OXFORD

Briefings in Bioinformatics, 17(4), 2016, 696–712

doi: 10.1093/bib/bbv066
Advance Access Publication Date: 17 August 2015
Paper

Drug–target interaction prediction: databases, web servers and computational models

Xing Chen,* Chenggang Clarence Yan,* Xiaotian Zhang, Xu Zhang, Feng Dai, Jian Yin and Yongdong Zhang

Corresponding author: Xing Chen, National Center for Mathematics and Interdisciplinary Sciences, Chinese Academy of Sciences, Zhongguancun East Road, Haidian District, Beijing 100190, China. E-mail: xingchen@amss.ac.cn
*The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

Abstract

Identification of drug–target interactions is an important process in drug discovery. Although high-throughput screening and other biological assays are becoming available, experimental methods for drug–target interaction identification remain to be extremely costly, time-consuming and challenging even nowadays. Therefore, various computational models have been developed to predict potential drug–target associations on a large scale. In this review, databases and web servers involved in drug–target identification and drug discovery are summarized. In addition, we mainly introduced some state-of-the-art computational models for drug–target interactions prediction, including network-based method, machine learning-based method and so on. Specially, for the machine learning-based method, much attention was paid to supervised and semi-supervised models, which have essential difference in the adoption of negative samples. Although significant improvements for drug–target interaction prediction have been obtained by many effective computational models, both network-based and machine learning-based methods have their disadvantages, respectively. Furthermore, we discuss the future directions of the network-based drug discovery and network approach for personalized drug discovery based on personalized medicine, genome sequencing, tumor clone-based network and cancer hallmark-based network. Finally, we discussed the new evaluation validation framework and the formulation of drug–target interactions prediction problem by more realistic regression formulation based on quantitative bioactivity data.

Key words: drug–target interactions prediction; drug discovery; computational models; biological networks; machine learning

Drug discovery

Despite many advances in the past decades, drug discovery is still a costly and inefficient process [1–3], with costs for each new molecular entity (NME) estimated at $1.8 billion [2]. Furthermore, new drugs often take nearly a decade to reach market; for instance, only approximately 20 new drugs have

Xing Chen, PhD, is an assistant research fellow of the National Center for Mathematics and Interdisciplinary Sciences, Chinese Academy of Sciences. His research interests include network pharmacology, disease and non-coding RNAs and machine learning.
Chenggang Clarence Yan, PhD, is an assistant research fellow of the Department of Automation, Tsinghua University. His research interests include network pharmacology, parallel computing, video coding and image processing.
Xiaotian Zhang is a student of the School of Mechanical, Electrical & Information Engineering, Shandong University. Her research interests include network pharmacology, disease and non-coding RNAs and machine learning.
Xu Zhang is a student of the School of Mechanical, Electrical & Information Engineering, Shandong University. Her research interests include network pharmacology, disease and non-coding RNAs and machine learning.
Feng Dai, PhD, is an associate research fellow of the Institute of Computing Technology, Chinese Academy of Sciences. His research interests include network pharmacology, video coding and parallel computing.
Jian Yin, PhD, is an associate professor of the Department of Computer, Shandong University. His research interests include network pharmacology, parallel computing and image processing.
Yongdong Zhang, PhD, is a research fellow of the Institute of Computing Technology, Chinese Academy of Sciences. His research interests include network pharmacology, video coding, parallel computing and image processing.
Submitted: 21 May 2015; **Received (in revised form):** 16 July 2015

been approved by US Food and Drug Administration (FDA) as NMEs each year [4]. In recent years, the rate of successful drug development has decreased. In this light, new uses for existing or abandoned drugs are urgent [5]. Such a new strategy is called drug repositioning or drug repurposing [6].

Traditional drug discovery primarily followed the 'one molecule-one target-one disease' paradigm, which aimed to seek the most specific drugs to act on individual targets for individual diseases [7]. In this approach, a specific protein is studied *in vitro*, in cells and in whole organisms, and evaluated as a drug target for a specific therapeutic indication. This historical paradigm has resulted in the identification of some effective chemical molecules that affect specific proteins. However, the major limiting factor of this traditional drug discovery paradigm is that pharmaceuticals are designed to target individual factors in a disease system, but complex diseases are multifactorial in nature and vulnerable at multiple attacks. The disease symptom is a progression accumulation of mutations and interventions of different genes and pathways. Multiple stages along the disease pathway may need to be manipulated simultaneously for an effective treatment of the diseases. Because this traditional paradigm ignores the complex interactions between drugs and their target proteins and the important fact that many complex diseases tend to be associated with multiple target proteins, this paradigm has not accelerated the new drug discovery rate as expected [8–10]. Recently, there is an increasingly accepted concept 'polypharmacology', i.e. drugs often work by targeting not a single target protein, but multiple ones [11, 12]. In addition, multiple targets are often involved in the same disease [9]. Therefore, to increase the drug efficacy and overcome drug resistance and toxicity, much attention has been paid to multiple-target drug development and drug combination research [10, 13–15]. Because of the 'polypharmacological' property of a drug, 'off-target' is an unintended occurring activity for a drug. On one hand, the off-target activities might result in some undesired side effects [16]. On the other hand, they can also occasionally be beneficial for some new or unexpected therapeutic effects for old drugs [16]. Such polypharmacological features could help us find new uses of drugs, namely drug repositioning, which has been mentioned above [17].

Nowadays, a critical phase to accelerate the progression of drug discovery is to confirm whether a drug could interact with a target [18]. Drug discovery requires the accurate identification of the complex interactions between drugs and a wide variety of protein targets. All of these highlight the critical role of the identification of drug–target interactions in drug discovery.

Drug–target interactions

The majority of drug targets are cellular proteins, which aim to treat or diagnose a disease by selectively interacting with chemical compounds [19]. Current studies have shown that classical therapeutic drug targets contain ~130 protein families [20, 21], such as enzymes, G-protein-coupled receptors (GPCRs), ion channels and transporters, nuclear hormone receptors [21, 22]. Many efforts have been made to estimate the total number of drug targets [19, 20, 23, 24]. There are estimated about 6000–8000 targets in the human genome that have pharmacological interest, but only a small part of these targets have been involved in approved drugs so far [20, 22, 25]. A large number of putative drug targets remains to be validated.

From the view point of drug, although it is estimated that PubChem database contains 35 million compounds, only <7000 compounds have the information of their corresponding target proteins [26, 27]. Furthermore, it is estimated that the set of all possible small molecules has already consisted of >10 60 compounds [28]. Among these drugs with corresponding target proteins, most of them are small chemical compounds, which interact with an appropriate target protein involved in a disease of interest and inhibit or activate the biological behavior of the target proteins. Besides the selective targets, drugs may also interact with additional proteins, which are not their primary therapeutic targets, i.e. off-target effects.

Correct identification and validation of drug–target interactions is the first step on drug discovery pipeline. Until now, there are many potential drug–target interactions that have not been discovered [29]. The identification of novel drugs and their targets is still an extremely difficult goal owing to the relatively limited knowledge about the complex relationship between chemical space and genomic space [28, 30, 31]. There are many factors that affect the establishment of the interactions between a drug and its targets, such as various chemical bonds that are related to the affinity of the drug for its targets [25]. However, a number of factors make the identification of drug–target interactions more urgent than ever before. Firstly, although over the past decade, a growing number of compounds were synthesized, their drug effects and target proteins are still unclear [32]. Secondly, there are still a variety of diseases that cannot be cured and many new diseases emerge every year [33, 34]. Finally, large-scale data sets on various properties of compounds [35], features of target proteins [36] and responses in the human physiological system [37] have been collected by researchers. However, these high-dimensional data sets present great challenges to researchers owing to their high dimensionality, complex structure and distinct types [38]. Considering the existence of multiple drugs and various target proteins and complicated associations between them, experimental verification of drug–target associations remains to be time-consuming and expensive and limited to small-scale research even nowadays [39, 40]. Therefore, there is urgent need for appropriate and powerful computational prediction methods that could detect the complex drug–target associations effectively on a large scale. Computational drug–target interaction identification could benefit both better understanding of complex biological interactions and important biological processes and the acceleration of novel drug discovery and human medical improvement. Especially, predicting potential drug–target interactions from heterogeneous biological data has been the hot topic of computational biology, which could provide new potential drug–target interaction candidates for biological experimental validation and decrease the time and cost of biological experiments [10].

Databases and web servers

DrugBank

(http://www.drugbank.ca) [41]

The DrugBank database is a richly annotated bioinformatics and cheminformatics resource that combines detailed drug data (e.g. chemical, pharmacological and pharmaceutical) with comprehensive target information (e.g. sequence, structure and pathway). The database is updated frequently. So far, it has contained 7759 drug entities and 15 199 drug–target interactions (see Table 1 for the statistics of the number of drugs, target proteins and drug–target interactions in some of the databases covered in this review. Some databases do not provide these statistics in their databases and published paper.).

Table 1. The statistics of the number of drugs, target proteins and drug–target interactions in some of the databases covered in this review

Databases	The number of compound/ligand–target interactions	The number of compounds or ligands	The number of targets
DrugBank	15 199	7759	4104
TTD		20 667	2360
SuperTarget	332 828	195 770	6219
MATADOR		775	
STITCH	367 000	390 000	3 600 000
TDR Targets		968	448
PDTD			841
ChEMBL		1 463 270	10 774
SIDER		996	
ChemBank		1 700 000	
The IUPHAR/BPS Guide to PHARMACOLOGY	12 829	7586	2726
CancerDR		148	116
BindingDB	1 132 739	489 416	7020
DCDB		904	805
ASDCD	1225	105	

TTD: Therapeutic target database

(http://bidd.nus.edu.sg/group/ttd/ttd.asp) [42]

Therapeutic Target Database (TTD) provides the information about known and explored therapeutic protein and nucleic acid targets, the targeted diseases, pathway information and corresponding drugs directed at each of these targets. Knowledge of these targets and corresponding drugs, especially those in clinical uses and trials, is highly useful for accelerating drug discovery. Recently, the information of 1755 biomarkers for 365 disease conditions and 210 drug scaffolds for 714 drugs and leads has been further added into this database.

SuperTarget

(http://bioinf-apache.charite.de/supertarget_v2/) [43]

SuperTarget is an extensive database for analyzing 332 828 drug–target interactions. This database allows querying by drugs, targets, drug-target-related pathways, drug-target-related ontologies and cytochromes P450s.

MATADOR

(http://matador.embl.de) [44]

Manually Annotated Targets and Drugs Online Resource (MATADOR) is a database resource for protein–chemical interactions, including multiple direct and indirect modes of drug–target interactions. The manually annotated list of direct (binding) and indirect interactions between proteins and chemicals was assembled by automated text mining followed by manual collection. It allows searching by drugs or target proteins.

STITCH

(http://stitch.embl.de/) [45]

STITCH is a database of known and predicted chemical–protein interactions, which integrates the evidence derived from experiments, other databases and literatures. Compared with the previous version, recently, the number of high-confidence chemical–protein interactions in human has increased by 45% in the latest version of STITCH.

