Briefings in Bioinformatics, 2017, 1–15

doi: 10.1093/bib/bbx017
Paper

A review of network-based approaches to drug repositioning

Maryam Lotfi Shahreza, Nasser Ghadiri, Sayed Rasoul Mousavi, Jaleh Varshosaz and James R. Green

Corresponding author. Nasser Ghadiri, Department of Electrical and Computer Engineering, Isfahan University of Technology, Isfahan 84156-83111, Iran. Tel.: +98-31-3391-9058; Fax: +98-31-3391-2450; E-mail: nghadiri@cc.iut.ac.ir; nghadiri@gmail.com

Abstract

Experimental drug development is time-consuming, expensive and limited to a relatively small number of targets. However, recent studies show that repositioning of existing drugs can function more efficiently than *de novo* experimental drug development to minimize costs and risks. Previous studies have proven that network analysis is a versatile platform for this purpose, as the biological networks are used to model interactions between many different biological concepts. The present study is an attempt to review network-based methods in predicting drug targets for drug repositioning. For each method, the preferred type of data set is described, and their advantages and limitations are discussed. For each method, we seek to provide a brief description, as well as an evaluation based on its performance metrics.
We conclude that integrating distinct and complementary data should be used because each type of data set reveals a unique aspect of information about an organism. We also suggest that applying a standard set of evaluation metrics and data sets would be essential in this fast-growing research domain.

Key words: drug repurposing; drug–target interaction; biological networks; network analysis; machine learning

Introduction

Drug research and development is a complicated, lengthy and expensive process. It often takes 10–15 years of research and 0.8–1.5 billion dollars to bring a drug from abstract concept to market-ready product [1]. Every year, ∼90% of drugs fail during FDA evaluations, preventing their use in actual therapy [2].

A drug typically involves a particular impact on one or more target proteins. Such effects often change the pharmaceutical function of the targets [3]. The rapid identification of targets can play a pivotal role during drug development because it can elucidate potential therapeutic properties [4]. Drug repositioning (DR) seeks to find new uses for existing drugs, with established

Maryam Lotfi Shahreza is a member of Data and Knowledge Research Laboratory and a PhD candidate in the Department of Electrical and Computer Engineering, Isfahan University of Technology, Isfahan, Iran. Her research interests include issues related to bioinformatics and computational biology, especially systems biology, complex networks and pharmacology sciences.
Nasser Ghadiri is an assistant professor and the director of Data and Knowledge Research Laboratory in the Department of Electrical and Computer Engineering, Isfahan University of Technology. His research interests include bioinformatics, complex networks, protein–protein interaction networks, medical text mining, data mining and graph mining. He has published >20 peer-reviewed papers in journals and conference proceedings.
Sayed Rasoul Mousavi earned his PhD in computing from Imperial College London. His research interests include bioinformatics, computational genomics, haplotyping, optimization and algorithms for computationally hard problems.
Jaleh Varshosaz is a full professor of Pharmaceutics and the director of Drug Delivery Systems Research Center of Isfahan University of Medical Sciences. She has been selected as one of the top 1% of the most-cited scientists in the world according to Essential Science Indicators ranking at 2015 and 2016 with >200 articles indexed by Thomson-Reuters ISI.[WorldCat]
James Green is an associate professor in the Department of Systems and Computer Engineering at Carleton University. His research interests include pattern classification challenges in biomedical informatics. Current research projects include the prediction of protein structure, function and interaction; the identification of microRNA in unique species; the design of novel assistive devices for persons with disabilities; unobtrusive patient monitoring; and the acceleration of scientific computing using parallel computing. Green has published >60 peer-reviewed manuscripts in journals and conference proceedings.
Submitted: 15 November 2016; **Received (in revised form):** 21 January 2017

and demonstrated human safety. In technical terminology, DR is the process by which one finds new indications for approved drugs [5].

The repositioning approach bypasses many of the pre-approval tests essential for newly developed therapeutic compounds. Such agents have already been proven to be safe for humans, and the DR process can shorten the drug development cycle to 3–12 years for a repositioned drug [6, 7]. In recent years, DR is receiving increased interest from governments, nongovernmental agencies and academic researchers.

DR is also known as drug repurposing, drug redirecting, drug retasking, drug reprofiling and therapeutic switching. DR is mainly applied to find new uses for existing or failed drugs, which have established safety and pharmacokinetic profiles [8]. A number of *in silico* tools have been developed to reduce the number of potential candidate target molecules, which should be screened and to discover, otherwise, unlikely novel targets [9]. We will use 'drug repositioning methods' hereafter to refer to *in silico* methods for DR.

The present article attempts to provide a survey of network-based DR approaches. The methods can be categorized into three general groups based on their principle source of biological data and core methodology, including gene regulatory networks, metabolic networks and molecular interaction networks. Additionally, the methods that are based on integrated networks are discussed in detail.

The next subsection describes the most significant benefits of DR, followed by a brief description of the different classifications of DR methods.

Advantages of DR

DR has demonstrated many advantages; chief among them is the drastic reduction in the costs and risks of drug development. Repositioned drugs can often enter clinical phases sooner, as there is reduced risk of unexpected toxicity, often leading to shorter overall development time lines [10–12].

DR has found important applications in disease and related therapeutic areas. For example, as the demand for anticancer drugs continues to increase, discovering novel anticancer therapies from available drugs has become increasingly popular [11]. DR is also helpful in orphan diseases, where only a limited number of people are affected. In such cases, there is an insufficient financial benefit to warrant full *de novo* drug development by the pharmaceutical industry [13]. DR seeks to identify novel therapies with significantly reduced development costs, thereby benefitting these orphan diseases. Additionally, drug resistance is one of the main reasons to reduce the drug efficacy; DR turns out to be a good approach to overcome drug resistance, for example the use of non-antibiotic drugs to overcome antimicrobial resistance [14]. Finally, DR provides new opportunities to advance personalized medicine [11]; it helps to improve drug efficacy by subtype-specific drugs [15], preventing drug failure due to lack of efficacy. The last example shows a synergy between DR and personalized medicine.

Classification of DR methods

There are different classifications for DR methods; each of which seeks to categorize the existing methods depending on some important metrics. Two relevant major *in silico* DR approaches are docking simulation and machine learning. 'Molecular docking' methods try to simulate and model physical interactions between drugs and targets [6] and are used in structural molecular biology and computer-assisted drug design. Successful docking methods can efficiently search high-dimensional conformation spaces and accurately rank the candidate dockings using a scoring function [16].

An excellent literature review of molecular docking and relevant theories is provided in [16]. There are some limitations, however, in the use of molecular docking in DR. The requirement of known three-dimensional (3D) structure of chemical ligands and protein targets severely limits the application of docking because the structures of many physiologically important proteins are not fully resolved [6]. Moreover, molecular docking methods demand significant computational resources resulting in extended runtimes [3]. Additionally, because of errors in the determined protein structure, and the incomplete modeling of atomic and molecular interactions, the results of molecular docking have high false-positive rates [6]. Machine learning methods appear more favorable than docking simulation, as they can examine a larger number of promising candidates for further experimental screening [3].

As described in [6], machine learning methods can be classified as either 'drug-based' or 'disease-based' methods. Drug-based methods try to discover repositioning opportunities by chemical or pharmaceutical perspective investigation, while disease-based methods focus on disease management, symptomatology or pathology. When more precise identification of pharmacological properties is required, drug-based approaches may be preferred, for which we need pharmacological or chemical data on drugs. In contrast, when there is insufficient knowledge of the drug pharmacology, disease-based approaches may be preferred. This is also true when our focus is on a particular disease or therapeutic category. Each approach presents unique informatics challenges, making it often necessary to incorporate elements from both drug- and disease-based methods for a successful process.

In another classification, Gonen [17] grouped traditional computational DR methods into three categories: (i) docking simulations, (ii) ligand-based approaches and (iii) literature text mining. Previous classifications were based on the data type used in each method. Another classification, which focuses on the models and methodologies used during DR [12], has categorized DR methods into data-driven and hypothesis-driven methods. This general classification could be applied to any computational approach. In the context of systems biology, the two approaches are referred to as top-down and bottom-up modeling, respectively.

Data-driven approaches analyze large-scale '-omics' data sets using statistical modeling techniques. The hypothesis-driven approaches, conversely, are applied to relatively small systems, which often have fewer molecular components. A significant challenge to hypothesis-driven methods is that the quantitative details of the interactions are required. We need to hypothesize the appropriate forms of the governing equations in the interactions. One of the major hypothesis-driven approaches is dynamical modeling which tries to find and describe the relationships between different components of an entity and its actions and reactions [12].

The type of data being used and the level that the biological system is under study can dictate which modeling approaches are most appropriate. Thanks to advances in biological sciences, we have access to a lot of '-omics' molecular data in different levels such as the genome, transcriptome, proteome and

metabolome; therefore, using data-driven approaches is an increasingly viable option.

One of the most frequently used data-driven approaches is network modeling. A network-based strategy to identify a drug target first reconstructs a biological network and then simulates its interactions. The resulting interaction relationships between drug targets can reveal potential drug targets. For example, an otherwise promising potential drug target may be excluded from consideration if it participates in many biological pathways, as any adjustment to its function will potentially affect its other activities leading to side effects [4].

