



In silico Methods for Identification of Potential Therapeutic Targets

Xuting Zhang¹ · Fengxu Wu² · Nan Yang¹ · Xiaohui Zhan³ · Jianbo Liao³ · Shang kang Mai³ · Zunnan Huang^{1,4} 

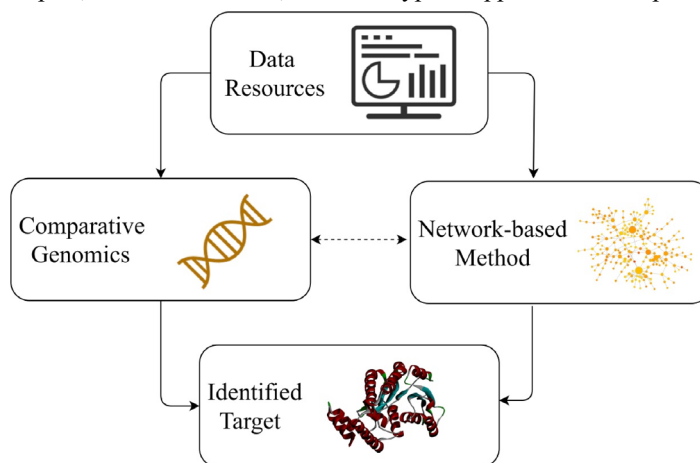
Received: 19 May 2021 / Revised: 19 October 2021 / Accepted: 1 November 2021 / Published online: 26 November 2021
© The Author(s) 2021

Abstract

At the initial stage of drug discovery, identifying novel targets with maximal efficacy and minimal side effects can improve the success rate and portfolio value of drug discovery projects while simultaneously reducing cycle time and cost. However, harnessing the full potential of big data to narrow the range of plausible targets through existing computational methods remains a key issue in this field. This paper reviews two categories of in silico methods—comparative genomics and network-based methods—for finding potential therapeutic targets among cellular functions based on understanding their related biological processes. In addition to describing the principles, databases, software, and applications, we discuss some recent studies and prospects of the methods. While comparative genomics is mostly applied to infectious diseases, network-based methods can be applied to infectious and non-infectious diseases. Nonetheless, the methods often complement each other in their advantages and disadvantages. The information reported here guides toward improving the application of big data-driven computational methods for therapeutic target discovery.

Graphical abstract

We provide a review on two categories of in silico methods for potential therapeutic target identification: comparative genomics and network-based methods; the contents include the basic principles, available services, tools and typical application examples.



Keywords Therapeutic target · Drug discovery · Target identification · Comparative genomics · Network

Xuting Zhang, Fengxu Wu and Nan Yang have contributed equally to this work.

✉ Zunnan Huang
zn_huang@gdmu.edu.cn; zn_huang@yahoo.com

Extended author information available on the last page of the article

1 Introduction

Target identification and validation is the top priority in drug discovery [1]. Molecules or drugs that interact with a rational target or selected combinations of targets have

improved odds of therapeutic success. An analysis of Astra-Zeneca's drug research and development programs showed that 82% of program terminations in preclinical studies were due to safety issues, of which 25% were target-related [2]. Meanwhile, 48% of safety failures in clinical trials are target-related. Therefore, guidance on the appropriate selection of candidate targets can help improve the success rate and portfolio value of drug discovery projects while also reducing time and cost [3].

Traditionally, target discovery has relied on wet experiments, a process that is time-consuming, expensive, and low in accuracy. With the development of bioinformatics, chemical informatics, and omics, computer-aided therapeutic target discovery methods or *in silico* methods have come to the fore [4–6]. By integrating big data with computational methods, computer-aided therapeutic target discovery greatly reduces the scope of experimental targets, shortens the drug discovery and development cycle, and reduces the experimental cost. At present, the two main categories of *in silico* methods for potential therapeutic target identification are comparative genomics [7] and network-based methods [8]. One of many important characteristics differentiating these methods is that comparative genomics is mostly used in infectious diseases, whereas network-based methods can be used not only in infectious diseases, but also in non-infectious diseases. Nonetheless, these categories of methods often complement each other in their advantages and disadvantages.

With the completely sequenced human genome, in addition to the completed genome sequences of many model organisms, there are increasing research-focused efforts to understand the function of a genome and molecular evolution. Finding potential therapeutic targets among cellular functions based on understanding their related biological processes in pathogens and their hosts has become imperative as antimicrobial resistance continues to spread rapidly. To identify therapeutic targets, comparative genomics combines the information contained in genome database resources and software to reveal fatal weaknesses of pathogens that affect their growth and reproduction in the host, such as genes essential for the survival, growth, and important functions of pathogens [9]. In addition, comparative genomics can also filter out homologs by comparing genomes of pathogens and hosts, avoiding the toxic and side-effects of newly designed drugs on the host, in turn, increasing the success rate of drug design [9].

With many pathogenic variants associated with disease in non-coding regions or difficult to target genes, the number of associations that are candidates for development into drugs is limited. Approaches that combine data from pathway databases or biological networks can broaden the number of potential targets to increase the number of associations that lead to effective treatments. As such, network-based

strategies are among the state-of-the-art computation models for target identification and are also an important bridge connecting network pharmacology [10], network medicine [11], network biology [12], systems biology [13], and multi-omics data. By combining pathway analysis and the network graph theory concept, network-based strategies not only focus on the interactions (edges) between individual molecules (nodes) and coordinated pathways but also enable a systematic visual exploration of the biological (or biomedical) networks to identify the components of functional importance in the network. In this regard, network-based methods are invaluable in identifying biomarkers, discovering disease diagnosis targets, and finding potential therapeutic targets [14]. The main concept of network-based methods is to map all the relevant data to a visual network. Highly connected nodes (central nodes) that act as bridges between consecutive network components in a single network are predicted as essential proteins or genes of the pathogen (or biological process) and shown to be related to the modular structure of the physical and functional interaction network. Such nodes are hypothesized to be ideal therapeutic targets in the network because they maintain the network integrity [8]. Meanwhile, by searching for highly differential nodes in different networks, those nodes that specifically exist in disease cells can also be hypothesized as potential therapeutic targets [15].

Here, we provide a detailed review of the rationales of comparative genomics and network-based methods for the *in silico* identification of potential therapeutic targets (Fig. 1). We describe the commonly used databases, software, and applications and discuss these methods in the context of their advantages and disadvantages, contrasts and similarities, comparison with related target identification methods, and relevant published reviews and prospective studies. The information provided in this review will help readers and researchers quickly understand the rationales of *in silico*

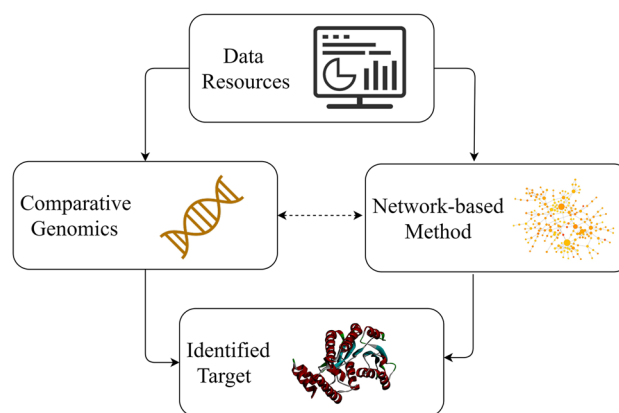


Fig. 1 Simplified workflow of *in silico* methods for identification of potential therapeutic targets

therapeutic target identification methods that could further advance research in this area.

2 In silico Comparative Genomics and Network-Based Methods

2.1 Comparative Genomics Methods

In the past two decades, whole-cell screening (including large numbers of genetic screening) and in vitro screening of synthetic libraries have been used to identify novel lead compounds with powerful antimicrobial properties [16]. With the completely sequenced human genome, in addition to the completed genome sequences of numerous bacteria and fungi, the number of genes has been rapidly growing. Probing and comparing sequence characteristics between and within species have become a part of most biological queries [17]. Comparative genomics [7] and the recently emerged subtractive genomics (described later in Sect. 6.5) [18] are useful tools for the identification of potential therapeutic targets, such as conserved genes [17] and putative essential genes [9] that affect cell viability in pathogens. Comparative genomics approaches are based on the hypothesis that potential targets are critical in the survival of pathogens and constitute a key component of their metabolic pathways [19]. Moreover, to eliminate

deleterious host responses, the target should have no conserved homolog in the human host [20]. Spaltmann et al. proposed two criteria for a gene to be considered a therapeutic target. First, the gene must be necessary for the survival and growth of the pathogen, thereby improving the therapeutic effect of the drug acting on the target. Second, the gene should exist in pathogens but not in mammals; in this way, the drug would have the potential to become a broad-spectrum antimicrobial agent [21]. A gene that meets these criteria can be found using a comparative genomics approach.

In Fig. 2, we have summarized the three main steps involved in comparative genomics-based identification of therapeutic targets [22]. The first step is the collection of metabolic pathway enzymes or essential genes of pathogens. It involves obtaining all the metabolic pathways that exist both in the host and pathogen from the Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathway Database [23]. Then, all pathogen pathways are compared with host pathways to determine any overlap [22]. Next, the metabolic pathways are classified. Pathways existing in both the pathogen and the host are removed and named shared pathways, while those existing in the pathogen but not in the host are pooled and named unique pathways [19]. Finally, the gene names and identification of all involved enzymes in the shared and unique pathways are identified and collected from the KEGG Genes Database [22].

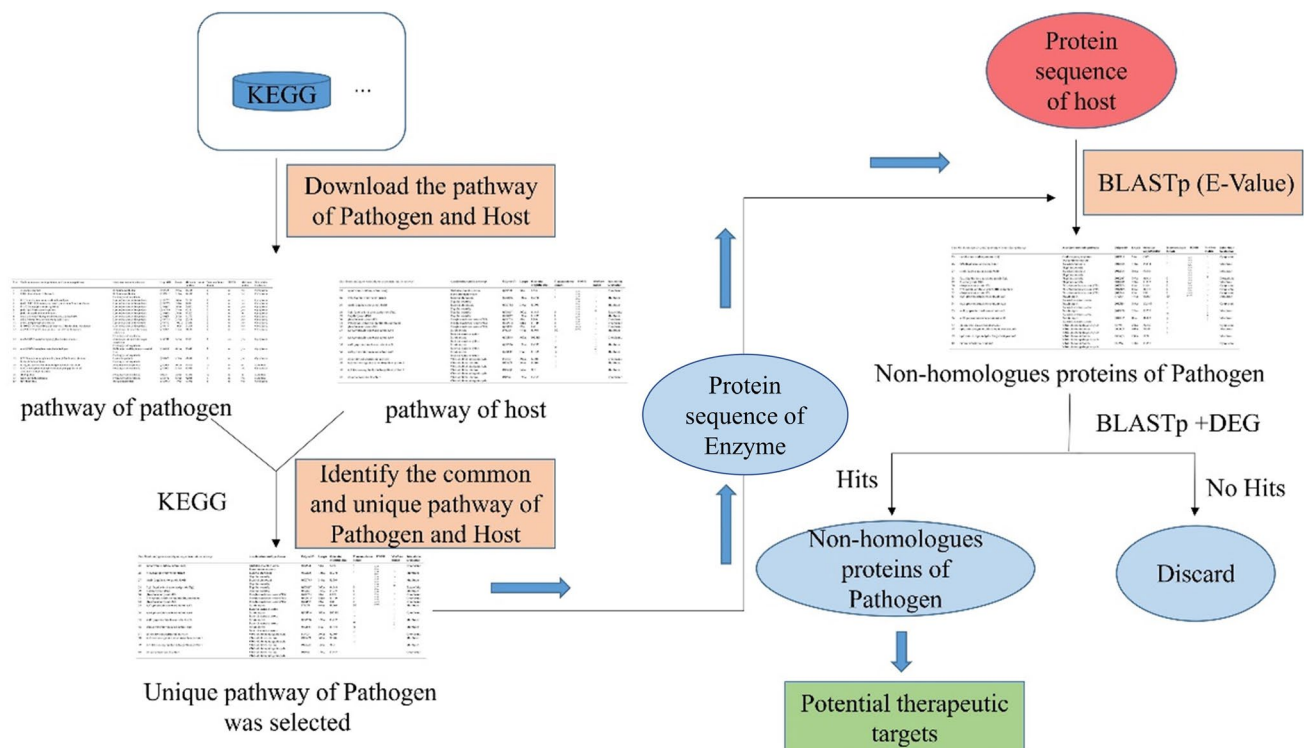


Fig. 2 The workflow of identifying potential therapeutic targets by comparative genomics

Step two is the retrieval analysis of the protein sequences and the use of the basic local alignment search tool (BLAST). First, the protein sequences of all enzymes involved in unique pathways are retrieved from the Universal Protein Resource (UniProt) database [24] in FASTA format. Then, each protein sequence is submitted to a BLASTp analysis (a protein–protein analysis that compares an amino acid sequence against a protein sequence database; discussed in further detail in Sect. 4.1) against the sequences of enzymes in the host metabolic pathways at a set E-value cutoff, the threshold to define a BLAST “hit.” BLAST results with no hits with host enzymes are identified as non-homologous enzymes of the pathogen [9].

The third and final step in the comparative genomics-based identification of therapeutic targets is the identification of essential non-homologous enzymes in the pathogen. To achieve this, the BLASTp analysis is carried out in the database of essential genes (DEG). The protein sequences with significant homology in the DEG database are described as protein sequences vital to the pathogen's survival [18].

Therapeutic targets identified by comparative genomics methods have two essential characteristics. One, the selected targets have significant impacts on some important physiological functions of the pathogen, ensuring the effectiveness of the newly designed drug. Two, by comparing the protein sequences between potential therapeutic targets and the host to identify whether there is homology, any toxic side effects on the human body when the drug interacts with the target can be avoided, in turn, improving the safety of the pharmacological effects of new drugs [20].