TDR targets

(http://tdrtargets.org/) [46]

The TDR Targets Database is a chemogenomics resource for neglected tropical diseases, which is aimed at facilitating the identification and prioritization of drugs and drug targets in neglected disease pathogens. The database includes pathogen genomic information with functional data (e.g. expression, phylogeny and essentiality) for genes, the addition of new genomes and integration of chemical structure, property and bioactivity information for biological ligands, drugs and inhibitors.

PDTD

(http://www.dddc.ac.cn/pdtd/) [47]

PDTD (Potential Drug Target Database) is a dual-function database, which integrates an informatics database and a structural database of known and potential drug targets. The database focuses on those drug targets with known 3D structures, and the drug targets in this database were categorized into 15 and 13 types according to the criteria of therapeutic areas and biochemical criteria.

ChEMBL

(https://www.ebi.ac.uk/chembldb) [48]

ChEMBL contains binding, functional and ADMET (i.e. assessment of *in vivo* absorption, distribution, metabolism, excretion and toxicity properties) information for a larger number of drug-like bioactive compounds. These data are manually collected from the published literature on a regular basis. Currently, the database contains 5.4 million bioactivity measurements, which are useful for drug discovery.

Integrity

(http://integrity.thomson-pharma.com) [49]

This database contains a large number of drugs that are annotated with their corresponding drug targets, associated diseases and the information on clinical phases of the drugs.

FAERS

(http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/default.htm)

The FDA Adverse Event Reporting System (FAERS) is a database that contains the information obtained from adverse event and medication error reports submitted to FDA on side effect keywords (adverse event keywords) for drugs.

SIDER

(http://sideeffects.embl.de/) [50]

A public and computer-readable database that contains information on marketed medicines and their recorded side effects (i.e. adverse drug reactions), including side-effect frequency, drug and side-effect classifications as well as links to further information, such as drug–target associations.

JAPIC

(http://www.japic.or.jp/)

Japan Pharmaceutical Information Center (JAPIC) database manages all package-insert information of pharmaceutical products in Japan, which contains side effects information for drugs (pharmaceutical molecules).

ChemBank

(http://chembank.broadinstitute.org/) [51]

A database contains freely available collection of data derived from small molecules and small-molecule screens, and resources for studying their properties so that biological and medical insights can be gained. ChemBank is unique among small-molecule databases in the following three ways: its dedication to the storage of raw screening data, its rigorous definition of screening experiments in terms of statistical hypothesis testing and its hierarchical metadata-based organization of related assays into screening projects.

The IUPHAR/BPS Guide to PHARMACOLOGY

(http://www.guidetopharmacology.org/) [52]

The IUPHAR/BPS Guide to PHARMACOLOGY is an open knowledgebase that provides the information of approved targets and experimental drugs. Specifically, the data of pharmacological, chemical, genetic, functional and pathophysiological are included in this database.

CancerDR

(http://crdd.osdd.net/raghava/cancerdr/) [53]

The CancerDR provides comprehensive information of 148 anti-cancer drugs, and their pharmacological profiling across 952 cancer cell lines. Comprehensive information of all the 116 drug targets has been provided, such as 1356 unique mutations in the cancer cell lines and the information of gene ontology, pathways, phylogeny about the drug targets. CancerDR makes full use of the information of mutations in drug targets to provide effective personalized cancer therapies and will be useful for identification of genes encoding drug targets based on genetic alterations as well as the residual resistance.

BindingDB

(http://www.bindingdb.org/bind) [54]

The BindingDB is a binding database, which holds 1 132 739 experimentally determined protein–ligand binding affinities among 489 416 small molecule ligands and 7020 protein targets. It has become one of the most extensive public databases of protein–ligand binding affinities.

ZINC

(http://zinc.docking.org) [55]

ZINC is the largest database for ligand discovery, which is especially important for those investigators seeking chemical matter for their biological targets. ZINC contains >20 million commercially available compounds for ligand discovery and virtual screening and allows known compound to be looked up by the target they bind.

canSAR

(https://cansar.icr.ac.uk) [56]

With the growing publicly available biological data, such as biological annotations, chemical screening, RNA interference screening, expression, amplification and 3D structural, canSAR is developed to integrate these data sets to facilitate cancer research and drug discovery in a more comprehensive view of relevant data from various sources. canSAR is one of the largest freely available integrated resource, which enables users to obtain information in a more efficient way.

PDSP

(http://pdsp.med.unc.edu) [57]

The Psychoactive Drug Screening Program (PDSP) could screen the compounds that have previous reports of pharmacological, biochemical or behavioral activities. The program is mainly used for the identification of novel targets (genes or molecules) for clinical research or treatment of mental disorders.

DCDB

(http://www.cls.zju.edu.cn/dcdb/) [58]

DCDB (Drug Combination Database) is a database offering information about drug combinations developed by researchers from Zhejiang University. It is the first database devoted to the research and development of drug combinations. Its current version comprises 1363 approved or investigational drug combinations, including 237 unsuccessful drug combinations, involving 904 individual drugs, from >6000 references. DCDB summarizes patterns of beneficial drug interactions and provides a basis for theoretical modeling and simulation of drug interactions. The drug combinations in the database are manually collected from PubMed and the US FDA Orange-Book [59]. The information about drugs and their targets are manually annotated based on the literature and relevant databases such as Drugbank [41], PubChem, UniProt and Drugs.com. The web interface of DCDB allows search by drug name, drug combination, disease and drug target, respectively. In addition, drugs can also be searched by chemical similarity. For each drug combination in DCDB, its intended activity, indication, potential interaction mechanisms, classification, status, related references and a number of external links are available.

ASDCD

(http://asdcd.amss.ac.cn/) [60]

Owing to the resistance of existing antifungal drugs and the limitation of available new drugs, it is urgent to develop new

antifungal synergistic drug combinations. Considering the ever-growing demand for effective antifungal drugs, Chen *et al.* developed ASDCD, which is the first DCDB devoted to antifungal drug research, aiming to facilitate drug combination analysis and new antifungal drug development. Its current version includes 210 antifungal drug combinations and 1225 drug–target interactions involving 105 individual drugs from >12 000 references.

DINIES

(http://www.genome.jp/tools/dinies/) [61]

Drug–target interaction network inference engine based on supervised analysis, DINIES, is a web server to infer potential drug–target interaction network. DINIES can accept flexible input data, such as chemical structure, side effects, amino acid and protein domains. Furthermore, each data set will be transformed into a kernel similarity, and various state-of-the-art machine learning methods will be used to realize the drug–target interactions prediction.

SuperPred

(http://prediction.charite.de/) [62]

SuperPred is web server for predicting the Anatomical Therapeutic Chemical (ATC) code and targets of small molecules. In SuperPred, the ATC code prediction is based on the similarity search pipeline, which integrated 2D, fragment and 3D similarity. Drug target prediction is based on the similarity distribution, which can estimate individual thresholds and probabilities for a specific target by four input options, including searching the compound's name in PubChem database, creating compound's structure by SMILES, drawing the structure using ChemDoodle editor and uploading the molecule's file, respectively.

SwissTargetPrediction

(http://www.swisstargetprediction.ch/) [63]

SwissTargetPrediction is a web server to infer the targets of bioactive small molecules based on the combination of 2D and 3D similarity values with known ligands. Besides, it can provide the predicting results by five different organisms, including *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Bos taurus* and *Equus caballus*.

Computational models

Nowadays, although high-throughput screening and other biological assays are becoming available, experimental methods for drug–target interaction prediction remain extremely challenging and time-consuming [39, 40, 64]. Therefore, it is important to develop new effective non-experimental method to infer drug–target associations. Molecular docking has been widely applied to virtually screen the compounds against target proteins when 3D structure of compounds are available [65–67]. However, the 3D structures of the majority of drugs are difficult to be obtained; thus, this important limitation has limited the wide use of molecular docking on a large scale. Nowadays, a number of computational models have been developed to address the drug-target prediction problem. In this review, these models have been divided into three main categories, including network-based model, machine learning-based model and other models.

Network-based model

Recently, various network-based methods have been proposed. Network has become an effective tool to predict underlying drug–target associations.

MTOI

Yang *et al.* [13] developed a computational algorithm to infer potential drug targets by systematically analyzing the transformation between the disease state and the desired state in a disease network. The aim of MTOI is to find multiple target optimal intervention (MTOI) solutions that give the best disease state transformation. Therefore, the output of MTOI includes not only plenty of potential drug–target interactions, but also optimal combinatorial intervention solutions. The method was applied to an inflammation-related network—the arachidonic acid (AA) metabolic network (AAnetwork) for the identification of optimal multi-target anti-inflammatory intervention solutions. Particularly in stage 2, Monte Carlo-simulated annealing was performed to find the desired state, and the objective function was defined as follows:

$$\left|\left(\frac{C_{T,net}}{C_{T,disease}}\right)_{LTB4}-0.1\right|+\left|\left(\frac{C_{T,net}}{C_{T,disease}}\right)_{PGE2}-0.1\right|+\left|\frac{\left(\frac{C_{T,PGI2}}{C_{T,TXA2}}\right)_{net}}{\left(\frac{C_{T,PGI2}}{C_{T,TXA2}}\right)_{disease}}-1\right|$$

where $C_{T,net}$ and $C_{T,disease}$ were the 1 h cumulative production of the metabolite in the present network and disease states, respectively.

Drug side-effect similarity-based method

Campillos *et al.* [68] proposed a method by using drug side-effect similarity to identify whether two drugs interact with the same target. In this article, the author proposed a sigmoid function (P_{2D}) that modeled the probability of sharing the same target based on chemical similarity information (2D Tanimoto coefficient) and a linear function (P_{SE}) that modeled the probability of sharing the same target depending on the side-effect similarity measure, respectively.

$$P_{2D}(y)=\left(1+e^{\frac{B-y}{A}}\right)^{-1}\quad(A=6.19,\quad B=0.68)$$

$$P_{SE}(x)=A\cdot x+B\quad(A=-0.084,\quad B=0.047)$$

P_{2D} and P_{SE} are the probability of sharing the same target as a function of chemical similarity and side-effect similarity for all the drug pairs, respectively. A and B are parameters of the function. Therefore, a drug side-effect network has been constructed with 1018 drug side-effect relations and 746 marketed drugs. This method required the detailed information of drug side effects. Therefore, it can be only applied to marketed drugs that have package inserts to demonstrate their side effects. This important limitation has seriously limited its wide application.

NRWRH

Based on the assumption that similar drugs often interact with similar target proteins and the integration of drug–drug similarity network, protein–protein similarity network and known drug–target interaction networks into a heterogeneous network, Chen *et al.* [10] developed an effective model of Network-based Random Walk with Restart on the Heterogeneous network (NRWRH) to predict potential drug–target interactions by implementing random walk on the heterogeneous network (Figure 1).

