Chapter 23

Recommendation Techniques for Drug–Target Interaction Prediction and Drug Repositioning

Salvatore Alaimo, Rosalba Giugno, and Alfredo Pulvirenti

Abstract

The usage of computational methods in drug discovery is a common practice. More recently, by exploiting the wealth of biological knowledge bases, a novel approach called drug repositioning has raised. Several computational methods are available, and these try to make a high-level integration of all the knowledge in order to discover unknown mechanisms. In this chapter, we review drug–target interaction prediction methods based on a recommendation system. We also give some extensions which go beyond the bipartite network case.

Key words Drug–target interaction prediction, Drug combination prediction, Drug repositioning, Hybrid methods network-based prediction, Recommendation systems

1 Introduction

Historically, some proteins have been chosen as druggable [1] and it has been shown that drugs with very different chemical structures target the same proteins and the same protein is druggable from different drugs. This gives the intuition that drugs are not specifically designed to diseases [2]. Recently, the trend in the pharmaceutical industry, thanks to the bioinformatics predictions methods, has changed. The new experimental drugs have a wider variety of target proteins and analysis on drug–target and gene–disease networks highlighted that few of them are essential proteins and they are correlated with tissue specificity and are more disease-specific [3].

Following this trend, one of the very attractive drug discovery techniques is drug repositioning [4]. The usage of known drugs for new therapeutically scope represents a fast and costly effective strategy for drug discovery. The prevalence of studies has raised a wide variety of models and computational methods to identify new therapeutic purposes for drugs already on the market and sometimes even in disuse. Computational methods try to make a high level of integration of all the knowledge in order to discover any

Oliviero Carugo and Frank Eisenhaber (eds.), *Data Mining Techniques for the Life Sciences*, Methods in Molecular Biology, vol. 1415, DOI 10.1007/978-1-4939-3572-7_23,

unknown mechanisms. In [5], a compressive survey on the techniques and models is given. These models using tools available in chemoinformatics [1, 6, 7], bioinformatics [8–11], network and system biology [1] allow the development of methods that can speed up the design of the drug. Following [5], repositioning methods can be grouped into the following categories: blinded, target-based, knowledge-based, signature-based, pathway- or network-based, and targeted-mechanism-based.

The basic approach to repositioning is known as blinded. Blind methods do not include biological information or pharmaceutical discoveries and commonly relies on serendipity and depend on random tests on specific diseases [12, 13].

Target-based repositioning includes high-throughput experiments on drug and biomarkers of interest in connection with in-silico screening for the extraction of compounds from libraries based, for example, on docking [2–4] or on comparisons of the molecular structures [5, 6]. This approach compared to the blind one is more effective as different targets link directly to the mechanisms of the disease. Therefore, these methods in a short time (i.e., a few days) are used to do the screening of all molecules for which the chemical structure is known. In [1], authors designed a framework for drug repositioning based on the functional role of novel drug targets. They proceeded by detecting and annotating drug-induced transcriptional modules in cell-specific contexts which allowed also to detect novel drug mechanism of action. In silico results were confirmed by in vitro validation of several predicted genes as modulators of cholesterol homeostasis.

Knowledge-based drug repositioning takes into account information concerning drugs, drug–target interaction networks [7–9], drug chemical structure, the structure of its targets (including also their similarity), side effects, and affected metabolic pathways [10]. This knowledge enables the development of integrated high-performance predictive models [11]. In [8], a bipartite graph linking US Food and Drug Administration-approved drugs to proteins by drug target binary associations is exploited. In [10], the authors identified new drug–target interactions (DTI) using side effect similarity. In [14], the authors make use of transcriptional responses, predicted and validated new drug modes of action, and drug repositioning. Furthermore, in [15], the authors presented a bipartite graph learning method to predict DTI by integrating chemical and genomic data. In [16], Cheng et al. (2012) presented a technique based on network-based inference (NBI) implementing a naive version of the algorithm proposed in [17]. In [18], Alaimo et al. extended the approach of [17] presenting a hybrid approach for the network-based inference drug–target interaction prediction and drug repositioning. In [19], the authors used a machine learning method to predict new ones with high accuracy. In [12], the

authors introduced a Network-based Random Walk with Restart on the Heterogeneous network (NRWRH) algorithm predicting new interactions between drugs and targets by means of a model dealing with an "heterogeneous" network. In [13], the authors proposed the Bipartite Local Model-Interaction-profile Inferring (BLM-NII) algorithm. Interactions between drugs and targets are deduced by training a classifier.

Signature-based methods use expression data to discover off-target related to known molecules for the treatment of other pathologies [20]. Some of these methods also incorporate time-course quantitative data showing that a drug can give the survival outcome in connection to the clinical conditions [21]. This allows to stratify patients. Furthermore, these methods by integrating the quantitative information are able to discover additional mechanisms of action not yet known to molecules and known compounds. In [22], the authors predicted therapeutic relationship drug–disease so far not described by combining publicly available disease microarray data of human cell lines treated with drugs or small molecules obtained from Gene Expression Omnibus (GEO) of the National Center for Biotechnology Information (NCBI). With this approach they identified about 16,000 pairs of possible drug–disease, in which 2664 are statistically significant and more than half suggest a therapeutic relationship. To validate the hypothesis, the authors tested the cimetidine as a therapeutic approach for lung adenocarcinoma (LA). Cancer cells exposed to cimetidine showed a dose-dependent reduction in growth and proliferation (experiments performed on mice implanted with human cell lines of LA). Furthermore, to test the specificity of this proposal, a similar experiment was carried out in mice with cell lines of ACHN renal cell carcinoma (the score of the signature was not significant for cimetidine), and in agreement with the computational analysis there has been no effect. In [23] by integrating publicly available gene expression data, the authors discovered that anticonvulsant topiramate is a hypothetical new therapeutic agent for inflammatory bowel diseases (IBD). They experimentally validated the topiramate's efficacy in ameliorating atrinitrobenzenesulfonic (TNBS)-induced rodent model of IBD even though the exact pharmacodynamics mechanism of action is not known.

