7 Computational Methods for Drug Repurposing

Rosaria Valentina Rapicavoli, Salvatore Alaimo, Alfredo Ferro, and Alfredo Pulvirenti

Abstract

The wealth of knowledge and multi-omics data available in drug research has allowed the rise of several computational methods in the drug discovery field, resulting in a novel and exciting strategy called drug repurposing. Drug repurposing consists in finding new applications for existing drugs. Numerous computational methods perform a high-level integration of different knowledge sources to facilitate the discovery of unknown mechanisms. In this chapter, we present a survey of data resources and computational tools available for drug repositioning.

Introduction

Systematic drug repurposing, also known as drug repositioning, is the re-evaluation of known, pharmaceutically relevant compounds to identify new therapeutic applications.

Finding alternative uses for old drugs has the advantage of optimizing the discovery and development research process, yielding high cost and time savings in drug development. Since in vitro and in vivo screening, chemical optimization, toxicology, mass production, and clinical trials have already been completed and can be bypassed, substantial risks and "overheads" are removed from the path to market (these are known as Bioavailability and Absorption, Distribution, Metabolism, Excretion and Toxicity—ADMET profiles) [1].

Ideal drug candidates for repurposing are those that have passed Phase III, in terms of the American Food and Drug Administration (FDA) system, as this implies that they have proven to be effective in larger populations and verified to be safe. In this way, clinical trials can proceed at a much faster rate [1].

A repurposed drug does not need the initial 6–9 (or more) years, neither 2–3 billion dollars typically required for new drug development [2, 3], but it will proceed directly to preclinical testing and clinical trials, resulting in reduced risks and costs.

R. V. Rapicavoli
Department of Physics and Astronomy,
University of Catania, Catania, Italy

Department of Clinical and Experimental Medicine,
Bioinformatics Unit, University of Catania,
Catania, Italy

S. Alaimo · A. Ferro · A. Pulvirenti (✉)
Department of Clinical and Experimental Medicine,
Bioinformatics Unit, University of Catania,
Catania, Italy
e-mail: alfredo.pulvirenti@unict.it

A. Laganà (ed.), *Computational Methods for Precision Oncology*, Advances in Experimental Medicine and Biology 1361, https://doi.org/10.1007/978-3-030-91836-1_7

Among the best-known examples, sildenafil citrate (brand name: Viagra) [3] has been repurposed from a common hypertension drug to therapy for erectile dysfunction.

There are many successes in repositioning old drugs, and what was initially driven by serendipity is now operated by focused and systematic computational explorations that precede shorter experimental design cycles [1].

In a world where thousands of therapeutic molecules are known, drug repositioning is becoming an attractive form of drug discovery with a significant impact on personalized medicine.

Customizing or optimizing repositioning methods into efficient drug repositioning pipelines requires a comprehensive understanding of the available methods obtained by evaluating both biological and pharmaceutical knowledge and the mechanism of action of drugs [3].

In addition, the advent of high-throughput technologies to explore biological systems (drug-related data, high-throughput genomic screens, protein structures) resulted in the generation of an impressive amount of data that requires computational analysis and mining tools to be explored and used. Methods and tools available in chemoinformatics, bioinformatics, network biology, and systems biology play a key role in making full use of known targets, drugs, and biomarkers or disease pathways, thus leading to the development of proof-of-concept methods and accelerated timeframe clinical trial design.

Repositioning involves a deep synergy of investigators and computational scientists to develop relevant and realistic exploration tools. However, advanced computational tools are often difficult to understand or use, limiting their accessibility to scientists without a solid computational background [1]. For example, life scientists will find it challenging to use many of the computational tools that require data preparation, installation, and execution of packaged software; computer scientists, on the other hand, will not be able to make experimental validations of predictions.

In this chapter, we describe how to choose a proper drug repositioning approach based on information and knowledge, focusing on prioritizing the methods. Then, we discuss some of the tools built to facilitate the approach to this research field for both life scientists and computer scientists, bridging the gap due to different cultural backgrounds.

Drug Repositioning Methods and Approaches

Over the past few years, the number of drug repositioning methods has increased dramatically. Applying an efficient drug repositioning pipeline to a specific study requires identifying suitable methods based on available information about the drugs or diseases of interest [1, 3]. Therefore, it becomes essential to understand these existing methods better and prioritize them based on specific studies.

Computational drug repositioning methods can be classified as target-based, knowledge-based, signature-based, pathway- or network-based, and mechanism-targeted methods.

According to the information available and the elicited mechanisms, methods can be defined as drug-oriented, disease-oriented, and treatment-oriented. Therefore, these computational drug repositioning methods allow researchers to screen almost any drug candidate and test it on a large number of diseases in a relatively short time [1].

Because repositioning studies are tied to prior knowledge and available information, this will guide the choice of a drug repositioning methodology, and therefore the prioritization: (i) when there is limited information available for the studied disease, phenotypic screening or off-label FDA use would be the best option; (ii) if a protein biomarker exists for the studied disease, target-based or knowledge-based methods should be prioritized; (iii) when disease information is available, knowledge-based or signature-based methods can be used to integrate available disease pathways or disease-related omics data into the drug repositioning process; (iv) when omics data related to drug treatment are available, signature-based or mechanism-targeted methods can be used to elucidate unknown targeted mechanisms, such as off-targets and targeted signaling pathways [1, 3].

Screening Methods or Blinded Research

Blind drug repositioning methods mainly depend on serendipitous identification from targeted disease- and drug-specific assays and do not involve pharmaceutical or biological information. These methods can be applied to a large number of drugs or diseases [3].

Target-Based Methods

Target-based drug repurposing methods involve in vitro and in vivo high-throughput or high-content screening (HTS/HCS) of drugs for a protein or biomarker of interest and an in silico screening of drugs or compounds from drug libraries. The use of target information in drug repurposing ensures a greater chance of finding valuable drugs than blind methods. These methods allow researchers to screen almost any drug or compound with known chemical structure information within days [3].

Knowledge-Based Methods

Knowledge-based drug repositioning methods apply bioinformatics or cheminformatics approaches to integrate available drug information, drug-target networks, chemical structures of targets and drugs, clinical trial information (including adverse effects), FDA approval labels, and signaling or metabolic pathways. This knowledge is then used to predict unknown mechanisms, unknown drug similarities, and new biomarkers for diseases [3].

Signature-Based Methods

These methods rely on the use of gene signatures derived from disease omics data (i.e., microarray, RNA-seq), with or without treatments, to uncover unknown off-targets or unknown disease mechanisms. This type of data is now available on various databases, including NCBI Gene Expression Omnibus (GEO), Connectivity Map (CMap), and Cancer Cell Line Encyclopedia (CCLE). Signature-based methods can support discovering unknown mechanisms of action of molecules and drugs because they are supported by molecular information from which valuable information can be extracted (i.e., differential expression of genes concerning disease or drug administration). This method is advantageous when, for example, drugs need to be repurposed for a large number of diseases. Since the required knowledge (biomarkers, targets) may not be available or may be difficult to derive from available literature or databases, deriving gene signatures for those diseases from publicly available genomic data becomes the best option [3].

Pathway- or Network-Based Methods

Pathway or network-based drug repositioning methods use available disease omics data, signaling or metabolic pathways, and protein interaction networks to reconstruct disease-specific pathways that provide key targets for repositioned drugs. These methods are beneficial in identifying, within extensive pathways, subnetworks, or a small number of crucial, targetable proteins [3].

Targeted Mechanism-Based Methods

Targeted mechanism-based methods use treatment omics data, known signaling pathway information, or protein interaction networks to delineate unknown drug action mechanisms. The application of these approaches involves the use of sophisticated computational models that are characteristic of Systems Biology. Such models find vast space in the era of precision medicine and can also be a valuable support in clinical practice [3]. One potential application is studying the molecular mechanisms that lead cancer patients to drug resistance after a few months of treatment [3]. The methods described above demonstrate that the success of drug repositioning is closely related to the complexity and richness of the available information [3].

Drug Repurposing Tools: Web-Based Solutions

The field of drug repositioning requires the close collaboration of scientists belonging to different fields. Life scientists, experimental and clinical scientists, evaluate and interpret data and results. Computer scientists and bioinformaticians provide powerful computational software and systems to model the intrinsic complexity of biological models and make predictions to acquire novel knowledge.

The correct use of this type of software may be complex when the appropriate bioinformatics skills are lacking. For this reason, in the last few years, several tools available on the web have emerged. These provide easy-to-use computational solutions to bridge the gap between wet-lab scientists and the software tools available for drug discovery.

It is possible to consider three main categories of web-based platforms that help in drug repurposing based on the type of interaction used to perform repositioning studies: predicting drug-target interactions and using drug-induced gene expression to predict new connections and link drugs to disease.