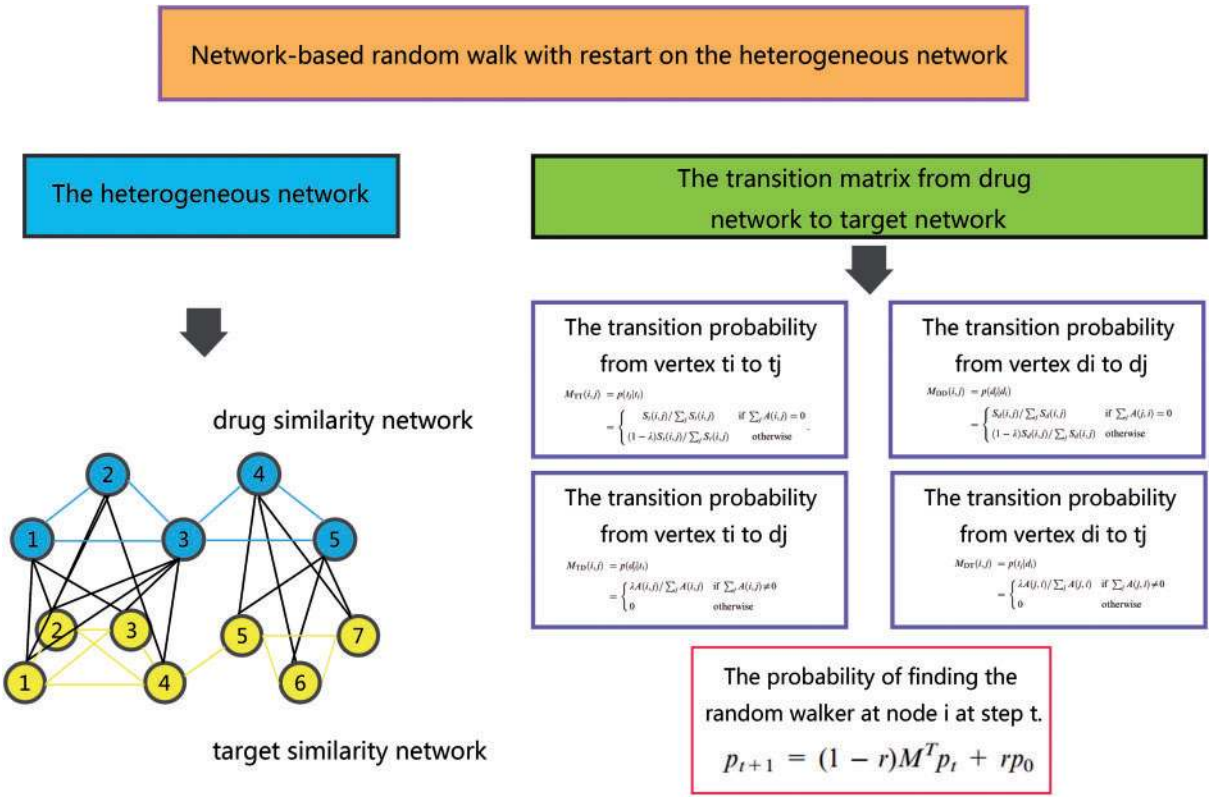

Figure 1. The basic idea of predicting drug–target interaction by implementing random walk on the heterogeneous network consisting of known drug–target interaction network, drug similarity network and target similarity network.

NRWRH makes full use of the network tool for data integration and drug–target interactions prediction, which is different from traditional random walk with restart. Here, random walk is implemented on the heterogeneous network, which consists of three different networks, i.e. drug–target interactions network, drug chemical structure similarity network and target protein sequence similarity network. There, even if the investigated drug has no known target, potential target of this given drug can still be predicted based on known targets of drugs, which are similar to this given drug. In NRWRH, the transition matrix on the heterogeneous network could be defined as

$$M = \begin{bmatrix} M_{TT} & M_{TD} \\ M_{DT} & M_{DD} \end{bmatrix}$$

where M_{TT} and M_{DD} are intertransition matrix indicating the probability from one target (drug) to other target (drug) in the random walk, respectively; $M_{TD}$is the transition matrix from target network to drug network, and M_{DT} is the transition matrix from drug network to target network. These four transition matrices could be further defined based on drug similarity network, target similarity network and known drug–target interaction network. Finally, random walk could be implemented based on the following iteration equations, and the probability of finding the random walker at node i at step t can be decided.

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

Here, the parameter r is the restart probability. NRWRH was applied to four classes of important drug–target interaction data sets (enzymes, ion channels, GPCRs and nuclear receptors), and significant performance improvement has been demonstrated in the terms of both cross validation and case studies.

DBSI, TBSI and NBI

Cheng *et al.* [69] developed three supervised inference models to predict drug–target interactions, namely, drug-based similarity inference (DBSI), target-based similarity inference (TBSI) and network-based inference (NBI) (Figure 2). For the DBSI, the score to predict the association between d_i and t_j is defined as follows:

$$v_{ij}^{D} = \frac{\sum_{l=1,l\neq i}^{n} S_c(d_i, d_l) a_{il}}{\sum_{l=1,l\neq i}^{n} S_c(d_i, d_l)}$$

where $S_c(d_i, d_l)$ is 2D chemical similarity between drugs d_i and d_l calculated by SIMCOMP [70], and $a_{ij} = 1$ if drugs d_i and target t_j have been linked in known drug–target interaction network. For the TBSI, the score to predict the association between d_i and t_j is defined as follows:

$$v_{ij}^{T} = \frac{\sum_{l=1,l\neq i}^{n} S_g(t_j, t_l) a_{jl}}{\sum_{l=1,l\neq i}^{n} S_g(t_j, t_l)}$$

where $S_g(t_j, t_l)$ is the genomic sequence similarity between targets t_j and t_l calculated based on normalized Smith–Waterman Score [71]. For the NBI, the final resource (score) f(i) of drug d_i after two-step diffusion is defined as follows:

$$f(i) = \sum_{l=1}^{m} \frac{a_{il}}{k(t_l)} \sum_{o=1}^{n} \frac{a_{ol} f_0(o)}{k(d_o)}$$

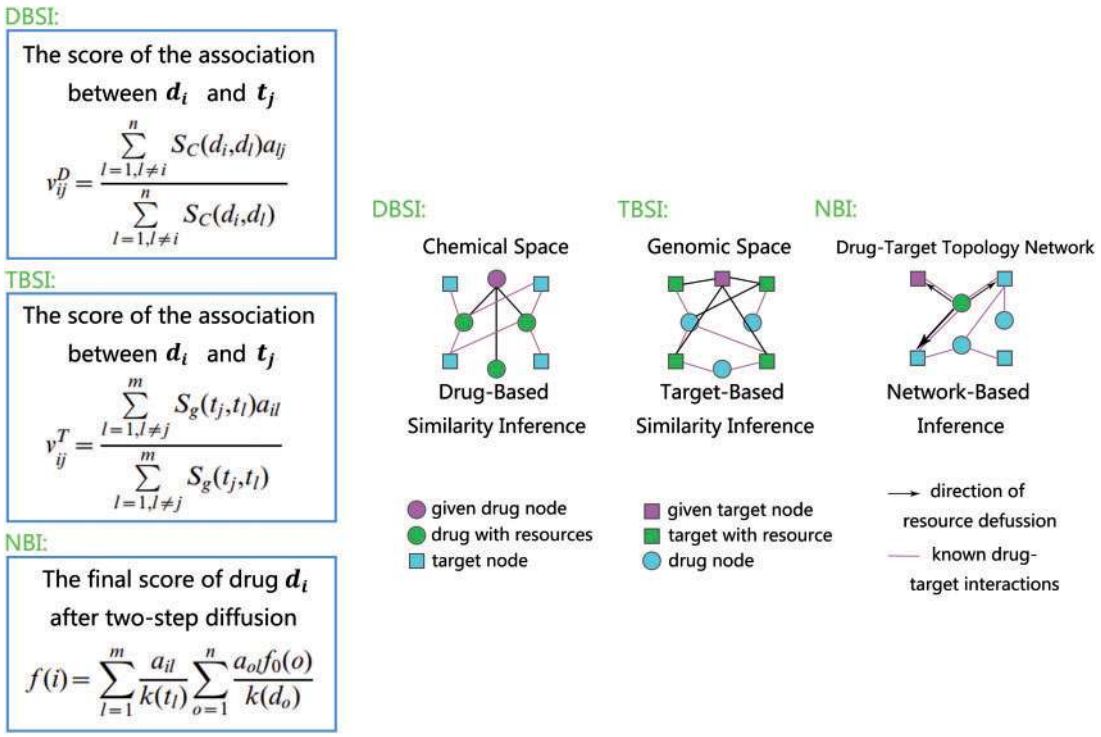

Figure 2. The basic idea of three supervised inference models to predict drug–target interactions: DBSI, TBSI and NBI.

where $f_0(o) = a_{oj}, o \in \{1, 2, \ldots, n\}$, $k(d_o) = \sum_{s=1}^{m} a_{os}$ and $k(t_1) = \sum_{s=1}^{n} a_{sl}$ denote the initial resource of drug d_0, the number of the targets that interact with d_o and the number of the drugs that interact with t_1, respectively. The difference among these three models lies in the similarity adaptation. DBSI and TBSI rely on chemical structural similarity and target sequence similarity, respectively, whereas NBI is only based on drug–target bipartite network topology similarity. Cross validation demonstrated that although NBI ignores drug chemical structure and target protein sequence information, it has shown the best performance among all of these three methods. However, NBI could not be applied to the new drugs without any known target information in the training set.

Within scores and between scores

Shi *et al.* [72] presented an approach to predict drug–target interactions by characterizing each drug–target pairs as a feature vector composed of within scores and between scores. The within score $C_t^w(t_x, d_y)$ and between score $C_t^b(t_x, d_y)$ in targets view are defined as follows:

$$C_t^w(t_x, d_y) = \max\left(\left\{S_t\left(t_x, t_p^i\right)\right\}\right), \qquad p = 1, 2, \ldots, T_i$$

$$C_t^b(t_x, d_y) = \max\left(\left\{S_t\left(t_x, t_q^i\right)\right\}\right), \qquad q = 1, 2, \ldots, T - T_i$$

where $S_t\left(t_x, t_p^i\right)$ is the similarity between target t_x and t_p^i, $S_t\left(t_x, t_q^i\right)$ is the similarity between target t_x and t_q^i. The formulas in drugs view are defined in the similar way as targets view. The advantage of this method is that it treated all types of drug–target pairs in a same form and the relationship between known drug–target interactions, and unapproved drug–target pairs could be investigated in the same visualized space.

Machine learning-based method

Nowadays, a number of machine learning-based methods have been developed to identify associations between drugs and target proteins on a large scale. In this review, we mainly introduce supervised learning method and semi-supervised learning method. In most of learning-based methods, different types of biological data sets have been integrated, such as drug chemical structures, target protein sequences and known drug–target interactions.

Supervised learning method

In the supervised learning method, drug–target pairs are labeled as positive or negative samples according to whether the known interaction between corresponding drug and target has been confirmed. However, the selection of negative samples is a common problem of all the supervised learning methods owing to the fact that we can not obtain drug–target pairs without interactions. Published experimental literatures only reports successful drug–target interactions obtained based on their biological experiments. Therefore, the unknown drug–target interactions have been regarded as negative samples in the supervised learning methods for drug–target interactions prediction. Inaccurate negative sample selection has largely influenced the predictive accuracy.