The pathway/network-based approaches use omic data, signaling pathways, and networks of protein–protein interaction to build disease-specific pathways containing end-point targets of repositioned drugs [24–26]. These methods have the advantage to identify signaling mechanisms hidden within the pathway and the signatures of the genes. The above approaches together with large-scale drug sensitivity screening led to predict combinations of drugs for therapeutically aims. In [27], the straight of the inference model is to use druggable targets resulting from taking into account

drug treatment efficacies and drug–target binding affinities information. They validated the model in breast and pancreatic cancers data by using siRNA-mediated target silencing highlighting also the drug mechanism of action in cancer cell survival pathways.

More recently, the development of multi-target drugs or drug combinations has been considered crucial to deal with complex diseases [28, 29]. Effective methods to improve the combinations prediction include: the choke point analysis [30, 31], a reaction that either uniquely consumes a specific substrate or produces a specific product in a metabolic network, and the comparison of metabolic networks of pathogenic and non-pathogenic strains [32]. These approaches commonly share the identification of nodes having a high ratio of incident k-shortest paths [32, 33]. On the other hand, it has been shown that the co-targeting of crucial pathway points [34, 35] is efficient against drug resistances both in anti-infective [36] and anti-cancer [37, 38] strategies. Two relevant examples are RAS [39, 40] and Survivin [41]-associated diseases.

In practice, a fundamental question is if the chosen drug is effective to the treated patient. A large amount of money is spent on drugs that have no beneficial effects on patients causing dangerous side effects. It is known that this is due to the genetic variants of individuals that influence metabolism, drug absorption, and pharmacodynamics. Although this, frequently GWAS for drugs are not replicated in either the same or different populations. Genomic and epigenomic profiling of individuals should be investigated before prescription, and a database of such profiling should be maintained to design new drugs and understand the correct use of the existing ones for the specific individual. Such profiling should exist for each individual and not as in the current era related only to publications which are sample case-specific and results are in some case difficult to replicate [42].

In this chapter, we review drug–target prediction and drug repositioning techniques based on hybrid recommendation methods. We give an in-depth review of our systems DT-Hybrid [18] and DT-web [43]. Then we present some generalization of our models that goes beyond the bipartite case.

2 Materials and Methods

In what follows, we introduce recommendation techniques especially focusing on those named Network-Based Inference Methods. These have been successfully applied in the prediction of drug–target interaction prediction and drug repositioning. We then describe our methodology DT-Hybrid and its application also on drug combination.

2.1 Recommendation Techniques for DTI Prediction

2.1.1 Background on Recommendation Algorithms

Recommendation algorithms are a class of systems for information filtering whose main objective is the prediction of users' preferences for some objects. In recent years, they have become commonly used and applied in various fields. Their main application lies in e-commerce in the form of web-based software. However, they have been successfully employed in other areas related, for example, to bioinformatics [16, 45].

A recommendation system consists of users and objects. Each user collects some objects, for which he can also express a degree of preference. The purpose of the algorithm is to infer the user's preferences and provide scores to objects not yet owned, so that the ones, which most likely will appeal the user, will be rated higher than the others.

In a recommendation system, we denote the set of objects as $O=\{o_1,o_2,\ldots,o_n\}$ and the set of users as $U=\{u_1,u_2,\ldots,u_m\}$. The whole system can be fully described by a sparse matrix $T=\{t_{ij}\}_{n\times m}$ called utility matrix. In such a matrix, t_{ij} has a value if and only if the user u_j has collected and provided feedback on the object o_i. In the event that users can only collect objects without providing any rating, the system can be described by a bipartite graph $G(O,U,E)$ where $E=\{e_{ij}:o_i\in O,u_j\in U\}$ is the set of edges. Each edge indicates that a user has collected an object. This graph can be described in a more compact form by means of an adjacency matrix $A=\{a_{ij}\}_{n\times m}$, where $a_{ij}=1$ if u_j collected o_i, and $a_{ij}=0$ otherwise. A reasonable assumption in this case is that the objects collected by a user corresponds to his preferences, and the recommendation algorithm aims to predict users' views on other items.

Up to now, the algorithm mostly applied in this context is collaborative filtering (CF) [44, 46]. It is based on a similarity measure between users. Consequently, the prediction for a particular user is computed employing information provided by similar ones. A Pearson-like evaluation is typically used to evaluate similarity between two users:

$$s_{ij}=\frac{\sum_{l=1}^{n}a_{li}a_{lj}}{\min\{k(u_i),k(u_j)\}}, \tag{1}$$

where $k(u_i)$ is the number of items collected by the user u_i. For any user–object pair (u_i-o_j), if not already collected $(a_{ij}=0)$, a predicted score v_{ij} can be computed as:

$$v_{ij}=\frac{\sum_{l=1,l\neq i}^{m}s_{li}a_{jl}}{\sum_{l=1,l\neq i}^{m}s_{li}}. \tag{2}$$

Two factors influence positively v_{ij}: objects collected from a large number of users, and objects collected frequently from users very similar to u_i. The latter correspond to the most significant predictions. All items are then sorted in descending order using their prediction score, and only those at the top will be recommended.

Verifying the reliability of a recommender system result is typically a complex phase. A basic evaluation strategy considers the system as a classification algorithm that distinguishes, for each user, liked objects from un-liked ones. We can then apply traditional metrics such as mean squared error or receiver operating characteristic curves to evaluate results. Another strategy is to define new metrics specifically designed to assess performances of a recommendation system [17].

In common between the two approaches is the application of a k-fold cross-validation to obtain a more accurate estimate of methods reliability. The set of all user–object preferences is randomly partitioned into k disjoint subsets. One is selected as a test set, and the recommendation algorithm is applied to the others. Evaluation metrics are then computed using the test set as a reference. The process is repeated until all the partitions have been selected as test set, and the results of each metric are averaged in order to obtain an unbiased estimate of the quality of the methodology.

Four metrics have been specifically developed to assess the quality of a recommender algorithm: two measure performances in terms of predictions accuracy, by measuring the capability of recovering interactions in the test set, whereas the other two measure recommendation diversity:

(a) Recovery of deleted links, r.

An accurate method typically will place potentially preferable objects higher than non-preferable ones. Assuming that a user has collected only liked items, the pairs present in the test set, in principle, should have a higher score than the others. Therefore, by applying the recommendation algorithm and computing the sorted set of predictions for a user u_j, we can compute a relative rank for an uncollected object o_i, whose position in the list is p, as:

$$r_{ij} = \frac{p}{o - k_j}, \tag{3}$$

Such a rank should be smaller if the pair $u_j - o_i$ is part of the test set. The recovery (r) corresponds to the average of such relative ranking for all user–object pairs in the test set. The lower its value, the greater is the ability of the algorithm to recover deleted interactions, and therefore to achieve accurate results.