The rest of the article is organized as follows. The section Molecular interaction networks discusses network-based approaches to DR. Methods are categorized according to the biological networks used. A numerical comparison of some of the major DR and drug–target prediction methods is provided in Section Comparison. Finally, the Conclusion and discussion section summarizes the advantages, limitations and future perspectives of network-based approaches to DR.

Molecular interaction networks

Networks are simple data structures from which different informative associations can be discovered via statistical and computational means [1]. A wide variety of concepts in biology are represented in the form of networks, and there has also been considerable interest in the investigation of the structure of such networks and the relationship between the networks and their underlying biological properties [18].

In biological networks, which are usually used to model interactions between many different biological concepts, nodes represent various components, from primitive ones, like genes and proteins, to complex phenotypes such as diseases. On the other hand, edges can be used to show any paired biological concept, such as the relationship between a gene and a protein or the functional similarity between genes. Edges may also represent multiple types of relationships simultaneously [12, 19]. These nodes and edges can be weighted with qualitative or quantitative information to emphasize specific concepts [20].

Biological systems consist of various molecular interactions, which can be represented as distinct molecular networks, depending on the nature of the interactions. The molecular networks can provide insights into the context in which the drug target works and can, therefore, help understand the drug mechanisms of action [21]. Previous studies have proved the usefulness of network structures in the identification of drug–target interactions (DTIs). Thus, they can also be used for DR [11]. Network algorithms can readily accomplish such tasks as visualizing various existing interactions, adding newly discovered relationships, and superimposing additional properties over primary components and their known interactions [20]. Various kinds of data from different data sets could be represented in one network. Therefore, its analysis has become a versatile platform for integrating multiple sources of high-throughput data and link data sets [20, 22]. A survey of the use of graph theoretical techniques in biology is provided in [17].

Some of the most significant limitations of these molecular origin-based repositioning strategies can be summarized as follows: our current knowledge about the molecular interactomes at different levels is far from complete, and corresponding profiles are noisy [8, 21]. Furthermore, these interactomes provide only static snapshots of the biological systems, whereas, in reality, we have dynamic systems [21]. The gathered information on the detailed interaction kinetics is limited, and there is no simple mapping between casual molecular origin and the living organism's response. It is, thus, the interplay between physiological environment and casual molecular origin that finally determines the outcomes of the disease development and drug treatment [8, 21].

Although these limit the application of the molecular network, previous studies have demonstrated the efficiency of networks and their analysis in DR. This led us to investigate network-based DR methods in greater detail. In the rest of this section, some notable aspects of the useful biological networks in DR are provided, and a review of different network-based DR methods is presented.

Gene regulatory networks (DNA–protein interaction networks)

Transcriptomic data can capture the dynamic properties of a cell and can provide insights into the mechanisms of a drug's function [21]. Gene expression patterns are known to change systematically in response to disease. The amount of messenger RNA transcripts for some (dysregulated) genes varies substantially between disease and control samples, which can be detected by differential analysis of gene expression [23]. Thanks to well-established technologies, such as gene microarrays, gene expression profiles are readily available and provide a rich source of disease expression data applicable to a variety of purposes. Observing as many drug targets serve as transcription factors (the proteins that interact with DNA to modify gene expression), drug targets are likely to correspond to expression regulators. Therefore, differential gene expression profiles may be used as input to prioritize potential drug targets [24]. The methods that are based on transcriptome data assume that drugs with similar gene expression signatures would target the same proteins.

Table 1 provides a brief review of some major network-based methods, which use gene expression data in DR.

Although the methods provided in Table 1 have all been demonstrated to be effective, some limitations remain for those based on gene expression signature comparison. First, one may face difficulty in defining a robust gene signature because of the existence of noise in some gene expression data, resulting in biased extracted responsive networks [8, 21]. Second, the genes used as drug targets and the genes regulated by a target (in the case that drug target is a transcription factor) may not always have significant expression changes. Thus, the assumption that the mathematical optimum of the responsive networks equals the maximal biological relevance is not necessarily correct. Third, previous studies of network topology characteristics would reveal that there is no significant correlation between potential target proteins and critical points in a network [25]. Fourth, because of the complexity of mapping between a responsive network and a living organism's response, the methods in this category often rely on estimated gene networks [4, 8].

Such limitations lead us to conclude that the integration of different information from different sources, such as molecular interaction networks, with gene expression profiles, would be required for robust DR.

Table 1. A summary of network-based DR methods using gene expression data

Name [Reference]	**Methods**[description]	Data sets	Case studies	Evaluation criteria
[24]	Neighborhood scoring, interconnectivity, network propagation, random walks[a]	Gene Expression Omnibus (GEO) repository and integrity	Scleroderma, different types of cancer and diabetes type 1	AUC[b]
[25]	Maximum flow[c]	DrugBank, Online Mendelian Inheritance in Man (OMIM), KEGG and PGDB	Prostate cancer	Mean average precision and average position (AP)
NFFinder [13]	Statistical analysis[d]	GEO, CMap and DrugMatrix	Neurofibromin	–
[26]	Bayesian networks[e]	Genetic interactions	Mammary epithelial carcinoma cell proliferation and breast cancer	–
[27]	Virtual gene technique, Bayesian networks[f]	Gene disruptant microarray data and time-course drug response microarray data	*Saccharomyces cerevisiae*	–
ksRepo (Kolmogorov–Smirnov enrichment testing for computational DR) [9]	Kolmogorov–Smirnov enrichment[g]	Comparative Toxicogenomics Database (CTD) and GEO	Prostate cancer	–
MFM (method of functional modules) [28]	Functional linkage network[h]	A drug response expression data set (The Library of Network-Based Cellular Signatures (LINCS) profiles), CMap, DrugBank, OMIM, GEO and The Cancer Genome Atlas (TCGA) portal	Breast, prostate and leukemia cancers	AUC

[a]They ran four mentioned methods separately for 30 different diseases and then combined their predictions by using a logistic regression model. Analyzed diseases were clustered two once by predicted drug targets and differentially expressed genes, and it is indicated that diseases with similar gene expressions will have similar drug targets.
[b]Area under the curve (AUC) of the ROC curve.
[c]This method is based on this assumption that regulatory adjustments actions and induced changes in cells have direct effects on gene expression.
[d]NNFinder is an online tool, which tries to identify drug–disease relations by transcriptomic data. It is accessible from http://nffinder.cnb.csic.es. Users should define lists of desirable upregulated and downregulated genes; NNFinder will compare them with its internal data set consisted of drugs- and diseases-related gene signatures.
[e]This method proposed a knowledge-driven systems biology method to construct a Dynamic Bayesian network. Chang *et al.* believe that biological network topology could be preserved by this model in addition to its ability to capture joint and conditional probabilities in the Bayesian network.
[f]The proposed model is a multilevel directed acyclic graph, which consisted of genes and drugs regarded as virtual genes.
[g]It is an open-source platform for DR. It could be run with any arbitrary pairs of expression data (drug exposure expression and disease expression), which have a common identifier system.
[h]Chen *et al.* tried to score correlations of upregulated and downregulated gene expressions in disease state and drug target exposure and used this score to predict drug candidates for repositioning.

Metabolic networks

In a metabolic network, the nodes represent chemical compounds and metabolites. Directed edges denote the reactions catalyzed by one or more enzymes. In an alternate representation of metabolic networks, each edge accounts for a reaction between two physical entities (nodes). In this way, a directed bipartite graph with two types of nodes (i.e. enzymes and metabolites) will be provided. If a metabolite is a direct product of a reaction, there will be a directed edge from that reaction to the metabolite. An edge from a metabolite to a reaction means that the metabolite activates the reaction. A reversible reaction corresponds to an undirected edge (or two complementary directed edges). This representation provides us with the opportunity to use topological features to describe different kinds of relations [29].

In other words, excessive concentration (mass flow) of a compound, caused by certain enzymes, may lead to a disease. These enzymes could be considered as drug targets of this disease, as their manipulation by drugs will adjust the concentration of disease-causing compound [29].

Flux balance analysis (FBA) is an important class of methods to identify drug targets. Such methods usually try to predict essential enzymes, which are critical to the survival and growth of pathogens. FBA is a constraint-based approach to optimize an objective function by linear programming. FBA consists of four steps: (1) system identification that includes modeling and determination of all reactions and metabolite of the corresponding system, (2) conversion of reactions to matrix formalism (this will facilitate the investigation of the properties of network), (3) definition of objective function and its related constraints (the objective function is often a biomass production determined by critical metabolites, and constraints should cover different aspects such as mass balance, environmental and topological features) and (4) finding the optimal solution of objective function [30, 31]. For example, to use FBA to identify antimicrobial drug targets, the objective function should follow the conditions related to growth of a pathogen [31].