Bipartite graph learning method

Yamanishi *et al.* [9] proposed a kernel regression-based method to infer drug–target interactions by integrating the chemical structure information of compounds, the sequence information of target proteins and the topology of known drug–target interactions network to investigate the four classes of drug–target interactions in human, including enzymes, ion channels, GPCRs and nuclear receptors (See Figure 3). A supervised learning framework has been developed based on a bipartite graph,

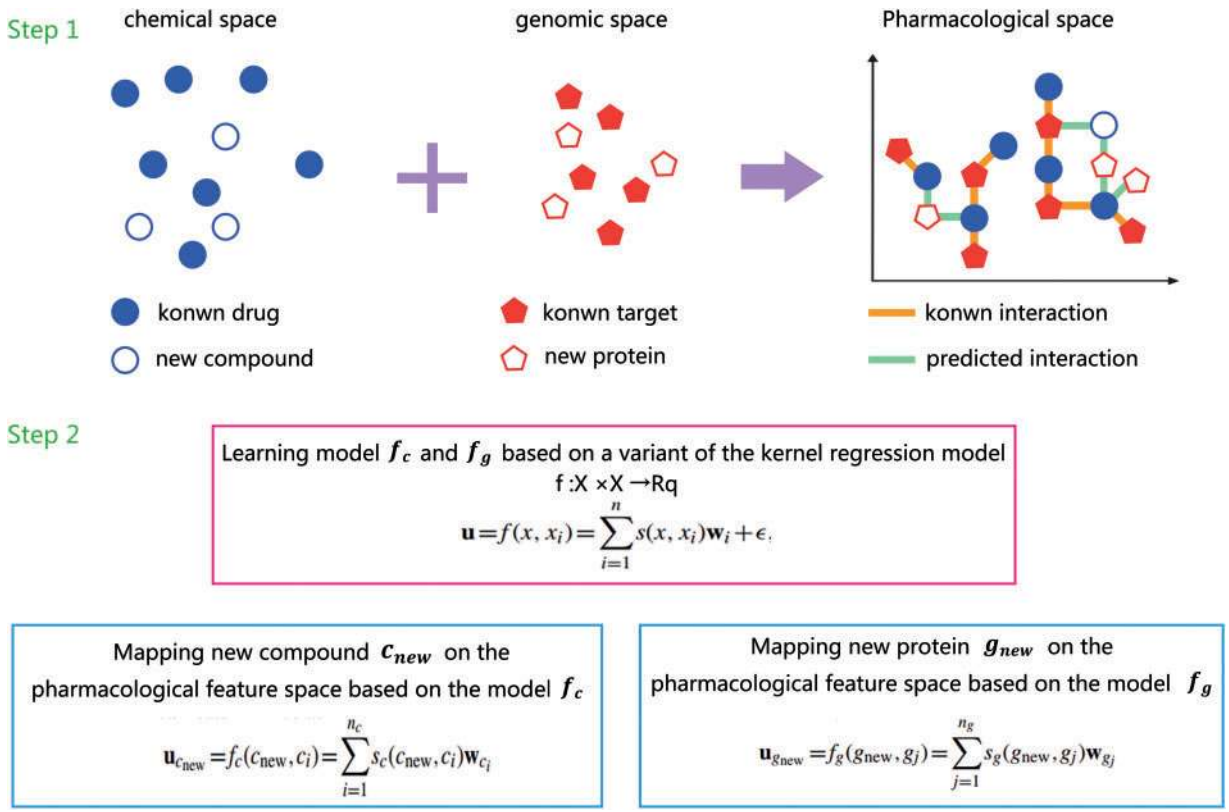

Figure 3. The basic idea of a kernel regression-based method developed based on a bipartite graph, which mapped drugs and targets in chemical space and genomic space into a unified space called pharmacological space and integrated the chemical structure information of compounds, the sequence information of target proteins, and the topology of known drug–target interactions network to investigate the drug–target interactions.

which mapped drugs and targets in chemical space and genomic space into a unified space called pharmacological space. In this study, two models (f_c and f_g) could be learned based on a variant of the kernel regression model to demonstrate the correlation between the chemical space/genomic space and the pharmacological feature space, respectively. For the new compound c_{new} and new proteing$_{new}$, we can map them on the pharmacological feature space based on these two models.

$$u_{c_{new}}=\sum_{i=1}^{n_c}s_c(c_{new},c_i)w_{c_i}$$

$$u_{g_{new}}=\sum_{j=1}^{n_g}s_g(g_{new},g_i)w_{g_j}$$

Where w_{c_i} and w_{g_j} are the weight vector, $s_c(\cdot,\cdot)$ and $s_g(\cdot,\cdot)$ indicate chemical structure similarity score and sequence similarity score. In this pharmacological space, known drug–target pairs that interact with each other are close, and drugs that have high structure similarity tend to interact with similar targets, and targets that have high sequence similarity tend to interact with similar drugs. Then, potential drug–target interactions are predicted by calculating the closeness between drugs and targets.

BLM

Bleakley and Yamanishi [73] developed a new supervised learning method, Bipartite Local Model (BLM), for the prediction of unknown drug–target interactions by transforming edge prediction problems into binary classification problems (Figure 4). Firstly, target proteins of a given drug are predicted based on the sequence structure similarities between targets. Then, potential associated drugs of a given target protein are predicted based on the chemical similarity between drugs. Therefore, independent target-based and drug-based prediction results for each putative drug–target interaction are obtained, respectively. Finally, a definitive prediction for each interaction is obtained based on the average of these two independent predictions.

Bipartite graph learning method by introducing the pharmacological data

Under the assumption that pharmacological effect similarity is more correlated with drug–target interactions compared with chemical structure similarity, Yamanishi *et al.* [74] further proposed a correlation-based model to infer the relationships of unknown drug–target pairs based on chemical structure information, genomic sequence information and pharmacological effect information on a large scale (Figure 5). For any drug candidate compounds, the pharmacological effect similarity is predicted from chemical structures of given compounds. Then, the pharmacological effect similarity is introduced into the supervised bipartite graph inference model [9] to identify unknown drug–target interactions. The prediction score for new compound y and protein z is defined as follows:

$$g(y,z)=\sum_{i=1}^{n_y}\sum_{j=1}^{n_z}a_{ij}S_{phar}(y_i,y)S_{geno}(z_j,z)$$

where $s_{phar}(\cdot,\cdot)$ and $s_{geno}(\cdot,\cdot)$ are pharmacological similarity function for compounds and sequence similarity function for proteins, n_y (n_z) is the number of compounds (proteins) and αij is the

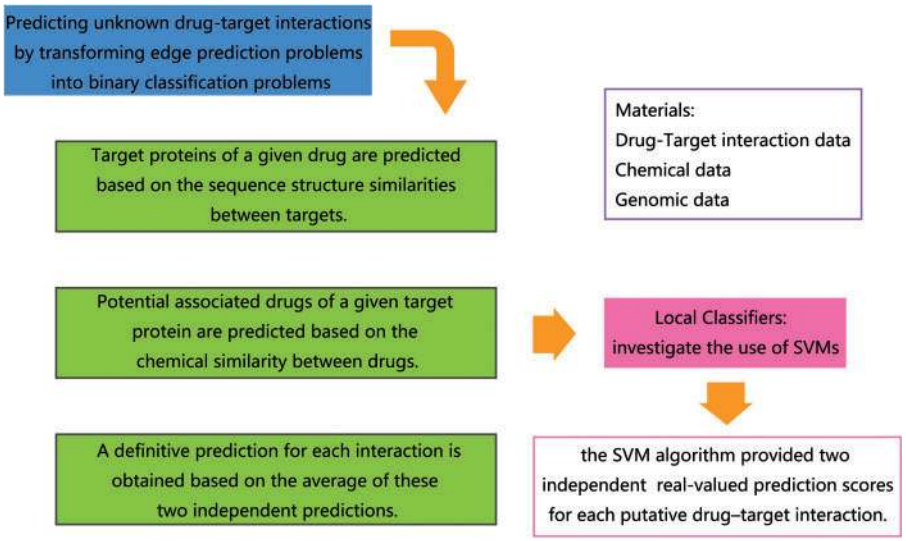

Figure 4. The basic idea of a new supervised learning method, BLM, was developed to predict unknown drug–target interactions by transforming edge prediction problems into binary classification problems.

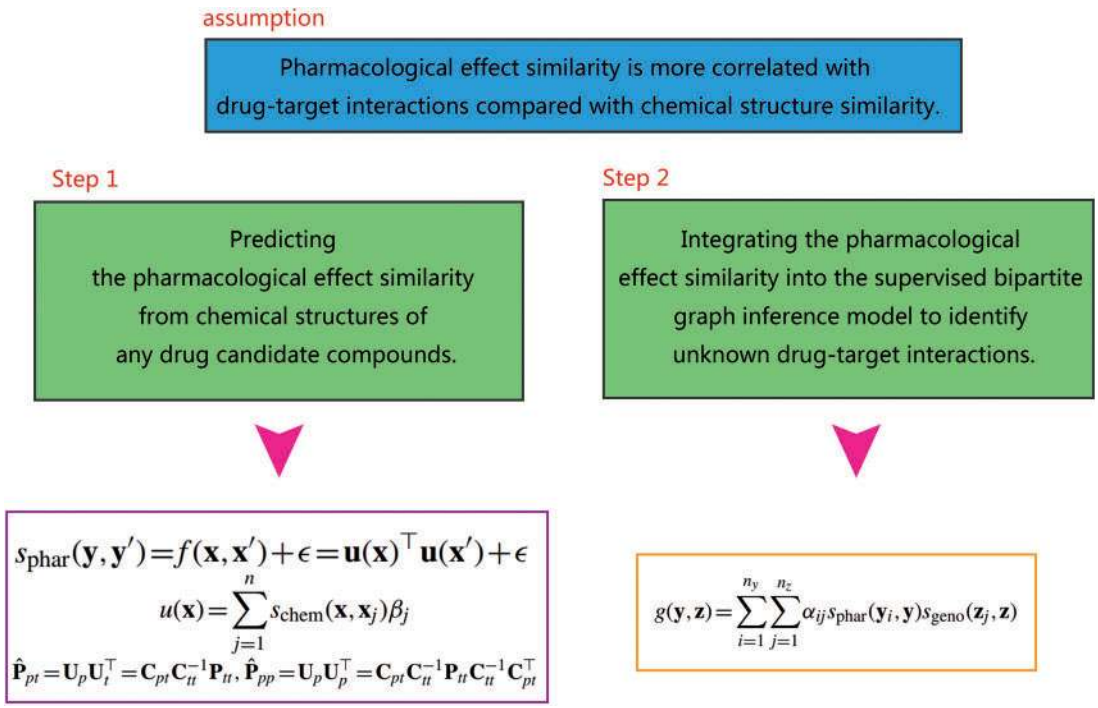

Figure 5. The basic procedure of a correlation-based model to infer the relationships of unknown drug–target pairs on a large scale, including the prediction of pharmacological effect similarity from chemical structures and the introduction of pharmacological effect similarity into the supervised bipartite graph inference model.

parameters learned from the model. The originality of this method is that it predicts potential pharmacological similarity for any drug candidate compounds and further integrates chemical, genomic and pharmacological information into a unified framework.