(b) Precision and recall enhancement, $e_P(L)$ and $e_R(L)$.

Typically, only the highest portion of the recommendation list of a user is employed for further purposes, which is why a more practical measure of the reliability of a recommendation system may consider only the Top-L predictions. For a user u_i, let D_i be the number of deleted object for user u_i, and $d_i(L)$ the ones predicted in the Top-L places. An average of the ratios $d_i(L)/L$ and $d_i(L)/D_i$ for all users with at least one object in the test set, correspond, respectively, to the precision $P(L)$ and recall $R(L)$ for the recommendation process [46, 47].

We can get a better perspective by considering these values with respect to random model. Let $P_{rand}(L)$ and $R_{rand}(L)$ be, respectively, the precision and the recall of a recommendation algorithm that randomly assign scores to user–object pairs. If the user u_i has a total of D_i objects in the test set, then $P^i_{rand}(L) = D_i / (o - k_i) \approx D_i / o$, since the total number of objects is much greater than the number of collected ones. Averaging for all users, we obtain $P_{rand}(L) = D/ou$, where D is the size of the test set. By contrast, the average number of deleted objects in the Top-L positions is given by $L \cdot D_i / (o - k_i) \approx L \cdot D_i / o$ and, therefore, $R_{rand}(L) = L/o$. We can now define precision and recall enhancement as:

$$e_P(L) = \frac{P(L)}{P_{rand}(L)} = P(L) \cdot \frac{ou}{D}, \tag{4}$$

$$e_R(L) = \frac{R(L)}{R_{rand}(L)} = R(L) \cdot \frac{o}{L}, \tag{5}$$

A high value of precision enhancement indicates that the fraction of relevant predictions made by the algorithm is substantially higher than a completely random one. A high recall enhancement indicates that the percentage of correct predictions is significantly higher than the null model.

(c) Personalization, $h(L)$.

A first measure of diversity to consider when evaluating a recommendation algorithm is the uniqueness of the predictions made for different users, namely the inter-user diversity. Given two users u_i and u_j, a measure of inter-list distance can be computed as:

$$h_{ij}(L) = 1 - \frac{q_{ij}(L)}{L}, \tag{6}$$

where $q_{ij}(L)$ is the number of common Top-L predictions between the two users. It follows immediately that this distance has a value 0 if the two users have the same prediction, 1 in the case of completely different lists. The average distance calculated for all possible pairs of users corresponds to the personalization

metric. Higher, or lower, values correspond, respectively, to a greater, or lesser, diversity of recommendations.

(d) Surprisal/novelty, $I(L)$.

Evaluating the ability of a recommendation system to generate novel and unexpected predictions is a key measure. In this context, we define as unpredictability of results, the ability to suggest items for which it is very unlikely that a user may already know them. To measure this, we use the concept of self-information or "surprisal" [52], which determines how unexpected is an object with respect to its global popularity. Given an object o_j, the probability that a user has collected it is given by $k(j)/m$. Its self-information is therefore $I_j = \log_2\left(m / k\left(j\right)\right)$. The average of such values for the Top-L predictions of a user u_i correspond to its self-information, $I_i(L)$. By averaging for all users, we get a measure of the global surprisal $I(L)$.

In classical applications, a value L equal to 30 is chosen a priori. In any case, no variations in the relative performances of the algorithms can be observed by varying L, as long as its value is significantly smaller than the number of objects in the system.

Typically, drug–target interaction (DTI) prediction methods are divided into two main classes:

- Traditional methods, in which new drugs are predicted for a specific target;
- Chemical biology methods, where new potential targets are predicted for a given drug [15].

Recommendation algorithms have the advantage of using both strategies at the same time: they can simultaneously assess new drug candidate for a specific target, and new potential targets for a given drug [17].

In order to use recommendation systems for the prediction of DTI, targets may be considered as objects, drugs as users, and experimentally validated DTI as the set of known user–object preferences. In such a system, only information about the presence or absence of an interaction will be available. Hence, it is easily possible to represent the entire knowledge in the form of a bipartite network. The prediction of user preferences, and their subsequent ranking, can be seen as the usage of the bipartite network to infer common features between drugs, and the employment of such characteristics in order to predict novel biologically significant DTIs. In this sense, it prevails the idea that structurally similar drugs will have similar target and vice versa.

The four metrics previously presented radically change meaning in the application to the DTI prediction. Recovery, precision,

and recall enhancement are directly related to the ability of the algorithm to predict biologically significant interactions. This is derived from the fact that, in the *k*-fold cross-validation procedure, test set elements should be ranked higher with respect to others. The recall provides information on the ability of the algorithm to find the real unknown interactions, while the precision indicates the ability to discern biologically meaningful interactions from untrue ones. The other two metrics (personalization and surprisal) are less important even if the capability of predicting unexpected interactions, combined with the ability to identify only significant results, can be critical for the purposes of producing novel biological knowledge previously totally ignored.

2.1.2 The DT-Hybrid Algorithm

In this section we will introduce the DT-Hybrid algorithm, a recommender system whose purpose is predicting DT interactions. To this end, we will initially describe graph-based recommendation methods, their versatility and main limitations. This will help understanding the idea behind DT-Hybrid and how it has been developed.

Graph-based recommendation algorithm is a class of collaborative filtering (CF)-like techniques, which use a network representation of user–object preferences to infer predictions. They apply a network projection technique to compress the information contained in the preferences network. Given a bipartite graph that represents a recommendation system $G(U, O, E)$, an object-projection corresponds to a new graph where:

- Nodes are only objects,
- Edges between two nodes are present if there is at least one path that connects two objects through a user in G,
- Weights in each edge are proportional to the probability that a user who has collected an object will want to collect another one.

More generally, a quantity of resource is associated with each object node, and the weight w_{ij} of the projection is the portion of the resource that j would distribute to i. In these terms, the calculation of weights may be associated with a two-step resource allocation process. In a first phase, the resource is transferred from object nodes to user ones. In the second step the resource now present in the user nodes is transferred back to object ones. Since the bipartite network is unweighted, the resource of a node should be equally distributed to its neighborhood.