Web-Based Tools: Predicting Drug-Target Interactions

Within this category, the various tools are classified into five subcategories based on the data used to do repositioning and how they are parameterized [1]:

1. Ligand similarity using fingerprint encoding
2. 3D structures of drugs and targets
3. Network-based approaches
4. Binding site parameterization
5. Other

Ligand Similarity Using Fingerprint Encoding

The paradigm underlying ligand-centered predictions is that the structural similarity implies comparable biological functions or properties. Similar compounds will therefore be likely to bind the same target, which is why a priori knowledge of query-binding targets is used to uncover previously unknown leads. In this sense, it becomes essential to know the fingerprint of molecules, be it 1D, 2D, or 3D.

Some tools belonging to this subcategory are indicated below.

ChemMapper

To find similar molecules and target annotations to identify candidate targets for a given query, ChemMapper uses a 3D similarity algorithm called SHAFTS (SHApe-FeaTure Similarity). The usage of 3D similarity metrics has been shown to improve off-target prediction accuracy [4].

SHAFTS relies on a triplet hashing technique for rapid alignment of molecular conformations and uses shape and chemotype to assess similarity [4].

ChemMapper uses drug information and target annotations from various sources such as ChEMBL, DrugBank, BindingDB, KEGG, and the Protein Database (PDB) [4].

ChemMapper offers the possibility to choose the most appropriate application depending on the final goal of the user (list of plausible proteins, related compounds) and the type of input available (protein ID, protein sequence, list of compounds) [4].

ChemProt 3.0

ChemProt 3.0 is a publicly available collection of chemical-protein disease annotation resources enabling the study of systems pharmacology for a small molecule at different levels of complexity (from molecular to clinical level) [5].

The platform allows users to navigate various data and make assessments from the global scale to specific analyses.

ChemProt 3.0 includes several computational approaches: Similarity Ensemble Approach—SEA, Quantitative Structure-Activity Relationship—QSAR, and network biology-based enrichment analysis [5].

These approaches support generating new hypotheses for bioactivity of novel and already annotated compounds and identifying other genes that may play significant roles in modulating chemical perturbations in humans [5].

The user can search for information about a compound, protein, or clinical outcome or can choose to perform a QSAR prediction for a specific compound. Each molecule can be imported as a SMILES (Simplified Molecular-Input Line-Entry System) code, or it can be drawn or uploaded from a compound structure file via the SD file format [5].

Through the "Heatmap" feature, ChemProt 3.0 allows the user to have a general overview of chemical-protein interactions, providing a global map linking bioactivities of compounds and proteins based on more than 7 million stored interactions collected from multiple databases annotating compounds, proteins, and diseases. ChemProt has one of the most extensive databases for each category (drugs, proteins, interactions, diseases) [5].

QSAR prediction can have two types of application cases: (i) Comparison of the query molecule with the drug set and thus the map will provide a method to navigate through known interactions. (ii) Prediction of new interactions. In this case, the similarity of the molecule fingerprints is used to generate similar drugs and predict the activity of the new compound [5].

HitPick

HitPick is a web server for identifying hits in high-throughput chemical screenings and predicting their molecular targets. It is currently the only resource that can process hits from chemical biology screening experiments and provide target prediction. Indeed, the user can upload the results of the biological assay [6].

HitPick applies the B-score method for identifying high-quality hits based on a statistical evaluation of many screening parameters and an integrative approach that combines 1-nearest-neighbor (1NN) similarity metrics and Laplacian-modified naïve Bayesian target models to predict the targets of identified hits [6].

Targets are predicted based on 2D molecular fingerprints.

The most similar compound from the compound-target interactions is identified using the pairwise Tanimoto coefficient. A ranking of target predictions will then be performed based on the Laplacian-modified Naive Bayesian method-based score.

iDrug-Target

iDrug-Target comprises four subpredictors: iDrug-GPCR, iDrug-Chl, iDrug-Ezy, and iDrug-NR, focusing, respectively, on the identification of drug interactions with G protein-coupled receptors (iDrug-GPCR), ion channels (iDrug-Chl), enzymes (iDrug-Ezy), and nuclear receptors (iDrug-NR) based on KEGG data. The predictions attempt to avoid oversampling due to non-interacting drug-target pairs. The Neighborhood Cleaning Rule and the Synthetic Minority Over-Sampling Technique are used to eliminate redundant negative samples, and some hypothetical positive samples are also added [7]. The Neighborhood Cleaning Rule (NCL) method is among the most popular under-sampling methods. All the samples of the class of interest are maintained, whereas those from the rest of the data are reduced [8]. On the other hand, Synthetic Minority Over-sampling TEchnique (SMOTE) is an over-sampling method that addresses this problem by creating synthetic minority samples to balance the data set. The minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any/all of the *k* minority class nearest neighbors [9].

iDrug-Target combines protein sequence encoding, using pseudo amino acid composition, with a 256-component 2D fingerprint representation of the ligand. This molecular signature must also be generated to construct a query. iDrug-Target uses a Support Vector Machine (SVM) to classify inputs as interactive or non-interactive [7].

Polypharmacology Browser—PPB

Merged footprints combining features between different footprints can also be generated. PPB searches through 4613 groups of at least ten annotated targets of bioactive molecules from ChEMBL and returns a list of predicted targets ranked by consensus voting scheme and *p*-value [10].

Targets can be ranked by their *p*-values. Indeed, it was found that the pairwise overlap between high confidence (low *p*-value) targets of different fingerprints was significantly higher than low confidence (high *p*-value) targets. In PPB, the similarity is calculated using city-block distances. This tool reports better performance from fusion and pairwise combination fingerprints than single fingerprints [10].

Similarity Ensemble Approach—SEA

SEA was the first tool that used ligand similarity to cluster proteins. The protein clusters thus formed represented functional themes that were potentially useful in predicting the polypharmacology of ligands [11].

The ligands were grouped according to the minimum coverage tree, while Tanimoto coefficients (TC) were used to determine similarity and Daylight 2D fingerprints. The encoding of the ligands is done through 2D fingerprints. The pipeline suggested by SEA, which leads to the identification of new suggestions of repositioned drugs, ultimately provides for validation with experimental techniques [11].

SuperPred

SuperPred is a prediction web server able to connect the chemical similarity of compounds to drugs with molecular targets and a therapeutic approach based on the principle of similar property [11, 12].

The ligand-target interactions are first aggregated by SuperTarget, ChEMBL, and BindingDB, then the set of ligands is normalized/cleaned using JChem to obtain a single set of ligands [11, 12].

Among them, only molecular targets will be extracted through the use of stringent binding affinity thresholds.

Drug-target prediction is achieved by considering the 2D Tanimoto similarity between a query compound and the ligands associated with their respective targets (target sets) [11, 12].

The specificity of each prediction is done through the calculation of two parameters called Z scores and *E*-values. The *E*-value is used as a threshold value. An *E*-value >1 means that the prediction is random. In order to evaluate the similarity between ligands, a weighting factor is calculated. The use of weight improves the accuracy of the predictions.

Thanks to the presence of these thresholds, SuperPred has a prediction success rate of 94.1% [11, 12].

SwissTargetPrediction

SwissTargetPrediction is a web server that has been online since 2014 [13, 14] and whose rationale is based on the observation that similar bioactive molecules are more likely to share similar targets. Thus, identifying proteins with known ligands similar to the query molecule can predict the targets of a given molecule.

This tool combines 2D and 3D similarity metrics to predict targets of bioactive molecules to improve target prediction accuracy. Query molecules can be inputted either as SMILES or drawn in 2D using the javascript-based molecular editor of ChemAxon. This system uses ChEMBL version 23 (the old version was based on ChEMBL version 16) as a data source. The dataset includes 376,342 unique compounds (580,496 binding activities on 3068 protein targets) [13, 14].

SwissTargetPrediction offers the possibility to perform predictions in different organisms, and mapping predictions by homology within and between different species is enabled for close paralogs and orthologs. The updated version can choose among humans, rats, and mice [13, 14].

The similarity quantification consists of calculating a pairwise comparison of 1D vectors describing molecular structures. The 2D measure uses the Tanimoto coefficient between path-based binary footprints (FP2), while the 3D measure is based on a Manhattan similarity distance between Electroshape 5D float vectors (ES5D) [13, 14].

Targets are prioritized based on a logistic regression of the 2D–3D similarity values [13, 14].

TarPred

TarPred is an online computational model based on a reference library containing 533 individual targets with 179,807 active ligands [13, 15].

Given a query compound, TarPred provides the first 30 ranked interacting targets. For each of them, the structure of the three most similar ligands is displayed, along with the disease indications associated with each target. This information helps understand the mechanisms of action and toxicity of active compounds and may offer new inputs for drug repositioning [13, 15].

To calculate the similarity of the query with the set of drug-related targets, TarPred also uses a combination of ECFP4 (Extended-Connectivity Fingerprints), designed for molecular characterization, similarity searching, and structure-activity modeling, together with the Tanimoto coefficient. The prioritized list of targets produced is closely associated with FDA-approved drugs [13, 15].