BLM-NII

Mei *et al.* [75] presented a new method named BLM-NII, which integrated Neighbor-based Interaction-profile Inferring (NII) and existing BLM model so that BLM model has been further extended to solve the limitation that it cannot predict the interactions for new drug or target candidates. In the NLL model, for a new drug d_i, the j-th dimension of its interaction profile $I^d(i)$ is defined as follows:

$$I_j^d(i) = \sum_{h=1}^{m_d} s_{ih}a_{hj}$$

where s_{ih} is the chemical similarity between drug d_i and d_h, m_d is the number of drugs and A is the adjacency matrix of known drug–target interaction network ($A(i,j) = a_{ij} = 1$ if drug d_i and

target t_j has known interaction). Furthermore, linear scale is implemented by the following operation:

$$I_j^d(i) = \frac{I_j^d(i) - min_h I_h^d(i)}{max_h I_h^d(i) - min_h I_h^d(i)}$$

Then, the model such as support vector machine or regularized least squares could be used to learn a certain classifier. Therefore, the model of drug d_i could be learned in the following way:

$$Mod_d^{'}(i) = train(S^t, I^d(i))$$

where S^t is target similarity matrix, which indicates the observed data of targets, and $I^d(i)$ is considered to be label information of each target to show whether the interaction exists between this target and investigated drug. Similarly, for the target t_j, we also could obtain its interaction profile $I^t(j)$ and could learn the model for this target as follows:

$$Mod_t^{'}(j) = train(S^d, I^t(j))$$

where S^d is drug similarity matrix, which indicates the observed data of drugs. Finally, the above NII strategy is integrated with BLM framework. BLM-NII model can make the predictions for new drug and target candidates and achieves good performance for inferring interactions between drugs and four classes of targets, especially for nuclear receptors. The most significant improvement has obtained when data set contains many drugs/targets with no known interactions.

RBM

Wang and Zeng [6] proposed the first learning method to predict not only the binary interactions between drugs and targets but also different types of interactions (i.e. how they interact with each other) in the framework of restricted Boltzmann machines (RBM) based on a multidimensional drug–target network (Figure 6). This method cast the novel drug–target interaction problem into a two-layer RBM model (hidden unit layer and visible unit layer) and there are no intra-layer connections in these two layers. Furthermore, Contrastive Divergence (CD) algorithm is applied to train RBM model and make predictions. It has been demonstrated that multiple types of interactions integration could significantly improve the prediction accuracy of previous methods, which only used a single interaction type. One limitation of this method is that the prediction results could be obtained only based on the known drug–target interactions network. Drug similarity and target protein similarity network could not be introduced into this model.

Random forest

Based on the random forest (RF) learning algorithm, Cao *et al.* [76] proposed a large-scale computational method to predict potential drug–target interactions by integrating the information of chemical properties (e.g. compound fingerprints or substructures), the information of biology properties (e.g. biomedical and physicochemical properties of protein targets) and the drug–target association network information. However, because of the use of network features, this method cannot be applied to identify new interactions for a drug or a target that does not have any known drug–target interactions.

Two strategies for negative sample selection

As has been mentioned above, a common problem for supervised methods of drug–target interactions prediction is the lack of a negative data set. To address this problem, Wang *et al.* [77] proposed two strategies to help general machine learning method better select the negative training samples. These two strategies aim at increasing the prediction accuracy in cross-validation and filtering out as many non-drug-target proteins as possible, respectively. In the first strategy, the drug protein's deviation is defined as follows:

$$\xi(X_i) = \sum_j \left| \frac{\mathrm{mean}(x_j)(x_{ij} - \mathrm{mean}(x_j))}{\mathrm{var}(x_j)\sum\left(\mathrm{mean}(x_j)^2/\mathrm{var}(x_j)\right)} \right|$$

Here, for the ith drug proteins, vector$X_i = (x_{i1}, x_{i2}, \ldots, x_{im})$ represented m properties (attributes) of this protein, where x_{ij} denoted the jth property's value of the ith drug proteins. Furthermore, $x_j = (x_{1j}, x_{2j}, \ldots, x_{nj})$denoted the vector of jth attribute. In the author's experiments, they chose the proteins with $\xi(X_i) > 0.42)$ as the negative samples because they found the cumulative distribution indicated that $\xi(X_i \leq 0.42) > 0.95$. As for the second strategy, the probability of each unknown protein i to be negative sample was defined as:

$$P(X_i \in NT) = \frac{\left(\xi(X_i) - \mathrm{mean}\left(\xi_{\mathrm{positive}}\right)\right)^2}{\sum_i \left(\xi(X_i) - \mathrm{mean}\left(\xi_{\mathrm{positive}}\right)\right)^2}$$

In the author's experiments, they supposed each protein has a probability of 0.5 to be considered as the negative sample. Therefore, for the 3834 proteins in the test data set, they obtained 1917 proteins to function as a negative data set.

Semi-supervised learning method

To face the challenge of negative samples selection problem, some semi-supervised methods were presented in which both few labeled data and many unlabeled data are integrated to make the prediction.

NetLapRLS

Xia *et al.* [78] developed a semi-supervised learning method, NetLapRLS, which combines chemical space, genomic space as well as known drug–protein interaction network information into a heterogeneous biological space to predict potential drug–target interactions (Figure 7). This manifold regularization method uses labeled and unlabeled information instead of using labeled data alone to realize better prediction results for each chemical–protein pair. Firstly, the cost function of NetLapRLS is defined as follows:

$$F_d^* = \min \quad J(F_d) = \| Y - F_d \|_F^2 + \beta_d \mathrm{Trace}(F_d^T L_d F_d)$$

where $\| \cdot \|_F$ is Frobenius norm, Trace is the trace of a matrix, Y is the adjacency matrix of the known drug–target interaction network and β_d is the trade-off parameter in the drug space. L_d could be obtained by implement Laplacian operation to normalize the drug similar matrix W_d as follows:

$$L_d = D_d^{-1/2}(D_d - W_d)D_d^{-1/2}$$

where D_d is defined such that $D_d(i,i)$ are the sum of the ith row of drug similar matrix W_d. Then, we could obtain the solution of this optimization problem:

$$F_d^* = W_d(W_d + \beta_d L_d W_d)^{-1} Y$$

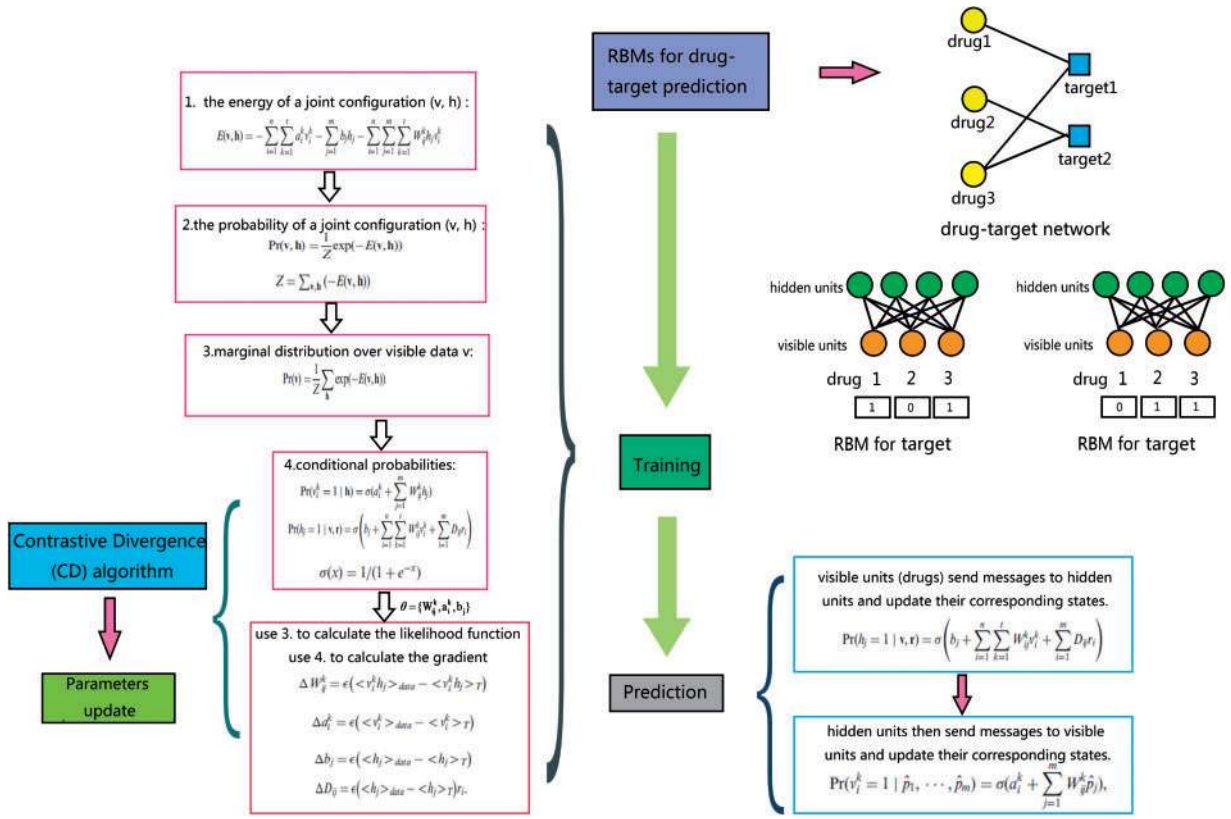

Figure 6. The flow chart of predicting drug–target interactions using RBM is shown here, including the construction of RBMs from a drug–target interaction network, the train of RBM by CD algorithm and prediction implementation.

Similarly, the cost function and its corresponding classifier could be constructed in the target space as follows:

$$F_p^* = W_p\left(W_p + \beta_p L_p W_p\right)^{-1} Y^T$$

where β_p is the trade-off parameter in the target protein space, and L_p could be obtained by implement Laplacian operation to normalize the target similar matrix W_p. Therefore, two classifiers have been constructed in the drug space and target space, respectively. Final prediction results are obtained by implementing a mean operation for the prediction result from drug and target spaces.

Kron-RLS

In previous studies, plenty of machine learning models have been developed for the drug–target interaction prediction models based on drug chemical structure and target genomic sequence similarity information. However, these traditional models are developed based on on/off interaction data, which do not reflect the real-life problem in practical cases of drug–target interactions. In that case, most machine learning models treated the drug–target interaction prediction as a binary classification problem [79]. Drug–target interaction is not a simple binary on/off relationship. Pahikkala *et al.* [79] illustrated the effects of four factors that may improve the prediction performance of drug–target interaction prediction, including the problem formulation, evaluation data set, evaluation procedure and experimental setting. Especially, for the problem formulation, more realistic prediction results should be obtained by formulating the prediction problem as regression prediction, rather than traditional binary classification. Therefore, they further proposed the Kronecker RLS (Kron-RLS) model based on quantitative bioactivity data for kinase inhibitors [such as the dissociation constant (k_d) or inhibition constant (k_i) to reflect the whole spectrum of interaction affinity between a ligand molecule and a target molecule] to predict potential drug–target interactions. The systematic mapping provides broader insights into the interaction patterns between drug and targets. Given training drug–target pairs input data x_i and their real-valued labels y_i (interaction affinities), the prediction function could be obtained by finding a minimizer of the following objective function:

$$J(f) = \sum_{i=1}^{m} (y_i - f(x_i))^2 + \lambda \| f \|_k^2, \quad (\lambda > 0)$$

where λ is the user-provided regularization parameter controlling the compromise between prediction error on the training samples and the model complexity, $\| f \|_k$ is the norm of f measured in the Hilbert space, and k is a kernel function obtained based on the drug chemical structure similarity and target protein sequence similarity, respectively.