Therefore, given a bipartite graph $G(U, O, E)$, which represents the set of user–object preferences, $A = \left\{a_{ij}\right\}_{n\times m}$ is its adjacency matrix. Now, let $f(s) \geq 0$ be the initial resource allocated in the node o_j. After the first pass, all the resource flows from O nodes to

U nodes. The amount of resource allocated in node u_l can be calculated as:

$$f(u_l) = \sum_{i=1}^{n} \frac{a_{il} f(o_i)}{k(o_i)}, \tag{7}$$

where $k(x)$ is the degree of node x in the bipartite network. In the subsequent phase, the resource is transferred back to object nodes and its final amount in node o_i can be assessed as:

$$f'(o_i) = \sum_{l=1}^{m} \frac{a_{il} f(u_l)}{k(u_l)} = \sum_{l=1}^{m} \frac{a_{il}}{k(u_l)} \sum_{j=1}^{n} \frac{a_{jl} f(o_j)}{k(o_j)}, \tag{8}$$

which can be further rewritten as:

$$f'(o_i) = \sum_{j=1}^{n} w_{ij} f(o_j), \tag{8a}$$

where

$$w_{ij} = \frac{1}{(i,j)} \sum_{l=1}^{m} \frac{a_{il} a_{jl}}{k(u_l)}, \tag{9}$$

and

$$(i,j) = k(o_j). \tag{10}$$

The matrix $W = \{w_{ij}\}_{n \times n}$ is the object-projection of the bipartite network, and the whole set of predictions will be computed as:

$$R = W \times A. \tag{11}$$

This methodology, called network-based inference (NBI), can be easily adapted to any bipartite network. In [16], it has been successfully used to predict possible novel DTI interactions. Let $D = \{d_1, d_2, \ldots, d_m\}$ denote the set of drugs and $T = \{t_1, t_2, \ldots, t_n\}$ the set of targets. The DTI network can be fully represented by a bipartite graph $G(D, T, E)$ as previously described. An adjacency matrix $A = \{a_{ij}\}_{m \times n}$ can also be associated with the bipartite network, where $a_{ij} = 1$ if drug d_i and target t_j interacts, $a_{ij} = 0$ otherwise. Therefore, by applying the NBI methodology, putative DTI may be computed.

The recommendation algorithm previously described is extremely versatile and practical for the production of possible novel DTIs. However, it does not include any knowledge on the application domain. DT-Hybrid [18] is a recommendation algorithm that extends [16] by adding information on the similarity between drugs and targets. Despite its simplicity, the technique provides a comprehensive and practical framework for the in silico prediction of DTIs.

Let $S = \{s_{ij}\}_{n\times n}$ be a targets similarity matrix (i.e., BLAST bit scores [48] or Smith–Waterman local alignment scores [49]), and $S^1 = \{s'_{ij}\}_{m\times m}$ a drug structural similarity matrix (i.e., SIMCOMP similarity score [50]). In order to be able to introduce such a similarity in the recommender model, it is necessary to build a processed similarity matrix $S^2 = \{s''_{ij}\}_{n\times n}$, where each element s''_{ij} describes the similarity between two targets t_i and t_j based on the common interactions in the network, weighting each one by drugs similarity. In other words, if two targets t_i and t_j are linked by many highly similar drugs then s''_{ij} will be high. S^2 can be computed as:

$$s''_{ij} = \frac{\sum_{k=1}^{m}\sum_{l=1}^{m}\left(a_{il}a_{jk}s'_{lk}\right)}{\sum_{k=1}^{m}\sum_{l=1}^{m}\left(a_{il}a_{jk}\right)}. \tag{12}$$

By linearly combining the matrices S and S^2, it is possible to obtain the final similarity matrix $S^{(1)} = \{s^{(1)}_{ij}\}_{n\times n}$:

$$S^{(1)} = \alpha \cdot S + (1-\alpha)\cdot S^2, \tag{13}$$

where α is a tuning parameter.

It is now possible to modulate the weights w_{ij} of the resource-allocation procedure by using the matrix $S^{(1)}$ and suitably modifying the Eq. 10:

$$(i,j) = \frac{k(t_i)^{1-\lambda}\cdot k(t_j)^{\lambda}}{s^{(1)}_{ij}}, \tag{14}$$

where λ is a fundamental parameter that mediates between two different resource distribution processes: an equal distribution among neighbors (as the NBI algorithm) and a nearest-neighbor averaging process. This aspect has been added to DT-Hybrid to ensure greater reliability in the presence of very sparse networks, for which it is necessary to be less conservative when producing predictions.

Finally, by means of Eqs. 9, 11 and 14, it is possible to compute candidate DTI interactions. For each drug, DT-Hybrid will return the Top-L predicted targets sorted by score in descending order.

In order to fairly evaluate and compare the methodologies described before, common data sets and protocols are needed. For this purpose, each algorithm has been evaluated using five datasets that contain experimentally verified interactions between drugs and targets.

Four data sets were built by grouping all possible experimentally validated DTIs based on their main target type: enzymes, ion channels, G-protein-coupled receptors (GPCRs), and nuclear receptors (Table 1). Another data set was built by taking all information on drug and targets available in DrugBank.

Table 1
Description of the dataset: number of biological structures, targets, and interactions together with a measure of sparsity

Dataset	Structures	Targets	Interactions	Sparsity
Enzymes	445	664	2926	0.0099
Ion channels	210	204	1476	0.0344
GPCRs	223	95	635	0.0299
Nuclear receptors	54	26	90	0.0641
Complete DrugBank	4,398	3,784	12,446	0.0007

Note: The sparsity is obtained as the ratio between the number of known interactions and the number of all possible interactions

To assess the similarity between drugs, a SIMCOMP 2D chemical similarity has been chosen [50]. SIMCOMP represents the two-dimensional structure of a compound through a graph of connections between molecules. The similarity is obtained by seeking the maximum common sub-graph between two drugs. This is obtained by seeking the maximal cliques in associated graphs.

The similarity between targets has been assessed through the Smith–Waterman local sequence alignment algorithm [49]. The idea behind this choice is to find common docking sites between two targets, namely similar portions of the target sequence. Although this assumption is not always valid, such a choice was made also for performance reasons.