2.2 Network-Based Methods

The reason the network-based method can be used for therapeutic target identification is based on the assumption that the influence of specific locations in a biological network can spread along the edges (interactions) of the network [11]. The rationales of network-based methods for predicting therapeutic targets are centrality and differentia. Centrality refers to the analysis of network topological parameters when building a single network. A node in a more central position indicates that it plays a more integral role in the network. For example, it may be an essential protein for pathogen survival and thus identified as a potential therapeutic target [8]. However, centrality sometimes cannot be applied directly to normal human protein networks because of the toxicity of acting on such critical nodes [10, 25]. To solve this problem, the direct screening and elimination process of homologous proteins involved in metabolism can be complemented with differential network analysis in which two or more networks are compared, such as normal cell and disease (mostly cancer) cell networks, different subtype networks

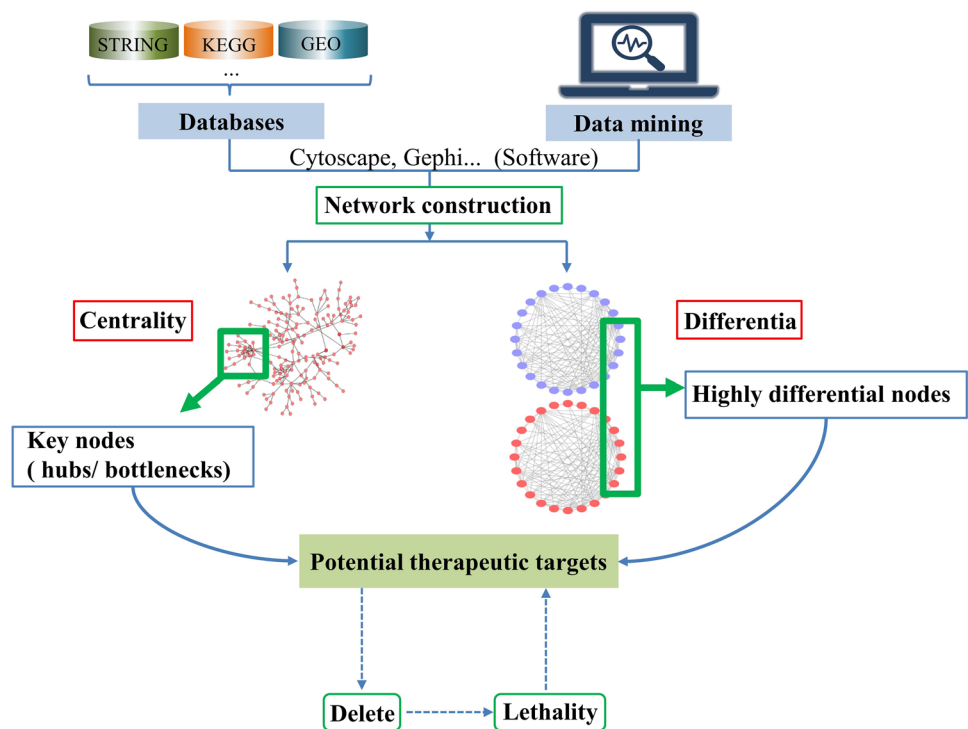
of cancer, and tissue-specific networks. In this way, the node sets specific to disease cells or highly differential between networks are obtained and identified as potential therapeutic targets [26]. Differential network analysis can also screen out targets that exist in disease cells but not in normal cells or targets connected differentially in different networks to make the identified targets more selective, thereby improving therapeutic security. The highly differential nodes obtained in this way can be further analyzed using network topology to obtain highly centralized nodes that have been double-screened, increasing the reliability of the identified nodes [27].

According to the rationales of centrality and differentia, network-based methods can be divided into two approaches: the centrality-based approach and the differentia-based approach (Fig. 3). The first step in both approaches is network construction. Network construction refers to obtaining a large number of relevant data sets through data mining [28] or from various databases, websites, and experimental data and carrying out attribute mapping through network visualization tools, namely comprehensive data visualization [29]. Some types of constructed networks are protein–protein interaction (PPI) networks [30], gene interaction networks [31], and miRNA–mRNA interaction networks [32]. After the network is built, the processes of the two approaches diverge.

The centrality-based approach uses some network analysis tools to (i) analyze the topological parameters of nodes in networks and (ii) select nodes with high degree centrality (hub nodes) and high betweenness centrality (bottlenecks), which are often integral in networks and thus can be selected as potential therapeutic targets [33]. The degree centrality of a node refers to the number of direct connections the node has with other nodes in the network [11], while the betweenness centrality of a node refers to the number of shortest paths that pass through the node in the network [34]. The centrality-based approach is most suitable for rapidly growing cells, such as pathogens and cancer cells [8]. In addition to the widely-used degree centrality and betweenness centrality, other parameters, such as closeness centrality, clustering coefficient, average shortest path, eigenvector centrality, and spectral gap centrality, can also be used as centrality indices to predict the importance of nodes, and thus to identify potential therapeutic targets [35, 36]. For further understanding of the definitions of the parameters mentioned above, two references are recommended [35, 36].

As mentioned above, differential network analysis requires the construction of two or more networks, including normal and disease cell networks [30] or networks of different subtypes of cancer [37]. After the construction of networks is completed, some algorithms can be applied to identify differential components between networks, to select nodes that exist in disease cell networks but not in normal

Fig. 3 Simplified rationale and approaches of network-based methods for potential therapeutic target identification



cell networks, or to select nodes that are highly differentially connected between or among networks, as predicted potential therapeutic targets [26, 38].

Potential targets identified through centrality and differentia can be further prioritized by observing the lethality of the network when those nodes are removed [39]. Generally, network lethality after removal of a node is positively correlated with the connectivity of the node. When nodes with high degree centrality are deleted, the network diameter will increase rapidly [40]. When nodes with high betweenness centrality are deleted, (i) the average path length will decrease rapidly [41]; (ii) network topology, such as the characteristic path length, will change significantly; (iii) the ability of the remaining nodes to communicate with each other will be weakened, and (iv) the network will disintegrate [42]. Therefore, the more lethal the removal of a node to the network, the more important the node's role, and the greater its potential as a therapeutic target [39].

3 Databases

Data acquisition is indispensable to any research work. Therefore, we summarized the databases useful in comparative genomics and network-based methods for identifying potential therapeutic targets. Although some databases can be used for both types of *in silico* methods, we placed them in separate tables because the most popular features of these databases differ between the two approaches.

3.1 Comparative Genomics

The relevant databases for comparative genomics can be roughly divided into two categories: (i) general databases; those usually used in comparative genomics, such as DEG, KEGG [23], and UniProt; and (ii) specific databases, which mainly provide pathogenic gene sequences of bacteria and fungi, such as the Tuberculosis Database (TBDB), WormBase, and the Virulence Factors of Pathogenic Bacteria Database (VFDB). Table 1 lists the general and specific databases with brief descriptions, including the coverage, availability, latest update, and URL.

DEG is a commonly used database in comparative genomics that contains 53,885 essential genes and 786 non-coding essential sequences critical to the survival and growth of bacteria, archaea, and eukaryotes for homology analyses [44]. DEG 15 is the most recent version of this database. It is worth noting that DEG has multiple built-in tools for data analysis and display, such as a subcellular location and distribution analysis tool, a pathway and genomics enrichment analysis tool, and a Venn maps generation tool for comparing genomes between experiments [54].

TBDB is an online platform for basic scientific research on tuberculosis and drug and vaccine discovery and development research. It contains genome sequence data and microarray and RT-PCR expression data, including over 3,000 *Mycobacterium tuberculosis* (*Mtb*) microarrays (2,700 from humans and mice and 260 for *Streptomyces coelicolor*) and 95 RT-PCR datasets, for numerous strains of *Mtb*, as well as data for more

Table 1 General and specific databases for comparative genomics

Database	Description	Coverage	Availability	Latest update	URL	References
UniProt ^a	A comprehensive resource for protein sequence and annotation data	305,529 proteomes	Free	2020	https://www.uniprot.org/	[24]
UniProtKB/Swiss-Prot ^a	A high-quality annotated and non-redundant protein sequence database	564,277 proteins	Free	2021	https://www.uniprot.org/uniprot/?query=reviewed:yes	[43]
DEG ^a	A database hosting records of currently available essential genomic elements	53,885 essential genes and 786 essential non-coding sequences	Free	2017	http://tubic.tju.edu.cn/deg/	[44]
pDEG ^a	A database contains many details of the predicted essential genes of 16 <i>Mycobacterium</i> genomes	5,880 essential genes	Free	2011	https://origin.tubic.org/pdeg	[45]
OGEE ^a	An essentiality database that includes essential and non-essential genes from large-scale experiments	127 gene essentiality experiments for 91 species, 38,822 genes covered by multiple experiments	Free	2021	http://ogeedb.embl.de	[46]
KEGG ^a	A database resource that integrates genomic, chemical and systemic functional information	Four categories (systems, genomic, chemical, and health information) from 18 databases	Free	2021	https://www.kegg.jp/	[47]
PMDB ^a	A public resource aimed at storing manually built three-dimensional models of proteins	Contains > 74,000 models for approximately 240 proteins	Free	2011	http://www.caspar.it/PMDB	[48]
ModBase ^a	A database of comparative protein structure models	Almost 30 million reliable models for domains in 4.7 million unique protein sequences	Free	2013	https://modbase.compbio.ucsf.edu/modbase-cgi/index.cgi	[49]
RefSeq ^a	A comprehensive, integrated, non-redundant, well-annotated database of sequences, including genomic DNA, transcripts, and proteins	197,232,209 proteins, 36,514,168 transcripts, and 108,257 organisms	Free	2021	https://www.ncbi.nlm.nih.gov/refseq/	[50]
Tuberculosis Database ^b	An online database provides integrated access through a single portal to sequence data and annotation, expression data, and literature curation for tuberculosis	Genome sequences of 20 strains	Free	2021	https://www.tbdb.org/	[51]
WormBase ^b	A database about <i>Caenorhabditis elegans</i> genome	<i>C. elegans</i> and other <i>Caenorhabditis</i> genomes	Free	2017	https://wormbase.org/#012-34-5	[52]
VFDB ^b	A database of virulence factors of various medically significant bacterial pathogens	Virulence factors of bacterial pathogens	Free	2021	http://www.mgc.ac.cn/VFs/main.htm	[53]

^aDenotes general database^bDenotes specific database

than 20 *Mtb*-related strains from in vitro tuberculosis-related experiments and tuberculosis-infected tissues. A wide range of tools is incorporated in the database for browsing, analyzing, searching, and downloading the data [51].

3.2 Network-Based Methods

There are many databases used in network-based methods. We roughly divided the databases into two categories: direct databases and indirect databases. Direct databases cover the interaction data and can be directly imported into network visualization software for network construction. Examples are the Search Tool for Retrieval of Interacting Genes/Proteins (STRING) [55] and the Molecular INTeraction (MINT) database [56]. Indirect databases do not directly cover interaction data but provide detailed annotation of network nodes allowing an in-depth exploration of the network. Some examples include the gene expression omnibus (GEO) [57] and Drug-Bank [58]. Table 2 (direct databases) and Table 3 (indirect databases) list the databases commonly used in network-based methods, with brief descriptions, including the coverage, availability, latest update, and URL.

STRING [55] is the most commonly used direct database in network-based methods. It houses a large number of known and predicted PPIs, including both physical and functional interactions. The data come from the following five main sources: genomic context analysis, high-throughput experimental data, conserved co-expression, artificial text mining, and known information in databases [55]. At the time of writing, STRING covers 24,584,628 proteins from 5090 organisms [55]. This database provides an intuitive and fast viewer for online use, supports online network visualization, and provides a user-friendly platform for data integration with knowledge from other public resources [55].

The GEO database [57] is the most commonly used indirect database in network-based methods. It is a universal public repository for archiving and freely distributing high-throughput microarray, next-generation sequencing, and other forms of high-throughput functional genomic data, with complete and clear annotations from the research community [57]. To date, the GEO database covers 162,671 series comprising 4,777,869 samples. It provides a powerful search engine for users to identify, analyze, and visualize related data of interest. It also supports sophisticated field queries, sample comparison applications, and gene expression profiles [57].

4 Software and Tools

4.1 Comparative Genomics

Table 4 lists the software and tools used in comparative genomics to identify targets. Brief descriptions, availability,

the latest update, and the URL are also provided. In comparative genomics, the BLAST suite (BLASTn, BLASTp, BLASTx, tBLASTn, and tBLASTx) is widely used to analyze the functional and evolutionary relationship between nucleic acid and protein sequences [73]. BLAST is a free online tool that can also be downloaded offline from the National Center for Biotechnology Information (NCBI) website. BLASTn is for nucleic acid sequence alignment; BLASTp is for protein sequence alignment; BLASTx compares the six-frame conceptual translation products of a nucleotide query against a protein sequence database; tBLASTn compares a protein query sequence against a sequence database dynamically translated in all six reading frames, and tBLASTx compares the six-frame translation of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database [73, 74]. There are many specific search modules in NCBI besides those regular modules. For example, smartBLAST [75] can be used to query highly similar proteins, GlobalAlign module to compare two sequences in the entire sequence, CD-search [76] to find conservative domains in a sequence, and CDART to query sequences with similar conservative domain architecture [77]. Moreover, NCBI provides an independent program BLAST+ for users that dramatically accelerates the speed of long sequences query and chromosome length databases query to address the problem of slow-speed BLAST online comparison [78]. Recently, Du et al. designed a cross-platform local BLAST visualization software developed in Python using the in-built graphical user interface (GUI) module TKinter [79]. BlastGUI, as it is known, utilizes BLAST+ as a comparison tool to perform the local operation and sequence comparison visualization. This user-friendly tool allows users without familiarity in computational coding and basic computer skills to compare a sequence directly without additional formatting efforts [79]. BlastGUI preprocesses the input sequence, so the computational complexity of sequence comparison is low. To carry out the comparison, the user enters the file in FASTA format into the search box of BLAST. The maximum acceptable length of nucleotide and protein sequences is generally 1000–2000, and the maximum molecular weight of the protein is 10 to 100 kD. The sequence information can be obtained from NCBI free of charge. Alternatively, the NCBI BLAST uses the indirect BLAST algorithm to run a large number of BLAST searches without using a browser, and the comparison results are returned by e-mail [73].

