing and adverse effect detection (Guney et al., 2016).

Another important measure based on the shortest path is efficiency, which is defined as

$$E(G) = \frac{1}{N(N-1)} \sum_{i \neq j \in G} \frac{1}{d_{ij}}$$

where N is the number of nodes, and d_{ij} is the shortest path for nodes i and j (Latora and Marchiori, 2001). Efficiency measures the traffic capacity of a network and how efficiently it exchanges information (Latora and Marchiori, 2001). Csermely et al. found it is possible to efficiently inhibit targets through a small number of inhibitors instead of a complete inhibition of a single target (Csermely

et al., 2005). In addition, the concept of network efficiency rationalizes multitarget strategy in drug design, which is useful in the development of drug combinations (Cheng et al., 2019; Csermely et al., 2005; Vazquez, 2009).

Based on the network shortest path, it is possible to measure the importance of a node, which is characterized through betweenness in the following equation:

$$B(v) = \sum_{i \neq j} \frac{\delta_{ij}(v)}{\delta_{ij}}$$

where δ_{ij} is the number of the shortest paths from i to j, $\delta_{ij}(v)$ is the number of the shortest paths that travel through node v. It should be noted the degree and betweenness of a node is not correlated, which means nodes with a small degree could have large betweenness (Guimerà et al., 2005; Joy et al., 2005; Yu et al., 2007). As mentioned, the betweenness characterizes the importance of a node. A node with high betweenness is known as a bottleneck in a network. Bottlenecks control the flow of information in a network and improve network efficiency. It has shown that proteins with high betweenness are essential and tend to be highly pleiotropic (Ahmed et al., 2018; Estrada and Ross, 2018; Zou et al., 2008).

18.3.3.3 Module and motifs

In complex networks, a dense subgraph or subnetwork is referred to as a module. The modularity of a network is defined as (Clauset et al., 2004; Newman and Girvan, 2004; Newman, 2012):

$$M = \frac{1}{2E} \sum [A_{ij} - P_{ij}] \delta_{C_i, C_j}$$

where E is the number of network edges, A is an adjacent matrix, P is the expected number of edges from node i to j, δ_{ij} is a Kronecker function which equals 1 only if node i and j belongs to the same community. A number of methods have been developed to identify modules and communities in a network (Ahn et al., 2010; Palla et al., 2005; Palla et al., 2007; Rosvall and Bergstrom, 2007; Rosvall and Bergstrom, 2008). Module analysis provided an effective approach to investigating complex networks by identifying specific modules instead of unfolding the entire network. Increasing results showed that modules are important in uncovering new drug targets and promoting drug development (Derry et al., 2012).

Motifs in a network are defined as connection patterns with a high occurring number in a network than in randomized networks (Alon, 2007; Milo et al., 2002; Shen-Orr et al., 2002). Universal classes of networks can be defined through motifs. Some of the basic motifs, including 13 3-node directed motifs and 30 undirected motifs (also named graphlets) with node numbers ranging from 2 to 5, are shown in Fig. 18.4 (Milo et al., 2002; Pržulj, 2007). Analysis of network motifs is useful to identify druggable targets. Tan et al. used network motifs analysis to uncover basic principles of cellular target druggability, which describes the capacity of a target