The similarities calculated by the two algorithms were normalized using the equation introduced in [15]:

$$S_{norm}(i,j) = \frac{S(i,j)}{\sqrt{S(i,i) \cdot S(j,j)}}. \tag{15}$$

In this way, resulting similarity matrices will hold the main properties of distances (positivity, symmetry, triangle inequality).

For the evaluation of the results a tenfold cross-validation procedure was applied and the four metrics defined previously were computed, focusing mainly on the two that are synonymous with the biological reliability of results. Everything was repeated 30 times in order to obtain more unbiased results. It is important to note that the random partitioning method associated with the cross-validation can cause the isolation of some nodes in the network on which the tests are being performed. A main limitation of recommendation algorithms just described is the inability to predict new interactions for drugs or targets for which no information is available. This implies that in the presence of isolated nodes a bias is introduced in the evaluation of results. For this reason,

Table 2
Optimal values of λ and α parameters for the data sets used in the experiments (Enzymes, ion channels, GPCRs, nuclear receptors, complete DrugBank)

Data set	λ	α
Enzymes	0.5	0.4
Ion channels	0.5	0.3
GPCRs	0.5	0.2
Nuclear receptors	0.5	0.4
Complete DrugBank	0.8	0.7

during the computation of each partition it must be ensured that each node in the bipartite network has at least a link to another node. Finally, the algorithms were compared by choosing only the Top-30 predictions in descending order of score for each drug.

To better assess the impact of adding information about the application domain, an additional algorithm called Hybrid was evaluated. Hybrid can be considered as a variation of DT-Hybrid that does not include any similarity.

DT-Hybrid and Hybrid depend on the λ parameter, while DT-Hybrid also on the α parameter. For this reason, an a priori analysis of the two is needed to understand their behavior. Table 2 shows their values, which allow best performance in terms of biological reliability of predictions. No law regulating their behavior has been discovered, as they depend heavily on the specific characteristics of each data set. For this reason, a prior analysis is necessary in order to select the best ones according to each specific situation.

An evaluation of the algorithms in terms of precision and recall enhancement (Tables 3 and 4) shows that DT-Hybrid is able to surpass both NBI and Hybrid in terms of interactions recovery.

Table 3
Comparison between DT-Hybrid, Hybrid, and NBI

Algorithm	$e_P(30)$	$e_R(30)$	$AUC(30)$
NBI	538.7	55.0	0.9619 ± 0.0005
Hybrid	861.3	85.7	0.9976 ± 0.0003
DT-Hybrid	**1141.8**	**113.6**	$\mathbf{0.9989 \pm 0.0002}$

Note: For each algorithm the complete DrugBank dataset was used to compute the precision and recall metrics, and the average area under ROC curve (AUC). Bold values represent best results

Table 4
Comparison of DT-Hybrid, Hybrid, and NBI through the precision and recall enhancement metric, and the average area under ROC curve (AUC) calculated for each of the four datasets listed in Table 1

	$e_P(30)$			$e_R(30)$			*AUC*(30)		
Data set	**NBI**	**Hybrid**	**DT-Hybrid**	**NBI**	**Hybrid**	**DT-Hybrid**	**NBI**	**Hybrid**	**DT-Hybrid**
Enzymes	103.3	104.6	**228.3**	19.9	20.9	**32.9**	0.9789 ± 0.0007	0.9982 ± 0.0002	**0.9995 ± 0.0001**
Ion channels	22.8	25.4	**37.0**	9.1	9.7	**10.1**	0.9320 ± 0.0046	0.9929 ± 0.008	**0.9973 ± 0.0006**
GPCRs	27.9	33.7	**50.4**	7.5	**8.8**	5.0	0.9690 ± 0.0015	0.9961 ± 0.0007	**0.9995 ± 0.0006**
Nuclear receptors	28.9	31.5	**70.2**	0.3	**1.3**	**1.3**	0.9944 ± 0.0007	0.9986 ± 0.0004	**1.0000 ± 0.0000**

Note: The results were obtained using the optimal values for λ and α parameters as shown in Table 2. Bold values represent best results

A significant improvement has been achieved mainly in the recall (e_R), which measures the ability of a recommendation algorithm to recover the true significant interactions, so it is synonymous with the biological quality of the results. The use of receiver operating characteristic curves (ROC) to evaluate the performance of the algorithm further demonstrates that the integration of specific information of the application domain is crucial to achieve results that are more significant. This is reflected further by analysis of the average areas under the ROC curves (AUC) which show an increase in performance (Tables 3 and 4). A more comprehensive analysis and comparison of DT-Hybrid is available in [18].

2.1.3 An Extension to DT-Hybrid: p-Value-Based Selection of DTI Interactions

One of the main limitations of the approaches described above lies in the selection of significant predictions. A classic methodology used for recommendation algorithm consists of ordering the predictions for each drug in descending order, and collecting only the Top-L ones. This however is not always a good choice when predicting interactions between drugs and targets. A more objective methodology based on statistical criteria is required [43].

A good idea might be calculating an additional similarity between targets that take into account their function. Therefore, such a similarity can be used to build a correlation measure between subsets of targets, and evaluate, for each drug, which subset of predicted targets has a similarity unexpectedly high with respect to the validated ones. All this can be achieved using a similarity based on ontological terms (i.e., GO terms), and the computation of a p-value score.

First, after applying DT-Hybrid and computing an initial list of predictions for the drugs, each target is annotated with the corresponding ontological terms. Using, then, the ontology DAG (Directed Acyclic Graph), a similarity between terms can be defined on the basis of their distance. A DAG can be constituted by a set of disconnected trees, which could make impossible to obtain a finite similarity value for each pair of nodes. For this reason, all the root nodes of the trees that make up the DAG have been connected to a new single dummy root node. This does not alter the properties of the network but allows the computation of a similarity for each possible pair of ontological terms.

Now, for each predicted target of a drug, a correlation measure can be defined as the maximum similarity between the ontological terms associated with its predicted target and the validated ones. The correlation of a subset of predicted targets can be defined as the minimum correlation calculated for each target within the subset. Therefore, let M_i be a subset of predicted targets for the drug d_i, m be the total number of targets, and q_i be the number of targets having a correlation greater than that of M_i. The p-value, $p(M_i)$, is the probability of drawing by chance $|M_i| = k_i$ terms whose correlation is greater than the observed minima.