Li *et al.* [29] developed a FBA model on metabolic networks to predict DTIs. Their solution is based on two-stage linear programming. They retrieved the required data from Kyoto Encyclopedia of Genes and Genomes (KEGG) and tried to detect drug targets of human hyperuricemia. They also presented their computational analysis on a simulated metabolic network. Folger *et al.* [32] used a greedy search heuristic approach on

Table 2. A summary of the network-based DR methods used PPINs

Name [Reference]	Methods	Data sets	Case studies	Evaluation criteria
[33]	Support vector machines (SVMs), L1-regularized logistic regression, k-nearest neighbors[a]	Human PPIN data set UniHI, DrugBank and GeneCards database	–	Z-score and SD
[23]	Set-cover based formulation of coordinate dysregulation in complex phenotypes[b]	GEO	Predicting metastasis of colon cancer	Harmonic mean of PR
[34]	Similarity comparison[c]	Genotator, DrugBank and STRING	Hypertension, diabetes mellitus, Crohn disease, autism	–
[35]	Cross talk by analysis of betweenness centrality[d]	KEGG, OMIM and iRef Index database	Parkinson's disease	–

[a]Zhang *et al.* compared some PPIN's topological features and introduced best ones suitable for DR purposes.
[b]Koyuturk defined different disease dysregulations and focused on identifying dysregulated subnetworks through the integration of gene expression and PPI data.
[c]It has two steps. In the first step, it built a PPIN by use of shared disease genes, and then found related drugs to each disease; in this step, each drug that has at least one target in constructed PPIN will be considered as a DR candidate. The second step applies a similar process for drugs with no known targets; this step is based on similarity of drugs.
[d]First a PPIN is created by betweenness centrality, and then try to find key points for decreasing neurotoxicity.

metabolic networks to predict drug targets of lung cancer. Chavali *et al.* [30] reviewed methods for identification of antimicrobial drug targets that used metabolic network modeling based on FBA. They also discussed the advantages and disadvantages of the methods.

Protein–protein interaction networks

Protein–protein interaction networks (PPINs) are an important category of molecular interaction networks that represent interactions between known drug targets and other proteins, or between proteins that have indirect interaction with targets [22]. The central assumption of most methods that use PPIN to predict DTIs is that the proteins targeted by similar drugs are functionally related and are 'close' in the PPIN [21]. On the other hand, it is believed that topological analysis of PPINs could facilitate drug–target prediction, as proteins do function in the context of interaction networks. It means that proteins rarely work in isolation [33]. PPINs could be used to analyze the global organization of cells by providing a comprehensive map of functional interactions in the cell [23].

Table 2 demonstrates some primary PPIN-based methods in DR. Studies on several complex diseases show that dysregulated genes in similar diseases have more probability to interact with each other in PPINs [4]. Thus, researchers have developed methods to identify dysregulated subnetworks. Dysregulated subnetworks are connected sub-graphs of the human PPIN that exhibit collective differential expression with respect to the disease phenotype [23].

As searching the entire subnetwork space of a PPIN leads to intractable computational problems, development of sophisticated computational algorithms would be required to discover dysregulated subnetworks efficiently [23].

PPINs for repositioning drugs suffer from some limitations, despite the great successes they have achieved. As PPINs are derived from a diverse range of experimental sources, they contain potential functional links, which have not been yet characterized in detail. Furthermore, the required data are noisy and incomplete, resulting in a biased extracted network. Moreover, as in gene regulatory networks, there is no simple mapping between a responsive network and the living organism's response [8]. Again, we conclude that an integration of data derived from various sources would be necessary for effective DR.

Drug–target interactions

Drug discovery and design are primarily based on DTI. Many drugs are nonspecific and show reactivity to additional targets besides their primary targets. Clearly, the DR task is simplified if the drug targets are accurately predicted. Experimental determining of DTI is both time-consuming and resource-consuming. Hence, it is necessary to develop computational methods to predict the potential DTIs [8, 36].

Different methods exist for predicting potential DTIs, many of which use a network representation. In network-based models, a bipartite interaction network is constructed where nodes represent drugs and targets, and edges denote interactions. These interactions are based on aggregation of multiple types of pharmacological and clinically relevant associations.

Table 3 describes some important DR methods that make use of DTIs. Some of these methods are based on DTI, and some are extended by using, for example, protein–protein similarity and drug–drug similarity. These extended models tend to outperform the earlier models. For example, when a given drug has known targets, the candidate targets could be ranked by measuring the similarity between them and known targets based on protein similarity. If the drug has no known target, it is not sufficient to only base our judgment on target similarity; hence, drug analogy must also be taken into account. In this case, the potential targets of the given drug are selected based on the target information of drugs similar to the given drug. In this regard, Yamanishi *et al.* [39] demonstrated that if two drugs have similar structure, their chance to interact with similar target proteins will be higher. Likewise, two target proteins with high sequence similarity are more likely to interact with similar drugs.

Despite good performance observed in several methods introduced in Table 3, most suffer from a significant limitation: they are based on training data and, hence, are not able to accurately predict drug/target candidates, which are entirely new. A drug-candidate compound is called new if it does not have any known targets, and a candidate target protein is called new if it is not targeted by any known drugs/compounds. It would be problematic in practice, as a significant number of compounds and proteins are considered new by these definitions.

Some difficulties of developing computational methods for the prediction of such potential interactions can be summarized

Table 3. A summary of network-based DR methods based on DTI

Name [Reference]	Using methods	Using data sets	Case studies	Evaluation criteria
PSL (probabilistic soft logic) [37], [38]	Probabilistic soft logic[a]	Data sets of [39][b], DrugBank, KEGG Drug, DCDB and Matador	–	AUC, AUPR[c], P@n (average precision of the top *n* interaction predictions)
DBSI (drug-based similarity inference), TBSI (target-based similarity inference) and NBI and (network-based inference)[d] [36] [41]	Recommendation algorithms[e]	Data sets of [39], DrugBank	Compound purchase, dipeptidyl peptidase-IV inhibition assay, yeast two-hybrid system-based assay and Microculture Tetrazolium (MTT) assays	AUC, PR. (NBI performed best.)
DT-Hybrid[f] [42, 43]	Recommendation algorithms[g]	Data sets of [39], DrugBank	Compound purchase, dipeptidyl peptidase-IV inhibition assay, yeast two-hybrid system-based assay and MTT assays	ROC and AUC
KBMF2K (kernelized Bayesian matrix factorization with twin kernels) [17]	Bayesian formulation[h]	Data sets of [39]	–	AUC
KRM (kernel regression-based method) [39]	Supervised bipartite graph learning approach[i]	KEGG BRITE, BRENDA, DrugBank, KEGG LIGAND, KEGG GENES and SuperTarget	Predicted enzyme interaction network and predicted GPCRs interaction network	ROC (performed 10-fold cross-validation procedure), AUC, sensitivity, specificity and PPV (positive predictive value)
BLM (bipartite local model) [44]	Supervised inference method[j]	Data sets of [39]	–	ROC curve, PR curve, AUC >97% in some cases and AUPR scores of up to 84%
BLM-NII (BLM with Neighbor-based Interactionprofile Inferring) [45]	Neighbor-based inferring integrated with BLM[k]	Data sets of [39]	–	AUC and AUPR
[46]	Statistical analysis of topological measures[l]	MalaCard online system and STITCH	Cystinosis	–
NetCBP (Network-Consistency-based Prediction method) [47]	A semi-supervised inference method[m]	Data sets of [39]	–	ROC and AUC
[48]	Bipartite network projection, kernelized score functions[n]	DrugBank and STITCH	–	ROC and AUC
WNN (weighted nearest neighbor) [49]	Simple WNN procedure[o]	Data sets of [39]	–	AUC and AUPR
WNN-GIP (weighted nearest neighbor-Gaussian interaction profile) [50]	Classification by kernel (regularized least squares)[p]	Data sets of [39]	–	AUC and AUPR
LPMIHNs (label propagation with mutual interaction information derived from heterogeneous networks) [51]	Label propagation algorithm[q]	Data sets of [39]	–	AUC and AUPR

(continued)

Table 3. Continued

Name [Reference]	Using methods	Using data sets	Case studies	Evaluation criteria
LapRLS and NetLapRLS (Laplacian regularized least square and Net LapRLS) [52]	A semi-supervised learning method based on Laplacian operation[r]	Data sets of [39]	–	AUC, AUPR, sensitivity, specificity and PPV
MSCMF (multiple similarities collaborative matrix factorization) [53]	Weighted low-rank approximation in conjunct with nonnegative matrix factorization[s]	Data sets of [39], one synthetic clustering data	–	AUPR

[a]PSL is based on a drug–target bipartite graph augmented with drug similarity and target similarity. It provided a kind of weight learning based on approximate maximum likelihood. By using of similarity predicates, the number of rules was adjusted.

[b]A data set contains drug, protein targets and their interactions, which are categorized by four groups of protein targets (enzyme, GPCR, ion channel and nuclear receptor). The primary resources of these data are KEGG, BRITE, BRENDA, SuperTarget and DrugBank.

[c]AUPR curve.

[d]In fact, NBI is first proposed by Zhou *et al.* [37], then Cheng *et al.* [36] introduced it and after that Alaimo *et al.* [40] applied NBI algorithm and extended it as Domain Tuned-Hybrid (DT-Hybrid) by adding similarity among drugs and targets.