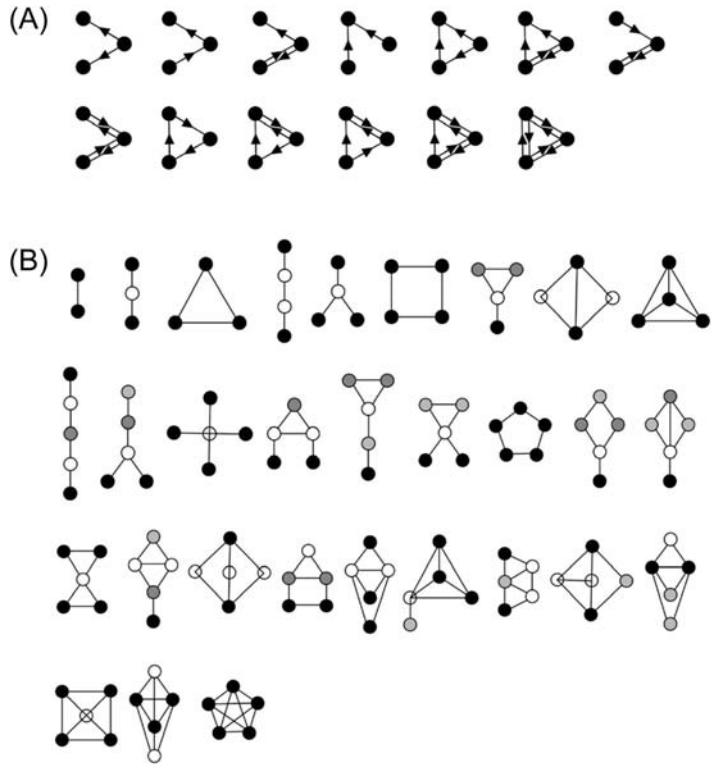

Figure 18.4 An illustration of (A) 13 3-node directed motifs and (B) 30 graphlets with node numbers ranging from 2 to 5.

modulated by a drug (Wu et al., 2016). They found that highly druggable motifs share a consensus topology of a negative feedback loop without any positive feedback loops (Wu et al., 2016). On the opposite, the motifs of low druggability consist of multiple positive direct regulations and positive feedback loops. In addition, Tan et al. showed druggability can be reduced by adding direct regulations to a drug−target network (Wu et al., 2016).

18.4 Conclusion and perspectives

The paradigm has shifted from one-bullet-one-target to a network view in drug discovery and development. As demonstrated by Yildirim et al., the complex network nature of drug−target interaction imposes a holistic philosophy in drug discovery and drug repurposing in the next decades (Yildirim et al., 2007). As high-throughput experimental data is expanding rapidly and dramatically, novel databases and data management methodologies are emerging, especially when considering chemicals derived from complicated herbal plants and their related targets. A uniform data

format or data transformation platform will facilitate data utilization more efficiently, which is also fundamental to the construction, prediction, and analysis of drug−target interaction networks.

Various computational algorithms have been proposed for drug−target interaction prediction and analysis. Through the framework of network science, it is possible to reconstruct drug−target interaction networks without concerning the three-dimensional structures of drugs and targets. These methods showed high performance in accuracy and speed, which are important in real applications including target prediction and mechanism elucidation. However, these methods still need to be further completed, including drug−target interaction prediction method without a prior knowledge of ligand, as well as the development of quantitative methods for the analysis of drug−target interaction. Nevertheless, the coupling of big data and network science in drug−target interaction has opened a new era in drug discovery and development.

Acknowledgments

This research was funded by the National Natural Science Foundation of China (31872723), the Priority Academic Program Development (PAPD) of Jiangsu Higher Education Institutions, Zunyi Science and Technology Project (2018(21)), and Guizhou Provincial Natural Science Foundation (QKH-J[2020]1Y045).

References

Ahmed, H., Howton, T.C., Sun, Y., Weinberger, N., Belkhadir, Y., Mukhtar, M.S., 2018. Network biology discovers pathogen contact points in host protein-protein interactomes. Nat. Commun. 9, 2312.

Ahn, Y.-Y., Bagrow, J.P., Lehmann, S., 2010. Link communities reveal multiscale complexity in networks. Nature 466, 761−764.

Alon, U., 2007. Network motifs: theory and experimental approaches. Nat. Rev. Genet. 8, 450−461.

Basak, S.C., Gute, B.D., Mills, D., Hawkins, D.M., 2002. Quantitative molecular similarity methods in the property/toxicity estimation of chemicals: a comparison of arbitary versus tailored similarity spaces. J. Mol. Struct.: Theochem 622, 127−145.

Basak, S.C., Gute, B.D., Mills, D., 2006. Similarity methods in analog selection, property estimation and clustering of diverse chemicals. Arch. Org. Chem. 9, 157−210.