Protein sequences that interact with FDA-approved drugs (FDA-approved drug targets) are retrieved from DrugBank and subjected to BLAST against BindingDB proteins [13, 15].

TarPred calculates ECFP4 similarity scores between the query compound and ligand sets, producing a ranked list of targets [13, 15].

TargetHunter

TargetHunter is a web-based tool that uses an algorithm based on the Tanimoto similarity index, called TAMOSIC (Targets Associated with its MOst SImilar Counterparts) [16].

The similarity to the query compound is calculated using three different 2D fingerprints. The targets associated with the first "N" most similar compounds are shown as possible targets. The data for the compounds are retrieved from ChEMBL [16].

3D Structures of Drugs and Targets

Structure-based design is founded on the knowledge of the three-dimensional structure of the molecular target for the drug. The methods to derive the 3D structure are X-ray crystallography and NMR solution. Alternatively, homology models based on related proteins are commonly used.

This type of approach focuses on exploring the similarity of binding sites from PDB crystal structures.

Structure-based design predominantly uses molecular modeling techniques such as docking and pharmacophore models to calculate binding affinities of leads.

From the computational point of view, these techniques are more expensive. In fact, most of the computational research in this area is used to create predictive software rather than building real-time web-based applications.

Below are some web services that use this type of approach.

idTarget

idTarget can predict possible binding targets of a small chemical molecule via a *divide et impera* docking approach combined with scoring functions based on regression analysis and quantum chemical charge models. The affinity profiles of the protein targets are used to provide the confidence levels of the prediction. The *divide et impera* docking approach uses small overlapping grids adaptively constructed to limit the search space, thus achieving better efficiency in terms of time. idTarget performs screening on almost all protein structures deposited in the Protein Data Bank (PDB) [17].

The search engine of the idTarget web server is MEDock, which generates initial docking poses of the small ligand [17].

Protein-Drug Interaction Database (PDID)

PDID can be used to systematically catalog protein-drug interactions and facilitate various studies related to drug polypharmacology and drug repurposing.

PDID queries the binding sites within the PDB's drug-protein complexes based on stringent filters against all other proteins on the PDB to find likely off-targets of the original drugs [18].

PDID uses experimentally curated interactions present in DrugBank, BindingDB, and Protein Data Bank.

PDID is based on nearly 1.1 million all-atom predictions on the entire human structural proteome (10,000 structures for over 3700 proteins) and provides access to all putative targets (between 4444 and 7184, depending on the prediction method used) of several popular drugs. Therefore, it represents a valid starting point for drug repositioning [18].

TARget FIShing DOCKing (TarFisDock)

TarFisDock is a tool created to automatize the procedure of searching for small molecule-protein interactions on an extensive repertoire of protein structures. It provided a database of potential drug targets (PDTDs) containing 698 protein structures covering 15 therapeutic areas and was one of the first online tools to offer a reverse ligand-protein docking program. Reverse ligand-protein docking aims to search for potential protein targets by examining an appropriate protein database [19].

TarFisDock requests as input the small molecule to be tested in standard mol2 format and performs the docking through the DOCK 4.0 algorithm using protein structures present in PDTD. Targets can be provided by the user or retrieved from PDTD. The ligand-protein interaction energy terms of the DOCK program are adopted to classify proteins [19].

Network-Based Approaches

Many databases store annotations on system-wide biological networks, including information on various entities that interact with drugs (e.g., targets). Integrating these types of biological networks can help understand the pharmacological properties of specific molecules and thus in drug repositioning. However, working with this kind of data poses new challenges related to managing multidimensional interaction networks.

BalestraWeb

BalestraWeb is an online service that allows users to make predictions about potential interactions between a chosen drug and target or predict the most likely interaction partners of any drug or target listed in DrugBank. It also enables to perform similarity search between drugs or determine the most similar targets based on their interaction patterns [20]. The system uses active learning (AL) techniques relying on probabilistic matrix factorization (PMF) to calculate the statistical weight of each approved drug for all targets associated with the entire set of approved drugs. The server allows three types of queries to be submitted: drug-target interaction, drug-drug similarity, and target-target similarity [20].

Predictions made by BalestraWeb are not dependent on structural or chemical similarities [20].

CSNAP

CSNAP (Chemical Similarity Network Analysis Pull-down) is a computational tool for target identification based on network similarity.

The method combines chemical similarity networks (CSNs) and chemical consensus that results in chemotype-based subnetworks, which predict targets for a set of drug classes [21].

The compounds and their information (e.g., bioactivity) are stored in databases such as ChEMBL and PubChem. Such compounds are grouped by CSN, and target prediction will be based on a consensus statistic determined by the target frequency shared by the first neighbors centered on the compound in the query. The resulting subnetwork will consist of nodes representing compounds and edges representing similarity [21].

The S score is used to rank the targets of the first neighbor compounds, and the significance of each composite protein pair is calculated using an H score. CSNAP appears to have greater predictive ability than the SEA approach [21].

DASPfind

DASPfind is a web service for identifying novel drug-target interactions using “simple paths” of particular lengths inferred from a heterogeneous graph composed of three types of subgraphs: drug-target interactions, drug-drug similarities, and similarities between drug-protein targets [22].

The various known interactions were extracted from the KEGG BRITE, BRENDA, SuperTarget, and DrugBank databases. The chemical structures of the drugs were extracted from the KEGG LIGAND database, and the similarities between the drugs were calculated using SIMCOMP. Target similarity scores are calculated using a normalized version of the Smith-Waterman algorithm [22].

DASPfind performs best when a subjective test using only the "top 1 candidate" is used [22].

nAnnoLyze

nAnnoLyze is a web-based tool for target identification centered on the hypothesis that structurally similar binding sites associate with similar ligands and is based on network-based comparative docking called Annolyze. nAnnoLyze integrates structural information into a bipartite network of interactions and similarities to predict compound-protein structural interactions on a proteomic scale [23].

This network consists of compounds found in PDB, protein-binding sites from LigBase, human proteome structure from ModBase, and DrugBank compounds [23].

Then, the protein subnetwork is constructed using targets that bind ligands above a threshold of drug similarity. The network is connected using the structural similarity of the binding sites calculated by ProBis. The two subnetworks are joined if a resolved PDB structure validates a known ligand-target interaction. Only proteins that have a resolved 3D structure are used for nAnnolyze predictions [23].

PROMISCUOUS

PROMISCUOUS is one of the first public network-based Web servers for drug repurposing [24, 25]. The network employed consists of nodes representing drugs, proteins, side effects, and edges representing drug-target, drug-drug, target-target, and drug-side effect interactions. The information to support the network comes from publicly available databases such as SuperDrug, DrugBank, ChEMBL, SIDER, TTD, SuperTarget, and SuperPred [24, 25].

In the updated version of the Promiscuous 2.0 Database, the number of drugs and drug-like compounds has been significantly increased from 25,000 to nearly 1 million (side effects ~110,000, drug-target interactions ~3 million), compared to the first version. Promiscuous 2.0 also includes a section devoted to potential treatments for COVID-19 [24, 25].

Promiscuous is an easy-to-use resource that allows users to interactively create complex interaction networks and infer new indications for existing compounds. Users can also submit new molecular structures and be presented with suggested application areas or, vice versa, get potential drug candidates for disease indications of interest [24, 25].

SLAP

Semantic Link Association Prediction (SLAP) is a web-based tool that predicts associations between drugs and targets through semantic database integration and statistical modeling. SLAP predicts associations using "path models," predefined association paradigms that include nodes and edges [26]. These are part of a semantic network constructed using drug-drug and protein-protein similarity and drug-target interactions from Chem2Bio2RDF during semantic annotations from the Chem2Bio2OWL ontology. The drug-target pairs used to construct the association network are taken from DrugBank [26].

SLAP uses the Heap-based Dijkstra algorithm to find the shortest path length between two nodes (shortest path length < 3). The predicted values are associated with a *p*-value, calculated as the sum of the Z-scores of all valid paths between two nodes, that allows their ranking based on significance [26].

Three types of input can be given to SLAP: drug-pair and predict association; targets predicted by drugs and drugs with similar biological function; proteins alone and get associated ligands/for and obtain associated ligands [26].

The performance of SLAP is comparable to SEA for drug-target predictions and CMap for drug-association predictions [26].

STITCH

STITCH, a search tool for interacting chemicals, is a web service focused on providing the user a comprehensive map of drug-target interactions with sophisticated filters and visualization [27–30].

We can consider this platform as an interface that integrates drug-target interaction data resources derived from high-throughput, manually curated database experiments and many predictive algorithms.