[e]DBSI is a kind of item-based collaborative filtering (CF) used chemical structural similarity. TBSI is a kind of user-based CF used genomic sequence similarity. NBI method used mass diffusion on known drug–target bipartite network.

[f]DT-Web is a Web interface to the DT-Hybrid.

[g]DT-Hybrid is available as an R package at http://sites.google.com/site/ehybridalgo/. In addition to DTIs, it is augmented by structural similarity of drugs and sequence similarity of targets. The recommendation technique used in DT-Hybrid is one-mode projection.

[h]As the first step, KBMF2K constructed a pharmacological space by projection of drugs and targets into low-dimensional spaces separately using two distinct kernel matrices. In the next step, it would predict DTIs in three different scenarios.

[i]KRM also provided a pharmacological space by integration of drugs and targets data into a unified space by using of Gaussian functions and then tried to predict drug–target relations by using of a kind of kernel regression. Yamanishi *et al.* gathered and used four classes of DTIs (enzymes, ion channels, GPCRs and nuclear receptors), which are used as gold standard data sets in a lot of next DR approaches.

[j]BLM is a supervised learning method based on KRM. It in fact used a classification rule two times, once to predict targets of a drug and other once to predict related drugs of a target and then combined results of both.

[k]It is a semi-supervised version of BLM. BLM could not predict drugs and targets of new ones, BLM-NII tried to solve this problem by merging of BLM with a neighbor-based method. Neighbors are identified by similarity profiles of drugs and targets separately.

[l]It tried to identify potential targets of lysosomal diseases by using of some graph-based analysis tools.

[m]It is a semi-supervised learning method based on graph Laplacian. A vector of graph Laplacian score was computed for each drug and target and then potential DTIs would be predicted by measuring the network consistency with known DTIs.

[n]Re *et al.* first integrated different heterogeneous data into a homogenous network and then used kernelized score functions to rank drugs. In this regard, drug similarity is measured by three different resources, and various network integration methods are used.

[o]It defined two kinds of interaction profiles for drugs and targets and then applied a normalized Gaussian kernel on them and finally tried to find potential DTIs by different classifier methods such as Regularized Least Squares-Kronecker product kernel (RLS-Kron).

[p]The main purpose of this method is to extend Gaussian interaction profile by ability of interaction detection for new drugs and new targets.

[q]LPMIHN is a semi-supervised learning method. It is constructed of two independent label propagations on a heterogeneous drug and target network, and thus it could use topology information of the network effectively.

[r]LapRLS is a semi-supervised learning method. First, it constructed two similarity domains for drugs and targets separately. Drug similarity domain is a linear combination of chemical similarity and number of common targets between two drugs. Target similarity domain is consisted of sequence similarity and number of common drugs between two targets. In second step, two classifiers were used for prediction of DTI, and their results were integrated. NetLapRLS is different from LapRLS in classifying section; it tried to minimize a cost function by a continuous classification function.

[s]It is in fact a kind of CF. First of all, it combined linearly different matrices of drug similarity and target similarity to achieve one-drug similarity matrix and one-target similarity matrix, then defined an objective function based on squared error minimization and finally provided a matrix of DTIs by inner production of two estimated matrices.

as follows: first of all, known DTIs are relatively rare, and the performance of prediction should be assessed using new drugs without any known target interaction information. Second, the selection of negative samples is difficult or even impossible because of the scarcity of experimentally verified negative DTI exemplars [47]. Possible solutions could be found for both problems. For solving the first problem, an acceptable yet simple methodology is presented by Alaimo *et al.* [54]. They proposed to use targets of highly similar drugs to a new drug as its initial targets, and in the same way, to use drugs of highly similar targets to a new target as its initial drugs. Then, they applied the intended method. A limitation of this methodology is how to define 'highly similar' drugs or targets. For solving the second problem, a solution is a random selection of negative samples from unverified relations [3]. Many of these unverified data are, in fact, undiscovered DTIs, so it could not be a reasonable solution.

Drug–drug interactions

Drug–drug interactions in this context are used to clarify the association between two drugs based on their similarity. As mentioned earlier, drug–drug and target–target similarities can augment both sides of the drug–target network. The similarities between drugs signify different meanings, such as chemical similarities or biological effect similarities. These similarities may also be established by many different methods using different structural features, including two-dimensional (2D) topological fingerprints or 3D conformations.

For example, if two drugs are shown to induce similar molecular profiles, for example gene expression patterns in cultured cells, or if the drugs have been reported to have similar side effects, an interaction could be said to exist between them. The drug chemical structures can also point toward repositioning opportunities. The central hypothesis of these approaches is that the molecules with similar chemical structures often affect the proteins and biological systems in similar ways. However, similar structures do not always lead to the same biological behavior. In general, the degree of similarity between drugs can be exploited using computational approaches for DR [6]. Incorporating the similarities between drugs into the repositioning inferences is a profound research domain to be deeply investigated in future.

In this regard, Hattori *et al.* [55] compared two chemical compounds by a 2D graph. The nodes of this graph were atoms, and the edges corresponded to covalent bonds. Compounds with a higher mutual chemical similarity have greater chance to enjoy the same bioactivity [50]. The methods rooted in chemical similarity often extract a set of chemical features for each drug in a set of drugs and then interrelate the drugs using the extracted features. DR opportunities can, thus, be uncovered through simple chemical association or search for some common biological features, such as the known drug targets [6].

In addition to the methods in Table 3, other repositioning methods have leveraged drug–drug similarity, including similarity-based Inference of drug-TARgets (SITAR) [56] and Mode of Action by NeTwoRk Analysis (MANTRA) [57].

SITAR is a logistic regression classifier. It used five drug similarities and three target similarities. Drug similarity measures are based on chemical structure, side effects of the drug, gene expression profiles and the Anatomical, Therapeutic and Chemical classification system. Target similarity measures were computed by sequence similarity, closeness in PPIN and semantic gene ontology similarities. It tried to identify which pairs of similarities could achieve the best prediction results. Its data resources are KEGG Drug, DrugBank, DCDB and Matador.

MANTRA is a DR approach based on community detection. Its main idea is to predict drug affects similarities and their mechanism of action (MOA). Connectivity Map (cMap), DrugBank and ChemBank are its basic data resources. It provided a case study on anticancer compounds.

One of the main limitations of DR based on drug–drug similarity is the untrustworthy state of many structures and chemical properties of known drug compounds. Furthermore, many physiological effects cannot be predicted by structural features alone [6]. A multitude of similarity measures, like the one used in [58], can be applied to overcome these limitations.

The similarity-based approach is a typical strategy for drug-drug interaction (DDI) prediction. However, most similarity-based DDI prediction algorithms rely solely on immediate similarities while overlooking the transitivity of similarity.

Drug–disease associations

Most studies have focused on the association between a single gene and a single disease. In contrast, the systems biology approaches apply network-based tools to enable a better understanding of the relationships among multiple genes and diseases [12].

Computational assessment of similarities in molecular profiles is a way through which drugs can be linked to disease states to reposition [6, 59]. If a pharmacologically active compound is exposed to a biological system, the system will be perturbed through the MOA of the compound [6].

The availability of many public repositories, such as the Drug versus Disease, the Database for Annotation, Visualization and Integrated Discovery and the Gene Set Enrichment Analysis, provides the opportunity to compare the drug and disease signatures in gene expression profiles [59].

So, it is possible to construct a 'signature' of the molecular activity of a pharmaceutical compound acting in a biological system. A comparison of these signatures of molecular activity reveals therapeutic relationships between drugs and diseases, even when a drug's MOA or a primary target is unknown [6]. Table 4 lists some of the most important methods in this regard.

Drug–adverse effect associations

In addition to their primary desired effects, many drugs induce some unintended effects on the living organism, which together constitute a drug's overall effect profile. Those wanted or unwanted behavioral, or physiological changes can be evaluated as the drug's indications and side effects, respectively [8, 62]. When predicting a drug's targets, as opposed to the therapy information, little attention has been paid to their adverse effects. It is known that binding of drugs generates side effects to off-targets, which perturb unexpected metabolic or signaling pathways. These off-targets may also help to predict therapeutic targets [11, 21].

Drugs with similar side-effect profiles may share similar therapeutic properties in their relevant mechanisms; thereby, we can connect drugs to diseases [6, 11]. Analogous to the approaches based on gene expression profiles, the drug effect-based approaches assume that the drugs with similar therapeutic effect may target the same protein(s) [21].