This can be computed through a hypergeometric distribution in the following way:

$$p(M_i) = \frac{\binom{q_i}{k_i}\binom{m-q_i}{k_i-k_i}}{\binom{m}{k_i}} = \frac{\binom{q_i}{k_i}}{\binom{m}{k_i}}. \quad (16)$$

The *p*-value is used to provide a quality score for the association between predicted targets and validated ones of a single drug. No correction for multiple testing was applied, as each *p*-value is considered independent of the others. The subset of predictions chosen as a result of the algorithm is the one that simultaneously maximizes the correlation and minimizes the *p*-value.

At this point, it is essential to establish a criterion for selecting subsets of targets. An objective assessment would occur by calculating all possible subsets of predicted targets. However, this is not feasible given their large number. A strategy that is based on the classic Top-*L* selection can be employed. Divide the range of correlation values for a drug in *L* parts, and use the minimum in each partitions as the lower bound used for the selection of targets to put in a subset.

2.1.4 Applying DT-Hybrid for Drug Combinations Prediction

Because of the complexity of diseases, the development of multi-target drugs or combinations of existing drugs is a crucial problem in today's medicine. In particular, existing drugs have a huge number of targets still unknown, and the use of DTI prediction techniques is essential in order to elucidate their functioning. This can pave the way to the production of more effective drug combinations with fewer side effects than in the past. The idea, which is at the basis of the prediction of drug combinations, is the discovery of the minimum set of targets that can influence a set of genes of interest [43]. In order to do so, it is necessary to work in a multi-pathway environment in which all the chains of interactions between genes are taken into account simultaneously. The genes of interest for a disease must not be directly targeted in order to minimize side effects.

First, from the most common databases, a single multi-pathway environment should be built. This can be achieved by merging metabolic and signaling pathways (Reactome, PID, and KEGG). In this phase, it is essential to normalize entity names in each pathway, as different databases may use different types of nomenclature. To do so, a reference identifier is needed. Entrez identifiers are associated in this phase with each entity, where available. The environment so built can be queried for information about the best targets for a combined therapy.

Starting from a set of genes associated with a particular condition, all pairs that are within a specified range (Direct–Indirect Range) are selected. Such a range may be chosen in order to

minimize side effect. The potential targets are filtered, to avoid further side effects, by removing targets that lie outside a pair range. At this point, it is necessary to apply a heuristic to select the minimum list of targets needed to affect all genes of interest [51]. We select the targets that reach the largest number of genes of interest, and remove them from our list. The process is repeated until all genes of interest are reached. Each gene thus selected is, then, connected to predicted or validated drugs by means of DT-Hybrid and the results thus obtained can be used for subsequent experiments. In this way, we seek to obtain the minimum set of drugs that allows acting on the genes of interest, minimizing possible side effects, thus reducing toxicity associated with combined therapy.

2.2 Beyond Hybrid Methods and Drug Repositioning

The DTI prediction algorithms should also work in case new compounds or new targets, for which no information is yet known, are introduced into the system. The main problem of recommendation algorithms is that, despite their accuracy, they fail to produce predictions in presence of these conditions.

2.2.1 Limitations of Recommendation Algorithms

Consider, for example, the addition of a new compound for which only structure is known, but no specific targets are available. The initial resource $f(o_i)$ to be assigned to known target nodes would be zero. Therefore, Eqs. 8 and 8a would always return null values, and no prediction can be made.

This situation is not unrealistic, there are many drugs designed for a specific purpose, which, however, fail the early trial stages because they do not work on the targets for which they were developed. In this case, finding possible targets is fundamental in order to predict new uses for them. The process described here is an example of drug repurposing.

A simple and natural strategy to formulate predictions of drugs for which no known information is available can exploit a CF-like approach. Let d_t be a drug for which there is no known target, but only structural information is available. We can compute the similarity of such a drug with the others, and select those that have a high similarity (i.e., greater than 0.8). Such targets can be exploited as possible initial knowledge for d_t, filtering out those that do not appear in the majority of cases.

This also applies in the presence of new targets for which no known drug is known to work. In this case, suggesting possible novel therapies is important if they represent key molecular elements in disease processes.

The CF-like strategy described above presents some problems: the main choices, such as the similarity threshold and the selection of the initial targets, are arbitrary and depend strongly on the user. Recommendation applied on tripartite networks is a way to reduce the number of arbitrary choices, leaving to the user only the selection of the initial number of predicted targets to use in the DTI prediction phase.

Consider, for example, the problem of predicting an initial set of target for a new drug. A drug–drug–target tripartite network can be built, and, by means of a tripartite network recommendation algorithm, an initial set of targets can be predicted and exploited for the real DTI inference phase. Let $G(D, D, T, E, w)$ be such a tripartite graph, where D is the set of drug, T is the set of targets, E is a set of edges, and $w : E \to \mathbb{R}$ a weight function. The last two entities in the graph can be built as follows:

- Take all the experimentally verified DTI and assign them a weight equal to 1;
- Take all possible drug–drug pairs (avoiding self-connection) and assign them a weight equal to their similarity, computed as described previously.

In particular, this tripartite network can be compactly described by means of two adjacency matrices: the similarity matrix between drugs (S^1) and the original DTI adjacency matrix (A). The application of a tripartite network recommendation algorithm will return a list of drug–target predictions. Inferences will be available also for each drug for which there was no initial information. By taking the Top-L predictions of such drugs, we can build an initial set of targets to employ in a subsequent DTI prediction phase.

In [45], a methodology that extends DT-Hybrid to tripartite networks was defined. It uses a multi-level resource allocation process, which in each step takes into account the resource of the previous one. For simplicity, we call D' the first partition in our network, D the second one, and T the third. In the first level of the allocation process, an initial amount of resource is moved from D nodes to D' node and vice versa. In the second level, the resource is initially transferred from T nodes to D nodes, where it is combined with the previous level amount and, then, moved back to T nodes. In this way, we can define a procedure for the computation of predictions.