STITCH has been updated many times and, over the many years of development, has been connected to many databases such as DrugBank, GLIDA, MATADOR, TTD, CTD, KEGG, PID, Reactome, BioCyc, ChEMBL, PDSP Ki Database, and PDB. Over time, STITCH has also been implemented with automated text mining algorithms that predict interactions based on co-occurrence in PubMed, MEDLINE, and NIH Re-PORTER [27–30].

A confidence score is given for each interaction indicating its level of significance and certainty.

As input, it is possible to give a chemical name, a gene name, a chemical structure, or a protein sequence, from which a network of interactions with related chemicals and proteins will be generated [27–30]. STITCH is a well-established and widely used resource by many research groups that directly use its results [27–30].

DT-Web

DT-Web [31] is a web-based application to the Domain Tuned-Hybrid (DT-Hybrid) [32], which extends a well-established recommendation technique from domain-based knowledge that includes drug and target similarity.

This method, together with domain-specific knowledge expressing drug-target similarity, is used to calculate recommendations for each drug.

DT-Web can consider different matrices as input: known drug-target matrix, drug-drug similarity matrix, and target-target similarity matrix.

The drug-target interactions are taken from DrugBank, and from this data, an adjacency matrix is constructed. The drug-drug similarity is assessed using SIMCOMP, and then a similarity matrix is constructed. The target similarity matrix can be obtained by performing BLAST or using the Smith-Waterman local alignment technique.

Then, using these three matrices, a drug-target interaction network is constructed. Each target is mapped to its Entrez Identifier and annotated with Gene Ontology (GO) terms in this interaction network. For each pair of GO terms, the similarity score is calculated. Therefore, a *p*-value is calculated to evaluate the association between the predicted and validated targets.

Another potential of DT-Web is that, given a set of candidate disease genes as input, it can predict drug combinations whose targets are at an optimal distance from those genes. DT-Web shows better results than NBI and Hybrid, network-based interaction prediction algorithms.

Searching off-lAbel dRUg aNd NEtwoRk—SAveRUNNER

SAveRUNNER is a freely available network-based algorithm for drug repurposing to detect potential new indications for existing drugs that could be used for other purposes [2, 33].

Starting from a list of drug-target interactions and disease-gene associations, this tool predicts drug-disease associations by computing a new network-based similarity measure that prioritizes associations between drugs and diseases located in the same neighborhoods [2, 33].

The SAveRUNNER pipeline consists of two macro steps [2, 33]:

1. The construction of the proximity-based drug-disease network
2. The construction of a similarity-based bipartite drug-disease network

The construction of the proximity-based drug-disease network comprises three phases:

Computation of network proximity (*p*) to measure how close the disease and drug modules are in the human interactome. Given two modules T and S that, respectively, represent the drug module, containing all *t* targets of the drug, and the disease module, comprising all *s* genes of the disease, we can describe this measure as the average length of the shortest path between the elements of T and S [2, 33].

Computation of z-score proximity and p values. SAveRUNNER calculates z-scores and their *p*-values by building a reference distance distribution corresponding to the expected distance between two randomly selected sets of proteins with the same size and degree distribution as the original sets of disease proteins and drug targets in the human interactome. The procedure is repeated 1000 times, and the z-score and its *p*-value are calculated through the mean and standard deviation of the reference distance distribution [2, 33].

Selection of statistically significant drug-disease associations by filtering *p*-values (generally, *p*-value ≤0.05) [2, 33].

Next, the pipeline involves the construction of a similarity-based bipartite drug-disease network that comprises the following steps:

Computation of Network Similarity

The similarity measure is calculated from the network proximity measure *p* through the equation

$$\text{Similarity} = \frac{\max(p) - p}{\max(p)} \quad p$$

$$= \text{network proximity.}$$

This measure assumes a value between 0 and 1 [2, 33].

Cluster Detection

SAveRUNNER uses a clustering algorithm based on greedy optimization of the modularity network to define drug and disease groups. Each identified cluster is evaluated by the cluster quality score (QC) [2, 33].

Adjustment of Network Similarity

If a drug and a disease are part of the same cluster, the drug can probably be repurposed for the disease. Thus, the drug-disease pair should have a higher similarity [2, 33].

Therefore, the similarity of a drug-disease pair belonging to the same cluster is increased proportionally to the cluster's QC score. On the other hand, if two nodes do not fall into the same cluster, QC is set to zero and the similarity value does not change [2, 33].

Normalization of Network Similarity by Applying a Sigmoid Function

SAveRUNNER outputs a list of predicted and prioritized drug-disease associations in a weighted bipartite network format, in which nodes represent drugs and diseases. A link between a drug and a disease occurs if the corresponding drug targets and disease genes are close in the interactome with a significant *p*-value ($p \leq 0.05$). Their interactions are represented by weighted edges in which the weight corresponds to the adjusted and normalized similarity value [2, 33].

Binding Site Parametrization

Binding sites are structural regions of macromolecules that bind ligands through interactions that are almost always reversible and can often be accompanied by conformational changes in the molecules. These are often conserved regions that can be used to search for other ligand-binding proteins that generally bind to other molecules by exploiting the structural similarity of these binding regions. Below, we explore some of the methods designed to predict targets based on the binding sites of query molecules.

ProBis

The ProBiS-ligands Web server predicts the binding of ligands to a protein structure. Starting with a protein structure or binding site, ProBiS-ligands identify model proteins in the Protein Data Bank (PDB) that share similar binding sites to the query [34].

The algorithm uses the structure and physicochemical properties of the constituent amino acids and their backbones to compare two protein-binding sites [34].

Then, it detects structures sharing similar 3D amino acid motifs to the searched protein within the PDB [34].

ProBiS-Database is a repository of non-redundant-binding sites and associated PDB structures, which is updated weekly. ProBiS can be used through pre-calculated data to get results faster or by starting from scratch by looking for a specific protein [34].

PoSSuM

Pocket Similarity Search using Multiple-sketches, PoSSuM, searches the entire PDB database for binding similarity of all coupling molecules. PoSSuM accepts three types of input: a protein structure; a ligand-binding site; and a ligand [35, 36].

Given a protein query, PoSSuM will search for all known ligand-binding sites with a structure similar to the input. PoSSuM can search for any known ligand-binding site or putative-binding site [35, 36]. It uses a neighbor-searching algorithm called SketchSort. The similarity measure is determined based on cosine similarity and a *p*-value indicating significance [35, 36]. On the other hand, dissimilarity values are given by the mean square deviation [35, 36].

Other Web-Based Tools

This section is dedicated to tools that use disease association-dependent annotations. Disease-based approaches are used when drug pharmacology is not present or not considered.

MeSHDD

MeSHDD is a literature-based repositioning methodology that leverages drug-drug similarity based on the MeSH term co-occurrence [37]. MeSHDD clusters drugs based on disease-centered Medical Subject Heading (MeSH) terms found in the MEDLINE Baseline Repository, which contains manually annotated MeSH terms for over 20 million biomedical articles, to predict shared indications [37].

MeSHDD uses drugs from DrugBank, including manually curated information on approved, investigational, and illicit drugs and their targets, mechanisms of action, and indications. Co-occurrence of drug-MeSH terms is calculated using a hypergeometric *P*-value, followed by a Bonferroni correction [37]. The drug-drug similarity is measured by calculating the bitwise distance from converting *p*-values to a binary representation. Drugs are clustered based on pairwise distances and bootstrap-means clustering techniques (implemented in *R*), and the Jaccard index was used to compare the clustering of various *k*-values [37].

RE:fine Drugs

RE:fine drugs is a freely available interactive dashboard for integrated search and discovery of drug repurposing candidates from GWAS and PheWAS repurposing datasets constructed using previously reported methods in Nature Biotechnology [38].

Given a disease as input to the web server, users receive a list of drugs that can potentially treat that disease [38].

The output predictions are classified as known/discovered if present in DrugBank, strongly supported if present in the NIH clinical trial registry and biomedical literature, probable if the evidence is in the NIH clinical trial registry or biomedical literature, and novel if not present in either [38].

Bayesian ANalysis to Determine Drug Interaction Targets—BANDIT

BANDIT is a machine learning algorithm that uses a Bayesian approach to integrate multiple data types to predict possible interactions with therapeutic effects. The rationale for this approach is integrating multiple data types to significantly improve the accuracy of target prediction [39].

BANDIT integrates over 20,000,000 data points from six distinct data types (drug efficacy, post-treatment transcriptional responses, drug structures, reported adverse effects, bioassay results, and known targets) [39]. The tool is based on a database containing approximately 2000 different drugs with 1670 different known targets and over 100,000 compounds without known targets (orphans) [39].

For each data type, a similarity score is calculated for all drug pairs with known targets. For each pair, BANDIT converts the similarity score into a likelihood ratio. These ratios are then combined to obtain a total likelihood ratio (TLR) proportional to the probability that two drugs share a target, given all available evidence [39].