4.2 Network-Based Methods

Table 5 provides brief descriptions, availability, latest update, and URL of software and tools for network-based methods used in previous target identification studies over the past 5 years. Among them, Cytoscape is the most widely

Table 2 Frequently used direct databases in network-based methods

Database	Description	Coverage	Availability	Latest update	URL	References
STRING	A database of known and predicted protein–protein interactions	24,584,628 proteins from 5090 organisms	Free	2019	http://string-db.org	[55]
MINT	A database designed to store data on functional interactions between proteins	647 organisms and 131,695 interactions	Free	2012	https://mint.bio.uniroma2.it/	[56]
HPRD	A centralized platform to visually depict and integrate information pertaining to domain architecture, post-translational modifications, interaction networks, and disease association for each protein in the human proteome	41,327 protein–protein interactions	Free	2010	http://hprd.org/	[59]
IntAct	Provides an open-source database system and analysis tools for molecular interaction data	119,281 interactors and 1,130,596 interactions	Free	2021	https://www.ebi.ac.uk/intact/	[60]
BioGRID	A comprehensive biomedical resource of curated protein, genetic, and chemical interactions	2,005,220 protein and genetic interactions, 29,093 chemical interactions	Free	2021	https://thebiogrid.org/	[61]
DIP	Catalogs experimentally determined interactions between proteins	28,850 proteins and 81,923 interactions	Free	2020	https://dip.doe-mbi.ucla.edu/dip/Main.cgi	[62]
STITCH	A database of known and predicted interactions between chemicals and proteins	9,643,763 proteins from 2,031 organisms	Free	2016	http://stitch.embl.de/	[63]
miRTarBase	A database of comprehensively annotated, experimentally validated miRNA–target interactions	422,517 curated miRNA–target interactions from 4076 miRNAs and 23,054 target genes	Free	2018	http://miRTarBase.cuhk.edu.cn/	[64]
TarBase	A database of experimentally supported miRNA–gene interactions with detailed information for each interaction	56 tissues, 516 cell types, and 665,843 interactions	Free	2017	http://www.microna.gr/tarbase	[65]

Table 3 Frequently-used indirect databases in network-based methods

Database	Description	Coverage	Availability	Latest update	URL	References
GEO	A public functional genomics data repository supporting microarray experiment (MIAME)-compliant data submissions	162,671 series and 4,777,869 samples	Free	2021	https://www.ncbi.nlm.nih.gov/geo/	[57]
DrugBank	A comprehensive database containing information on drugs and drug targets	14,315 drugs, 4885 targets, and 18,866 drug-target associations	Free	2021	https://go.drugbank.com/	[58]
KEGG	A database resource that integrates genomic, chemical, and systemic functional information	Four categories (systems, genomic, chemical, and health information) from 18 databases	Free	2021	https://www.kegg.jp/	[47]
UniProt	A comprehensive resource for protein sequence and annotation data	305,529 proteomes	Free	2021	https://www.uniprot.org/	[24]
GO	A database source of information on the functions of genes	7,934,369 annotations	Free	2021	http://geneontology.org/	[66]
DAVID	Provides a comprehensive set of functional annotation tools for investigators to understand the biological meaning behind a large list of genes	Information on > 1.5 million genes from > 65,000 species	Free	2020	https://david.ncifcrf.gov	[67]
OMIM	A comprehensive, authoritative compendium of human genes and genetic phenotypes	6,799 phenotypes for which the molecular basis is known and 4,370 genes with phenotype-causing mutation	Free	2021	https://omim.org/	[68]
CARD	A bioinformatic database of resistance genes, their products, and associated phenotypes	88 pathogens and 222,011 alleles	Free	2021	https://card.mcmaster.ca/	[69]
DO	A standardized ontology for human diseases providing descriptions of human disease terms, phenotype characteristics, and related medical vocabulary disease concepts	> 10,500 disease terms and > 7,500 disease terms defined	Free	2021	http://www.disease-ontology.org/	[70]
TTD	A database to provide information about the known and explored therapeutic protein and nucleic acid targets, the targeted disease, pathway information, and the corresponding drugs directed at each of these targets	37,316 drugs and 3,419 targets	Free	2020	http://db.idrblab.net/ttd/	[71]
GeneCards	Provides comprehensive, user-friendly information on all annotated and predicted human genes	42,087 HUGO Gene Nomenclature Committee (HGNC)-approved, 20,916 protein-coding, 219,587 RNA genes	Free	2020	https://www.genecards.org/	[72]
DEG	A database hosting records of currently available essential genomic elements	53,885 essential genes and 786 essential non-coding sequences	Free	2017	http://tubic.tju.edu.cn/deg/	[44]
NDARO	A database of antimicrobial resistance data	> 300,000 pathogen isolates	Free	2020	https://www.ncbi.nlm.nih.gov/pathogens/antimicrobial-resistance/	∧
PATRIC	Provides integrated data and analysis tools to support biomedical research on bacterial infectious diseases	Genome metadata from > 60 fields, three types of protein families	Free	2020	https://www.patricbrc.org/	[31]

Table 4 Software and tools of comparative genomics in the identification of potential therapeutic targets

Software and tools	Description	Availability	Latest update	URL	References
ClustalW/ ClustalX	Multiple alignment of nucleic acid and protein sequences	Free	2010	http://www.clustal.org/clustal2/	[80]
Clustal Omega	Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega	Free	2016	http://www.clustal.org/omega/	[81]
MUSCLE	One of the best-performing multiple alignment programs according to published benchmark tests, with accuracy and speed that are consistently better than ClustalW	Free	2020	http://www.drive5.com/muscle/	[82]
Jalview	A program for multiple sequence alignment editing, visualization, and analysis	Free	2020	http://www.jalview.org/	[83]
KAAS	A web-based server automatically assigns <i>K</i> -numbers to genes in the genome, enabling reconstruction of KEGG pathways and BRITE hierarchies	Free	2015	http://www.genome.jp/kegg/kaas/	[84]
CD-HIT	A program for clustering and comparing protein or nucleotide sequences	Free	2015	http://weizhongli-lab.org/cd-hit/	[85]
PGAT	A prokaryotic-genome analysis tool focused particularly on comparing different strains of the same species	Free	2011	http://nwrce.org/pgat	[86]
ESSENTIALS	Software for predicting essential genes by utilizing transposon insertion sequencing analysis	Free	2012	https://trac.nbic.nl/essentials/	[87]

Table 5 Software and tools of network-based methods for identification of potential therapeutic targets

Software and tools	Description	Availability	Latest update	URL	References
Cytoscape	An open-source platform for complex network visualization and analysis	Free	2020	https://cytoscape.org/	[88]
Gephi	Open-source software for network visualization and analysis	Free	2017	https://gephi.org/	[94]
NetworkAnalyst	A comprehensive network visual analytics platform for gene expression analysis	Free	2021	https://www.networkanalyst.ca/	[95]
HIPPIE	A web tool to generate reliable and meaningful human protein–protein interaction networks	Free	2019	http://cbdm-01.zdv.uni-mainz.de/~mschaefer/hippie/	[96]
PathwayLinker	Assembles validated physical and genetic interaction data with pathway information	Free	/	http://PathwayLinker.org	[97]
KOBAS	A webserver for gene/protein functional annotation and gene set enrichment	Free	2020	http://kobas.cbi.pku.edu.cn/kobas3	[98]
BioCyc	A webserver containing a collection of 18,030 Pathway/Genome Databases (PGDBs), plus software tools for exploring them	Free	2021	https://biocyc.org/	[99]
Cfinder	Software for finding and visualizing overlapping dense groups of nodes in networks	Free	2014	http://cfinder.org/	[100]
Pajek	A program package for the analysis and visualization of large networks	Free	2021	http://mrvar.fdv.uni-lj.si/pajek/	[101]
NetworkX	A Python package for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks	Free	2021	https://networkx.org/	\

used and representative software. Therefore, we chose it as an example for further description of the network-based methods. Cytoscape is a general-purpose platform to analyze and visualize complicated molecular interaction networks. It can be used for integrating massive molecular interaction data. Dynamic states and molecular interactions are mapped as attributes on nodes and edges, and static hierarchical data (such as protein function ontology) are supported by annotations [88]. The Cytoscape Core is the code that organizes, displays, reads, and writes networks but contains no biology-related functionality. It is equipped with basic functionality to lay out and query the network, visually integrate the network with expression profiles, phenotypes, and other molecular states, and link the network to databases of functional annotations [88]. This core functionality is extended by Cytoscape apps. Cytoscape allows users to import attributes from tables whose simplest format are tab-delimited text files containing one column of primary identifiers of network nodes and auxiliary columns of attributes needed mapping to the nodes [89]. To reduce the complexity of a large interaction network, users can create filters based on the attributes as needed and use the Cytoscape built-in function to search [89]. In addition to directly filtering nodes using the built-in topological parameters in Cytoscape, users can also use apps (formerly called plugins), such as stringApp [90], the Biological Networks Gene Oncology (BiNGO) tool [91], Molecular Complex Detection (MCODE) [92], and cytoHubba, a user-friendly interface to explore key nodes and subnetworks [93]. StringApp combines the resources of the STRING database and Cytoscape in the same workflow and facilitates the import of STRING molecular networks into Cytoscape for executing STRING analysis in the script file [90]. BiNGO provides a comprehensive set of annotation tools for Gene Ontology (GO)-level annotations of a variety

of organisms. It enables the extraction of information about overexpression of a gene in biological networks and supports user-defined annotations and ontologies [91]. MCODE enables searches for densely connected regions within large PPI networks that may reflect molecular complexes. The method is based on connectivity data [92]. CytoHubba provides a one-stop calculation of 11 topological analysis methods to help users explore hub objects from complex biological networks [93]. These useful apps are freely available from the Cytoscape App Store (<http://apps.cytoscape.org/>).

5 Applications

5.1 Comparative Genomics

With the arrival of the post-genome era, target-based drug design strategy has gradually become the focus [102]. Both the improvement of the sequencing technology and the exponential explosion of the number of fully sequenced genomes has made it possible to select reasonable new therapeutic targets and vaccine candidates throughout the genome. Drug resistance is becoming increasingly widespread due to the continuous evolution of bacterial strains, such as *Streptococcus pneumoniae* and *Mtb*. Knowledge of therapeutic targets and drug candidates is useful for enhanced drug discovery and is becoming increasingly reliant on comparative genomics technology [103]. Table 6 lists recent applications of comparative genomics in finding therapeutic targets. We selected some specific examples to describe in this section.

Determining essential genes of pathogens is a common method to identify potential therapeutic targets. For example, Tilahun et al. [104] retrieved the protein-coding genes of *Mtb* from the *Mtb* database and identified the essential genes

Table 6 Examples of prediction of potential therapeutic targets by comparative genomics in recent years

Databases	Software and tools	Comparative types	Related pathogens/diseases	Predicted targets	References
UniProt, DEG, Swiss-Prot, TIGR	BLASTx	Genes	<i>Helicobacter pylori</i>	<i>H. pylori</i> essential genes	[20]
PDTD, DSSP	ClustalW, DOCK4.0, TarFisDock	Proteins	<i>H. pylori</i>	PDF	[109]
WormBase	BLASTp	Genes	Human fungal	589 essential genes	[110]
NCBI Proteins	ConSurf server, MUSCLE, Jalview	Proteins	Influenza A virus	NS1 protein, NS2 protein	[111]
<i>Pseudomonas</i> , PDB	BLAST, ExPASy server, ClustalW, ESPrpt	Genes	<i>Pseudomonas</i>	DAHPS sequence	[48]
COGS, DEG, Pathema-JCVI, STRING	BLASTx, BLASTp	Proteins	<i>Clostridium botulinum</i>	39% essential proteins	[112]
KEGG, DEG, Swiss-Prot	BLASTp	Proteins	<i>Actinobacillus pleuropneumoniae</i>	rpoA, metG, gltX	[113]
NCBI Genome, Drug-Bank, DEG	BLAST	Proteins	Nontuberculous mycobacteria	15 candidate proteins	[114]

by a BLAST search of the retrieved protein-coding genes against DEG. Then, the corresponding protein sequences, obtained by searching in DEG, were used to perform a BLASTp search of human protein sequences to avoid host toxicity in the subsequent drug development. Finally, 572 essential genes with no homology to human genes were selected from 3958 genes of *Mtb*. Discovering potential therapeutic targets from the proteins encoded by essential genes can refine the search scope of therapeutic targets. The existence of homologous genes is a powerful predictor of biological importance [105] and a breakthrough in therapeutic target identification. For example, Satya et al. [48] sequenced the gene encoding 3-deoxy-D-arabinoheptulosonate-7-phosphate synthase (*DAHPS*) in *Pseudomonas fragilis* (*Pf*). Sequence analysis showed high homology (84%) of *Pf-DAHPS* with other *Pseudomonas DAHPS*, indicating that it was possible to design a broad-spectrum drug for the genus by targeting the *DAHPS* sequence. By analyzing the homology between the protein sequence encoded by *DAHPS* and human protein sequences, *DAHPS*, which does not exist in humans, was proposed to be an important potential antibacterial target. The predicted three-dimensional structure of *Pseudomonas DAHPS* may provide an option for reasonable drug design [48].

Comparative genomics can be used to understand the molecular mechanism of disease and predict targets for new drug design. For example, Zumla et al. [106] discovered that the sequence homology of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) genome with SARS-CoV and Middle East respiratory syndrome coronavirus (MERS-CoV) was about 82%, and the homology of structural proteins was over 90%. The high sequence homology revealed their common pathogenic mechanism. Therefore, the authors of the study designed and developed direct-acting antiviral drugs that target highly conserved enzymes in SARS-CoV-2, such as the main protease (M_{PRO}) or 3C-like protease (3C_{pro}), the papain-like protease (P_{Lpro}), non-structural protein 12 (Nsp12), and RNA-dependent RNA polymerase (RdRP). Among them, ganciclovir and maraviroc, the drugs against M_{PRO}, were considered effective for the treatment of coronavirus disease 2019 (COVID-19) [107].