Other methods

Chemical similarities

Based on the assumption that similar drugs tend to interact with similar targets, Keiser *et al.* [80] predicted underlying interactions between drugs and targets according to the chemical similarities between drugs and ligand sets, which are known to

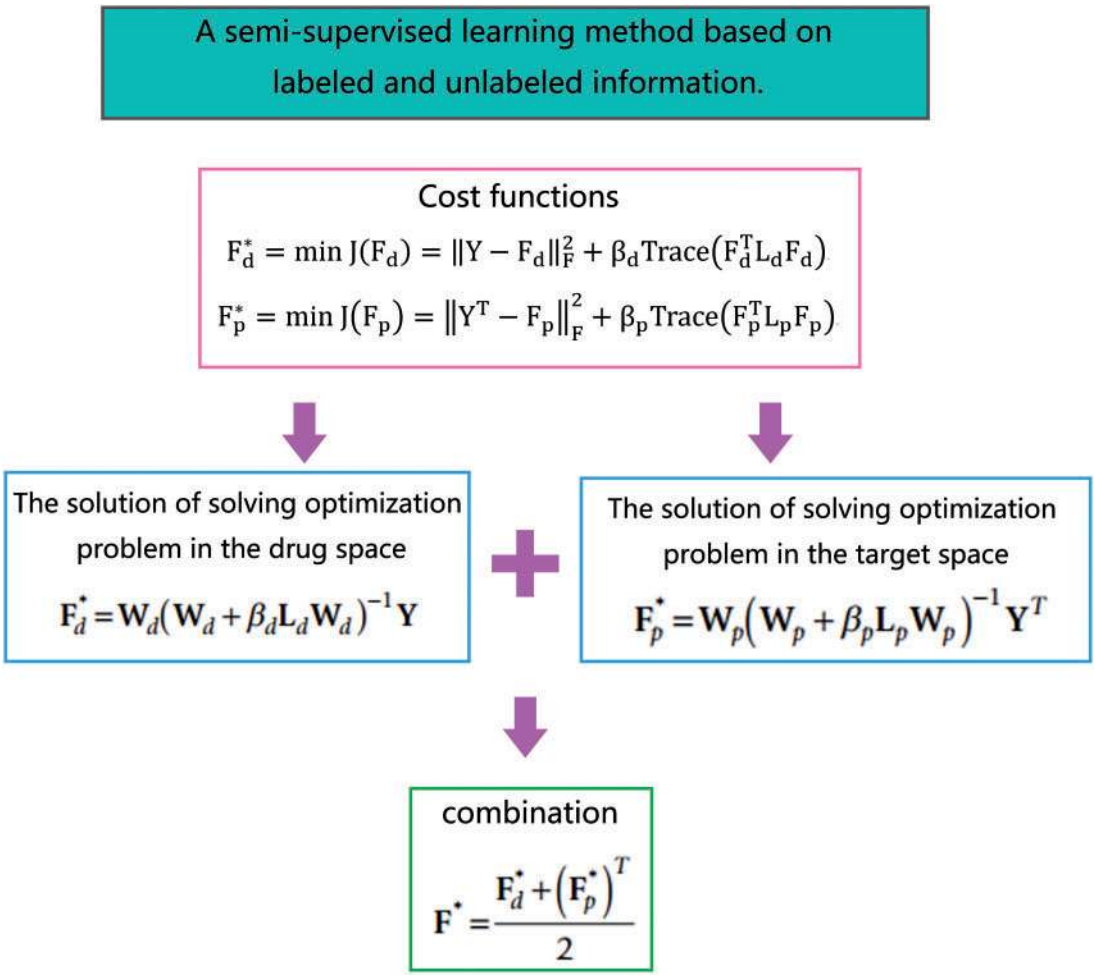

Figure 7. The flow chart of a semi-supervised learning method, NetLapRLS, which combines chemical space, genomic space as well as known drug–protein interaction network information into a heterogeneous biological space to predict potential drug–target interactions based on labeled and unlabeled information instead of using labeled data alone.

modulate the function of a few kinds of protein receptors. This method uses 2D structure similarities of ligands to infer a number of drug–target connections. Furthermore, 30 predicted interactions were tested experimentally and 23 new drug–target associations were confirmed. However, the available sequence information of protein targets is not taken into consideration here. Therefore, this method is limited to predict to the interactions between known ligands and different protein families.

A two-step similarity-based method

Chen and Zeng [32] predicted target group of drugs using a two-step similarity-based method, which integrated graph-representation-based method and chemical-functional-group-representation-based method. Firstly, drugs were encoded by their corresponding graph representation. Five target groups were denoted by $C = [C_1, C_2, C_3, C_4, C_5]$ to indicate 'G Protein-coupled Receptors', 'Cytokine Receptors', 'Nuclear Receptors', 'Ion Channels' and 'Enzymes', respectively. The target groups of any drug d_i can be described as

$$C(d_i) = [c_{i,1}, c_{i,2}, c_{i,3}, c_{i,4}, c_{i,5}]^T$$

where $c_{i,j} = 1$ if C_j is a target group of d_i, otherwise, $c_{i,j} = 0$. Toward a query drug compound d, the possibility that C_j is a target group of drug d was calculated by:

$$S_1(d \Rightarrow C_j) = \max \quad S_g(d, d_i) \cdot c_{i,j}$$

where S_g is the drug similarity matrix based on graph representations. Then, a similarity score was obtained based on chemical functional group representation for a query drug that does not get the prediction for its candidate target groups in the first step. Based on the assumption that compounds with the same functional group tend to react in a similar way, authors selected 28 major functional groups and represented each drug d by a 28-dimension vector $F(d) = (g_1, g_2, \cdots, g_{28})^T$. Towards a query drug compound d, the possibility that C_j is a target group of drug d was further calculated by:

$$S_2(d \Rightarrow C_j) = \max \quad S_f(d, d_i) \cdot c_{i,j}$$
$$S_f(d, d_i) = \frac{F(d_1) \cdot F(d_2)}{\| F(d_1) \| \cdot \| F(d_2) \|},$$

where $F(d_1) \cdot F(d_2)$ is the dot product of two vectors $F(d_1)$ and $F(d_2)$, and $\| F(d_1) \|$ and $\| F(d_2) \|$ are the modulus of vector $F(d_1)$ and $F(d_2)$, respectively. Although these two kinds of representation methods were applied to encode query drugs, there are also some query drugs that cannot obtain prediction results,

such as the drug without any similarity at all to all the drugs in training data set.

PMF

Cobanoglu *et al.* [81] developed a Probabilistic Matrix Factorization (PMF)-based active learning methodology to predict drug–target interactions, which does not rely on chemical/target similarity or external data collection. The method is shown to group drugs according to their therapeutic effects instead of their 3D structure similarity.

Discussion and conclusion

Identification of drug–target interactions is the important foundation of drug discovery. In this article, databases and web servers involved in drug-target identification and drug discovery are summarized. In addition, we mainly introduced some state-of-the-art computational models for drug–target interactions prediction, such as network-based models and machine learning-based models. Furthermore, we categorized machine learning-based model into supervised and semi-supervised model, which has essential difference in the adoption of negative samples. Most of these models take advantages of different types of biological data sets to implement prediction, but they only could predict the binary relationships between drugs and target. Wang and Zeng [6] successfully developed the first learning method to predict the types of drug–target interactions in the multidimensional network.

Network-based and machine learning-based models have their advantages and disadvantages. The key advantage of most of these approaches is that they are applicable to compounds without known 3D structures. Furthermore, almost all the models can effectively predict novel drug–target interactions for drugs, which have at least one known associated target proteins. More importantly, some models could be further applied to the new compounds without any known associated target proteins by integrating drug similarity, target similarity and known drug–target interactions.

As for the supervised machine learning method, it has important limitations as follows. Firstly, there are no experimental validated non-drug–target interactions so that it is difficult to select negative samples. Most of supervised learning method regard the unknown drug–target interactions or randomly select the unconfirmed drug–target pairs as negative samples, which would largely influence the predictive accuracy of the method. Secondly, two different classifiers from drug and target space are constructed in some methods, such as BLM; hence, the final result is the average of these two predictions, which will result in biases. Although semi-supervised learning method NetLapRLS makes use of the unlabeled information and overcome the difficulty of selecting negative samples, it also has the same limitation of classifier combination.

Nowadays, network has become an effective tool in potential drug–target interaction identification and drug repurposing. Focusing on the limitations of current network-based drug–target interactions prediction models, the future direction of the network-based drug discovery could be summarized as follows. Firstly, more heterogeneous network about drugs and target proteins should be integrated, such as drug chemical structure similarity network, protein sequence similarity network, known drug–target interaction networks, drug side-effect network, metabolic network related to specific disease and target–protein interaction network. Furthermore, new developed network-based computational models should be implemented on this heterogeneous network rather than the single network. In this way, even if there is no known target protein for the investigated drug, we still can obtain potential target of this drug based on the known targets of drugs without high similarity with this given drug. Then, the main limitation for almost all the network-based approaches is they cannot implement prediction for any drug–target pairs without known reachable paths in the network. As what has been pointed out in the literature [72], drug–target interaction network usually is composed of several isolated subnetworks, and most of current network-based methods cannot predict the interactions between the drug in one subnetwork and the target in another. Furthermore, existing network-based models tends to bias to the targets with more known associated drugs. Therefore, network tools should be further developed to solve these two critical problems in the future. Another future direction of network-based drug discovery is adopting global network information to capture the potential association between drugs and target proteins, whose advantage over local network information-based models has been demonstrated in many previous computational biology research. Finally, we are entering the era of personalized medicine and high-throughput genome sequencing, especially in cancer studies. The ultimate goal of biology in the future is to provide personalized treatment regimens for cancer patients, and it is improper to use a single or several drug targets for all the patients [82]. Therefore, the network approach should be applied to discovery personalized drugs by integrating the tumor clone-based network, cancer hallmark-based network and sequencing technologies [82–84]. Wang *et al.* [83] have described the models with hallmark-based network to study the tumor clones and significantly accelerate the understanding of tumor evolution and tumorigenesis. Using this important and novel framework, they proposed many valuable drug discovery strategies. More importantly, network-based models could be constructed in this framework to solve many important problems as follows: (i) the prediction of personalized drug targets; (ii) drug resistance prediction; (iii) personalized drug effect prediction; (iv) personalized molecular signature identification for therapeutic evaluation after cancer treatment; (v) personalized cancer risk prediction for healthy individuals [83]. Successful network-based models for these important problems would have critical impact on timely diagnosis, personalized treatment, prognosis and personalized prevention of cancer [83].