The integrative approach of BANDIT can accurately identify drugs that share targets, discern the mechanisms of approved drugs, explain existing but not fully known clinical phenotypes, and repurpose drugs for new therapeutic indications [39]. Finally, BANDIT is a dynamic system that can be continuously updated [39]. BANDIT showed high accuracy in identifying shared target interactions and discovering novel targets for cancer treatment [39]. The use of this tool led to the identification of 14 novel microtubule inhibitors, including 3 with activity on resistant cancer cells [39].

Using Drug-Induced Gene Expression to Predict New Connections and Link Drugs to Disease

Drug-induced gene expression refers to the differential mRNA expression profiles in a cell line before and after drug treatment. This repurposing approach is accomplished by comparing disease-associated expression signatures with these drug-induced expression signatures, looking for drugs that have opposite effects on the disease and may be effective.

CMap

Connectivity Map (CMap) relies on a database of pre- and post-gene expression profiles from cellular samples in response to various types of perturbation, e.g., genetic perturbations in response to drug administration. CMap provides mRNA expression data from DNA microarrays for researchers who want to monitor differential expression to identify drugs that produce reverse signatures to query expression signatures. Connectivities are measured using the Kolmogorov-Smirnov statistical test. To date, CMap has generated a library containing over 1.5M gene expression profiles from ~5000 small molecule compounds and ~3000 gene reagents, tested in multiple cell types [40, 41]. CMap has profoundly impacted therapeutic research and has opened new challenges in scientific investigations in drug repurposing, MoA elucidation, biological understanding, and systems biology [40, 41]. It provides one of the most valuable and direct methods to investigate the alternative therapeutic potential of drugs [40, 41].

DeSigN

DeSigN (differentially expressed gene signatures—inhibitors) associates disease signatures with drug response signatures based on IC50 (quantitative measure of drug efficacy often used to prioritize compounds in vitro) data. Unlike CMap, which uses pre- and post-gene expression profiles, DeSigN uses only baseline gene expression profiles. DeSigN is constructed using GDSC [42].

GoPredict

GoPredict uses gene expression data integrated with heterogeneous public information, such as signaling pathways and drug-target information. It takes gene expression data as input and returns drug predictions as output. The reference databases used in GoPredict are TCGA, KEGGDrug, DrugBank, and Gene Ontology [43].

MANTRA 2.0

MANTRA 2.0 predicts molecular drug targets from gene expression profiles before and after drug perturbation in a collaborative and additive learning environment [44].

An automated pipeline of MANTRA 2.0 transforms the gene expression profiles into a single drug “node” in the network and allows users to explore their neighbors to find new indications and interactions. They calculate a prototype ranked list (PRL) for each drug, followed by a method to compare two PRLs using a Gene Set Ensemble Approach (GSEA) based method [44].

NFFinder

NFFinder uses the MARQ method to compare molecular signatures. Performing this analysis requires two sets of expression data, up- and downregulated genes compared to GEO, CMap, and DrugMatrix data [45].

PDOD

The online server Prediction of Drugs with Opposing Effects on Disease Genes—PDOD uses gene expression data and associates to them information regarding "effect-type" and "effect-direction" using pathway information (KEGG) and drug-target information from DrugBank [46]. It uses case/control expression datasets published in GEO to determine which gene expression changes happen due to a specific disease and looks for a drug that can counteract them [46].

To extract the gene signature, PDOD draws differentially expressed genes from the expression data by applying Limma and a function that evaluates the drug-disease score based on the parameterization of relationships [46].

RGES

The Reverse Gene Expression Score—RGES is a system providing a predictive measure on how a given drug could reverse the gene expression profile for a given disease. The principle consists of contrasting overexpressed while increasing weakly expressed ones, thus restoring gene expression to levels closer to normal tissue [33].

First, the computational pipeline needs to compute disease gene expression signatures and drug-induced gene expression signatures [33]. From these two molecular signatures, it can calculate the Reverse gene expression score (RGES) between disease and drug. This score ranges from −1 to 1, and it represents a measure of how much the drug under consideration can counteract the changes in expression due to disease. A low RGES value indicates higher potency to reverse disease gene expression and vice versa [33].

RGES is hence dependent on biological conditions. It is also reported that it is positively correlated with drug efficiency and, therefore, the IC50. RGES could also be used to provide insights into drug candidates' mechanisms [33].

The required data to perform the analysis can be taken from various publicly available databases such as TCGA, which includes gene expression profiles of tissue samples, LINCS, which includes perturbagen-mediated gene expression profiles, ChEMBL, which includes drug activity in cancer cells, and CCLE, which includes gene expression profiles of cancer cells [33].

Thanks to the progressively decreasing cost of many profiling technologies, large volumes of gene expression profiles of drugs in different biological conditions can be produced and made available to apply various drug repositioning and compound screening techniques such as RGES [33].

Data Sources for Drug Repurposing

During the past decade, the rapid collection of genomic data has brought an explosion of new insights into the genetic basis of diseases. It is enough to mention the numerous studies through which the association of gene loci with the risk of developing certain diseases has been discovered or the sequencing of human tumors, thanks to which somatic mutations underlying many types of cancer have been identified.

The acquisition of new knowledge about some disease phenotypes and drug-induced perturbations has increased the interest in new computational methods that can analyze and integrate large amounts of data to uncover new disease targets.

In general, applying these approaches on drug perturbation datasets has helped improve the understanding of the connection between genes, drugs, and diseases, as these methodologies can lead to the generation of novel hypotheses.

Drug repurposing tools: web-based solutions

Categories and core concepts	Tools	Features	Web link
Ligand similarity using fingerprint encoding Core concept: structural similarity implies comparable biological function or properties	*ChemMapper*	3D similarity algorithm SHAFTS (SHApe-FeaTure Similarity). It uses shape and chemotype to assess similarity	http://lilab.ecust.edu.cn/chemmapper/
	ChemProt 3.0	2D similarity-based algorithm. It includes several computational approaches: Similarity Ensemble Approach—SEA, Quantitative Structure-Activity Relationship—QSAR, and network biology-based enrichment analysis	http://potentia.cbs.dtu.dk/ChemProt/
	HitPick	Prediction based on 2D molecular fingerprints. B-score method based on a statistical evaluation of screening parameters for hits identification and integrative approach combining 1-nearest-neighbor (1NN) similarity metrics and Laplacian-modified naïve Bayesian target models to predict the targets of identified hits	http://mips.helmholtz-muenchen.de/hitpick
	iDrug-Target	2D molecular fingerprint-based approach. It uses Support Vector Machine (SVM) to classify inputs as interactive or non-interactive	http://www.jci-bioinfo.cn/iDrug-Target/
	Polypharmacology browser—PPB	Multi fingerprint-based approach. Similarity is calculated using city-block distances	http://gdbtools.unibe.ch:8080/PPB/
	Similarity ensemble approach—SEA	2D molecular fingerprint-based approach	http://sea.bkslab.org/
	SuperPred	2D molecular fingerprint-based approach (2D Tanimoto strategy)	http://prediction.charite.de
	SwissTargetPrediction	Combination of 2D (2D Tanimoto strategy) and 3D (Manhattan similarity distance) similarity approach. It is possible to perform predictions in different organisms (human, rat, and mouse)	http://www.swisstargetprediction.ch
	TarPred	Combination of ECFP4 (Extended-Connectivity Fingerprints), designed for molecular characterization, similarity searching, and structure-activity modeling, and Tanimoto coefficient (2D fingerprint similarity)	http://www.dddc.ac.cn/tarpred
	TargetHunter	Tanimoto similarity index—TAMOSIC (Targets Associated with its MOst SImilar Counterparts). 2D molecular fingerprint-based approach	http://www.cbligand.org/TargetHunter/
3D structures of drugs and targets knowledge Core concept: knowledge of the three-dimensional structure of the molecular target for the drug	*idTarget*	Divide et impera docking approach in combination with scoring functions based on regression analysis and quantum chemical charge models	http://idtarget.rcas.sinica.edu.tw/
	Protein-Drug Interaction Database—PDID	It is based on all-atom predictions on the entire human structural proteome and provides access to all putative targets (depending on the prediction method used: ILbind, SMAP, and eFindSite) of several popular drugs	http://biomine.ece.ualberta.ca/PDID/
	TARget FIShing DOCKing—TarFisDock	Reverse ligand-protein docking approach. The docking is performed through the DOCK 4.0 algorithm using protein structures present in PDTD	http://www.dddc.ac.cn/tarfisdock/