Comparative genomics is used to find potential therapeutic targets for the development of human drugs and animal drugs. Damte et al. [108] selected five unique pathways of *Mycoplasma hyopneumoniae* strains in KEGG. They then used BLASTp in NCBI to compare the only two protein sequences in the unique pathways with the porcine protein sequences. It was found that the two protein sequences in the unique pathways were not homologous to the porcine protein sequences. Therefore, those essential proteins, which exist in *M. hyopneumoniae* but not in the host (pig), may be useful in drug design and vaccine production against *M.*

hyopneumoniae. For more examples of comparative genomics used to identify potential targets, readers can refer to the list of references provided in Table 6.

5.2 Network-Based Methods

Different types of biological networks can be used to predict potential therapeutic targets by network-based methods, such as PPI networks, gene interaction networks and miRNA–mRNA interaction networks. Table 7 lists almost all applications since 2015 of network-based methods to predict potential therapeutic targets, including the databases, software and tools, network types, related pathogens/diseases/processes, and the identified targets. Some of the targets in Table 7 have been verified or used for drug design. Here, we select several examples of previous studies that have used different network types for further description.

PPI networks are the most widely used molecular networks in target discovery. For example, Huo et al. predicted proteins FGG, SLC9A3, MAPK14, FGF1, FGB, F13A1, and CASR as potential therapeutic targets for the treatment of coronary heart disease (CHD) by combining the centrality-based and differentia-based approaches [30]. They extracted PPIs related to Danshensu (one of the main active ingredients of *Salvia miltiorrhiza*, known as Danshen) from the STRING database, then integrated the data with the CHD gene expression profile and microarray data obtained from the GEO database to construct a non-CHD state co-expression protein interaction network (CePIN) and a CHD state CePIN on Cytoscape [30]. The non-CHD network contained 91 nodes and 98 edges, and the CHD state CePIN contained 99 nodes and 110 edges [30]. Then, topological analysis and network comparison were performed along with the calculation of network connectivity after the removal of candidate nodes. Finally, two bottleneck proteins, FGG and SLC9A3, existing only in the CHD state CePIN, were selected as the targets of Danshensu in the treatment of CHD and as the potential targets for new drug design [30]. In addition, MAPK14, FGF1, FGB, F13A1, and CASR, obtained through the differentia-based approach, also represented potential therapeutic targets for the treatment of CHD and had been confirmed to be related to CHD to some extent [30].

There are also examples of the use of the centrality based approach alone to identify potential therapeutic targets. For example, Moon et al. generated a list of 1089 differentially expressed genes from patients with diffuse systemic sclerosis by a literature search in Google Scholar and PubMed using specific keywords [125]. Then, using the centrality-based approach to build a PPI network, they identified 1068 interactions of those 1089 genes. Finally, a network centrality analysis identified four hub genes (CTGF, HCK, LYN, PDGFRB) as potential therapeutic

Table 7 Applications of network-based methods for potential therapeutic target identification

Databases	Software and tools	Network types	Related pathogens/diseases/processes	Predicted targets	References
STRING, MIIP	\	PPIN	<i>Plasmodium falciparum</i>	PF10_0232, PF11475w, PF13_0228	[115]
UniProt, STRING, MINT	Cytoscape	PPIN	Sperm-egg interaction defect	FNI, EGFR, ITGAV, ITGB3, COL1A1, ITGB5	[116]
UniProt, STRING	Cytoscape	PPIN	Cancer	SAGE1, SPO11, MAGEC2, FTHL17, DDX53, BAGE2	[117]
GEO, STITCH, STRING, HPRD, KEGG, SignalLink, OMIM, GAD, FunDO, NHGRI GWAS Catalog	PathwayLinker, Cytoscape	PPIN	Hyperlipidemia	COTL1, VASP, HHAT	[118]
HPRD, GEO, GAD, DO, OMIM, DrugBank, GO, KEGG	\	PPIN	Polycystic ovary syndrome	ESR1, RXRA, NCOA1, NR1P1, ESR2, THRB, RARA, NR0B2, NCOA3, HNF4A, PPARA, PPARG, PPARGC1A, MED1, NR2F1, PNRC2, PGR, ESRRA, ESRRG, RXRB, RARG, VDR	[119]
GO, Swiss-Prot, STRING, BioGRID, DIP, IntAct, BIND, MINT	Cytoscape	PPIN	<i>Clostridium difficile</i>	CD2787, CD0237, CD1214, CD2629, CD2643	[120]
STITCH, STRING, GEO	Cytoscape	PPIN	Coronary heart disease	FGG, SLC9A3, MAPK14, FGF1, FGB, F13A1, CASR	[30]
GEO, STRING, GEO, DAVID	Gephi	PPIN	Japanese encephalitis virus	STAT1	[121]
STRING, KEGG, GO	Cytoscape	PPIN	Wound healing	Rela, Nfkb1, Tnfrsf1a	[122]
CellMiner, GEO, HPRD, miRTarBase, TarBase, starBase, SM2miR	CFinder, Cytoscape	PPIN	Ovarian cancer	miR-24-3p, miR-192-5p, miR-139-5p, miR-155-5p	[123]
HPRD, BioGRID, MINT, IntAct, STRING, GEO, CRG	Cytoscape	PPIN	Rheumatoid arthritis	JUN, SYK, LCK	[124]
Google Scholar, PubMed	\	PPIN	Systemic sclerosis	CTGF, HCK, LYN, PDGFRB	[125]
GEO, STRING, JASPAR, TarBase, miRTarBase	Cytoscape, NetworkAnalyst	PPIN	Alzheimer's disease	PPARG	[126]
DIP, KEGG, DEG, DrugBank	Cytoscape, Gephi	PPIN	<i>Streptococcus suis</i>	GlnQ3, GlnQ4, GlnQ1, GlnQ5, PstB1, SSU05_1769	[127]
STRING, UniProt, Pfam	\	PPIN	SARS-CoV-2	MIB1, TBK1, VPS11, AP3B1, GORASP1, GOLGA2	[128]
STRING	Cytoscape	PPIN	Schizophrenia	MAPK1, MAP2K1, CDC42, HSPA1, HSPA8, HRAS, CLTB, SNAP91	[129]
VirusMINT, IntAct, VirusMentha	Cytoscape	PPIN	Influenza A virus	LNx2, MEOx2, TFPC2, PRKRA, DVL2, POLR3F, SNAPC4, GLYR1, ATP6V1G1, PCBPI, EEF1D, DVL3, CREB3	[130]
GEO, STRING	\	PPIN	Pancreatic ductal adenocarcinoma	ITGAV, ITGA2	[131]
TCMSP, TCM, BATMAN-TCM, DRUGBANK, UniProt, GeneCards, OMIM, STRING, KEGG	Cytoscape	PPIN	Colorectal cancer	AKT1, JUN, CDKN1A, BCL2L1, NCOA1	[132]

Table 7 (continued)

Databases	Software and tools	Network types	Related pathogens/diseases/processes	Predicted targets	References
HuRI, HINT, STRING	Cytoscape	PPIN	Colon adenocarcinoma, glioblastoma multiforme, small cell lung cancer	FANCD2, NCOA4, IKBKB, RHOA	[133]
GEO, miRWalk3, DAVID, STRING	Cytoscape	PPIN	Breast cancer	MAPK1, PRKACA, miR-214-3p, miR-587, miR-4472, miR-4422	[134]
GEO, DAVID, STRING	Cytoscape	PPIN	Steroid-induced osteonecrosis of the femoral head	CXCR1, FPRI, TYROBP, MAPK1	[135]
GEO, DAVID, STRING	Cytoscape	PPIN	Pulmonary arterial hypertension	CCL5, CXCL12, VCAM1, CXCR1, SPP1	[136]
GEO, STRING	Cytoscape	PPIN	Oral squamous cell carcinoma	APP, EHMT1, ACACB, PCNA, PLAU, FST, HMG2, LAMC2, SPP1	[137]
PATRIC, ARDB, CARD, NDARO, STRING	Cytoscape	GIN	<i>Pseudomonas aeruginosa</i> PA01	oprJ, oprM, oprN, ampC, gyrA, mexA, oprD, mexB, nfxB	[31]
NCBI genome, PATRIC, ARDB, CARD, NDARO, STRING, DAVID, TTD	Cytoscape	GIN	<i>Proteus mirabilis</i>	rpoB, tufB, rpsL, fusA, rpoC, rpoA	[138]
ARDB, STRING	Cytoscape	GIN	<i>Enterococcus faecalis</i> V583	MraY, PbpC, MurE, MurG, MurD	[139]
ARDB, STRING	Cytoscape	GIN	<i>Salmonella enterica</i> serovar Typhimurium CT18	tolC, macB, acrA, acrB, mdfA	[140]
STRING	Cytoscape	GIN	<i>Klebsiella pneumoniae</i>	gyrA, parC, gyrB, parE, recA	[141]
TCGA, miRCancer, miR2Disease, HMDD, GeneCards, HPAD	\	MMIN	Tumorigenesis	ASPG, AQP2, CNOT8, CTFS1, IFNAR2, MOCS2, PRSS37, VCP	[32]

PPIN protein–protein interaction network, GIN gene interaction network, MMIN miRNA–mRNA interaction network

targets [125]. In another example, Fathima et al. used non-apoptotic cell death genes of colon adenocarcinoma (COAD), glioblastoma multiforme (GBM), and small cell lung cancer (SCLC) screened from their transcriptome profiles to build three PPI networks [133]. Through centrality analysis, 4 of the top 10 hub proteins, which were not found or only found in one target database, were considered as novel valid therapeutic targets (FANCD2 and NCOA4 for COAD, IKBKB for GBM, and RHOA for GBM and SCLC) [133].

As mentioned above, PPI networks, gene interaction networks, and miRNA–mRNA interaction networks) have applications in predicting potential therapeutic targets. For example, Miryala et al. [31] identified 337 functional interactions of 60 antimicrobial resistance genes of *Pseudomonas aeruginosa* PA01 from the PathoSystems Resource Integration Center (PATRIC) tool, The Antibiotic Resistance Genes Database (ARDB) [142], the comprehensive antibiotic resistance database (CARD), the National database of antibiotic-resistant organisms (NDARO), and the STRING database. By constructing and analyzing the gene interaction network in Cytoscape, nine hub genes were obtained as potential therapeutic targets for new drug development [31]. Xue et al. [32] constructed a miRNA–mRNA interaction network using miRNA and mRNA expression data, and the clinical data of three cancer types downloaded from The cancer genome atlas (TCGA) database [143]. The top 20 miRNAs with the highest degree in each data set were annotated via miR-Cancer (a microRNA–cancer association database) [144], miR2Disease (a microRNA–disease database) [145], and the Human microRNA Disease Database (HMDD) [146]. After mapping the genes predicted as the targets of more than three miRNAs in the subnetworks to the human protein atlas database (HPAD) [147], eight genes (ASPG, AQP2, CNOT8, CTPS1, IFNAR2, MOCS2, PRSS37, and

VCP) were finally identified as potential therapeutic targets [32].

5.3 Comprehensive Applications

In addition to using comparative genomics and network-based methods independently, they can also be combined for target identification. Table 8 lists recent applications of the combined methods for potential therapeutic target identification. We chose three of them as representatives for further description. Nayak et al. screened putative targets for pathogens causing bacterial pneumonia. By bit score, E-value threshold, and sequence length screening of the complete proteome of 13 pathogenic bacterial strains using comparative genomics, 74 proteins non-homologous to human and intestinal flora were identified [103]. An interaction network for the 74 proteins was constructed in Cytoscape, and 12 built-in central parameters of cytoHubba were used to prioritize the nodes, culminating in the identification of 20 genes as hub nodes. Among the 20 genes, 10 have been reported or confirmed as drug targets, and the remaining 10 were considered new potential therapeutic targets for the treatment of bacterial pneumonia [103]. Melak and Gakkhar used BLAST to perform comparative analysis for the H37RV protein-coding genes obtained from the TBDB against DEG and identified 572 essential genes non-homologous with humans [104]. Then, they prioritized the resulting proteins based on centrality measurement in the PPI network, resulting in the identification of 137 central proteins. Combining flux balance analysis of the reactome and structural assessment of targetability, secY (Rv0732), katG (Rv1908c), gltB (Rv3859c), and sirA (Rv2391) were identified as potential therapeutic targets against *Mtb* H37RV [104]. Gupta et al. [148] performed subtractive genomic and comparative genomics of 16 pathogenic *Leptospira* strains retrieved from NCBI against DEG and the Cluster of Essential Genes (CEG) [149] using the Cluster Database at High Identity

Table 8 Applications of comprehensive methods for potential therapeutic target identification

Databases	Software and tools	Related pathogens/diseases	Predicted targets	Ref
DEG, STRING	BLASTp, Cytoscape	<i>Listeria monocytogenes</i> strain EGD-e	dnaN, lmo0162, polC	[150]
TBDB, DEG, STRING	BLASTp	<i>Mycobacterium tuberculosis</i> H37Rv	Rv1908c, Rv3795, Rv3793, Rv3794	[104]
TBDB, DEG, STRING	BLAST, Cytoscape	<i>M. tuberculosis</i> H37Rv	secY, katG, gltB, sirA	[151]
STRING	Cytoscape	Multi-drug resistant <i>Clostridium difficile</i> strain 630	hom, asd, dapG	[152]
DEG, CEG, VFDB, Drug-Bank, UniProt, DAVID, HPIDB	BLASTp, CD-Hit, BioCyc webserver, Cytoscape	Bacterial pneumonia	manL, cps4L, recU, SP_0645, ezcA, prsA, tarJ, SP_1280, SP_1617, ptsG, dltD, hprK, pepF, coiA, fibB, acpS, manA, mvaK2, mtlD, mtlF	[103]
NCBI, DEG, CEG, UniProt	KAAS, BLASTp, Cytoscape	<i>Leptospira</i>	lpxB, lpxK, kdtA, fliN, cobA, metX, thiL, ubiA	[148]

with Tolerance (CD-Hit) and BLASTp to identify 34 common genes. After analyzing and comparing two extended PPI networks of two strains and multiple sequence alignment, eight proteins (lpxB, lpxK, kdtA, fliN, cobA, metX, thiL, and ubiA) were identified as putative therapeutic targets for drug design or vaccine development [148].