Furthermore, the phenotypic expression of a side effect can be akin to that of a disease, implying that the underlying

Table 4. A summary of network-based DR methods based on drug indications and drug side effects

Name [Reference]	Using methods	Using data sets	Case studies	Evaluation criteria
ProbS and HeatS (probabilistic spreading and heat spreading) [10]	Computational inference[a]	CTD	Three drugs (felodipine, aspirin and tamoxifen)	AUC and ROC
PREDICT [60]	Logistic regression classifier[b]	DrugBank, OMIM, DCDB, Matador, KEGG DRUG, DailyMed, SIDER, UniProt and Gene Expression Atlas of ArrayExpress	–	AUC
[61]	Two clustering algorithms (ClusterONE and Louvain's modularity)[c]	KEGG Medicus, DrugBank and NCBI's Entrez Gene	Gorlin syndrome, Alzheimer and hidradenitis suppurative	–
[11]	Statistical analysis[d]	Meyler's Side Effects of Drugs (15th edition), FDA drug approval package, Side Effects of Drugs Annuals (2007–12), SIDER, Citeline Pipeline, Thomson Reuters Partnering and GeneGo	Dynastat, Tasmar and Adamon	Normalized discounted cumulative gain, Expression Analysis Systematic Explorer score

[a]The main idea of these methods is prediction of drug–disease associations by means of network topology. They are based on the recommendation methods proposed by Zhou *et al.* [41]. ProbS has better performance than Heats, and both of them cannot work perfectly for new drugs.
[b]PREDICT tried to predict drug–disease associations for both drugs with known indications and new ones. The main idea is to investigate the similarity of potential drug–disease associations with known associations. In this regard, it used a gold standard of drug–disease associations constructed from UMLS, DrugBank and OMIM.
[c]This method is built on a weighted heterogeneous network, which represents different associations between drugs and diseases such as shared genes and phenotypes and then applied graph clustering methods on this network.
[d]In Step 1, it created a feature vector of side effects for each drug, and then in Step 2, it calculated similarity of drugs based on these vectors. In Step 3, calculated similarities are evaluated, and Step 4 optimized some needed thresholds for construction of drug–drug network. Based on neighborhood in this drug–drug network, new candidates for DR were proposed.

pathways or physiological systems may be perturbed in the same way by both the drug and the disease condition. This phenomenon underlies the interrelation of drugs or diseases by side-effect profiles, even when the precise pharmacological mechanism facilitating the side effect is not recognized [6].

The pharmacological information associated with drugs, then, introduces an alternative predictive way for the drug targets and has been proven to provide complementary information to molecular data, for example genome sequence or transcriptome data [21]. A drug network can be constructed based on the similarities in side effects. In this way, the indications of a drug may be predicted by the functional distribution of its neighboring drugs [11].

Only a few repositioning efforts to date have focused on physiological responses. Most of these methods leverage the side-effect data from SIDER [63] as a primary source of the known side effects. However, the limited quantity of drugs in SIDER may impact the DR process [11]. Furthermore, it is important to remember that clinical trials are conducted on relatively small patient populations, and the effects observed during the clinical trials may be incidental and not caused by the drug [64].

The method presented in the last row in Table 4, and the PREDICT method [65] used drug side effect data to rank drug–disease associations.

Unfortunately, the approaches mentioned above are limited to well-studied drugs because of the of adverse reaction information scarcity. In fact, the side-effect similarity methods require a distinct side-effect profile for each drug; however, the current disease and drug phenotype data are noisy and far from complete. The side-effect profile for a newly approved drug may only be fully discerned after years of clinical use and post-market surveillance [6, 21]. In the absence of such reporting, it is often necessary to predict a drug's side effects [8].

Also, in some cases, drugs with similar side effects do not have any common target proteins. This would give rise to prediction methods, which instead determine a drug's side effect based on the similarity of molecular structures of their targets. Also, there is no simple mapping between phenotype and MOA. In reality, the phenotypical outcomes of drug's mode of action are highly dependent on the organism's genetic map, medication history and other traits. Therefore, a similar phenotype does not necessarily correspond to the same mode of action [8, 21].

The other apparent limitation of the side-effect similarity approach is that partial side-effect information collected by spontaneous reporting systems during post-marketing surveillance is also confounded by individual patients' medication histories or traits and other hidden factors [8].

Disease–disease interactions

In DR, it is assumed that two drugs with similar molecular pathophysiology are interchangeable [6]. To reposition drugs, then, some computational strategies seek molecular relationships between distinct disease pathologies. In such approaches, repositioning opportunities arise when diseases that exhibit similarity at the molecular level are found, even when there is no apparent similarity at the phenotypic or clinical level. Therefore, the drugs might be shared among diseases with high similarity in their molecular activities.

The usefulness of these approaches will be limited by the ability to measure and represent the molecular pathology underlying the disease. The process of modeling the molecular state of diseases by computational methods is complicated because of the abundance of different incorporated molecular entities and organ systems in a disease pathology [6]. There are a few methods based on only disease–disease relations.

Multiple target optimal intervention [62] is one of these methods, based on Monte Carlo simulated annealing. It tries to identify effective points of intervention and the combination of interventions that can best restore the disease network to a desired normal state. It provides a case study on the arachidonic acid metabolic network.

Integrated network-based methods

Every type of link in a network represents specific information about an organism. For example, protein–protein interaction (PPI) data represent which proteins can physically interact, but it does not suggest any information about how an organism will react to stimuli. However, gene expression data can reveal how an organism responds to stimuli regarding the amount of RNA produced but does not contain information about the physical mechanisms by which changes in the organism's behavior are manifested. Therefore, the integration of distinct types and supplementary data would seem necessary [19, 21].

As detailed in Table 5, some methods are suggested that integrate different biological networks to reposition drugs. These approaches have a common shortcoming, which is the fact that each prediction is made separately based on separate metrics, and then these metrics are aggregated to arrive at the final results.

A class of integration methods is based on multiple relationships, including the methods used in literature mining-based DR. A well-known model in this class is the ABC model that provides the opportunity to detect the relationships between two concepts without any explicit communications. Suppose that through a data source, we know that C is a disease and B is one of its characteristics. Another data source provides us with the knowledge that A (e.g. an individual drug) has special effects on B. In this case, we might consider an implicit relation between A and C. There are two different processes for discovery based on ABC model, the open discovery model and the closed discovery model. In the closed discovery model, we have two known concepts A and C, and we try to find an implicit relation between them by a third concept like B. In the open discovery model, our desirable concept is A. In this case, the process consists of two stages. The first is to find the concepts related to A, named B, and in the second step, we try to find the concepts related to B, named C. It should be noted that in both cases, open and closed, it would be possible to have multiple relationships between concepts A, B and C. This will give rise to providing scores to measure the strength of relation between A and C [76–78]. Different methods were also created successfully by these models such as CoPub [76] that retrieves hidden relationships between drugs, genes, pathways and diseases to facilitate DR. Another method is proposed by Yang *et al.* [78] that finds disease–gene and gene–drug relations by using of ABC model with the aim of identification of anticancer drugs candidates for repositioning. Andronis *et al.* [77] also reviewed different methods based on two discovery models in the domain of DR.

Although there are many benefits to combining different heterogeneous data, it may reinforce bias in network models. For example, various research works on networks have reported that many biological networks appear to have scale-free topology. Networks inferred on this criterion will systematically overlook possible networks with alternative architectures. It has been shown that many of the observed scale-free topologies in biological networks are, in fact, not scale-free [19].

In addition to the network types described above, there are some other rarely used molecular interaction networks including target–ligand interaction used by Liu *et al.* [58] and Jacob *et al.* (pair kernel method) [40]. All in all, there are many methods based on verified molecular origins augmented with predicted molecular origins to reveal a drug's new roles. The assumption behind all these methods is that drugs with similar molecular origins can combat diseases with similar molecular origins. Therefore, indications transferred between drugs and medications can also be transferred among diseases. The key to these computational methods is, then, how to measure the similarity at a molecular origin level from the data sources at hand. In network-based methods, this translates into a question of how to capture similarities of molecular origins among drugs and diseases at a network level [8].

It is worth mentioning here that there are a few surveys that discuss different aspects of drug discovery. These include Ding *et al.* [3] who described the similarity-based machine learning methods, which have tried to predict DTIs. Dai *et al.* [21] presented different computational methodologies for identifying drug targets. To this end, they introduced some available data sets about drug target information. Finally, Lee *et al.* [79] investigate different aspects of drug development and DR for neuropsychiatric disorders.

Comparison

As discussed in Section Molecular interaction networks, most methods evaluate their prediction performance by using area under the curve (AUC) and area under the precision–recall (AUPR). A receiver operating characteristic (ROC) curve is the plot of the true positives (TPs) as a function of the false positives (FPs) based on various decision thresholds, where TPs are correctly predicted interactions and FPs are predicted by the method but are not true interactions. Precision is defined as the fraction of TPs among all targets predicted to be positive by the method, and recall is the fraction of truly predicted targets identified over the entire set of the all valid targets.

AUC usually represents the overall performance of the algorithm. However, previous studies have shown that in scale-free networks, such as biological networks, PR curves are more informative because of the impact of skewed edge distributions on the performance of prediction algorithms [8]. In the case of an uneven number of true and false examples, the accuracy of PR curves will be increased. Furthermore, a curve dominates in ROC space if and only if it dominates in PR space [47].