Drug repurposing tools: web-based solutions

Categories and core concepts	Tools	Features	Web link
Biological networks *Core concept: integrating these types of biological networks can be of great help in understanding the pharmacological properties of certain molecules and thus in drug repositioning*	*Balestra Web*	It uses active learning (AL) techniques based on probabilistic matrix factorization (PMF) to calculate the statistical weight of approved drugs for all targets associated with the entire set of approved drugs. Predictions made by BalestraWeb do not depend on structural or chemical similarities	http://balestra.csb.pitt.edu/
	CSNAP	Combination of chemical similarity networks (CSNs) and chemical consensus from chemotype-based subnetworks to predict targets for a set of drug classes	https://services.mbi.ucla.edu/CSNAP/index.html
	DASPfind	Network-based approach. Drug similarities are calculated using SIMCOMP. Target similarity is calculated using a normalized version of the Smith-Waterman algorithm	http://www.cbrc.kaust.edu.sa/daspfind/
	nAnnolyze	Network-based comparative docking approach. Structural information, interactions, and similarities are integrated to predict compound-protein structural interactions on a proteomic scale	http://www.marciuslab.org/services/nAnnoLyze
	PROMISCUOUS	Network-based approach wherein nodes represent drugs, proteins, and side effects, and edges represent drug-target, drug-drug, target-target, and drug-side effect interactions	https://bioinformatics.charite.de/promiscuous2/
	SLAP	Semantic Link Association Prediction. Drug-target associations are predicted through semantic database integration and statistical modeling. Semantic network is constructed using Chem2Bio2OWL ontology drug-drug and drug-target interactions using Chem2Bio2RDF and drug-target pairs used to construct the association network from DrugBank	http://chem2bio2rdf.org/slap
	STITCH	It provides a comprehensive map of drug-target interactions. It integrates drug-target interaction data resources derived from high-throughput, manually curated database experiments and predictive algorithms	http://stitch.embl.de/
	DT-Web	Recommendation-based algorithm. Domain-specific knowledge expressing drug-target similarity is used to calculate recommendations for each drug	http://alpha.dmi.unict.it/dtweb/
	SAveRUNNER	Network-based algorithm for drug repurposing. It provides an R code. From a list of drug-target interactions and disease-disease associations requested as input, it predicts drug-disease associations by computing a network-based similarity measure	https://github.com/giuliafiscon/SAveRUNNER.git
Binding site parametrization Core concept: methods designed to predict targets based on the binding sites of query molecules	*ProBis*	It identifies model proteins in the Protein Data Bank (PDB) that share similar binding sites to the query (3D structure). It uses the ProBiS algorithm	http://probis.cmm.ki.si/
	PoSSuM	All-pairs similarity	http://possum.cbrc.jp/PoSSuM/

Drug repurposing tools: web-based solutions

Categories and core concepts	Tools	Features	Web link
Using drug-induced gene expression to predict new connections Core concept: mRNA expression profiles in a cell line, before and after drug treatment	*Connectivity Map—CMap*	Predictions are based on a database of pre- and post-gene expression profiles from cellular samples in response to various types of perturbation (drug effects)	http://www.broad.mit.edu/cmap
	DeSigN	It associates disease signatures with drug response signatures based on IC50 data. It does not use pre- and post-gene expression profiles but only baseline gene expression profiles	http://design.cancerresearch.my/
	GoPredict	It integrates gene expression data with heterogeneous public information (signaling pathways, drug-target information, etc.).	http://csblcanges.fimm.fi/GOPredict/
	MANTRA 2.0	It uses gene expression profiles before and after drug perturbation to define the drug behavior into the network as a new "node." MANTRA 2.0 allows to explore the network to find new indications and interactions	http://mantra.tigem.it/
	NFFinder	It uses MARQ method to compare molecular signatures	http://nffinder.cnb.csic.es/
	PDOD	Prediction of Drugs with Opposing Effects on Disease Genes. It uses gene expression data (GEO) and associates to them pathway information from KEGG and drug-target information from DrugBank	http://gto.kaist.ac.kr/pdod/index.php/main
	RGES	Reverse Gene Expression Score involves using gene expression data to predict how a given drug might reverse the gene expression profile for a given disease by antagonizing genes that are overexpressed (underexpressed) due to the disease and thereby reverts gene expression closer to normal tissue levels	https://github.com/Bin-Chen-Lab/RGES
Others Core concept: disease association–dependent annotations	*MeSHDD*	Literature-based repositioning methodology	http://apps.chiragjpgroup.org/MeSHDD/
	RE:fine drugs	It integrates GWAS and PheWAS reposition datasets using drug-gene-disease triads	
	BANDIT	Machine learning algorithm using a Bayesian approach to integrate multiple data types to predict to which target (enzyme, receptor, or other) a drug may interact to have its therapeutic effect. The rationale for this approach is that the integration of multiple data types improves the accuracy of target prediction since each data type captures different aspects of a molecule's activity	Not publicly available. Select pieces of custom code can be made available upon request

Machine learning techniques and biomedical text mining approaches have been crucial in discovering hidden relationships between drugs and potential new therapeutic indications.

Systematic collection and analysis of gene expression data from human cell lines before and after drug treatment can be used to identify new opportunities for drug repurposing, discover new mechanisms of action for compounds, make small-molecule mimics of endogenous ligands, and predict side effects of such compounds [47].

This approach was initially enabled by the *Connectivity Map* that contains data on transcriptional responses of human cancer cell lines to various drugs/compounds and other small molecules.

The first version of this database had limitations due to its small scale, leading to the extension of the *Connectivity Map* project through the NIH *Library of Integrated Network–based Cellular Signatures* (LINCS) program. A new approach was introduced to increase the available experimental data. A cheaper technology than the classic RNA-seq, called L1000, was employed. The LINCS-L1000 provides the signatures of ~50 human cell lines in response to ~20,000 drugs (at various concentrations) for a total of over a million experiments [47].

In this section, we will provide an overview of CMap and its evolution LINCS L1000. These “big data” resources provide essential but straightforward platforms for characterizing small molecule–induced changes in gene expression and determining connections, similarities, or dissimilarities among diseases, drugs, genes, and pathways.

CMap

The *Connectivity Map* (CMap), introduced in 2006 by Lamb et al., is a database collecting gene-expression profiles of drug-treated human cell cultures, which has been used for investigation of polypharmacology and drug repurposing.

Gene expression profiles are a series of experiments conducted using a microarray platform (Affymetrix HT_HG_U133 and HG_U133A) and standardized preprocessing (MAS 5.0). Experiments were done on different cell lines at different vehicle concentrations and time points compared to controls [48].

In the original CMap study, the initial reference database (Build 1) included 455 treatment-control pairs, where treatment constitutes a selection of 165 drugs, 42 different concentrations, 2-time points, and four human cell lines (MCF7, PC3, SKMEL5, and HL60). Subsequently, the database was significantly extended (Build 2), adding 1309 drugs with 156 different concentrations for a total of about 7000 gene expression profiles [48].

An “instance identifier” uniquely identifies each instance within the database. Thus, there is an instance representation in the reference database for each drug corresponding to treatment and control conditions [48].

The Connectivity Mapping Methods

CMap’s rationale is to use a reference database containing disease-specific gene expression profiles and compare it to the gene signature of a given drug. This approach is aimed to predict potential therapeutic candidate drugs. It also allows the identification of connections between drugs, genes, and diseases.

The CMap workflow comprises an initial query consisting of a set of gene signatures highly representative of a given biological state (e.g., disease). Although there is no definite way to generate the optimal gene signatures, the conventional approach identifies and uses a statistically significant list of differentially expressed genes (DEGs) calculated from disease and control samples. This list of genes will delineate the characteristic phenotype for a particular disease [48].

This kind of approach is platform-independent, allowing users to create query signatures from any gene expression platform [40]. Then, the query is used to interrogate the CMap catalog.

Within the database, each of the signatures consists of a weighted average of the three biological replicate perturbations to mitigate the effects of unrelated replicates or outliers [40].

At this point, a connectivity score with a p-value is estimated using a non-parametric rank-ordered Kolmogorov-Smirnov (KS) test. The "*connectivity score*" is normalized through the random permutation described by Lamb et al., assuming values from 1 to −1 to reflect the closeness between expression profiles [40, 48].

A positive correlation indicates the degree of similarity between a query signature and a perturbation-derived profile after specific treatment, whereas a negative correlation denotes an inverse similarity. These correlations are used to determine how exposure to a particular chemical may mimic or reverse the signature of the biological sample of interest.

A false discovery rate (FDR), which adjusts the *p*-value considering multiple hypothesis testing, and a t-parameter, which compares an observed enrichment score to all others in the database, are also calculated [40]. These metrics allow a comprehensive assessment of the relationship between a query and a perturbation, rather than just sorting by similarity.

Since the methodology behind CMap involves using expression profiles to define molecular signatures, it does not require prior knowledge of the detailed mechanism of action (MoA) or drug targets [40, 48]. This advantage makes it a widely used method in drug discovery and repositioning.

The original CMap database had limited chemical and genetic perturbation data due to the high cost of commercial gene expression microarrays and RNA sequencing (RNA-seq). In addition, the expression profiles looked only at a few cell lines leaving the uncertainty of applicability to other cell lines, animal models, or human systems.