6 Discussion

6.1 Advantages and Prospects of the Two Categories of In silico Methods for Target Identification

Current trends in drug discovery focus on understanding disease mechanisms, followed by target identification and lead compound discovery [5]. Compared with wet experimental methods, in silico methods provide the technology to systematically explore all possible interactions and illuminate the pharmacological patterns [153]. Reliable target identification methods used in conjunction with drug discovery approaches will improve the efficiency of computer-aided drug discovery [5]. Here, we discuss the advantages and prospects of comparative genomics and network-based methods for identifying potential therapeutic targets.

One advantage of comparative genomics is that the definition of essential genes and unique metabolic pathways not only represents the essential issues of biology but is also of great significance in practical applications [111]. Furthermore, with the establishment of Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) and exome sequencing technology, the number of sequenced human essential genes has increased remarkably [54]. In addition, with the development of bioinformatics and computer science, algorithms have been continuously optimized, generating convenient analysis tools for scientific researchers, and enhancing the potential for comparative genomics in potential therapeutic target identification.

Network-based methods have the advantage of generating visual interactive networks through given databases and are not limited by the lack of quantitative mechanical data [154]. Furthermore, network-based methods do not depend on negative samples and the three-dimensional structure of targets [155], which is time-efficient in the early work of target research. There is also promise that network-based methods will predict more than one target with simultaneous actions, such as a pair of essential proteins [154]. Moreover, network-based methods may be beneficial in identifying candidate multi-target sets in the development of multi-target drugs [156]. Compared with traditional wet experimental methods, which always limit cellular processes to a single component or signaling pathway, network-based methods can be used to identify potential therapeutic targets systematically [15].

6.2 Disadvantages/Limitations of the Two Categories of In silico Methods for Target Identification and Potential Solutions

Although comparative genomics and network-based methods have unique advantages and promising prospects to identify potential therapeutic targets, there are still some drawbacks. For comparative genomics, although this approach is commonly used in the development of drugs against drug-resistant bacteria, the failure rate of old antibiotics is much faster than the development of new antibiotics. Moreover, antibiotics are short-term therapies for the treatment of infections. Additionally, their value is considerably less than the drugs for chronic diseases, so the use of comparative genomics in the development of antibiotics is a long-debated topic [157]. Another issue is that although comparative genomics can reduce the number of experimental targets, making some attractive proteins become potential therapeutic targets, the range of potential targets screened by this method is still very wide and is limited by time and cost. It seems that most of these potential targets screened by comparative genomics will not be used for experimental validation. Therefore, it may be profitable to combine comparative genomics with network-based methods to narrow the scope of experimental targets further and reduce the time and material resources, thereby saving costs in the early stage of drug research and development.

Network-based methods are highly dependent on the accuracy of the source data, potentially requiring a great deal of labor to ensure its accuracy [158]. A promising direction to resolve this problem will be integrating different types and complementary data in the future [6]. Other drawbacks of network-based methods are that they cannot predict proteins or genes without interaction data, and the interactions cannot be quantified [155]. Improved network construction and analysis algorithms or mathematical modeling methods [159] may be required to overcome these issues.

6.3 Comparison and Contrast of the Two Categories of In silico Methods for Target identification

Comparative genomics and network-based methods have unique advantages and disadvantages in predicting targets. Comparative genomics almost exclusively searches within the range of pathogen-associated sequences, limiting the scope to the proteomes closely related to the pathogen. Conversely, network-based methods can be used in pathogens and construct a network for human disease-related proteins or genes. In contrast to comparative genomics, network-based methods can connect long-distance relationships through interactions [160], permitting research into the interplay of evolutionary drivers on a larger scale. Conversely, comparative genomics is usually superior to network-based

methods in accuracy because comparative genomics directly compares sequences, which are always constant and almost have no deviation. However, there may be false positives and false negatives in the interaction data used in network-based methods [161], and the interactions are only qualitative [160], which may lead to bias. In summary, the combined use of comparative genomics and network-based methods may be more beneficial than either method alone to improve the accuracy and efficiency in target identification.

6.4 Previous Reviews and Prospective Studies on In silico Methods for Target Identification

We have collected five reviews on in silico methods for identifying potential therapeutic targets during 2016–2020, which will be briefly discussed in this section. Sekyere and Asante [7] reviewed comparative genomic analysis *trans*-complementation assays in the context of antibiotic resistance research and new drug discovery by describing the emergence of several new drug resistance genes, such as *lsa*(C), *erm*(44), *VCC-1*, *mcr-1*, *mcr-2*, *mcr-3*, *mcr-4*, *bla_{KLUC-3}*, and *bla_{KLUC-4}*. For readers interested in further understanding pathogen protein targets, Saha et al. reviewed the computational work and functional prediction from PPI networks applied to different infectious diseases with *Plasmodium falciparum* used as an example to analyze the process of protein target identification through the host–pathogen protein interactions [162]. Katsila et al. [5] surveyed chemical informatics and network-based methods for identifying therapeutic targets and introduced some databases and network computing tools for target identification. They also appraised the process of computer-aided drug design (CADD), including ligand-based drug design and structure-based drug design [5]. Readers interested in CADD can peruse their article for further understanding. Reisdorf et al. introduced database resources for identification, prioritization, and validation of disease targets, including emerging integrated bioinformatics platforms, such as Open Targets, and public resources, such as DrugBank and ChEMBL [163]. In comparison, the database resources we described focus more on classic or commonly used databases for applications. We also recommend the review by Agamah et al. [153], which examined current in silico methods for the identification of therapeutic targets and candidate drugs,

including network-based analysis approaches, data mining, reverse docking, biospectra analysis, and ligand-based in silico target prediction and compared the different approaches and propounded the benefits of hybrid approaches.

6.5 Related Methods for Target Identification

In silico subtractive genomics (first mentioned in Sect. 2.1), also known as differential proteome mining, is a comparative genomics-based method [164]. Subtractive genomics gradually subtracts proteins from the complete proteome of pathogens to find rational targets [18]. The difference between subtractive genomics and comparative genomics is in the range of application of the two methods. Subtractive genomics has been widely used for developing potential anti-pathogen infection drugs [18], whereas comparative genomics can be used not only to identify potential targets of pathogens but also to understand the molecular basis of disease [106].

For network-based methods, in addition to the centrality-based and differential-based approaches we reviewed above, there are also studies showing the use of network influence [165], controllability [166], and topological similarity strategy [167] in target identification, but the relevant applications are much fewer. Compared with network centrality, the network influence strategy focuses on the vulnerable nodes close to the central nodes in networks. Acting on these nodes may not be fatal but can have a major impact on the central nodes, so these nodes have the potential to be therapeutic targets [165]. The controllability strategy applies structural controllability theory to determine the minimum set of driver nodes in control of the entire network and identify indispensable nodes as prime targets for disease-causing mutations, viruses, and drugs [166]. The topological similarity strategy focuses on the nodes in the network with similar topological properties to the existing drug targets, which can be potentially developed as therapeutic targets [167].

Commonly used experimental methods for potential therapeutic target identification, especially for essential genes, include single-gene knockout, antisense RNA inhibition of gene expression, large-scale transposon mutagenesis, and CRISPR/Cas9 nuclease system knockout screening. The limitations of experimental methods in identifying essential genes are listed in Table 9 [168].

Table 9 Limitations of experimental methods in identifying essential genes

Experimental methods	Limitations
Single-gene knockout strategy	Requires detailed genome annotation
Antisense RNA inhibition method	Requires detailed genome annotation
Transposon mutagenesis	Missing low-abundance transcripts, low resolution in locating insertion sites, and narrow ranges in counting probe density

Current computational studies are based on the integration of prior knowledge, the sparseness of which is still limiting the integrality and accuracy of computational prediction [169]. Data reproducibility of *in silico* methods is also an essential issue but might be improved by external validation and detailed reports of experimental datasets [153]. It should be emphasized that computational methods complement laboratory-based methods and that the targets identified by *in silico* methods need to be experimentally validated.

6.6 Review and Prospection of Deep Learning Architecture in Target Identification

Deep learning (DL), a relatively new computational technique that has become a hot research topic, has been rapidly developed and widely used to predict potential therapeutic targets. DL is a subclass of machine learning (ML) algorithms. It uses artificial neural networks with many layers of nonlinear processing units for learning data representations [170]. Therapeutic target identification based on ML or DL is usually used to predict targets of drug repositioning, which means to predict new targets for existing drugs. There are two steps in the ML method to predict therapeutic targets. First, the compounds are transformed into an effective representation, a process called input features, followed by the construction of the feature vectors as input for the ML algorithm to learn the functional relationship between the input feature and the target property [171]. Compared with ML methods, DL reconstructs the original input information into a distributed representation through neurons in the hidden layer. Another characteristic of DL models is that they can automatically learn features upon completing classification and other tasks and learn more complex features when the number of layers increases. DL architectures are well-suited for target prediction because they allow for multitask learning and automatically construct complex features, which, for target prediction, are assumed to be pharmacophore descriptors. Multitask learning has the advantage of allowing for multi-label information and can, therefore, utilize relations between targets. It also permits hidden unit representations to be shared among prediction tasks, which is particularly valuable because some targets have very few measurements available, making single-target prediction ineffective. In addition, DL can boost the performance of tasks with a few training examples. The other advantage of deep networks is that they provide hierarchical representations of a compound, where higher levels represent more complex properties [172].

Convolutional neural networks (CNNs) are a representative DL architecture in potential target prediction. CNNs contain convolutional layers, pooling layers, and fully connected layers. Convolutional layers and pooling layers are

responsible for the feature extraction, and fully connected layers are used to construct the nonlinear relationship of the extracted features for obtaining the output [171]. Another DL architecture is deep neural networks (DNNs), which contain multiple hidden layers, with each layer comprising hundreds of nonlinear process units. DNNs can deal with many input features, and the neurons in different layers of a DNN can automatically extract features at different hierarchical levels [173]. The third main DL architecture is auto-encoders, which is a neural network used for unsupervised learning. Auto-encoders contain an encoder part that transforms the input information into a limited number of hidden units and then couples a decoder neural network with the output layer having the same number of nodes as the input layer [174].

Several studies have reported DL for therapeutic target prediction in recent years [175–177]. For example, Wang et al. [178] constructed a framework that combines a biased support vector machine and a stacked auto-encoder DL model to identify drug target proteins. The stacked auto-encoders were trained to extract properties from the original protein representations, and the biased support vector machine was used to perform the potential target identification task. The framework identified 23% of the original non-drug target proteins as possible therapeutic target proteins. Zeng et al. [179] developed a DL method, named deepDTnet, for novel target identification. A DNN algorithm was used to learn the relationships between drugs and targets. The model was used to predict the new target for topotecan (an approved topoisomerase inhibitor of human retinoic-acid-receptor-related orphan receptor-gamma t, ROR- γ t). Human ROR- γ t was predicted as the target, and bioassay experiments showed high inhibitory activity ($IC_{50} = 0.43 \mu\text{M}$) on ROR- γ t. Lee et al. [180] proposed a DL model named DeepConv-DTI (deep learning with convolution on protein sequences for prediction of drug–target interaction) based on CNN for drug–target interactions prediction, which can be used for target identification. The training dataset contained 11,950 compounds, 3,675 proteins, and 32,568 drug–target interactions. The CNN model is constructed to capture local residue patterns and concatenate protein features with drug features through the fully connected layers. The hyperparameters with an external validation dataset were then optimized. The possible drug–protein interactions are output.

Although DL has advantages in recognition, classification, and feature extraction from complex and noisy data, it still has limitations. First of all, DL is a “black box,” which makes it hard to explain the prediction result and inherent principles of why the compound is effectively targeted to the predicted target. Second, it needs a large number of experimental datasets of drug–target relationships for its training. However, there is currently a lack of experimental data of drug–target relationships [181]. Consequently,

there is a risk of overfitting when training the model, leading to low accuracy of the prediction result. Third, DL is usually computationally intensive, time-consuming, and often requires access to and programming knowledge for graphics processing units. DL has recently been applied successfully in therapeutic target identification. However, due to the lack of large-scale studies or experimental data and the hyperparameter selection bias that comes with the high number of potential DL architectures, DL still has scope for improvement and development in research to predict potential therapeutic targets [172, 182].

7 Conclusion

In this review, we introduced, in detail, the two categories of in silico methods for potential therapeutic target identification—comparative genomics and network-based methods—and summarized the databases and software commonly used for these approaches. We also collected and highlighted some previous applications of these methods for therapeutic target identification. Additionally, we analyzed the advantages and disadvantages of the methods and their application prospects. Finally, we accentuated the characteristics of our review in the context of previously published relevant reviews and methods. The purpose of this review was to help readers quickly understand the rationales of in silico methods for potential therapeutic target identification, and become familiar with the available tool resources and the applications of these methods, to harness the full use of the existing tools for target prediction. We strongly believe that more accurate predictions due to users' familiarity with existing resources will increase the importance of computational methods in the identification of potential therapeutic targets for future research. In turn, the failure rate due to target problems in drug development, the input–output ratio of drug discovery, and the cost of subsequent experiments can be expected to reduce and the drug development cycle time to shorten.

Acknowledgements We thank Wordvice for their help in revising the English grammar.

Author contributions ZNH, FXW, XTZ, NY and SKM contributed to the design and conception of the study. XTZ, NY, XHZ, JBL, and SKM performed information retrieval and analysis. XTZ, FXW, NY, XHZ and JBL wrote the manuscript. XTZ, FXW and NY created the tables and figures. ZNH and FXW guided the manuscript writing and revised the manuscript. ZNH provided financial support. All authors contributed to manuscript revision and have read and approved the submitted version.