Yamanishi *et al.* [39] collected a set of DTIs, drug similarities and protein–target similarities as a 'golden standard' data set. As shown in Table 3, many methods have used these data, thereby providing an opportunity for more accurate comparison between methods. Self-reported AUC and AUPR scores of some methods using this same gold standard data set are presented in Table 6. In this gold standard data set, interactions between drugs and target proteins are obtained from the KEGG BRITE (http://www.genome.jp/kegg/brite.html), BRENDA (http://www.brenda-enzymes.org/), SuperTarget (http://insilico.charite.de/supertarget/) and DrugBank (http://www.drugbank.ca) databases; drug similarity data are computed by SIMCOMP on their chemical structures; and protein–target sequence similarities have been calculated using a normalized version of Smith–Waterman scores. Chemical structure and protein sequence data were taken from KEGG (http://www.kegg.jp). Yamanishi *et al.* [39] categorized this gold standard data set into four groups based on protein target types as enzyme, G Protein–Coupled Receptor (GPCR), ion channel and nuclear receptor. Each group

Table 5. A summary of integrated network-based DR methods

Name [Reference]	Using methods	Using data sets	Case studies	Evaluation criteria
TL_HGBI (Triple Layer Heterogeneous Graph Based Inference) [66]	Information flow-based method[a]	DrugBank, OMIM and Sophic Integrated Druggable Genome Database	Huntington disease, non-small cell lung cancer, alcohol dependence, small cell lung cancer and polysubstance abuse	ROC and AUC
SLAMS (Similarity-based LArge-margin learning of Multiple Sources) [67]	Canonical correlation analysis, and large margin method[b]	DrugBank, PubChem, UMLS, UniProt Knowledgebase, SIDER and National Drug File-Reference Terminology	Rheumatoid arthritis disease	Precision, recall, *F*-score and AUC
PreDR (Predict Drug Repositioning) [68]	SVM classifier based on kernel fusion[c]	PubChem, KEGG BRITE, BRENDA, SuperTarget and DrugBank	–	ROC, AUC, accuracy and *F*-measure
[69]	Supervised bipartite graph inference[d]	KEGG DRUG, KEGG LIGAND, Japan Pharmaceutical Information Center, KEGG GENES and data sets of [39]	Some of the analgesic and antipyretic drugs	ROC, AUC, sensitivity, specificity and PPV
NRWRH (Network-based Random Walk with Restart on the Heterogeneous network) [70]	Random walk with restart[e]	Data sets of [39]	–	Fold enrichment, rank cutoff curves, ROC and AUC
DReSMin (Drug Repositioning Semantic Mining) [71]	Semantic subgraph detection[f]	DrugBank, ChEMBL, UniChem and DisGeNET	Propranolol drug	Average co-prediction
[72]	Bipartite network projection and prioritization	TDR Targets database[g], KEGG, OrthoMCL database, ChEMBL and PubChem	Kinetoplastid parasites and *Plasmodium falciparum*	AUC
[73]	Classification based on a logistic regression[h]	Chemical genomics data set (LINCS)[i], PubChem, DrugBank, KEGG, TCGA data portal[j] and MSigDB database[k]	Glioblastoma, lung cancer and breast cancer	Experimental analysis
[74]	Monte Carlo Multiple Minimum conformational analysis[l]	ChEMBL and SIDER	–	AUC
[75]	Ontology-based reasoning[m]	FDA-approved drugs, KEGG, PharmGKB[n], DrugBank and PINA database[o]	Colorectal cancer	Experimental analysis

[a]TL-HGBI is used a three-layer heterogeneous network to predict drug–disease associations. This method is an iterative algorithm based on information flow. In fact, it is an extension of a two-layered model proposed by the same authors previously.

[b]SLAMS assumed that similar drugs could be used for similar diseases. In this regard, it used three different measures for drug similarity, chemical structure, side effects and protein targets. Prediction scores are computed for each ones by using of neighboring classifier. A final score is calculated by means of large margin method.

[c]In the first step, PreDR tried to fuse chemical structures, target proteins and side-effect information by using of a kernel function. In the second step, it used known drug–disease relations as training set and tried to predict new drug–disease relation by applying a SVM-based algorithm.

[d]It has two steps. Step 1 is necessary for drugs without pharmacological information, which try to predict such information by using of chemical structures. In Step 2, drug–target relations were predicted by means of predicted pharmacological information and sequence similarity of targets. The proposed method is a kind of distance learning.

[e]NRWRH was applied a random walk algorithm on a heterogeneous network that consisted of drug similarity, target similarity and DTI information. In this regard, it defined an initial probability matrix and four transition matrices.

[f]Mullen *et al.* provided a framework, which user could introduce a semantic subgraph and a threshold for semantic distance. Then, DReSMIN would search desirable integrated network to find the subgraph according to specified threshold. By using a pruning step and a splitting method, it could reduce the runtime. It is noticeable that in semantic subgraph detection, both of topological and semantical similarities are important.

[g]http://tdrtargets.org.

[h]It is using data that consisted of chemical similarity of drugs and gene expression similarity of targets and drug–target relations. It used seven different classifier methods that are based on logistic regression: three of them used only one of mentioned data, three of them used two combinations of data and one of them used all of data.

[i]www.lincsproject.org.

[j]https://tcga-data.nci.nih.gov.

[k]http://www.broadinstitute.org/gsea/msigdb.

[l]The proposed method tried to predict both of drug–target and drug–side effect simultaneously. In this regard, it used different data integration including 3D structural similarity of drugs with target information, drug–side effect relations with drug–target relations, drug–side effect and target–phenotype predictors.

[m]It first created an ontology including different classes such as drug, disease, gene, SNP, pathway and their relationships. Then, potential drug targets were identified by semantic reasoning methods, and finally, these predicted targets were prioritized by using of PPIN and literature searching.

[n]https://www.pharmgkb.org.

[o]http://cbg.garvan.unsw.edu.au/pina.

Table 6. Self-reported AUC and AUPR based on golden standard data sets

Method	Data set: enzyme		Data set: ion channel		Data set: GPCR		Data set: nuclear receptor		Reference of results
	AUC	AUPR	AUC	AUPR	AUC	AUPR	AUC	AUPR	
Bipartite graph learning	0.904	–	0.851	–	0.899	–	0.843	–	[39]
BLM	0.973	0.841	0.970	0.779	0.953	0.667	0.858	0.600	[44]
BLM-NII	0.988	**0.929**	0.990	0.950	0.984	0.865	0.982	0.821	[45, 50, 51]
DBSI	0.807	–	0.803	–	0.802	–	0.758	–	[47]
DT-Hybrid	**0.999**	–	0.997	–	**0.999**	–	**1.000**	–	[42]
GIP	0.685	0.150	0.637	0.179	0.679	0.260	0.758	0.357	[50]
GIP-RLS	0.982	0.885	0.986	0.927	0.947	0.713	0.906	0.610	[49]
Hybrid	0.998	–	0.993	–	0.996	–	0.999	–	[42]
KBMF2K	0.832	0.287	0.799	0.245	0.857	0.347	0.824	0.354	[17, 47, 50]
KRM	0.967	0.831	0.969	0.778	0.947	0.664	0.867	0.61	[44]
LPMIHN	0.999	**0.929**	**0.998**	**0.961**	0.9986	**0.973**	0.996	**0.970**	[51]
MINProp	0.990	0.849	0.984	0.841	0.987	0.937	0.973	0.938	[51]
NBI	0.975	–	0.976	–	0.946	–	0.838	–	[36]
Nearest profile	0.767	–	0.751	–	0.729	–	0.71	–	[39]
NetCBP	0.825	–	0.803	–	0.823	–	0.839	–	[47]
NetLapRls	0.956	0.826	0.947	0.825	0.931	0.66	0.856	0.516	[52]
NN	0.93	0.638	0.917	0.538	0.885	0.485	0.851	0.536	[44]
NRWRH	0.953	0.634	0.971	0.591	0.945	0.674	0.821	0.160	[51]
SITAR	0.922	0.877	0.927	0.889	0.946	0.939	0.863	0.851	[56]
Weighted profile	0.864	0.63	0.819	0.172	0.765	0.109	0.749	0.171	[45]
WNN	0.819	0.299	0.757	0.249	0.848	0.308	0.788	0.434	[50]
WNN-GIP	0.861	0.280	0.775	0.233	0.872	0.311	0.839	0.456	[50]
Yamanishi 2010	0.845	–	0.731	–	0.812	–	0.830	–	[69]

The top values are shown in bold.

contains corresponding targets, drugs and interactions of drugs and targets.

Table 6 is sorted alphabetically by method column and the best AUC, and AUPR results are indicated in bold.

As one can observe in Table 6, Domain Tuned-Hybrid (DT-Hybrid) [42] and label propagation with mutual interaction information derived from heterogeneous network (LPMIHN) [51] tend to dominate others, and MINProp [80] provides the best results after them (it should be noted that Hybrid is a simpler version of DT-Hybrid). DT-Hybrid is a recommendation method based on projection. It is believed that integration of biological knowledge with topological information of bipartite interaction network is the primary cause of DT-Hybrid domination. LPMIHN is a label propagation algorithm on heterogeneous networks. It seems that the strategy of LPMIHN to integrate the similarity information with topological information of the interactions contributes to its success. MINProp is also a label propagation algorithm in heterogeneous networks exploited by Yan *et al*. [51] for DR to provide a comparison between LPMIHN and MINProp. No single method among those presented above dominates all others in every situation.