To improve the system and overcome these significant limitations, the same team of researchers developed a new simplified platform called L1000 to facilitate rapid and high-throughput gene expression profiles at a lower cost.

L1000

The L1000 platform, developed at the Broad Institute by the CMap team, is a method to facilitate high-throughput, low-cost gene expression profiling and is suitable for extending CMap at a large scale [40, 48]. The development of this method was part of the NIH LINCS (*Library of Integrated Cellular Signatures*) consortium, which funds the generation of expression profiles across multiple cell types and perturbations. To date, through L1000 technology, over 1 million gene expressions have been profiled and collected.

Its name, L1000, is because it contains several reference transcripts equal to 1000, used to estimate the signature of the whole genome gene expression generated by microarrays. Effectively, the basic idea is that it is possible to capture any cellular state by starting from a certain number of representative transcripts at a low cost.

The authors used a set (12,031) of Affymetrix HGU133A expression profiles available in the Gene Expression Omnibus (GEO) to define the threshold for the number of transcripts. From this analysis, it was estimated that 1000 landmarks were sufficient to recover 82% of the information in the entire transcriptome [40].

The L1000 platform combines ligation-mediated amplification, optically addressed and barcoded microspheres (beads), and a flow cytometric detection system for gene expression signature analysis [40]. The L1000 platform is based on hybridization, making the detection of non-abundant transcripts feasible and with a substantial degree of similarity to the profiles obtained with RNA-seq platforms while bypassing the problem of prohibitive costs inherent in this conventional technique.

CMap and its updated versions provide a hypothesis-generating tool to identify new therapeutic targets (drug repositioning), signaling pathways affected by a compound, and search for new mechanisms of action (MoA), including potential side effects. It allows identifying new or known disease-gene-drug connections, depending on the observed level of changes.

Among the most exciting uses is the functional annotation of previously uncharacterized small molecules. For example, using the new-generation CMAp, a new inhibitor of casein kinase CSNK1A1 (compound BRD-1868) was discovered. CSNK1A1 is a protein essential for the survival of some myeloid malignancies. It is also implicated in resistance to EGFR inhibitors [40].

To facilitate the fruition and use of this system, a platform called CLUE—CMap Linked User Environment has been developed. It can provide several analyses and allow access to all data at multiple levels of pre-processing via Gene Expression Omnibus (GEO: GSE92742) [40].

The L1000 LINCS currently includes over 1 million gene expression profiles of chemically disrupted human cell lines. Several resources and databases derived from L1000 LINCS data are available, for example, the L1000 Characteristic Direction Signature (L1000CDS2) search engine described below.

L1000CDS2

L1000CDS2 is a web-based search engine software designed to query gene expression signatures versus LINCS data to discover and prioritize small molecules that reverse or mimic the entered gene expression profile [47].

To compute the signatures, the L1000CDS2 uses a multivariate method called the Characteristic Direction (CD). Processing L1000 data with the Characteristic Direction (CD) method significantly improves the signature noise compared to the MODZ method used to calculate L1000 signatures [47]. The L1000CDS2 tool can be applied in many biological and biomedical contexts, improving knowledge extraction from the LINCS L1000 resource.

The L1000CDS2 search engine prioritizes thousands of small molecule signatures and their pairwise combinations predicted to mimic or reverse an input gene expression signature. The L1000CDS2 search engine also predicts drug targets for all small molecules profiled by the L1000 assay [47].

Rather than giving relevance to fold-change and assigning greater weight to single genes that show a big fold-change, the CD method assigns a higher weight to genes that move together in the same direction. Thus, a gene that changes less but "moves" along with a large group of other genes may have more weight than a single gene that has changed more in magnitude [47].

The method first identifies the linear hyperplane that best separates control samples from treatment samples using linear discriminant analysis and then uses the normal to this hyperplane to define the direction of change in expression space for each gene [47]. The CD method is more sensitive in identifying "correct" differentially expressed genes than the other alternative methods [47]. CD L1000 signatures can be accessed through an advanced web-based application called L1000CDS2 [47].

When accessing L1000CDS2, there are five sections on the application's home page [47]: the first section on the left consists of two text boxes to enter up- or downregulated genes. The application also gives the possibility to insert an input signature [47]. In this case, the signature should be pasted in the upregulated gene textbox and expression values. The search can be started by clicking on the "search" button once the text boxes are filled [47].

In the central part of the home page, there is a section dedicated to some examples, a configuration section, a section dedicated to metadata, and a section dedicated to recent searches [47].

Optional parameters provided in the configuration section offer several possibilities to customize a search process. For example, through the *mimic/reverse* cursor, it is possible to look for small molecules that mimic or reverse the input signature. The default search mode is *reverse*. The system also supports searching for paired combinations of small molecules [47].

In the *metadata* section, any metadata associated with the input signature can be entered. In the recent searches section, the last 20 queries are stored and are easily accessible by clicking on each entry [47].

Interestingly, there is a function that allows users to share their input signatures and metadata so that others can query those signatures [47].

After starting the search by clicking the Search button, the first 50 signatures are shown in a table on the results page (14 entries for each page) [47].

Each entry provides seven columns of signature information: rank, score, perturbation, cell line, dose, time point, and overlap with input [47].

Clicking the overlap button, the overlapped genes (and their values) will be shown in the two text boxes. If the user had given up/down genes as input, then the first box will show the overlapping genes between the up input and the up signature, while the second box will show the overlap between the down input and the down signature. If the input is a signature, then the first box will show the genes with a positive value from the input signature, and the second box will have negative values [47].

It is possible to download all the information about a signature as a JavaScript Object Notation file (JSON) by clicking on the download button. Through the tag button, it is possible to view the inserted metadata [47].

Clicking the diamond icon button, it is possible to execute the enrichment analysis on the substructures of the best classified small molecules. The enrichment analysis results are displayed as a table where each row provides three pieces of information: the substructure, the *p*-value (calculated using Fisher's exact test), and the perturbation count. The substructure is represented as a string in the SMARTS format [47].

The cloud icon is used to download the results in table format to a .csv file. Clicking on the share icon provides a permanent URL that can be used to share the enrichment analysis results through an email, publication, or other documentation [47].

If the user chooses to search for combinations of small molecules, then a table of signature combinations will appear below the table of single perturbation results. Each entry provides information about the identified combinations: rank, synergy score, and combinations [47].

When looking for combinations, L1000CDS2 compares each possible pair among the top 50 matching signatures and calculates the potential synergy between each pair by examining the level of orthogonality. The synergy score is calculated as the combined overlap of the differentially expressed genes of the two drug signatures with the input gene lists [47].

Clicking on a perturbation will highlight that perturbation in the single signature results table so that the user can learn more details about that particular perturbation. Clicking the cloud download button in the upper right will download the combination table in a .csv file [47].

In summary, L1000CDS2 is a computational method that potentially elevates the usefulness of a subset of the newly generated publicly available LINCS-L1000 data set to rapidly prioritize small molecules that could reverse or mimic expression in disease and other biological settings [47].

Thanks to L1000CDS2, kenpaullone has been identified as a small molecule that can potentially interfere with the infectious process caused by Ebola by inhibiting GSK3B. Kenpaullone induces the expression of immune response genes and, as such, is a potential antiviral candidate [47].

Conclusion

In this chapter we have reviewed data resources and computational tools available for drug repositioning with the aim of providing a comprehensive guide for researchers and practitioners interested in such a topic. The survey highlights the content and the limitations of each tool or database and compares their content.

References

1. Sam E, Athri P. Web-based drug repurposing tools: a survey. Brief Bioinform. 2019;20:299–316.
2. Fiscon G, Paci P. SAveRUNNER: an R-based tool for drug repurposing. BMC Bioinformatics. 2021;22:150.
3. Jin G, Wong STC. Toward better drug repositioning: prioritizing and integrating existing methods into efficient pipelines. Drug Discov Today. 2014;19:637–44.
4. Gong J, Cai C, Liu X, Ku X, Jiang H, Gao D, Li H. ChemMapper: a versatile web server for exploring pharmacology and chemical structure association based on molecular 3D similarity method. Bioinformatics. 2013;29:1827–9.
5. Kringelum J, Kjaerulff SK, Brunak S, Lund O, Oprea TI, Taboureau O. ChemProt-3.0: a global chemical biology diseases mapping. Database. 2016;2016:bav123. https://doi.org/10.1093/database/bav123.