Funding This work was supported by the National Natural Science Foundation of China (31770774), the Key Discipline Construction

Project of Guangdong Medical University (4SG21004G) and the Higher Education Reform Project of Guangdong Province (2019268).

Declarations

Conflicts of interest The authors confirm that this article content has no conflicts of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Tang Y, Zhu W, Chen K, Jiang H (2006) New technologies in computer-aided drug design: toward target identification and new chemical entity discovery. *Drug Discov Today Technol* 3:307–313. <https://doi.org/10.1016/j.ddtec.2006.09.004>
2. Cook D, Brown D, Alexander R, March R, Morgan P, Satterthwaite G, Pangalos MN (2014) Lessons learned from the fate of AstraZeneca's drug pipeline: a five-dimensional framework. *Nat Rev Drug Discov* 13:419–431. <https://doi.org/10.1038/nrd4309>
3. Morgan P, Brown DG, Lennard S, Anderton MJ, Barrett JC, Eriksson U, Fidock M, Hamren B, Johnson A, March RE, Matcham J, Mettetal J, Nicholls DJ, Platz S, Rees S, Snowden MA, Pangalos MN (2018) Impact of a five-dimensional framework on R&D productivity at AstraZeneca. *Nat Rev Drug Discov* 17:167–181. <https://doi.org/10.1038/nrd.2017.244>
4. Wooller SK, Benstead-Hume G, Chen X, Ali Y, Pearl FMG (2017) Bioinformatics in translational drug discovery. *Biosci Rep* 37:BSR20160180. <https://doi.org/10.1042/BSR20160180>
5. Katsila T, Spyroulias GA, Patrinos GP, Matsoukas MT (2016) Computational approaches in target identification and drug discovery. *Comput Struct Biotechnol J* 14:177–184. <https://doi.org/10.1016/j.csbj.2016.04.004>
6. Dai YF, Zhao XM (2015) A survey on the computational approaches to identify drug targets in the postgenomic era. *Biomed Res Int*. <https://doi.org/10.1155/2015/239654>
7. Sekyere JO, Asante J (2018) Emerging mechanisms of antimicrobial resistance in bacteria and fungi: advances in the era of genomics. *Future Microbiol* 13:241–262. <https://doi.org/10.2217/fmb-2017-0172>
8. Csermely P, Korcsmaros T, Kiss HJ, London G, Nussinov R (2013) Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review. *Pharmacol Ther* 138:333–408. <https://doi.org/10.1016/j.pharmthera.2013.01.016>
9. Zhang Z, Ren Q (2015) Why are essential genes essential? The essentiality of *Saccharomyces* genes. *Microb Cell* 2:280–287. <https://doi.org/10.15698/mic2015.08.218>
10. Hopkins AL (2007) Network pharmacology. *Nat Biotechnol* 25:1110–1111. <https://doi.org/10.1038/nbt1007-1110>

11. Barabasi AL, Gulbahce N, Loscalzo J (2011) Network medicine: a network-based approach to human disease. *Nat Rev Genet* 12:56–68. <https://doi.org/10.1038/nrg2918>
12. Barabasi AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. *Nat Rev Genet* 5:101–113. <https://doi.org/10.1038/nrg1272>
13. Xie L, Li J, Xie L, Bourne PE (2009) Drug discovery using chemical systems biology: identification of the protein-ligand binding network to explain the side effects of CETP inhibitors. *PLoS Comput Biol* 5:e1000387. <https://doi.org/10.1371/journal.pcbi.1000387>
14. Jain B, Raj U, Varadwaj PK (2018) Drug target interplay: a network-based analysis of human diseases and the drug targets. *Curr Top Med Chem* 18:1053–1061. <https://doi.org/10.2174/1568026618666180719160922>
15. Chu LH, Chen BS (2008) Construction of a cancer-perturbed protein-protein interaction network for discovery of apoptosis drug targets. *BMC Syst Biol* 2:56. <https://doi.org/10.1186/1752-0509-2-56>
16. Buysse JM (2001) The role of genomics in antibacterial target discovery. *Curr Med Chem* 8:1713–1726. <https://doi.org/10.2174/0929867013371699>
17. Abadio AK, Kioshima ES, Teixeira MM, Martins NF, Maigret B, Felipe MS (2011) Comparative genomics allowed the identification of drug targets against human fungal pathogens. *BMC Genomics* 12:75. <https://doi.org/10.1186/1471-2164-12-75>
18. Hosen MI, Tanmoy AM, Mahbuba DA, Salma U, Nazim M, Islam MT, Akhteruzzaman S (2014) Application of a subtractive genomics approach for in silico identification and characterization of novel drug targets in *Mycobacterium tuberculosis* F11. *Interdiscip Sci* 6:48–56. <https://doi.org/10.1007/s12539-014-0188-y>
19. Shanmugam A, Natarajan J (2010) Computational genome analyses of metabolic enzymes in *Mycobacterium leprae* for drug target identification. *Bioinformatics* 4:392–395. <https://doi.org/10.6026/97320630004392>
20. Dutta A, Singh SK, Ghosh P, Mukherjee R, Mitter S, Bandyopadhyay D (2006) In silico identification of potential therapeutic targets in the human pathogen *Helicobacter pylori*. *In Silico Biol* 6:43–47
21. Spaltmann F, Blunck M, Ziegelbauer K (1999) Computer-aided target selection—prioritizing targets for antifungal drug discovery. *Drug Discov Today* 4:17–26. [https://doi.org/10.1016/s1359-6446\(98\)01278-1](https://doi.org/10.1016/s1359-6446(98)01278-1)
22. Chawley P, Samal HB, Prava J, Suar M, Mahapatra RK (2014) Comparative genomics study for identification of drug and vaccine targets in *Vibrio cholerae*: MurA ligase as a case study. *Genomics* 103:83–93. <https://doi.org/10.1016/j.ygeno.2013.12.002>
23. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res* 38:D355–360. <https://doi.org/10.1093/nar/gkp896>
24. UniProt C (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res* 47:D506–D515. <https://doi.org/10.1093/nar/gky1049>
25. Lagunin AA, Ivanov SM, Glorizova TA, Pogodin PV, Filimonov DA, Kumar S, Goel RK (2020) Combined network pharmacology and virtual reverse pharmacology approaches for identification of potential targets to treat vascular dementia. *Sci Rep* 10:257. <https://doi.org/10.1038/s41598-019-57199-9>
26. Lichtblau Y, Zimmermann K, Haldemann B, Lenze D, Hummel M, Leser U (2017) Comparative assessment of differential network analysis methods. *Brief Bioinform* 18:837–850. <https://doi.org/10.1093/bib/bbw061>
27. Fadhal E, Mwambene EC, Gamielien J (2014) Modelling human protein interaction networks as metric spaces has potential in disease research and drug target discovery. *BMC Syst Biol* 8:68. <https://doi.org/10.1186/1752-0509-8-68>
28. Yang Y, Adelstein SJ, Kassis AI (2012) Target discovery from data mining approaches. *Drug Discov Today* 17(Suppl):S16–23. <https://doi.org/10.1016/j.drudis.2011.12.006>
29. Su G, Morris JH, Demchak B, Bader GD (2014) Biological network exploration with Cytoscape 3. *Curr Protoc Bioinform* 47:813–81324. <https://doi.org/10.1002/0471250953.bi0813s47>
30. Huo M, Wang Z, Wu D, Zhang Y, Qiao Y (2017) Using coexpression protein interaction network analysis to identify mechanisms of danshensu affecting patients with coronary heart disease. *Int J Mol Sci* 18:1298. <https://doi.org/10.3390/ijms18061298>
31. Miryala SK, Anbarasu A, Ramaiah S (2019) Systems biology studies in *Pseudomonas aeruginosa* PA01 to understand their role in biofilm formation and multidrug efflux pumps. *Microb Pathog* 136:103668. <https://doi.org/10.1016/j.micpath.2019.103668>
32. Xue J, Xie F, Xu J, Liu Y, Liang Y, Wen Z, Li M (2017) A new network-based strategy for predicting the potential miRNA-mRNA interactions in tumorigenesis. *Int J Genomics* 2017:3538568. <https://doi.org/10.1155/2017/3538568>
33. Farkas IJ, Korcsmaros T, Kovacs IA, Mihalik A, Palotai R, Simko GI, Szalay KZ, Szalay-Beko M, Vellai T, Wang S, Csermely P (2011) Network-based tools for the identification of novel drug targets. *Sci Signal* 4:pt3. <https://doi.org/10.1126/scisignal.2001950>
34. Yu H, Kim PM, Sprecher E, Trifonov V, Gerstein M (2007) The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS Comput Biol* 3:e59. <https://doi.org/10.1371/journal.pcbi.0030059>
35. Zhu M, Gao L, Li X, Liu Z, Xu C, Yan Y, Walker E, Jiang W, Su B, Chen X, Lin H (2009) The analysis of the drug-targets based on the topological properties in the human protein-protein interaction network. *J Drug Target* 17:524–532. <https://doi.org/10.1080/10611860903046610>
36. Peng Q, Schork NJ (2014) Utility of network integrity methods in therapeutic target identification. *Front Genet* 5:12. <https://doi.org/10.3389/fgene.2014.00012>
37. Zaman N, Li L, Jaramillo ML, Sun Z, Tibiche C, Banville M, Collins C, Trifiro M, Paliouras M, Nantel A, O'Connor-McCourt M, Wang E (2013) Signaling network assessment of mutations and copy number variations predict breast cancer subtype-specific drug targets. *Cell Rep* 5:216–223. <https://doi.org/10.1016/j.celrep.2013.08.028>
38. van Dam S, Vosa U, van der Graaf A, Franke L, de Magalhães JP (2018) Gene co-expression analysis for functional classification and gene-disease predictions. *Brief Bioinform* 19:575–592. <https://doi.org/10.1093/bib/bbw139>
39. Estrada E (2006) Protein bipartivity and essentiality in the yeast protein-protein interaction network. *J Proteome Res* 5:2177–2184. <https://doi.org/10.1021/pr060106e>
40. Jeong H, Mason SP, Barabasi AL, Oltvai ZN (2001) Lethality and centrality in protein networks. *Nature* 411:41–42. <https://doi.org/10.1038/35075138>
41. Hwang WC, Zhang A, Ramanathan M (2008) Identification of information flow-modulating drug targets: a novel bridging paradigm for drug discovery. *Clin Pharmacol Ther* 84:563–572. <https://doi.org/10.1038/clpt.2008.129>
42. Albert R, Jeong H, Barabasi AL (2000) Error and attack tolerance of complex networks. *Nature* 406:378–382. <https://doi.org/10.1038/35019019>
43. Boutet E, Lieberherr D, Tognolli M, Schneider M, Bansal P, Bridge AJ, Poux S, Bougueleret L, Xenarios I (2016) UniProtKB/Swiss-Prot, the manually annotated section of the UniProt