Bastian, M., Heymann, S., Jacomy, M. 2009. Gephi: an open source software for exploring and manipulating networks. In: International AAAI Conference on Weblogs and Social Media.

Batagelj, V., Andrej, M., Jünger, M.M.P., 2003. Pajek - analysis and visualization of large networks. Graph drawing software. Springer, Berlin.

Bleakley, K., Yamanishi, Y., 2009. Supervised prediction of drug−target interactions using bipartite local models. Bioinformatics 25, 2397−2403.

Blokh, D., Segev, D., Sharan, R., 2013. The approximability of shortest path-based graph orientations of protein−protein interaction networks. J. Comput. Biol. 20, 945−957.

Bolognesi, M.L., Cavalli, A., 2016. Multitarget drug discovery and polypharmacology. ChemMedChem 11, 1190−1192.

Borgatti, S.P., Halgin, D.S., 2011. On network theory. Organ. Sci. 22, 1168−1181.

Burley, S.K., Berman, H.M., Bhikadiya, C., Bi, C., Chen, L., Di Costanzo, L., et al., 2018. RCSB protein data bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. Nucleic Acids Res. 47, D464−D474.

Campillos, M., Kuhn, M., Gavin, A.-C., Jensen, L.J., Bork, P., 2008. Drug target identification using side-effect similarity. Science 321, 263.

Cao, D.-S., Zhang, L.-X., Tan, G.-S., Xiang, Z., Zeng, W.-B., Xu, Q.-S., et al., 2014. Computational prediction of drug-target interactions using chemical, biological, and network features. Mol. Inf. 33, 669−681.

Chaudhari, R., Tan, Z., Huang, B., Zhang, S., 2017. Computational polypharmacology: a new paradigm for drug discovery. Expert. Opin. Drug. Dis. 12 (3), 279−291.

Chen, C.Y.-C., 2011. TCM database@taiwan: the world's largest traditional chinese medicine database for drug screening in silico. PLoS One 6, e15939.

Chen, L., Zeng, W.-M., 2013. A two-step similarity-based method for prediction of drug's target group. Protein Pept. Lett. 20, 364−370.

Cheng, A.C., Coleman, R.G., Smyth, K.T., Cao, Q., Soulard, P., Caffrey, D.R., et al., 2007. Structure-based maximal affinity model predicts small-molecule druggability. Nat. Biotechnol. 25, 71−75.

Cheng, F., Liu, C., Jiang, J., Lu, W., Li, W., Liu, G., et al., 2012. Prediction of drug-target interactions and drug repositioning via network-based inference. PLoS Comput. Biol. 8, e1002503.

Cheng, F., Jia, P., Wang, Q., Lin, C.-C., Li, W.-H., Zhao, Z., 2014. Studying tumorigenesis through network evolution and somatic mutational perturbations in the cancer interactome. Mol. Biol. Evol. 31, 2156−2169.

Cheng, F., Kovács, I.A., Barabási, A.-L., 2019. Network-based prediction of drug combinations. Nat. Commun. 10, 1197.

Cheung, F., 2011. TCM: made in china. Nature 480, S82−S83.

Ciallella, H.L., Zhu, H., 2019. Advancing computational toxicology in the big data era by artificial intelligence: data-driven and mechanism-driven modeling for chemical toxicity. Chem. Res. Toxicol. 32, 536−547.

Clauset, A., Newman, M.E.J., Moore, C., 2004. Finding community structure in very large networks. Phys. Rev. E 70, 066111.

Cohen, R., Erez, K., ben-Avraham, D., Havlin, S., 2000. Resilience of the internet to random breakdowns. Phys. Rev. Lett. 85, 4626−4628.

Consortium, T.U., 2018. Uniprot: a worldwide hub of protein knowledge. Nucleic Acids Res. 47, D506−D515.

Csermely, P., Ágoston, V., Pongor, S., 2005. The efficiency of multi-target drugs: the network approach might help drug design. Trends Pharmacol. Sci. 26, 178−182.

Derry, J.M.J., Mangravite, L.M., Suver, C., Furia, M.D., Henderson, D., Schildwachter, X., et al., 2012. Developing predictive molecular maps of human disease through community-based modeling. Nat. Genet. 44, 127−130.