6. Liu X, Vogt I, Haque T, Campillos M. HitPick: a web server for hit identification and target prediction of chemical screenings. Bioinformatics. 2013;29:1910–2.
7. Xiao X, Min J-L, Lin W-Z, Liu Z, Cheng X, Chou K-C. iDrug-Target: predicting the interactions between drug compounds and target proteins in cellular networking via benchmark dataset optimization approach. J Biomol Struct Dyn. 2015;33:2221–33.
8. Abdouli NOA, Al Abdouli NO, Aung Z, Woon WL, Svetinovic D. Tackling class imbalance problem in binary classification using augmented neighborhood cleaning algorithm. In: Kim K, editor. Information science and applications. Lecture notes in electrical engineering. Berlin, Heidelberg: Springer; 2015. p. 827–34.
9. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling TEchnique. J Artif Intell Res. 2002;16:321–57.
10. Awale M, Reymond J-L. The polypharmacology browser: a web-based multi-fingerprint target prediction tool using ChEMBL bioactivity data. J Cheminform. 2017;9:11.
11. Keiser MJ, Roth BL, Armbruster BN, Ernsberger P, Irwin JJ, Shoichet BK. Relating protein pharmacology by ligand chemistry. Nat Biotechnol. 2007;25:197–206.
12. Nickel J, Gohlke B-O, Erehman J, Banerjee P, Rong WW, Goede A, Dunkel M, Preissner R. SuperPred: update on drug classification and target prediction. Nucleic Acids Res. 2014;42:W26–31.
13. Gfeller D, Grosdidier A, Wirth M, Daina A, Michielin O, Zoete V. SwissTargetPrediction: a web server for target prediction of bioactive small molecules. Nucleic Acids Res. 2014;42:W32–8.
14. Daina A, Michielin O, Zoete V. SwissTargetPrediction: updated data and new features for efficient prediction of protein targets of small molecules. Nucleic Acids Res. 2019;47:W357–64.
15. Liu X, Gao Y, Peng J, Xu Y, Wang Y, Zhou N, Xing J, Luo X, Jiang H, Zheng M. TarPred: a web application for predicting therapeutic and side effect targets of chemical compounds. Bioinformatics. 2015;31:2049–51.
16. Wang L, Ma C, Wipf P, Liu H, Su W, Xie X-Q. TargetHunter: an in silico target identification tool for predicting therapeutic potential of small organic molecules based on chemogenomic database. AAPS J. 2013;15:395–406.
17. Wang J-C, Chu P-Y, Chen C-M, Lin J-H. idTarget: a web server for identifying protein targets of small chemical molecules with robust scoring functions and a divide-and-conquer docking approach. Nucleic Acids Res. 2012;40:W393–9.
18. Wang C, Hu G, Wang K, Brylinski M, Xie L, Kurgan L. PDID: database of molecular-level putative protein–drug interactions in the structural human proteome. Bioinformatics. 2016;32:579–86.
19. Li H, Gao Z, Kang L, et al. TarFisDock: a web server for identifying drug targets with docking approach. Nucleic Acids Res. 2006;34:W219–24.
20. Cobanoglu MC, Oltvai ZN, Taylor DL, Bahar I. BalestraWeb: efficient online evaluation of drug-target interactions. Bioinformatics. 2015;31:131–3.
21. Lo Y-C, Senese S, Li C-M, Hu Q, Huang Y, Damoiseaux R, Torres JZ. Large-scale chemical similarity networks for target profiling of compounds identified in cell-based chemical screens. PLoS Comput Biol. 2015;11:e1004153.
22. Ba-Alawi W, Soufan O, Essack M, Kalnis P, Bajic VB. DASPfind: new efficient method to predict drug-target interactions. J Cheminform. 2016;8:15.
23. Martínez-Jiménez F, Marti-Renom MA. Ligand-target prediction by structural network biology using nAnnoLyze. PLoS Comput Biol. 2015;11:e1004157.
24. von Eichborn J, Murgueitio MS, Dunkel M, Koerner S, Bourne PE, Preissner R. PROMISCUOUS: a database for network-based drug-repositioning. Nucleic Acids Res. 2011;39:D1060–6.
25. Gallo K, Goede A, Eckert A, Moahamed B, Preissner R, Gohlke B-O. PROMISCUOUS 2.0: a resource for drug-repositioning. Nucleic Acids Res. 2021;49:D1373–80.
26. Chen B, Ding Y, Wild DJ. Assessing drug target association using semantic linked data. PLoS Comput Biol. 2012;8:e1002574.
27. Kuhn M, Szklarczyk D, Pletscher-Frankild S, Blicher TH, von Mering C, Jensen LJ, Bork P. STITCH 4: integration of protein-chemical interactions with user data. Nucleic Acids Res. 2014;42:D401–7.
28. Kuhn M, Szklarczyk D, Franceschini A, von Mering C, Jensen LJ, Bork P. STITCH 3: zooming in on protein-chemical interactions. Nucleic Acids Res. 2012;40:D876–80.
29. Szklarczyk D, Santos A, von Mering C, Jensen LJ, Bork P, Kuhn M. STITCH 5: augmenting protein–chemical interaction networks with tissue and affinity data. Nucleic Acids Res. 2016;44:D380–4.
30. Kuhn M, Szklarczyk D, Franceschini A, Campillos M, von Mering C, Jensen LJ, Beyer A, Bork P. STITCH 2: an interaction network database for small molecules and proteins. Nucleic Acids Res. 2010;38:D552–6.
31. Alaimo S, Bonnici V, Cancemi D, Ferro A, Giugno R, Pulvirenti A. 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(Suppl 3):S4.
32. Alaimo S, Pulvirenti A, Giugno R, Ferro A. Drug-target interaction prediction through domain-tuned network-based inference. Bioinformatics. 2013;29:2004–8.
33. Chen B, Ma L, Paik H, Sirota M, Wei W, Chua M-S, So S, Butte AJ. Reversal of cancer gene expression correlates with drug efficacy and reveals therapeutic targets. Nat Commun. 2017;8:16022.
34. Konc J, Janezic D. ProBiS-2012: web server and web services for detection of structurally similar binding sites in proteins. Nucleic Acids Res. 2012;40:W214–21.
35. Ito J-I, Tabei Y, Shimizu K, Tsuda K, Tomii K. PoSSuM: a database of similar protein-ligand binding and putative pockets. Nucleic Acids Res. 2012;40:D541–8.

36. Ito J-I, Ikeda K, Yamada K, Mizuguchi K, Tomii K. PoSSuM v.2.0: data update and a new function for investigating ligand analogs and target proteins of small-molecule drugs. Nucleic Acids Res. 2015;43:D392–8.
37. Brown AS, Patel CJ. MeSHDD: literature-based drug-drug similarity for drug repositioning. J Am Med Inform Assoc. 2017;24:614–8.
38. Moosavinasab S, Patterson J, Strouse R, Rastegar-Mojarad M, Regan K, Payne PRO, Huang Y, Lin SM. "RE:fine drugs": an interactive dashboard to access drug repurposing opportunities. Database. 2016;2016:baw083. https://doi.org/10.1093/database/baw083.
39. Madhukar NS, Khade PK, Huang L, Gayvert K, Galletti G, Stogniew M, Allen JE, Giannakakou P, Elemento O. A Bayesian machine learning approach for drug target identification using diverse data types. Nat Commun. 2019;10:5221.
40. Subramanian A, Narayan R, Corsello SM, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. Cell. 2017;171:1437–1452.e17.
41. Lamb J, Crawford ED, Peck D, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science. 2006;313:1929–35.
42. Lee BKB, Tiong KH, Chang JK, Liew CS, Abdul Rahman ZA, Tan AC, Khang TF, Cheong SC. DeSigN: connecting gene expression with therapeutics for drug repurposing and development. BMC Genomics. 2017;18:934.
43. Louhimo R, Laakso M, Belitskin D, Klefström J, Lehtonen R, Hautaniemi S. Data integration to prioritize drugs using genomics and curated data. BioData Min. 2016;9:21.
44. Carrella D, Napolitano F, Rispoli R, Miglietta M, Carissimo A, Cutillo L, Sirci F, Gregoretti F, Di Bernardo D. Mantra 2.0: an online collaborative resource for drug mode of action and repurposing by network analysis. Bioinformatics. 2014;30:1787–8.
45. Setoain J, Franch M, Martínez M, Tabas-Madrid D, Sorzano COS, Bakker A, Gonzalez-Couto E, Elvira J, Pascual-Montano A. NFFinder: an online bioinformatics tool for searching similar transcriptomics experiments in the context of drug repositioning. Nucleic Acids Res. 2015;43:W193–9.
46. Yu H, Choo S, Park J, Jung J, Kang Y, Lee D. Prediction of drugs having opposite effects on disease genes in a directed network. BMC Syst Biol. 2016;10:S2. https://doi.org/10.1186/s12918-015-0243-2.
47. Duan Q, Reid SP, Clark NR, et al. L1000CDS2: LINCS L1000 characteristic direction signatures search engine. npj Syst Biol Appl. 2016;2:16015. https://doi.org/10.1038/npjsba.2016.15.
48. Musa A, Ghoraie LS, Zhang S-D, Glazko G, Yli-Harja O, Dehmer M, Haibe-Kains B, Emmert-Streib F. A review of connectivity map and computational approaches in pharmacogenomics. Brief Bioinform. 2018;19:506–23.