It is well known that the data set used in the drug–target interactions prediction has great influences on the prediction performance. Nowadays, these constructed prediction models have shown their effectiveness in many network pharmacology applications, such as identifying new therapeutic indication of existing drugs and inferring potential synergistic drug combinations based on predicted drug–target interactions. However, they are often evaluated under overly simplified settings, and almost all the traditional machine learning prediction models treated the drug–target interaction prediction as a binary classification problem based on on/off interaction data, which do not reflect the real-life problem in practical cases of drug–target interactions. Pahikkala *et al.* [79] pointed out the following four important facts, which should be taken into consideration into the model development and evaluation because they can strongly influence the prediction performance: (i) problem formulation by more realistic regression formulation rather than standard binary classification; (ii) model prediction based on quantitative bioactivity data rather than on/off interaction data; (iii) model validation based on simple or nested cross-validation; (iv) model performance report based on different

experimental setting to find out whether training and test sets share common drugs and targets, only drugs or targets or neither.

Specially, the quantitative data would make the prediction results more accurate and provide broader insights into the interaction patterns of the drugs and targets, while the binary data sets usually ignore many important aspects of the drug–target interactions, including their dose-dependence and quantitative affinities. Some of the recently published data have involved information about the strength interaction binding affinities, so it is possible to consider the interaction prediction as a regression problem of the binding affinities rather than as a classification problem of whether there is an interaction. For example, Metz *et al.* (2011) [85] put forward to combine sequence-dependent and pharmacology-dependent networks in a harmonious way and further constructed a comprehensive kinome interaction network based on both sequence comparisons and multiple pharmacology parameters obtained from activity profiling data. They successfully collected >150 000 kinase inhibitory values of >3800 compounds tested against 172 different protein kinases. Furthermore, a robust statistical analysis of kinome profiling data was implemented. Davis *et al.* (2011) [86] tested the interaction of 72 known kinase inhibitors against a panel of 442 kinases assays covering >80% of the human catalytic protein kinome, which provides a comprehensive data set of kinase inhibitors across the kinome. These two large-scale data sets for clinically relevant kinase inhibitors provided the quantitative bioactivity spectra, including the kinase disassociation constant (K_d) and kinase inhibition constant (K_i) data sets. These data sets based on biochemical selectivity assays provided broader insights into the interaction patterns of kinase inhibitors. It is well known that plenty of regression models have been constructed and successfully applied to many important practical applications, so it is anticipated that regression model would play critical and important roles in the future drug–target interactions prediction.

For most of the aforementioned computational models, prediction performance is evaluated based on cross validation. However, the recent study demonstrated that the performance based on cross validation differences between in-sample and out-of-sample interactions [87]. In that study, they further performed experiments for protein–protein interactions prediction based on seven state-of-the-art methods and observed that the performance of each method differs significantly in different test classes [87]. While the study mainly concerns evaluation schemes for protein–protein interactions, the same principles apply for all the pair-input computational prediction problems. In our previous studies, we have developed the computational model of Laplacian Regularized Least Squares for LncRNA–Disease Association (LRLSLDA) to predict potential lncRNA–disease associations and further applied this new validation framework to evaluate the performance of LRLSLDA [88]. As a result, LRLSLDA have obtained an excellent predictive performance in different classes. Specially, for the drug–target interactions prediction problems, the paired nature of inputs leads to a natural partitioning of test pairs, and pair-input models may achieve significantly different predictive performances for distinct test classes [87]. According to the evaluation methods proposed in this article, the test samples of target–drug associations could be classified into four distinct classes: C1 is composed of the test samples sharing both drugs and targets with the training samples; C2 is composed of the test samples sharing only drugs with the training samples; C3 is composed of the test samples sharing only targets with the training samples; C4 is composed of the test samples sharing neither drugs nor targets with the training samples. Therefore, for the drug–target interaction prediction, it is important and necessary to report cross validation performance for all the four independent test classes.

Compared with other reviews such as [89], the research results covered in this article are more comprehensive. Sixteen state-of-the-art computational methods for drug–target interactions prediction have been included in this review and most of which with their core formulas and the figures demonstrating theirs basic ideas or procedures. For example, the model of Kron-RLS [79], which introduced a new research direction into drug–target interaction prediction research, is hardly mentioned in previous reviews. Then, we have collected plenty of databases and web servers. Furthermore, we made a discussion about the strength interaction binding affinities, drug–target interactions prediction based on regression rather than traditional on/off classification and the new cross validation framework, which would have a major impact on predictive research. After previous excellent review presented by Ding *et al.* [89], plenty of new computational models, databases, web servers have emerged. For example, Pahikkala *et al.* [79] illustrated the effects of four factors that may improve the prediction performance of drug–target interaction prediction, formulated the prediction problem as regression prediction and further developed the model of Kron-RLS based on quantitative bioactivity data for kinase inhibitors. Furthermore, the first learning method to predict not only the binary interactions between drugs and targets but also different types of interactions have been constructed in the framework of RBM based on a multidimensional drug–target network [6]. Considering the common problem for supervised models is the lack of negative samples, Wang *et al.* [77] proposed two strategies to help general machine learning method better select the negative training samples. Some new computational models for traditional drug–target on/off interactions also have been developed, such as the model of RF, within-scores and between-scores, BLM-NII, a two-step similarity-based method and PMF. Especially, the model of BLM-NII could solve the limitation of BLM that it cannot predict the interactions for new drug or target candidates by integrating NII and existing BLM model. Besides, plenty of new databases have been constructed, such as The IUPHAR/BPS Guide to PHARMACOLOGY (the data of pharmacological, chemical, genetic, functional and pathophysiological are included), CancerDR (provide comprehensive information of anti-cancer drugs, and their pharmacological profiling across 952 cancer cell lines) and ASDCD (the first DCDB devoted to antifungal drug research, aiming with antifungal drug combinations and drug–target interactions). Furthermore, some previous databases have been updated frequently, such as DrugBank (contain 7759 drug entities and 15 199 drug–target interactions), TTD (1755 biomarkers for 365 disease conditions, and 210 drug scaffolds for 714 drugs, and leads have been further added), and STITCH (the number of high-confidence chemical–protein interactions in human has increased by 45%). It is also worth noting that some easy-to-use web servers have been constructed recently, such as DINIES, SuperPred and SwissTargetPrediction.

Nowadays, a wide range of databases and web servers about drug–target interactions have been built, providing a variety of resources of drug space, target space, drug–target interaction network, side-effect network and other related networks. Therefore, making full use of different types of heterogeneous data sources, the computational predictive models can realize more accurate identification of new drug–target interactions.

Ideally, to overcome the limitations of the supervised models, both true-positive interactions and true-negative pairs should be reported in the databases and web servers. Experimentally measured negative samples would provide significant improvement in the performance of prediction models. Furthermore, although the source programs or software of some computational models are available, it may be difficult to use, and more easy-to-use web servers should be constructed in the future, which would benefit biologists to experimentally confirmed predicted drug–target interactions.

Key Points

- Developing effective computational models to predict potential drug–target interactions from heterogeneous biological data could benefit not only better understanding of the various interactions and biological processes, but also novel drugs discovery and human medicine improvement.
- Nowadays, a wide range of databases and web servers about drug–target interactions have been built, providing a variety of resources of drug space, target space, drug–target interaction network, side-effect network and other related networks.
- Many computational methods have been developed to predict potential drug–target interactions. Especially, network-based models and machine learning-based models have become the important and effective tools in computational drug–target interaction identification.
- Network-based and machine learning-based models for drug–target interaction prediction have their advantages and disadvantages.
- Making full use of different types of heterogeneous data sources could benefit more effective identification of new drug–target interactions based on computational models.

Funding

The National Natural Science of Foundation of China under Grant No. 11301517, 61472203, 61327902, the foundation from National Center for Mathematics and Interdisciplinary Sciences, CAS and State Key Laboratory of Intelligent Control and Decision of Complex Systems, Beijing Institute of Technology.

References

1. Kola I, Landis J. Can the pharmaceutical industry reduce attrition rates? *Nat Rev Drug Discov* 2004;**3**:711–15.
2. Paul SM, Mytelka DS, Dunwiddie CT, *et al*. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat Rev Drug Discov* 2010;**9**:203–14.
3. Dickson M, Gagnon JP. Key factors in the rising cost of new drug discovery and development. *Nat Rev Drug Discov* 2004;**3**:417–29.
4. Chen H, Zhang Z. A semi-supervised method for drug-target interaction prediction with consistency in networks. *PLoS One* 2013;**8**:e62975.
5. Booth B, Zemmel R. Prospects for productivity. *Nat Rev Drug Discov* 2004;**3**:451–6.
6. Wang Y, Zeng J. Predicting drug-target interactions using restricted Boltzmann machines. *Bioinformatics* 2013;**29**:i126–34.
7. Hopkins AL. Network pharmacology: the next paradigm in drug discovery. *Nat Chem Biol* 2008;**4**:682–90.
8. Iskar M, Zeller G, Zhao X-M, *et al*. Drug discovery in the age of systems biology: the rise of computational approaches for data integration. *Curr Opin Biotechnol* 2012;**23**:609–16.
9. Yamanishi Y, Araki M, Gutteridge A, *et al*. Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics* 2008;**24**:i232–40.
10. Chen X, Liu M-X, Yan G-Y. Drug–target interaction prediction by random walk on the heterogeneous network. *Mol Biosyst* 2012;**8**:1970–8.
11. MacDonald ML, Lamerdin J, Owens S, *et al*. Identifying off-target effects and hidden phenotypes of drugs in human cells. *Nat Chem Biol* 2006;**2**:329–37.
12. Xie L, Xie L, Kinnings SL, *et al*. Novel computational approaches to polypharmacology as a means to define responses to individual drugs. *Ann Rev Pharmacol Toxicol* 2012;**52**:361–79.
13. Yang K, Bai H, Ouyang Q, *et al*. Finding multiple target optimal intervention in disease-related molecular network. *Mol Syst Biol* 2008;**4**:228.
14. Zimmermann GR, Lehar J, Keith CT. Multi-target therapeutics: when the whole is greater than the sum of the parts. *Drug Discov Today* 2007;**12**:34–42.
15. Frantz S. Drug discovery: playing dirty. *Nature* 2005;**437**:942–3.
16. Allen JA, Roth BL. Strategies to discover unexpected targets for drugs active at G protein-coupled receptors. *Ann Rev Pharmacol Toxicol* 2011;**51**:117–44.
17. Ashburn TT, Thor KB. Drug repositioning: identifying and developing new uses for existing drugs. *Nat Rev Drug Discov* 2004;**3**:673–83.
18. Núñez S, Venhorst J, Kruse CG. Target–drug interactions: first principles and their application to drug discovery. *Drug Discov Today* 2012;**17**:10–22.
19. Imming P, Sinning C, Meyer A. Drugs, their targets and the nature and number of drug targets. *Nat Rev Drug Discov* 2006;**5**:821–34.
20. Overington JP, Al-Lazikani B, Hopkins AL. How many drug targets are there? *Nat Rev Drug Discov* 2006;**5**:993–6.
21. Hopkins AL, Groom CR. The druggable genome. *Nat Rev Drug Discov* 2002;**1**:727–30.
22. Drews J. Drug discovery: a historical perspective. *Science* 2000;**287**:1960–4.
23. Golden J. Prioritizing the human genome: knowledge management for drug discovery. *Curr Opin Drug Discov Dev* 2003;**6**:310–16.
24. Zheng C, Han L, Yap CW, *et al*. Progress and problems in the exploration of therapeutic targets. *Drug Discov Today* 2006;**11**:412–20.
25. Landry Y, Gies JP. Drugs and their molecular targets: an updated overview. *Fundam Clin Pharmacol* 2008;**22**:1–18.
26. Wheeler DL, Barrett T, Benson DA, *et al*. Database resources of the national center for biotechnology information. *Nucleic Acids Res* 2007;**35**:D5–12.
27. Sayers EW, Barrett T, Benson DA *et al*. Database resources of the national center for biotechnology information. *Nucleic Acids Res* 2011;**39**:D38–51.
28. Dobson CM. Chemical space and biology. *Nature* 2004;**432**:824–8.
29. Yu H, Chen J, Xu X, *et al*. A systematic prediction of multiple drug-target interactions from chemical, genomic, and pharmacological data. *PLoS One* 2012;**7**:e37608.