We believe that before offering any new method for DR or extending any of the presented methods, a set of tools would be necessary to provide a gold standard data set as well as a platform for evaluating the methods. In this regard, we dedicated a team at our research laboratory to extend a set of Web services that consist of three parts. The first component will provide a gold standard data set, including drug, target and disease ontology-based information. The second part is a meditation system, which links between different DR methods and physical data sources, like various biological ontologies. The third part is an evaluation framework that provides statistical and analytical analysis functions and comparisons between different methods. The researchers could use our provided data and upload their result files. The software then will compute their statistical and analytical parameters and draw the required diagrams for comparison with other methods.

Conclusion and discussion

A significant challenge in drug development is the discovery of novel drug targets. Network-based approaches are an effective approach to linking the molecular and the phenotype level for the identification of drug targets [1]. Increasing focus is being placed on DR because of the substantial economic incentives to reposition existing drugs, in particular for the treatment of orphan and rare disorders. Moreover, DR research is beneficial to human health, as it can facilitate discovering new uses for existing drugs.

Considerable attention is devoted to network-based approaches in DR. Furthermore, satisfactory outcomes have been achieved in network-based computational biology focusing on biomolecular interactions and '-omics' data integration.

Every repositioning approach comes with methodological advantages and limitations. A common problem in some supervised learning methods in this domain is that previously unknown DTIs are considered as negative training and testing exemplars. Considering that we seek to discover previously unknown (i.e. novel) DTIs, this inaccurate negative sample selection can negatively influence the predictive accuracy of a method. Other semi-supervised methods try to use unlabeled information, while their final results are based on two different classifiers in the drug and target protein spaces [70].

One important unregarded biological concept in almost all of reviewed approaches is that although the direct protein target for most inhibitors could be not necessarily the sensitizing aberrated protein itself, finding such direct protein targets is the main goal of the different approaches. It may be caused by

restricted ability of most of the methods to predict trivial associations only. Trivial associations are ones that could be recognized simply by a kind of similarity. For example, if two drugs have high structural similarity, targets of one of them are predicted as targets of the other one and so on. Unfortunately, we could not find any discussion about nontrivial associations in any reviewed papers.

It is worth noting that transformation of theoretical, computational models into practical use has not yet been realized because of some inevitable factors like missing data, data bias and other technical limitations of computational methods [11].

Finally, the most significant problem is the lack of structured gold standard data for DR, thereby complicating the comparison and performance evaluation of different computational methods [11]. Several recent studies have attempted to provide a comprehensive data set of drugs and targets, but many studies continue to evaluate novel methods on their data sets rather than a shared gold standard. It is noticeable that Yamanishi *et al.* [39] tried to provide such data sets to be practically applied by some others (Table 3), but these are both obsolete and incomplete. A renewed effort in this area is required.

DR is a complicated process where no single computational method achieves satisfactory performance. This motivates the integration of different biological information to improve the reach and reliability of new DR methods. As any single type of data presents a one-dimensional view of a biological system, evaluating the performance of an integrative method based on a single data type may not be reliable. On the other hand, it will be difficult to compare different methods, as they rely on different data types. Therefore, we need to create a reference body of data for the standardized evaluation of integrative network approaches. In this regard, some existing platforms like Ondex (Ondex is a data integration platform that integrates different biological data sets for drug discovery in the form of a semantic network. It assigns a concept class for each concepts and a relation type for each relation to encode various concepts and relationships in a single graph. Ondex also provides network visualization and automatic network searching features. Ondex is avaliable at: http://www.ondex.org/) [7] and ontologies like DReNIn (DReNIn is an application ontology, which tries to integrate various kinds of drug–target and drug–disease relations from different data sources. It is in fact a semantic network, which consisits of 25 relation types between different concepts related to DR. Its Web address: www.drenin.ncl.ac.uk. It is downloadable from: http://bitbucket.org/nclintbio/drenin_ontology) could be of use.

Key Points

- What was already known about computational DR methods:
- Previous research studies have made a strong case for the effectiveness of integrative network-based methods for predicting DTIs.
- Existing methods cannot predict interactions with new drugs (drugs without any known target) and new targets (targets without any known drug).
- What this study added to our knowledge:
- Most existing algorithms use only immediate similarities and do not consider the transitivity of similarity. It is necessary to use higher-order similarities to achieve more accurate predictions.
- The number of training data is much less than the number of unlabeled data and, more importantly, there is no high-quality negative sample data for DTI. Therefore, semi-supervised methods are proposed as a novel approach to DR, as their efficacy has been demonstrated in other disciplines.
- Despite the general agreement that the integration of different data sources seems necessary, there are few examples of methods leveraging this approach in practice.
- The disease is a major factor in drug discovery research, but most existing DR methods do not adequately take this into account.

References

1. Emmert-Streib F, Tripathi S, Simoes RM, *et al.* The human disease network. *Syst Biomed* 2013;**1**:20–8.
2. Weng L, Zhang L, Peng Y, *et al.* Pharmacogenetics and pharmacogenomics: a bridge to individualized cancer therapy. *Pharmacogenomics* 2013;**14**:315–24.
3. 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.
4. Jiang Z, Zhou Y. Using gene networks to drug target identification. *J Integr Bioinform* 2005;**2**:14.
5. Naylor S, Schonfeld JM. Therapeutic Drug Repurposing, Repositioning and Rescue - Part 1: Overview Drug Discovery World Winter 2014/15 49.
6. Dudley JT, Deshpande T, Butte AJ. Exploiting drug–disease relationships for computational drug repositioning. *Brief Bioinform* 2011;**12**:303–11.
7. Cockell SJ, Weile J, Lord P, *et al.* An integrated dataset for in silico drug discovery. *J Integr Bioinform* 2010;**7**:15–27.
8. Wu Z, Wang Y, Chen L. Network-based drug repositioning. *Mol Biosyst* 2013;**9**:1268–81.
9. Brown AS, Kong SW, Kohane IS, *et al.* ksRepo: a generalized platform for computational drug repositioning. *BMC Bioinformatics* 2016;**17**:78.
10. Chen H, Zhang H, Zhang Z, *et al.* Network-based inference methods for drug repositioning. *Comput Math Methods Med* 2015;**2015**:7.
11. Ye H, Liu Q, Wei J. Construction of drug network based on side effects and its application for drug repositioning. *PLoS One* 2014;**9**:e87864.
12. Zou J, Zheng MW, Li G, *et al.* Advanced systems biology methods in drug discovery and translational biomedicine. *Biomed Res Int* 2013;**2013**:8.
13. Setoain J, Franch M, Martinez M, *et al.* NFFinder: an online bioinformatics tool for searching similar transcriptomics experiments in the context of drug repositioning. *Nucleic Acids Res* 2015;**43**:W193–9.
14. Younis W, Thangamani S, Seleem MN. Repurposing non-antimicrobial drugs and clinical molecules to treat bacterial infections. *Curr Pharm Des* 2015;**21**:4106–11.
15. Li YY, Jones SJ. Drug repositioning for personalized medicine. *Genome Med* 2012;**4**:27.
16. Morris GM, Lim-Wilby M. Molecular docking. In: A Kukol (ed). *Molecular Modeling of Proteins*. Totowa, NJ: Humana Press, 2008, 365–82.
17. Gonen M. Predicting drug-target interactions from chemical and genomic kernels using Bayesian matrix factorization. *Bioinformatics* 2012;**28**:2304–10.
18. Mason O, Verwoerd M. Graph theory and networks in biology. *IET Syst Biol* 2007;**1**:89–119.