Eisenberg, E., Levanon, E.Y., 2003. Preferential attachment in the protein network evolution. Phys. Rev. Lett. 91, 138701.

Estrada, E., Ross, G.J., 2018. Centralities in simplicial complexes. Applications to protein interaction networks. J. Theor. Biol. 438, 46–60.
French, K.E., Harvey, J., McCullagh, J.S.O., 2018. Targeted and untargeted metabolic profiling of wild grassland plants identifies antibiotic and anthelmintic compounds targeting pathogen physiology, metabolism and reproduction. Sci. Rep. 8, 1695.
Gilson, M.K., Liu, T., Baitaluk, M., Nicola, G., Hwang, L., Chong, J., 2015. Bindingdb in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. Nucleic Acids Res. 44, D1045–D1053.
Guimerà, R., Mossa, S., Turtschi, A., Amaral, L.A.N., 2005. The worldwide air transportation network: anomalous centrality, community structure, and cities' global roles. Proc. Natl. Acad. Sci. 102, 7794.
Guney, E., Menche, J., Vidal, M., Barábasi, A.-L., 2016. Network-based in silico drug efficacy screening. Nat. Commun. 7, 10331.
Günther, S., Kuhn, M., Dunkel, M., Campillos, M., Senger, C., Petsalaki, E., et al., 2008. Supertarget and matador: resources for exploring drug-target relationships. Nucleic Acids Res. 36, D919–D922.
Hagberg, A., Swart, P., Chult, D. 2008. Exploring network structure, dynamics, and function using networkx. In: Proceedings of the 7th Python in Science Conference (SciPy 2008); Pasadena, CA, USA.
Haggarty, S.J., Koeller, K.M., Wong, J.C., Butcher, R.A., Schreiber, S.L., 2003. Multidimensional chemical genetic analysis of diversity-oriented synthesis-derived deacetylase inhibitors using cell-based assays. Chem. Biol. 10, 383–396.
Hahn, M.W., Kern, A.D., 2004. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. Mol. Biol. Evol. 22, 803–806.
He, X., Zhang, J., 2006. Why do hubs tend to be essential in protein networks? PLoS Genet. 2, e88.
Hopkins, A.L., 2008. Network pharmacology: the next paradigm in drug discovery. Nat. Chem. Biol. 4, 682–690.
Huang, L., Xie, D., Yu, Y., Liu, H., Shi, Y., Shi, T., et al., 2017. Tcmid 2.0: a comprehensive resource for tcm. Nucleic Acids Res. 46, D1117–D1120.
Jeong, H., Mason, S.P., Barabási, A.L., Oltvai, Z.N., 2001. Lethality and centrality in protein networks. Nature 411, 41–42.
Joy, M.P., Brock, A., Ingber, D.E., Huang, S., 2005. High-betweenness proteins in the yeast protein interaction network. J. Biomed. Biotechnol. 2005, 594674.
Keiser, M.J., Setola, V., Irwin, J.J., Laggner, C., Abbas, A.I., Hufeisen, S.J., et al., 2009. Predicting new molecular targets for known drugs. Nature 462, 175–181.
Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., et al., 2018. Pubchem 2019 update: improved access to chemical data. Nucleic Acids Res. 47, D1102–D1109.
Kuruvilla, F.G., Shamji, A.F., Sternson, S.M., Hergenrother, P.J., Schreiber, S.L., 2002. Dissecting glucose signalling with diversity-oriented synthesis and small-molecule microarrays. Nature 416, 653–657.
Latora, V., Marchiori, M., 2001. Efficient behavior of small-world networks. Phys. Rev. Lett. 87, 198701.
Lee, C.H., Yoon, H.-J., 2017. Medical big data: promise and challenges. Kidney Res. Clin. Pract. 36, 3–11.
Li, S., Zhang, B., Zhang, N., 2011. Network target for screening synergistic drug combinations with application to traditional chinese medicine. BMC Syst. Biol. 5, S10.
Liang, Z., Hu, G., 2016. Protein structure network-based drug design. Mini-Rev. Med. Chem. 16, 1330–1343.