The process just described can be summarized in a cascaded application of DT-Hybrid. DT-Hybrid is applied separately to the S^1 and A matrices, obtaining, respectively, the R^{S^1} and R^A matrices. The final result of the algorithm is the matrix $R' = \left\{r'_{ij}\right\}_{m \times n}$ computed as:

$$R' = R^{S^1} \cdot R^A. \tag{17}$$

The methodology described above can also be applied when no acting drug is known for some targets. In order to achieve this, we need to build a tripartite network $G(T, T, D, E, w)$ where D and T are, respectively, the set of drugs and targets, E is the set of edges, and $w : E \to \mathbb{R}$ is an edges weight function. As before, such a network can be described in a compact manner by two matrices: the targets

similarity matrix (S) and the DTI network adjacency matrix (A). Therefore, by applying our tripartite recommendation, the Top-L predictions, provided for a target of which no initial information is known, will constitute the list of drugs to be used for the subsequent DTI prediction phase.

The methodology described above is not a definitive solution to the problem of new drugs and targets, but it is a starting point to increase the usage of recommendation systems in this application field.

2.2.2 Tripartite Network Recommendation: An Approach to Drug Repositioning

An additional problem in the field of computational drug design is drug repositioning. It is the process of automating the discovery of new uses for existing drugs, resulting in a positive impact on time and cost for the discovery of such therapies.

In principle, knowing all possible targets of a drug allows researcher to check under which diseases it will work, and what will be its possible effect. Such knowledge is rarely available, but the use of DTI prediction techniques can have a positive influence in this type of study. Predicting unknown targets and associating them with the related diseases is a technique to guide the experimental work and define possible new uses for drugs already employed in clinical practice.

In this sense, the recommendation techniques applied on tripartite networks can automate the process previously described. Let $D = \{d_1, d_2, \ldots, d_n\}$ be a set of drugs, $T = \{t_1, t_2, \ldots, t_m\}$ be a set of targets, and $P = \{p_1, p_2, \ldots, p_k\}$ be a set of diseases. From experimentally validated information we can build a tripartite graph $G(D, T, P, E)$, where E is the set of all possible edges, namely all drug–target and target–disease interactions. The information contained in such a graph can be summarized in two matrices:

- $A^{DT} = \{a_{ij}^{DT}\}_{n \times m}$, where $a_{ij}^{DT} = 1$ if drug d_i acts on target t_j, $a_{ij}^{DT} = 0$ otherwise;
- $A^{TP} = \{a_{lo}^{TP}\}_{m \times k}$, where $a_{lo}^{TP} = 1$ if target t_l is associated with disease p_o, $a_{lo}^{TP} = 0$ otherwise.

The tripartite recommendation algorithm described above applied to graph G will result in the matrix $R' = \{r'_{io}\}_{n \times k}$, where r'_{io} indicates the degree of certainty with which we can associate the drug d_i with pathology p_o. Such a drug-disease score is computed simultaneously based on the number of predicted and validated targets that act on a drug, and the number of diseases associated with such targets. This implies that a drug that acts on many targets associated with the same disease will obtain high score.

The methodology described above allows us to infer possible novel connections between drugs and diseases that can make experimental research more focused, getting the most significant results in less time and with lower costs.

3 Conclusions

An important role in the reduction of the costly and time-consuming phases of drug discovery and design is played by bioinformatics. The usage of algorithms and systems for the prediction of novel drug–target interactions is a common practice. Be aware of the possible unknown effects on the proteome of a drug which can be crucial in exploiting its true potential or predicting side effects. Drug repositioning, drug combinations or substitutions reduce the need to develop new drugs. Drug repositioning identify new therapeutically purposes for drugs, while drug combination tries to modify or intensify the overall effect of two or more drugs. This is the context in which our approach DT-Web (available at http://alpha.dmi.unict.it/dtweb/) fits. Its main goal is to provide a simple system allowing users to quickly browse predictions of probable novel DTI, to produce new ones from their own data, or to simplify the experimental studies described above. This objective is achieved by using a database which combines our resource DT-Hybrid with data extracted from Drug-Bank and Pathway Commons. We also extended in a simple and natural way our DT-Hybrid algorithm to deal with compounds or molecules that are isolated within the bipartite networks (have not known target). Finally, we described a generalization of our methodology that goes beyond bipartite network and is able to deal with multipartite one.

References

1. Iskar M, Zeller G, Blattmann P et al (2013) Characterization of drug-induced transcriptional modules: towards drug repositioning and functional understanding. Mol Syst Biol 9:662
2. Li H, Gao Z, Kang L et al (2006) TarFisDock: a web server for identifying drug targets with docking approach. Nucleic Acids Res 34: W219–W224
3. Keiser MJ, Roth BL, Armbruster BN et al (2007) Relating protein pharmacology by ligand chemistry. Nat Biotechnol 25:197–206
4. Hopkins AL (2008) Network pharmacology: the next paradigm in drug discovery. Nat Chem Biol 4:682–690
5. González-Díaz H, Prado-Prado F, García-Mera X et al (2011) MIND-BEST: web server for drugs and target discovery; design, synthesis, and assay of MAO-B inhibitors and theoretical-experimental study of G3PDH protein from Trichomonas gallinae. J Proteome Res 10: 1698–1718
6. Keiser MJ, Setola V, Irwin JJ et al (2009) Predicting new molecular targets for known drugs. Nature 462:175–181
7. Kuhn M, Szklarczyk D, Pletscher-Frankild S et al (2014) STITCH 4: integration of protein–chemical interactions with user data. Nucleic Acids Res 42:D401–D407
8. Yildirim MA, Goh KI, Cusick ME et al (2007) Drug-target network. Nat Biotechnol 25: 1119–1126
9. Phatak SS, Zhang S (2013) A novel multimodal drug repurposing approach for identification of potent ACK1 inhibitors. Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing. NIH Public Access, 2013.
10. Campillos M, Kuhn M, Gavin AC et al (2008) Drug target identification using side-effect similarity. Science 321:263–266
11. Jin G, Wong STC (2014) Toward better drug repositioning: prioritizing and integrating existing methods into efficient pipelines. Drug Discov Today 19:637–644
12. Chen X, Liu MX, Yan G (2012) Drug-target interaction prediction by random walk on the heterogeneous network. Mol BioSyst 6: 1970–1978