- KnowledgeBase: how to use the entry view. *Methods Mol Biol* 1374:23–54. https://doi.org/10.1007/978-1-4939-3167-5_2
44. Luo H, Lin Y, Gao F, Zhang CT, Zhang R (2014) DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements. *Nucleic Acids Res* 42:D574–580. <https://doi.org/10.1093/nar/gkt1131>
 45. Lin Y, Zhang RR (2011) Putative essential and core-essential genes in *Mycoplasma* genomes. *Sci Rep* 1:53. <https://doi.org/10.1038/srep00053>
 46. Chen WH, Minguez P, Lercher MJ, Bork P (2012) OGEE: an online gene essentiality database. *Nucleic Acids Res* 40:D901–906. <https://doi.org/10.1093/nar/gkr986>
 47. Kanehisa M, Furumichi M, Sato Y, Ishiguro-Watanabe M, Tanabe M (2021) KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res* 49:D545–D551. <https://doi.org/10.1093/nar/gkaa970>
 48. Tapas S, Kumar Patel G, Dhindwal S, Tomar S (2011) In Silico sequence analysis and molecular modeling of the three-dimensional structure of DAHP synthase from *Pseudomonas fragi*. *J Mol Model* 17:621–631. <https://doi.org/10.1007/s00894-010-0764-y>
 49. Pieper U, Webb BM, Barkan DT, Schneidman-Duhovny D, Schlessinger A, Braberg H, Yang Z, Meng EC, Pettersen EF, Huang CC, Datta RS, Sampathkumar P, Madhusudhan MS, Sjölander K, Ferrin TE, Burley SK, Sali A (2011) ModBase, a database of annotated comparative protein structure models, and associated resources. *Nucleic Acids Res* 39:D465–474. <https://doi.org/10.1093/nar/gkq1091>
 50. Pruitt KD, Tatusova T, Maglott DR (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* 35:D61–65. <https://doi.org/10.1093/nar/gkl842>
 51. Galagan JE, Sisk P, Stolte C, Weiner B, Koehrsen M, Wymore F, Reddy TB, Zucker JD, Engels R, Gellesch M, Hubble J, Jin H, Larson L, Mao M, Nitzberg M, White J, Zachariah ZK, Sherlock G, Ball CA, Schoolnik GK (2010) TB database 2010: overview and update. *Tuberculosis* 90:225–235. <https://doi.org/10.1016/j.tube.2010.03.010>
 52. Lee RYN, Howe KL, Harris TW, Arnaboldi V, Cain S, Chan J, Chen WJ, Davis P, Gao S, Grove C, Kishore R, Muller HM, Nakamura C, Nuin P, Paulini M, Raciti D, Rodgers F, Russell M, Schindelman G, Tuli MA, Van Auken K, Wang Q, Williams G, Wright A, Yook K, Berriman M, Kersey P, Schedl T, Stein L, Sternberg PW (2018) WormBase 2017: molting into a new stage. *Nucleic Acids Res* 46:D869–D874. <https://doi.org/10.1093/nar/gkx998>
 53. Chen L, Yang J, Yu J, Yao Z, Sun L, Shen Y, Jin Q (2005) VFDB: a reference database for bacterial virulence factors. *Nucleic Acids Res* 33:D325–328. <https://doi.org/10.1093/nar/gki008>
 54. Luo H, Lin Y, Liu T, Lai FL, Zhang CT, Gao F, Zhang R (2021) DEG 15, an update of the Database of Essential Genes that includes built-in analysis tools. *Nucleic Acids Res* 49:D677–D686. <https://doi.org/10.1093/nar/gkaa917>
 55. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P, Jensen LJ, Mering CV (2019) STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* 47:D607–D613. <https://doi.org/10.1093/nar/gky1131>
 56. Licata L, Briganti L, Peluso D, Perfetto L, Iannuccelli M, Galeota E, Sacco F, Palma A, Nardoza AP, Santonico E, Castagnoli L, Cesareni G (2012) MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res* 40:D857–861. <https://doi.org/10.1093/nar/gkr930>
 57. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, Robertson CL, Serova N, Davis S, Soboleva A (2013) NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res* 41:D991–995. <https://doi.org/10.1093/nar/gks1193>
 58. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z, Assempour N, Iynkkaran I, Liu Y, Maciejewski A, Gale N, Wilson A, Chin L, Cummings R, Le D, Pon A, Knox C, Wilson M (2018) DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res* 46:D1074–D1082. <https://doi.org/10.1093/nar/gkx1037>
 59. Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, Surendranath V, Niranjan V, Muthusamy B, Gandhi TK, Gronborg M, Ibarrola N, Deshpande N, Shanker K, Shivashankar HN, Rashmi BP, Ramya MA, Zhao Z, Chandrika KN, Padma N, Harsha HC, Yatish AJ, Kavitha MP, Menezes M, Choudhury DR, Suresh S, Ghosh N, Saravana R, Chandran S, Krishna S, Joy M, Anand SK, Madavan V, Joseph A, Wong GW, Schiemann WP, Constantinescu SN, Huang L, Khosravi-Far R, Steen H, Tewari M, Ghaffari S, Blobel GC, Dang CV, Garcia JG, Pevsner J, Jensen ON, Roepstorff P, Deshpande KS, Chinnaiyan AM, Hamosh A, Chakravarti A, Pandey A (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res* 13:2363–2371. <https://doi.org/10.1101/gr.1680803>
 60. Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F, Campbell NH, Chavali G, Chen C, del-Toro N, Duesbury M, Dumousseau M, Galeota E, Hinz U, Iannuccelli M, Jagannathan S, Jimenez R, Khadake J, Lagreid A, Licata L, Lovering RC, Meldal B, Melidoni AN, Milagros M, Peluso D, Perfetto L, Porras P, Raghunath A, Ricard-Blum S, Roehert B, Stutz A, Tognolli M, van Roey K, Cesareni G, Hermjakob H (2014) The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res* 42:D358–363. <https://doi.org/10.1093/nar/gkt1115>
 61. Oughtred R, Rust J, Chang C, Breitkreutz BJ, Stark C, Willems A, Boucher L, Leung G, Kolas N, Zhang F, Dolma S, Coulombe-Huntington J, Chatr-Aryamontri A, Dolinski K, Tyers M (2021) The BioGRID database: a comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci* 30:187–200. <https://doi.org/10.1002/pro.3978>
 62. Xenarios I, Rice DW, Salwinski L, Baron MK, Marcotte EM, Eisenberg D (2000) DIP: the database of interacting proteins. *Nucleic Acids Res* 28:289–291. <https://doi.org/10.1093/nar/28.1.289>
 63. Szklarczyk D, Santos A, von Mering C, Jensen LJ, Bork P, Kuhn M (2016) STITCH 5: augmenting protein-chemical interaction networks with tissue and affinity data. *Nucleic Acids Res* 44:D380–384. <https://doi.org/10.1093/nar/gkv1277>
 64. Huang HY, Lin YC, Li J, Huang KY, Shrestha S, Hong HC, Tang Y, Chen YG, Jin CN, Yu Y, Xu JT, Li YM, Cai XX, Zhou ZY, Chen XH, Pei YY, Hu L, Su JJ, Cui SD, Wang F, Xie YY, Ding SY, Luo MF, Chou CH, Chang NW, Chen KW, Cheng YH, Wan XH, Hsu WL, Lee TY, Wei FX, Huang HD (2020) miRTarBase 2020: updates to the experimentally validated microRNA-target interaction database. *Nucleic Acids Res* 48:D148–D154. <https://doi.org/10.1093/nar/gkz896>
 65. Karagkouni D, Paraskevopoulou MD, Chatzopoulos S, Vlachos IS, Tastsoglou S, Kanellou I, Papadimitriou D, Kavakiotis I, Maniou S, Skoufos G, Vergoulis T, Dalamagas T, Hatzigeorgiou AG (2018) DIANA-TarBase v8: a decade-long collection of experimentally supported miRNA–gene interactions. *Nucleic Acids Res* 46:D239–D245. <https://doi.org/10.1093/nar/gkx1141>
 66. Carbon S, Douglass E, Good BM, Unni DR, Harris NL, Mungall CJ, Basu S, Chisholm RL, Dodson RJ, Hartline E, Fey P, Thomas PD, Albou L-P, Ebert D, Kesling MJ, Mi H, Muruganujan A, Huang X, Mushayahama T, LaBonte SA, Siegele DA, Antonazzo

- G, Attrill H, Brown NH, Garapati P, Marygold SJ, Trovisco V, dos Santos G, Falls K, Tabone C, Zhou P, Goodman JL, Strelets VB, Thurmond J, Garmiri P, Ishtiaq R, Rodríguez-López M, Acencio ML, Kuiper M, Lægread A, Logie C, Lovering RC, Kramarz B, Saverimuttu SCC, Pinheiro SM, Gunn H, Su R, Thurlow KE, Chibucos M, Giglio M, Nadendla S, Munro J, Jackson R, Duesbury MJ, Del-Toro N, Meldal BHM, Paneerselvam K, Peretto L, Porras P, Orchard S, Shrivastava A, Chang H-Y, Finn RD, Mitchell AL, Rawlings ND, Richardson L, Sangrador-Vegas A, Blake JA, Christie KR, Dolan ME, Drabkin HJ, Hill DP, Ni L, Sitnikov DM, Harris MA, Oliver SG, Rutherford K, Wood V, Hayles J, Bähler J, Bolton ER, De Pons JL, Dwinell MR, Hayman GT, Kaldunski ML, Kwitek AE, Laulederkind SJF, Plasterer C, Tutaj MA, VEDI M, Wang S-J, D'Eustachio P, Matthews L, Balhoff JP, Aleksander SA, Alexander MJ, Cherry JM, Engel SR, Gondwe F, Karra K, Miyasato SR, Nash RS, Simison M, Skrzypek MS, Weng S, Wong ED, Feuermann M, Gaudet P, Morgat A, Bakker E, Berardini TZ, Reiser L, Subramaniam S, Huala E, Arighi CN, Auchincloss A, Axelsen K, Argoud-Puy G, Bateman A, Blatter M-C, Boutet E, Bowler E, Breuza L, Bridge A, Britto R, Bye-A-Jee H, Casas CC, Coudert E, Denny P, Estreicher A, Famiglietti ML, Georghiou G, Gos A, Gruaz-Gumowski N, Hatton-Ellis E, Hulo C, Ignatchenko A, Jungo F, Laiho K, Le Mercier P, Lieberherr D, Lock A, Lussi Y, MacDougall A, Magrane M, Martin MJ, Masson P, Natale DA, Hyka-Nouspikel N, Orchard S, Pedruzzi I, Pourcel L, Poux S, Pundir S, Rivoire C, Speretta E, Sundaram S, Tyagi N, Warner K, Zaru R, Wu CH, Diehl AD, Chan JN, Grove C, Lee RYN, Muller H-M, Raciti D, Van Auken K, Sternberg PW, Berriman M, Paulini M, Howe K, Gao S, Wright A, Stein L, Howe DG, Toro S, Westerfield M, Jaiswal P, Cooper L, Elser J (2021) The Gene Ontology resource: enriching a GOLD mine. *Nucleic Acids Res* 49:D325–D334. <https://doi.org/10.1093/nar/gkaa1113>
67. Huang DW, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 4:44–57. <https://doi.org/10.1038/nprot.2008.211>
68. Amberger JS, Bocchini CA, Scott AF, Hamosh A (2019) OMIM.org: leveraging knowledge across phenotype-gene relationships. *Nucleic Acids Res* 47:D1038–D1043. <https://doi.org/10.1093/nar/gky1151>
69. Alcock BP, Raphenya AR, Lau TTY, Tsang KK, Bouchard M, Edalatmand A, Huynh W, Nguyen AV, Cheng AA, Liu S, Min SY, Miroshnichenko A, Tran HK, Werfalli RE, Nasir JA, Oloni M, Speicher DJ, Florescu A, Singh B, Faltyn M, Hernandez-Koutoucheva A, Sharma AN, Bordeleau E, Pawlowski AC, Zubyk HL, Dooley D, Griffiths E, Maguire F, Winsor GL, Beiko RG, Brinkman FSL, Hsiao WWL, Domselaar GV, McArthur AG (2020) CARD 2020: antibiotic resistance surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res* 48:D517–D525. <https://doi.org/10.1093/nar/gkz935>
70. Schriml LM, Mittra E, Munro J, Tauber B, Schor M, Nickle L, Felix V, Jeng L, Bearer C, Lichenstein R, Bisordi K, Campion N, Hyman B, Kurland D, Oates CP, Kibbey S, Sreekumar P, Le C, Giglio M, Greene C (2019) Human Disease Ontology 2018 update: classification, content and workflow expansion. *Nucleic Acids Res* 47:D955–D962. <https://doi.org/10.1093/nar/gky1032>
71. Wang Y, Zhang S, Li F, Zhou Y, Zhang Y, Wang Z, Zhang R, Zhu J, Ren Y, Tan Y, Qin C, Li Y, Li X, Chen Y, Zhu F (2020) Therapeutic target database 2020: enriched resource for facilitating research and early development of targeted therapeutics. *Nucleic Acids Res* 48:D1031–D1041. <https://doi.org/10.1093/nar/gkz981>
72. Safran M, Dalah I, Alexander J, Rosen N, Iny Stein T, Shmoish M, Nativ N, Bahir I, Doniger T, Krug H, Sirota-Madi A, Olender T, Golan Y, Stelzer G, Harel A, Lancet D (2010) GeneCards Version 3: the human gene integrator. Database (Oxford). <https://doi.org/10.1093/database/baq020>
73. Ladunga I (2002) Finding homologs to nucleotide sequences using network BLAST searches. *Curr Protoc Bioinform Chapter 3:Unit 3.3*. <https://doi.org/10.1002/0471250953.bi0303s00>
74. Hu G, Kurgan L (2019) Sequence similarity searching. *Curr Protoc Protein Sci* 95:e71. <https://doi.org/10.1002/cpps.71>
75. Manikyam HK, Joshi SK (2020) Whole genome analysis and targeted drug discovery using computational methods and high throughput screening tools for emerged novel coronavirus (2019-nCoV). *J Pharm Drug Res* 3:341–361
76. Marchler-Bauer A, Bryant SH (2004) CD-Search: protein domain annotations on the fly. *Nucleic Acids Res* 32:W327–W331. <https://doi.org/10.1093/nar/gkh454>
77. Geer LY, Domrachev M, Lipman DJ, Bryant SH (2002) CDART: protein homology by domain architecture. *Genome Res* 12:1619–1623. <https://doi.org/10.1101/gr.278202>
78. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. *BMC Bioinform* 10:421. <https://doi.org/10.1186/1471-2105-10-421>
79. Du Z, Wu Q, Wang T, Chen D, Huang X, Yang W, Luo W (2020) BlastGUI: a python-based cross-platform local BLAST visualization software. *Mol Inform* 39:e1900120. <https://doi.org/10.1002/minf.201900120>
80. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* 23:2947–2948. <https://doi.org/10.1093/bioinformatics/btm404>
81. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Soding J, Thompson JD, Higgins DG (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* 7:539. <https://doi.org/10.1038/msb.2011.75>
82. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32:1792–1797. <https://doi.org/10.1093/nar/gkh340>
83. Troshin PV, Procter JB, Barton GJ (2011) Java bioinformatics analysis web services for multiple sequence alignment—JABAWS:MSA. *Bioinformatics* 27:2001–2002. <https://doi.org/10.1093/bioinformatics/btr304>
84. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M (2007) KAAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res* 35:W182–185. <https://doi.org/10.1093/nar/gkm321>
85. Huang Y, Niu B, Gao Y, Fu L, Li W (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* 26:680–682. <https://doi.org/10.1093/bioinformatics/btq003>
86. Brittnacher MJ, Fong C, Hayden HS, Jacobs MA, Radey M, Rohmer L (2011) PGAT: a multistrain analysis resource for microbial genomes. *Bioinformatics* 27:2429–2430. <https://doi.org/10.1093/bioinformatics/btr418>
87. Zomer A, Burghout P, Bootsma H, Hermans P, van Hijum S (2012) ESSENTIALS: software for rapid analysis of high throughput transposon insertion sequencing data. *PLoS ONE* 7:e43012. <https://doi.org/10.1371/journal.pone.0043012>
88. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 13:2498–2504. <https://doi.org/10.1101/gr.1239303>