Loscalzo, J., Barabasi, A.-L., 2011. Systems biology and the future of medicine. WIRES Syst. Biol. Med. 3, 619–627.
Mei, J.P., Kwoh, C.K., Yang, P., Li, X.L., Zheng, J., 2013. Drug–target interaction prediction by learning from local information and neighbors. Bioinformatics 29, 238–245.
Mendez, D., Gaulton, A., Bento, A.P., Chambers, J., De Veij, M., Félix, E., et al., 2018. Chembl: towards direct deposition of bioassay data. Nucleic Acids Res. 47, D930–D940.
Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., Alon, U., 2002. Network motifs: Simple building blocks of complex networks. Science 298, 824.
Mohs, R.C., Greig, N.H., 2017. Drug discovery and development: role of basic biological research. Alzheimers Dement. 3 (4), 651–657.
Morris, G.M., Huey, R., Lindstrom, W., Sanner, M.F., Belew, R.K., Goodsell, D.S., et al., 2009. Autodock4 and autodocktools4: automated docking with selective receptor flexibility. J. Comput. Chem. 30, 2785–2791.
Mullard, A., 2020. $1.3 billion per drug? Nat. Rev. Drug. Discov. 19, 226.
Newman, M.E.J., 2003. The structure and function of complex networks. SIAM Rev. 45, 167–256.
Newman, M.E.J., Girvan, M., 2004. Finding and evaluating community structure in networks. Phys. Rev. E 69, 026113.
Newman, M.E.J., 2012. Communities, modules and large-scale structure in networks. Nat. Phys. 8, 25–31.
Palla, G., Derényi, I., Farkas, I., Vicsek, T., 2005. Uncovering the overlapping community structure of complex networks in nature and society. Nature 435, 814–818.
Palla, G., Barabási, A.-L., Vicsek, T., 2007. Quantifying social group evolution. Nature 446, 664–667.
Parkhe, A., Wasserman, S., Ralston, D.A., 2006. New frontiers in network theory development. Acad. Manage. Rev. 31, 560–568.
Pence, H.E., Williams, A., 2010. Chemspider: an online chemical information resource. J. Chem. Edu. 87, 1123–1124.
Pržulj, N., 2007. Biological network comparison using graphlet degree distribution. Bioinformatics 23, e177–e183.
Rosvall, M., Bergstrom, C.T., 2007. An information-theoretic framework for resolving community structure in complex networks. Proc. Natl. Acad. Sci. 104, 7327.
Rosvall, M., Bergstrom, C.T., 2008. Maps of random walks on complex networks reveal community structure. Proc. Natl. Acad. Sci. 105, 1118.
Ru, J., Li, P., Wang, J., Zhou, W., Li, B., Huang, C., et al., 2014. Tcmsp: a database of systems pharmacology for drug discovery from herbal medicines. J. Cheminform 6, 13.
Russell, C., Rahman, A., Mohammed, A.R., 2013. Application of genomics, proteomics and metabolomics in drug discovery, development and clinic. Ther. Deliv. 4, 395–413.
Santos, R., Ursu, O., Gaulton, A., Bento, A.P., Donadi, R.S., Bologa, C.G., et al., 2017. A comprehensive map of molecular drug targets. Nat. Rev. Drug. Discov. 16, 19–34.
Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., et al., 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 13, 2498–2504.
Shen-Orr, S.S., Milo, R., Mangan, S., Alon, U., 2002. Network motifs in the transcriptional regulation network of *Escherichia coli*. Nat. Genet. 31, 64–68.
Shi, J.-Y., Liu, Z., Yu, H., Li, Y.-J., 2015. Predicting drug-target interactions via within-score and between-score. BioMed. Res. Int. 2015, 350983.