13. Mei JP, Kwoh CK, Yang P et al (2013) Drug-target interaction prediction by learning from local information and neighbors. Bioinformatics 29:238–245
14. Iorio F, Bosotti R, Scacheri E et al (2010) Discovery of drug mode of action and drug repositioning from transcriptional responses. Proc Natl Acad Sci 107:14621–14626
15. Yamanishi Y, Araki M, Gutteridge A et al (2008) Prediction of drug--target interaction networks from the integration of chemical and genomic spaces. Bioinformatics 24:i232–i240
16. Cheng F, Liu C, Jiang J et al (2012) Prediction of drug-target interactions and drug repositioning via network-based inference. PLoS Comput Biol 8(e1002503)
17. Zhou T, Ren J, Medo M et al (2007) Bipartite network projection and personal recommendation. Phys Rev E 76:046115–046122
18. Alaimo S, Pulvirenti A, Giugno R et al (2013) Drug–target interaction prediction through domain-tuned network-based inference. Bioinformatics 29:2004–2008
19. van Laarhoven T, Nabuurs SB, Marchiori E (2011) Gaussian interaction profile kernels for predicting drug–target interaction. Bioinformatics 27:3036–3043
20. Lamb J, Crawford ED, Peck D et al (2006) The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science 313:1929–1935
21. Amelio I, Gostev M, Knight RA et al (2014) DRUGSURV: a resource for repositioning of approved and experimental drugs in oncology based on patient survival information. Cell Death Dis 5:e1051–e1055
22. Dudley JT, Sirota M, Shenoy M et al (2011) Computational repositioning of the anticonvulsant topiramate for inflammatory bowel disease. Sci Transl Med 3:96ra76
23. Sirota M, Dudley JT, Kim J et al (2011) Discovery and preclinical validation of drug indications using compendia of public gene expression data. Sci Transl Med 3:77
24. Li J, Lu Z (2012) A new method for computational drug repositioning using drug pairwise similarity. Presented at the IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2012, Ottawa, ON, Canada
25. Li J, Lu Z (2013) Pathway-based drug repositioning using causal inference. BMC Bioinformatics 14:S3
26. Li Y, Agarwal P (2009) A pathway-based view of human diseases and disease relationships. PLoS ONE 4:e4346
27. Tang J, Karhinen L, Xu T et al (2013) Target inhibition networks: predicting selective combinations of druggable targets to block cancer survival pathways. PLoS Comput Biol 9:e1003226–16
28. Ma J, Zhang X, Ung CY et al (2012) Metabolic network analysis revealed distinct routes of deletion effects between essential and non-essential genes. Mol BioSyst 8:1179–1186
29. Barve A, Rodrigues JFM, Wagner A (2012) Superessential reactions in metabolic networks. Proc Natl Acad Sci 109:E1121–E1130
30. Yeh I, Hanekamp T, Tsoka S et al (2004) Computational analysis of Plasmodium falciparum metabolism: organizing genomic information to facilitate drug discovery. Genome Res 14:917–924
31. Singh S, Malik BK, Sharma DK (2007) Choke point analysis of metabolic pathways in E. histolytica: a computational approach for drug target identification. Bioinformation 2:68
32. Perumal D, Lim CS, Sakharkar MK (2009) A comparative study of metabolic network topology between a pathogenic and a non-pathogenic bacterium for potential drug target identification. Summit on Translat Bioinforma 2009:100
33. Fatumo S, Plaimas K, Mallm J-P et al (2009) Estimating novel potential drug targets of Plasmodium falciparum by analysing the metabolic network of knock-out strains in silico. Infect Genet Evol 9:351–358
34. Zimmermann GR, Lehar J, Keith CT (2007) Multi-target therapeutics: when the whole is greater than the sum of the parts. Drug Discov Today 12:34–42
35. Savino R, Paduano S, Preianò M et al (2012) The proteomics big challenge for biomarkers and new drug-targets discovery. Int J Mol Sci 13:13926–13948
36. Bush K, Courvalin P, Dantas G et al (2011) Tackling antibiotic resistance. Nat Rev Microbiol 9:894–896
37. Kitano H (2004) Biological robustness. Nat Rev Genet 5:826–837
38. Logue JS, Morrison DK (2012) Complexity in the signaling network: insights from the use of targeted inhibitors in cancer therapy. Genes Dev 26:641–650
39. Nussinov R, Tsai C-J, Mattos C (2013) "Pathway drug cocktail": targeting Ras signaling based on structural pathways. Trends Mol Med 19:695–704
40. Holzapfel G, Buhrman G, Mattos C (2012) Shift in the equilibrium between on and off states of the allosteric switch in Ras-GppNHp affected by small molecules and bulk solvent composition. Biochemistry 51:6114–6126
41. van der Greef J, McBurney RN (2005) Rescuing drug discovery: in vivo systems

pathology and systems pharmacology. Nat Rev Drug Discov 4:961–967

42. Haibe-Kains B, El-Hachem N, Birkbak NJ et al (2013) Inconsistency in large pharmacogenomic studies. Nature 504:389–393
43. Alaimo S, Bonnici V, Cancemi D et al (2015) DT-Web: a web-based application for drug-target interaction and drug combination prediction through domain-tuned network-based inference. BMC Syst Biol 9:S4
44. Konstan JA, Miller BN, Maltz D et al (1997) GroupLens: applying collaborative filtering to Usenet news. Commun ACM 40:77–87
45. Alaimo S, Giugno R, Pulvirenti A (2014) ncPred: ncRNA-disease association prediction through tripartite network-based inference. Front Bioeng Biotechnol 2:71.
46. Herlocker JL, Konstan JA, Terveen LG et al (2004) Evaluating collaborative filtering recommender systems. ACM Transact Inform Syst 22:5–53
47. Swets JA (1963) Information retrieval systems. Science 141:245–250
48. Altschul SF, Gish W, Miller W et al (1990) Basic local alignment search tool. J Mol Biol 215:403–410
49. Smith TF, Waterman MS (1981) Identification of common molecular subsequences. J Mol Biol 147:195–197
50. Hattori M, Okuno Y, Goto S et al (2003) Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. J Am Chem Soc 125: 11853–11865
51. Chvatal V (1979) A greedy heuristic for the set-covering problem. Math Oper Res 4: 233–235
52. Tribus Myron Thermostatics and thermodynamics: an introduction to energy, information and states of matter, with engineering applications. van Nostrand, 1961.