30. Kanehisa M, Goto S, Hattori M, *et al*. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res* 2006;**34**:D354–7.
31. Stockwell BR. Chemical genetics: ligand-based discovery of gene function. *Nat Rev Genet* 2000;**1**:116–25.
32. Chen L, Zeng W-M. A two-step similarity-based method for prediction of drug's target group. *Protein Pept Lett* 2013;**20**:364–70.
33. Huang T, Tu K, Shyr Y, *et al*. The prediction of interferon treatment effects based on time series microarray gene expression profiles. *J Transl Med* 2008;**6**:44.
34. Huang T, Cui W, He Z-S, *et al*. Functional association between influenza A (H1N1) virus and human. *Biochem Biophys Res Commun* 2009;**390**:1111–13.
35. Giuliano KA, Haskins JR, Taylor DL. Advances in high content screening for drug discovery. *Assay Drug Dev Technol* 2003;**1**:565–77.
36. Hughes JE. Genomic technologies in drug discovery and development. *Drug Discov Today* 1999;**4**:6.
37. Petriz BA, Gomes CP, Rocha LA, *et al*. Proteomics applied to exercise physiology: a cutting-edge technology. *J Cell Physiol* 2012;**227**:885–98.
38. Ma H, Zhao H. Drug target inference through pathway analysis of genomics data. *Adv Drug Deliv Rev* 2013;**65**:966–72.
39. Haggarty SJ, Koeller KM, Wong JC, *et al*. Multidimensional chemical genetic analysis of diversity-oriented synthesis-derived deacetylase inhibitors using cell-based assays. *Chem Biol* 2003;**10**:383–96.
40. Kuruvilla FG, Shamji AF, Sternson SM *et al*. Dissecting glucose signalling with diversity-oriented synthesis and small-molecule microarrays. *Nature* 2002;**416**:653–7.
41. Law V, Knox C, Djoumbou Y, *et al*. DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res* 2014;**42**:D1091–7.
42. Qin C, Zhang C, Zhu F, *et al*. Therapeutic target database update 2014: a resource for targeted therapeutics. *Nucleic Acids Res* 2014;**42**:D1118–23.
43. Hecker N, Ahmed J, von Eichborn J, *et al*. SuperTarget goes quantitative: update on drug–target interactions. *Nucleic Acids Res* 2012;**40**:D1113–17.
44. Günther S, Kuhn M, Dunkel M, *et al*. SuperTarget and Matador: resources for exploring drug-target relationships. *Nucleic Acids Res* 2008;**36**:D919–22.
45. Kuhn M, Szklarczyk D, Pletscher-Frankild S, *et al*. STITCH 4: integration of protein–chemical interactions with user data. *Nucleic Acids Res* 2014;**42**:D401–7.
46. Magariños MP, Carmona SJ, Crowther GJ, *et al*. TDR Targets: a chemogenomics resource for neglected diseases. *Nucleic Acids Res* 2012;**40**:D1118–27.
47. Gao Z, Li H, Zhang H, *et al*. PDTD: a web-accessible protein database for drug target identification. *BMC Bioinformatics* 2008;**9**:104.
48. Gaulton A, Bellis LJ, Bento AP, *et al*. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res* 2012;**40**:D1100–7.
49. Emig D, Ivliev A, Pustovalova O, *et al*. Drug target prediction and repositioning using an integrated network-based approach. *PLoS One* 2013;**8**:e60618.
50. Kuhn M, Campillos M, Letunic I, *et al*. A side effect resource to capture phenotypic effects of drugs. *Mol Syst Biol* 2010;**6**:343.
51. Seiler KP, George GA, Happ MP, *et al*. ChemBank: a small-molecule screening and cheminformatics resource database. *Nucleic Acids Res* 2008;**36**:D351–9.
52. Pawson AJ, Sharman JL, Benson HE, *et al*. The IUPHAR/BPS Guide to PHARMACOLOGY: an expert-driven knowledgebase of drug targets and their ligands. *Nucleic Acids Res* 2014;**42**:D1098–106.
53. Kumar R, Chaudhary K, Gupta S, *et al*. CancerDR: cancer drug resistance database. *Sci Rep* 2013;**3**:1445.
54. Liu T, Lin Y, Wen X, *et al*. BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic Acids Res* 2007;**35**:D198–201.
55. Irwin JJ, Sterling T, Mysinger MM, *et al*. ZINC: a free tool to discover chemistry for biology. *J Chem Inf Model* 2012;**52**:1757–68.
56. Halling-Brown MD, Bulusu KC, Patel M, *et al*. canSAR: an integrated cancer public translational research and drug discovery resource. *Nucleic Acids Res* 2012;**40**:D947–56.
57. Roth BL. Psychoactive Drug Screening Program, Contract No. HHSN-271-2008-00025-C (NIMH PDSP) 2008. http://pdsp.med.unc.edu/.
58. Liu Y, Hu B, Fu C, *et al*. DCDB: drug combination database. *Bioinformatics* 2010;**26**:587–8.
59. Giordano S, Petrelli A. From single-to multi-target drugs in cancer therapy: when aspecificity becomes an advantage. *Curr Med Chem* 2008;**15**:422–32.
60. Chen X, Ren B, Chen M, *et al*. ASDCD: antifungal synergistic drug combination database. *PLoS One* 2014;**9**:e86499.
61. Yamanishi Y, Kotera M, Moriya Y, *et al*. DINIES: drug–target interaction network inference engine based on supervised analysis. *Nucleic Acids Res* 2014;**42**:W39–45.
62. Nickel J, Gohlke B-O, Erehman J, *et al*. SuperPred: update on drug classification and target prediction. *Nucleic Acids Res* 2014;**42**:W26–31.
63. Gfeller D, Grosdidier A, Wirth M, *et al*. SwissTargetPrediction: a web server for target prediction of bioactive small molecules. *Nucleic Acids Res* 2014;**42**:W32–8.
64. Whitebread S, Hamon J, Bojanic D, *et al*. Keynote review: in vitro safety pharmacology profiling: an essential tool for successful drug development. *Drug Discov Today* 2005;**10**:1421–33.
65. Cheng AC, Coleman RG, Smyth KT, *et al*. Structure-based maximal affinity model predicts small-molecule druggability. *Nat Biotechnol* 2007;**25**:71–5.
66. Donald BR. *Algorithms in Structural Molecular Biology*. Cambridge, MA: MIT Press, 2011.
67. Morris GM, Huey R, Lindstrom W, *et al*. AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. *J Comput Chem* 2009;**30**:2785–91.
68. Campillos M, Kuhn M, Gavin A-C, *et al*. Drug target identification using side-effect similarity. *Science* 2008;**321**:263–6.
69. Cheng F, Liu C, Jiang J, *et al*. Prediction of drug-target interactions and drug repositioning via network-based inference. *PLoS Comput Biol* 2012;**8**:e1002503.
70. Hattori M, Okuno Y, Goto S, *et al*. Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. *J Am Chem Soc* 2003;**125**:11853–65.
71. Smith TF, Waterman MS. Identification of common molecular subsequences. *J Mol Biol* 1981;**147**:195–7.
72. Shi J-Y, Liu Z, Yu H, *et al*. Predicting drug-target interactions via within-score and between-score. *BioMed Res Int* 2015;**2015**:350983.
73. Bleakley K, Yamanishi Y. Supervised prediction of drug–target interactions using bipartite local models. *Bioinformatics* 2009;**25**:2397–403.
74. Yamanishi Y, Kotera M, Kanehisa M, *et al*. Drug-target interaction prediction from chemical, genomic and

pharmacological data in an integrated framework. *Bioinformatics* 2010;**26**:i246–54.
75. Mei J-P, Kwoh C-K, Yang P, *et al*. Drug–target interaction prediction by learning from local information and neighbors. *Bioinformatics* 2013;**29**:238–45.
76. Cao DS, Zhang LX, Tan GS, *et al*. Computational prediction of drugtarget interactions using chemical, biological, and network features. *Mol Inform* 2014;**33**:669–81.
77. Wang JT, Liu W, Tang H, *et al*. Screening drug target proteins based on sequence information. *J Biomed Inform* 2014;**49**:269–74.
78. Xia Z, Wu L-Y, Zhou X, *et al*. Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces. *BMC Syst Biol* 2010;**4**:S6.
79. Pahikkala T, Airola A, Pietilä S, *et al*. Toward more realistic drug–target interaction predictions. *Brief Bioinform* 2015;**16**:325–37.
80. Keiser MJ, Setola V, Irwin JJ, *et al*. Predicting new molecular targets for known drugs. *Nature* 2009;**462**:175–81.
81. Cobanoglu MC, Liu C, Hu F, *et al*. Predicting drug–target interactions using probabilistic matrix factorization. *J Chem Inform Model* 2013;**53**:3399–409.
82. Wang E, Zou J, Zaman N, *et al*. Cancer systems biology in the genome sequencing era: part 2, evolutionary dynamics of tumor clonal networks and drug resistance. *Semin Cancer Biol* 2013;**23**:286–92.
83. Wang E, Zaman N, Mcgee S, *et al*. Predictive genomics: a cancer hallmark network framework for predicting tumor clinical phenotypes using genome sequencing data. *Semin Cancer Biol* 2015;**30**:4–12.
84. Wang E, Zou J, Zaman N, *et al*. Cancer systems biology in the genome sequencing era: part 1, dissecting and modeling of tumor clones and their networks. *Semin Cancer Biol* 2013;**23**:279–85.
85. Metz JT, Johnson EF, Soni NB, *et al*. Navigating the kinome. *Nat Chem Biol* 2011;**7**:200–2.
86. Davis MI, Hunt JP, Herrgard S, *et al*. Comprehensive analysis of kinase inhibitor selectivity. *Nat Biotechnol* 2011;**29**:1046–51.
87. Park Y, Marcotte EM. Flaws in evaluation schemes for pair-input computational predictions. *Nat Methods* 2012;**9**:1134–36.
88. Chen X, Yan G-Y. Novel human lncRNA–disease association inference based on lncRNA expression profiles. *Bioinformatics* 2013;**29**:2617–24.
89. Ding H, Takigawa I, Mamitsuka H, *et al*. Similarity-based machine learning methods for predicting drug–target interactions: a brief review. *Brief Bioinform* 2014;**15**:734–47.