Silverbush, D., Sharan, R., 2014. Network orientation via shortest paths. Bioinformatics 30, 1449–1455.
Szklarczyk, D., Santos, A., von Mering, C., Jensen, L.J., Bork, P., Kuhn, M., 2015. Stitch 5: Augmenting protein–chemical interaction networks with tissue and affinity data. Nucleic Acids Res. 44, D380–D384.
Szklarczyk, D., Gable, A.L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., et al., 2018. String v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 47, D607–D613.
Tao, W., Xu, X., Wang, X., Li, B., Wang, Y., Li, Y., et al., 2013. Network pharmacology-based prediction of the active ingredients and potential targets of chinese herbal radix curcumae formula for application to cardiovascular disease. J. Ethnopharmacol. 145, 1–10.
Vazquez, A., 2009. Optimal drug combinations and minimal hitting sets. BMC Syst. Biol. 3, 81.
Wang, Y., Zeng, J., 2013. Predicting drug-target interactions using restricted Boltzmann machines. Bioinformatics 29, i126–i134.
Wang, J.T., Liu, W., Tang, H., Xie, H., 2014. Screening drug target proteins based on sequence information. J. Biomed. Inf. 49, 269–274.
Wang, J., Zhang, C.-J., Chia, W.N., Loh, C.C.Y., Li, Z., Lee, Y.M., et al., 2015. Haem-activated promiscuous targeting of artemisinin in *Plasmodium falciparum*. Nat. Commun. 6, 10111.
Wang, F., Han, S., Yang, J., Yan, W., Hu, G., 2021. Knowledge-guided "community network" analysis reveals the functional modules and candidate targets in non-small-cell lung cancer. Cells 2021 (10), 402.
Whitebread, S., Hamon, J., Bojanic, D., Urban, L., 2005. Keynote review: In vitro safety pharmacology profiling: an essential tool for successful drug development. Drug. Discov. Today 10, 1421–1433.
Wishart, D.S., Feunang, Y.D., Guo, A.C., Lo, E.J., Marcu, A., Grant, J.R., et al., 2017. Drugbank 5.0: a major update to the drugbank database for 2018. Nucleic Acids Res. 46, D1074–D1082.
Wouters, O.J., McKee, M., Luyten, J., 2020. Estimated research and development investment needed to bring a new medicine to market, 2009–2018. JAMA 323 (9), 844–853.
Wu, F., Ma, C., Tan, C., 2016. Network motifs modulate druggability of cellular targets. Sci. Rep. 6, 36626.
Xia, Z., Wu, L.-Y., Zhou, X., Wong, S.T.C., 2010. Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces. BMC Syst. Biol. 4, S6.
Yamanishi, Y., Araki, M., Gutteridge, A., Honda, W., Kanehisa, M., 2008. Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. Bioinformatics 24, i232–i240.
Yamanishi, Y., Kotera, M., Kanehisa, M., Goto, S., 2010. Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. Bioinformatics 26, i246–i254.
Yan, W., Zhang, D., Shen, C., Liang, Z., Hu, G., 2018. Recent advances on the network models in target-based drug discovery. Curr. Top. Med. Chem. 18, 1031–1043.
Yang, K., Bai, H., Ouyang, Q., Lai, L., Tang, C., 2008. Finding multiple target optimal intervention in disease-related molecular network. Mol. Syst. Biol. 4, 228.
Ye, H., Ye, L., Kang, H., Zhang, D., Tao, L., Tang, K., et al., 2010. Hit: linking herbal active ingredients to targets. Nucleic Acids Res. 39, D1055–D1059.

Yildirim, M.A., Goh, K.I., Cusick, M.E., Barabasi, A.L., Vidal, M., 2007. Drug-target network. Nat. Biotechnol. 25, 1119−1126.

Yu, H., Kim, P.M., Sprecher, E., Trifonov, V., Gerstein, M., 2007. The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. PLoS Comput. Biol. 3, e59.

Zhu, H., 2020. Big data and artificial intelligence modeling for drug discovery. Annu. Rev. Pharmacol. Toxicol. 60, 573−589.

Zou, L., Sriswasdi, S., Ross, B., Missiuro, P.V., Liu, J., Ge, H., 2008. Systematic analysis of pleiotropy in C. elegans early embryogenesis. PLoS Comput. Biol. 4, e1000003.