19. Rider AK, Chawla NV, Emrich SJA. Survey of current integrative network algorithms for systems biology. In: A Prokop, B Csukás (eds). *Systems Biology: Integrative Biology and Simulation Tools*. Dordrecht: Springer Netherlands, 2013, 479–95.
20. Arrell DK, Terzic A. Network systems biology for drug discovery. *Clin Pharmacol Ther* 2010;**88**:120–5.
21. Dai YF, Zhao XM. A survey on the computational approaches to identify drug targets in the postgenomic era. *Biomed Res Int* 2015;**2015**:9.
22. Azuaje F. Drug interaction networks: an introduction to translational and clinical applications. *Cardiovasc Res* 2013;**97**:631–41.
23. Koyuturk M. Using protein interaction networks to understand complex diseases. *Computer* 2012;**45**:31–8.
24. Emig D, Ivliev A, Pustovalova O, *et al*. Drug target prediction and repositioning using an integrated network-based approach. *PLoS One* 2013;**8**:e60618.
25. Yeh SH, Yeh HY, Soo VW. A network flow approach to predict drug targets from microarray data, disease genes and interactome network—case study on prostate cancer. *J Clin Bioinforma* 2012;**2**:1.
26. Chang R, Shoemaker R, Wang W. A novel knowledge-driven systems biology approach for phenotype prediction upon genetic intervention. *IEEE/ACM Trans Comput Biol Bioinform* 2011;**8**:1170–82.
27. Imoto S, Tamada Y, Savoie CJ, *et al*. Analysis of gene networks for drug target discovery and validation. *Methods Mol Biol* 2007;**360**:33–56.
28. Chen HR, Sherr DH, Hu Z, *et al*. A network based approach to drug repositioning identifies plausible candidates for breast cancer and prostate cancer. *BMC Med Genomics* 2016;**9**:51.
29. Li Z, Wang RS, Zhang XS. Two-stage flux balance analysis of metabolic networks for drug target identification. *BMC Syst Biol* 2011;**5(Suppl 1)**:S11.
30. Chavali AK, D'Auria KM, Hewlett EL, *et al*. A metabolic network approach for the identification and prioritization of antimicrobial drug targets. *Trends Microbiol* 2012;**20**:113–23.
31. Raman K, Chandra N. Flux balance analysis of biological systems: applications and challenges. *Brief Bioinform* 2009;**10**:435–49.
32. Folger O, Jerby L, Frezza C, *et al*. Predicting selective drug targets in cancer through metabolic networks. *Mol Syst Biol* 2011;**7**:501.
33. Zhang J, Huan J. Analysis of network topological features for identifying potential drug targets. In: *Proceedings of 9th ACM International Workshop Data Mining Bioinformatics (BIOKDD 2010)*. Washington, DC, 2010.
34. Fukuoka Y, Takei D, Ogawa H. A two-step drug repositioning method based on a protein-protein interaction network of genes shared by two diseases and the similarity of drugs. *Bioinformation* 2013;**9**:89–93.
35. Keane H, Ryan BJ, Jackson B, *et al*. Protein-protein interaction networks identify targets which rescue the MPP+ cellular model of Parkinson's disease. *Sci Rep* 2015;**5**:17004.
36. 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.
37. Fakhraei S, Huang B, Raschid L, *et al*. Network-based drug-target interaction prediction with probabilistic soft logic. *IEEE/ACM Trans Comput Biol Bioinform* 2014;**11**:775–87.
38. Fakhraei S, Raschid L, Getoor L. Drug-target interaction prediction for drug repurposing with probabilistic similarity logic. In: *Proceedings of the 12th International Workshop on Data Mining in Bioinformatics*. ACM, Chicago, IL, 2013, 10–17.
39. 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.
40. Jacob L, Vert JP. Protein-ligand interaction prediction: an improved chemogenomics approach. *Bioinformatics* 2008;**24**:2149–56.
41. Zhou T, Ren J, Medo M, Zhang YC. Bipartite network projection and personal recommendation. *Phys Rev E Stat Nonlin Soft Matter Phys* 2007;**76**:046115.
42. Alaimo S, Pulvirenti A, Giugno R, *et al*. Drug–target interaction prediction through domain-tuned network-based inference. *Bioinformatics* 2013;**29**:2004–8.
43. Alaimo S, Bonnici V, Cancemi D, *et al*. DT-web: a web-based application for drug-target interaction and drug combination prediction through domain-tuned network-based inference. *BMC Syst Biol* 2015;**9**:S4.
44. Bleakley K, Yamanishi Y. Supervised prediction of drug-target interactions using bipartite local models. *Bioinformatics* 2009;**25**:2397–403.
45. 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.
46. McGarry K, Daniel U. Data mining open source databases for drug repositioning using graph based techniques. *Drug Discov World* 2015;**16**:64–71.
47. Chen H, Zhang Z. A semi-supervised method for drug-target interaction prediction with consistency in networks. *PLoS One* 2013;**8**:e62975.
48. Re M, Valentini G. Network-based drug ranking and repositioning with respect to DrugBank therapeutic categories. *IEEE/ACM Trans Comput Biol Bioinform* 2013;**10**:1359–71.
49. van Laarhoven T, Nabuurs SB, Marchiori E. Gaussian interaction profile kernels for predicting drug-target interaction. *Bioinformatics* 2011;**27**:3036–43.
50. van Laarhoven T, Marchiori E. Predicting drug-target interactions for new drug compounds using a weighted nearest neighbor profile. *PLoS One* 2013;**8**:e66952.
51. Yan XY, Zhang SW, Zhang SY. Prediction of drug-target interaction by label propagation with mutual interaction information derived from heterogeneous network. *Mol Biosyst* 2016;**12**:520–31.
52. Xia Z, Wu LY, Zhou X, *et al*. Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces. *BMC Syst Biol* 2010;**4**:S6.
53. Zheng X, Ding H, Mamitsuka H, *et al*. Collaborative matrix factorization with multiple similarities for predicting drug-target interactions. In: *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, Chicago, IL, 2013, 1025–33.
54. Alaimo S, Giugno R, Pulvirenti A. Recommendation techniques for drug-target interaction prediction and drug repositioning. *Methods Mol Biol* 2016;**1415**:441–62.
55. Hattori M, Okuno Y, Goto S, *et al*. Heuristics for chemical compound matching. *Genome Inform* 2003;**14**:144–53.
56. Perlman L, Gottlieb A, Atias N, *et al*. Combining drug and gene similarity measures for drug-target elucidation. *J Comput Biol* 2011;**18**:133–45.
57. Iorio F, Bosotti R, Scacheri E, *et al*. Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proc Natl Acad Sci USA* 2010;**107**:14621–6.
58. Liu X, Xu Y, Li S, *et al*. In silico target fishing: addressing a "Big Data" problem by ligand-based similarity rankings with data fusion. *J Cheminform* 2014;**6**:33.

59. Li J, Zheng S, Chen B, *et al*. A survey of current trends in computational drug repositioning. *Brief Bioinform* 2016;**17**:2–12.
60. Gottlieb A, Stein GY, Ruppin E, *et al*. PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Mol Syst Biol* 2011;**7**:496–6.
61. Wu C, Gudivada RC, Aronow BJ, *et al*. Computational drug repositioning through heterogeneous network clustering. *BMC Syst Biol* 2013;**7**:S6.
62. Yang K, Bai H, Ouyang Q, *et al*. Finding multiple target optimal intervention in disease-related molecular network. *Mol Syst Biol* 2008;**4**:228.
63. Kuhn M, Letunic I, Jensen LJ, Bork P. The SIDER database of drugs and side effects. *Nucleic Acids Res* 2016;**44**:D1075–9.
64. Zhang P, Wang F, Hu J, *et al*. Label propagation prediction of drug-drug interactions based on clinical side effects. *Sci Rep* 2015;**5**:12339.
65. Gottlieb A, Stein GY, Ruppin E, *et al*. PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Mol Syst Biol* 2011;**7**:496.
66. Wang W, Yang S, Zhang X, *et al*. Drug repositioning by integrating target information through a heterogeneous network model. *Bioinformatics* 2014;**30**:2923–30.
67. Zhang P, Agarwal P, Obradovic Z. Computational drug repositioning by ranking and integrating multiple data sources. In: H Blockeel, K Kersting, S Nijssen, F Železný (eds). *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, 579–94.
68. Wang Y, Chen S, Deng N, *et al*. Drug repositioning by kernel-based integration of molecular structure, molecular activity, and phenotype data. *PLoS One* 2013;**8**:e78518.
69. 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.
70. Chen X, Liu MX, Yan GY. Drug-target interaction prediction by random walk on the heterogeneous network. *Mol Biosyst* 2012;**8**:1970–8.
71. Mullen J, Cockell SJ, Tipney H, *et al*. Mining integrated semantic networks for drug repositioning opportunities. *PeerJ* 2016;**4**:e1558.
72. Berenstein AJ, Magariños MP, Chernomoretz A, *et al*. A multilayer network approach for guiding drug repositionng in neglected diseases. *PLoS Negl Trop Dis* 2016;**10**:e0004300.
73. Lee H, Kang S, Kim W. Drug repositioning for cancer therapy based on large-scale drug-induced transcriptional signatures. *PLoS One* 2016;**11**:e0150460.
74. Vilar S, Tatonetti NP, Hripcsak G. 3D pharmacophoric similarity improves multi adverse drug event identification in pharmacovigilance. *Sci Rep* 2015;**5**:8809.
75. Tao C, Sun J, Zheng WJ, *et al*. Colorectal cancer drug target prediction using ontology-based inference and network analysis. *Database* 2015;**2015**:1–9.
76. Frijters R, van Vugt M, Smeets R, *et al*. Literature mining for the discovery of hidden connections between drugs, genes and diseases. *PLoS Comput Biol* 2010;**6**:1–10.
77. Andronis C, Sharma A, Virvilis V, *et al*. Literature mining, ontologies and information visualization for drug repurposing. *Brief Bioinform* 2011;**12**:357–68.
78. Yang HT, Ju JH, Wong YT, *et al*. Literature-based discovery of new candidates for drug repurposing. *Brief Bioinform* 2016, doi:10.1093/bib/bbw030.
79. Lee HM, Kim Y. Drug repurposing is a new opportunity for developing drugs against neuropsychiatric disorders. *Schizophr Res Treatment* 2016;**2016**:12.
80. Hwang T, Kuang R. A heterogeneous label propagation algorithm for disease gene discovery. In: *Proceedings of the 2010 SIAM International Conference on Data Mining*, Columbus, OH, 2010, 583–94.