89. Bauer-Mehren A (2013) Integration of genomic information with biological networks using Cytoscape. *Methods Mol Biol* 1021:37–61. https://doi.org/10.1007/978-1-62703-450-0_3
90. Doncheva NT, Morris JH, Gorodkin J, Jensen LJ (2019) Cytoscape stringapp: network analysis and visualization of proteomics data. *J Proteome Res* 18:623–632. <https://doi.org/10.1021/acs.jproteome.8b00702>
91. Maere S, Heymans K, Kuiper M (2005) BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* 21:3448–3449. <https://doi.org/10.1093/bioinformatics/bti551>
92. Bader GD, Hogue CW (2003) An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinform* 4:2. <https://doi.org/10.1186/1471-2105-4-2>
93. Chin CH, Chen SH, Wu HH, Ho CW, Ko MT, Lin CY (2014) cytoHubba: identifying hub objects and sub-networks from complex interactome. *BMC Syst Biol* 8(Suppl 4):S11. <https://doi.org/10.1186/1752-0509-8-S4-S11>
94. Bastian M, Heymann S, Jacomy M (2009) Gephi: an open source software for exploring and manipulating networks. *ICWSM* 8:361–362
95. Zhou G, Soufan O, Ewald J, Hancock REW, Basu N, Xia J (2019) NetworkAnalyst 3.0: a visual analytics platform for comprehensive gene expression profiling and meta-analysis. *Nucleic Acids Res* 47:W234–W241. <https://doi.org/10.1093/nar/gkz240>
96. Schaefer MH, Fontaine JF, Vinayagam A, Porras P, Wanker EE, Andrade-Navarro MA (2012) HIPPIE: integrating protein interaction networks with experiment based quality scores. *PLoS ONE* 7:e31826. <https://doi.org/10.1371/journal.pone.0031826>
97. Farkas IJ, Szanto-Varnagy A, Korcsmaros T (2012) Linking proteins to signaling pathways for experiment design and evaluation. *PLoS ONE* 7:e36202. <https://doi.org/10.1371/journal.pone.0036202>
98. Xie C, Mao X, Huang J, Ding Y, Wu J, Dong S, Kong L, Gao G, Li CY, Wei L (2011) KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. *Nucleic Acids Res* 39:W316–322. <https://doi.org/10.1093/nar/gkr483>
99. Karp PD, Billington R, Caspi R, Fulcher CA, Latendresse M, Kothari A, Keseler IM, Krummenacker M, Midford PE, Ong Q, Ong WK, Paley SM, Subhraveti P (2019) The BioCyc collection of microbial genomes and metabolic pathways. *Brief Bioinform* 20:1085–1093. <https://doi.org/10.1093/bib/bbx085>
100. Adamcsek B, Palla G, Farkas IJ, Derenyi I, Vicsek T (2006) CFinder: locating cliques and overlapping modules in biological networks. *Bioinformatics* 22:1021–1023. <https://doi.org/10.1093/bioinformatics/btl039>
101. Mrvar A, Batagelj V (2016) Analysis and visualization of large networks with program package Pajek. *Complex Adapt Syst Model* 4:6. <https://doi.org/10.1186/s40294-016-0017-8>
102. Muller J, Hemphill A (2016) Drug target identification in protozoan parasites. *Expert Opin Drug Discov* 11:815–824. <https://doi.org/10.1080/17460441.2016.1195945>
103. Nayak S, Pradhan D, Singh H, Reddy MS (2019) Computational screening of potential drug targets for pathogens causing bacterial pneumonia. *Microb Pathog* 130:271–282. <https://doi.org/10.1016/j.micpath.2019.03.024>
104. Melak T, Gakkhar S (2015) Comparative genome and network centrality analysis to identify drug targets of *Mycobacterium tuberculosis* H37Rv. *Biomed Res Int* 2015:212061. <https://doi.org/10.1155/2015/212061>
105. Doyle MA, Gasser RB, Woodcroft BJ, Hall RS, Ralph SA (2010) Drug target prediction and prioritization: using orthology to predict essentiality in parasite genomes. *BMC Genom* 11:222. <https://doi.org/10.1186/1471-2164-11-222>
106. Zumla A, Chan JF, Azhar EI, Hui DS, Yuen KY (2016) Coronavirus—drug discovery and therapeutic options. *Nat Rev Drug Discov* 15:327–347. <https://doi.org/10.1038/nrd.2015.37>
107. Naqvi AAT, Fatima K, Mohammad T, Fatima U, Singh IK, Singh A, Atif SM, Hariprasad G, Hasan GM, Hassan MI (2020) Insights into SARS-CoV-2 genome, structure, evolution, pathogenesis and therapies: structural genomics approach. *Biochim Biophys Acta Mol Basis Dis* 1866:165878. <https://doi.org/10.1016/j.bbadis.2020.165878>
108. Damte D, Suh JW, Lee SJ, Yohannes SB, Hossain MA, Park SC (2013) Putative drug and vaccine target protein identification using comparative genomic analysis of KEGG annotated metabolic pathways of *Mycoplasma hypopneumoniae*. *Genomics* 102:47–56. <https://doi.org/10.1016/j.ygeno.2013.04.011>
109. Cai J, Han C, Hu T, Zhang J, Wu D, Wang F, Liu Y, Ding J, Chen K, Yue J, Shen X, Jiang H (2006) Peptide deformylase is a potential target for anti-*Helicobacter pylori* drugs: reverse docking, enzymatic assay, and X-ray crystallography validation. *Protein Sci* 15:2071–2081. <https://doi.org/10.1110/ps.062238406>
110. Kumar S, Chaudhary K, Foster JM, Novelli JF, Zhang Y, Wang S, Spiro D, Ghedin E, Carlow CK (2007) Mining predicted essential genes of *Brugia malayi* for nematode drug targets. *PLoS ONE* 2:e1189. <https://doi.org/10.1371/journal.pone.0001189>
111. Darapaneni V, Prabhaker VK, Kukol A (2009) Large-scale analysis of influenza A virus sequences reveals potential drug target sites of non-structural proteins. *J Gen Virol* 90:2124–2133. <https://doi.org/10.1099/vir.0.011270-0>
112. Muhammad SA, Ahmed S, Ali A, Huang H, Wu X, Yang XF, Naz A, Chen J (2014) Prioritizing drug targets in *Clostridium botulinum* with a computational systems biology approach. *Genomics* 104:24–35. <https://doi.org/10.1016/j.ygeno.2014.05.002>
113. Birhanu BT, Lee SJ, Park NH, Song JB, Park SC (2018) In silico analysis of putative drug and vaccine targets of the metabolic pathways of *Actinobacillus pleuropneumoniae* using a subtractive/comparative genomics approach. *J Vet Sci* 19:188–199. <https://doi.org/10.4142/jvs.2018.19.2.188>
114. Swain A, Gnanasekar P, Prava J, Rajeev AC, Kesarwani P, Lahiri C, Pan A (2021) A comparative genomics approach for shortlisting broad-spectrum drug targets in nontuberculous mycobacteria. *Microb Drug Resist* 27:212–226. <https://doi.org/10.1089/mdr.2020.0161>
115. Bhattacharyya M, Chakrabarti S (2015) Identification of important interacting proteins (IIPs) in *Plasmodium falciparum* using large-scale interaction network analysis and in-silico knock-out studies. *Malar J* 14:70. <https://doi.org/10.1186/s12936-015-0562-1>
116. Sabetian S, Shamsir MS (2015) Identification of putative drug targets for human sperm-egg interaction defect using protein network approach. *BMC Syst Biol* 9:37. <https://doi.org/10.1186/s12918-015-0186-7>
117. Kumar A, Sharma D, Aggarwal ML, Chacko KM, Bhatt TK (2016) Cancer/testis antigens as molecular drug targets using network pharmacology. *Tumour Biol* 37:15697–15705. <https://doi.org/10.1007/s13277-016-5333-2>
118. Rai S, Bhatnagar S (2016) Hyperlipidemia, disease associations, and top 10 potential drug targets: a network view. *OMICS* 20:152–168. <https://doi.org/10.1089/omi.2015.0172>
119. Huang H, He Y, Li W, Wei W, Li Y, Xie R, Guo S, Wang Y, Jiang J, Chen B, Lv J, Zhang N, Chen L, He W (2016) Identification of polycystic ovary syndrome potential drug targets based on pathobiological similarity in the protein-protein interaction network. *Oncotarget* 7:37906–37919. <https://doi.org/10.18632/oncotarget.9353>
120. Li CW, Su MH, Chen BS (2017) Investigation of the cross-talk mechanism in caco-2 cells during clostridium difficile infection

- through genetic-and-epigenetic interspecies networks: big data mining and genome-wide identification. *Front Immunol* 8:901. <https://doi.org/10.3389/fimmu.2017.00901>
121. Gupta MK, Behera SK, Dehury B, Mahapatra N (2017) Identification and characterization of differentially expressed genes from human microglial cell samples infected with Japanese encephalitis virus. *J Vector Borne Dis* 54:131–138
 122. Vitali F, Marini S, Balli M, Grosemans H, Sampaolesi M, Lusier YA, Cusella De Angelis MG, Bellazzi R (2017) Exploring wound-healing genomic machinery with a network-based approach. *Pharmaceuticals (Basel)* 10:55. <https://doi.org/10.3390/ph10020055>
 123. Liu W, Wang S, Zhou S, Yang F, Jiang W, Zhang Q, Wang L (2017) A systems biology approach to identify microRNAs contributing to cisplatin resistance in human ovarian cancer cells. *Mol Biosyst* 13:2268–2276. <https://doi.org/10.1039/c7mb00362e>
 124. Panga V, Raghunathan S (2018) A cytokine protein-protein interaction network for identifying key molecules in rheumatoid arthritis. *PLoS ONE* 13:e0199530. <https://doi.org/10.1371/journal.pone.0199530>
 125. Moon SJ, Bae JM, Park KS, Tagkopoulos I, Kim KJ (2019) Compendium of skin molecular signatures identifies key pathological features associated with fibrosis in systemic sclerosis. *Ann Rheum Dis* 78:817–825. <https://doi.org/10.1136/annrheumdis-2018-214778>
 126. Rahman MR, Islam T, Turanli B, Zaman T, Faruquee HM, Rahman MM, Mollah MNH, Nanda RK, Arga KY, Gov E, Moni MA (2019) Network-based approach to identify molecular signatures and therapeutic agents in Alzheimer's disease. *Comput Biol Chem* 78:431–439. <https://doi.org/10.1016/j.compbiolchem.2018.12.011>
 127. Tan MF, Zou G, Wei Y, Liu WQ, Li HQ, Hu Q, Zhang LS, Zhou R (2021) Protein-protein interaction network and potential drug target candidates of *Streptococcus suis*. *J Appl Microbiol* 131:658–670. <https://doi.org/10.1111/jam.14950>
 128. Nadeau R, Shahryari Fard S, Scheer A, Hashimoto-Roth E, Nygard D, Abramchuk I, Chung YE, Bennett SAL, Lavallee-Adam M (2020) Computational identification of human biological processes and protein sequence motifs putatively targeted by SARS-CoV-2 proteins using protein-protein interaction networks. *J Proteome Res* 19:4553–4566. <https://doi.org/10.1021/acs.jproteome.0c00422>
 129. Martins-de-Souza D, Guest PC, Reis-de-Oliveira G, Schmitt A, Falkai P, Turck CW (2021) An overview of the human brain myelin proteome and differences associated with schizophrenia. *World J Biol Psychiatry* 22:271–287. <https://doi.org/10.1080/15622975.2020.1789217>
 130. Farooq QUA, Shaikat Z, Aiman S, Zhou T, Li C (2020) A systems biology-driven approach to construct a comprehensive protein interaction network of influenza A virus with its host. *BMC Infect Dis* 20:480. <https://doi.org/10.1186/s12879-020-05214-0>
 131. Yan W, Liu X, Wang Y, Han S, Wang F, Liu X, Xiao F, Hu G (2020) Identifying drug targets in pancreatic ductal adenocarcinoma through machine learning, analyzing biomolecular networks, and structural modeling. *Front Pharmacol* 11:534. <https://doi.org/10.3389/fphar.2020.00534>
 132. Huang S, Zhang Z, Li W, Kong F, Yi P, Huang J, Mao D, Peng W, Zhang S (2020) Network pharmacology-based prediction and verification of the active ingredients and potential targets of zuojinwan for treating colorectal cancer. *Drug Des Devel Ther* 14:2725–2740. <https://doi.org/10.2147/DDDT.S250991>
 133. Fathima S, Sinha S, Donakonda S (2021) Network analysis identifies drug targets and small molecules to modulate apoptosis resistant cancers. *Cancers (Basel)* 13:851. <https://doi.org/10.3390/cancers13040851>
 134. Wu M, Zhao Y, Peng N, Tao Z, Chen B (2021) Identification of chemoresistance-associated microRNAs and hub genes in breast cancer using bioinformatics analysis. *Invest New Drugs* 39:705–712. <https://doi.org/10.1007/s10637-020-01059-1>
 135. Lin TY, Chen WJ, Yang P, Li ZQ, Wei QS, Liang D, Wang HB, He W, Zhang QW (2021) Bioinformatics analysis and identification of genes and molecular pathways in steroid-induced osteonecrosis of the femoral head. *J Orthop Surg Res* 16:327. <https://doi.org/10.1186/s13018-021-02464-9>
 136. Zeng Y, Li N, Zheng Z, Chen R, Peng M, Liu W, Zhu J, Zeng M, Cheng J, Hong C (2021) Screening of hub genes associated with pulmonary arterial hypertension by integrated bioinformatic analysis. *Biomed Res Int*. <https://doi.org/10.1155/2021/6626094>
 137. Yadav M, Pradhan D, Singh R (2021) Integrated analysis and identification of nine-gene signature associated to oral squamous cell carcinoma pathogenesis. *3 Biotech* 11:215. <https://doi.org/10.1007/s13205-021-02737-4>
 138. Miryala SK, Anbarasu A, Ramaiah S (2021) Gene interaction network approach to elucidate the multidrug resistance mechanisms in the pathogenic bacterial strain *Proteus mirabilis*. *J Cell Physiol* 236:468–479. <https://doi.org/10.1002/jcp.29874>
 139. Naha A, Kumar Miryala S, Debroy R, Ramaiah S, Anbarasu A (2020) Elucidating the multi-drug resistance mechanism of *Enterococcus faecalis* V583: a gene interaction network analysis. *Gene* 748:144704. <https://doi.org/10.1016/j.gene.2020.144704>
 140. Debroy R, Miryala SK, Naha A, Anbarasu A, Ramaiah S (2020) Gene interaction network studies to decipher the multi-drug resistance mechanism in *Salmonella enterica* serovar Typhi CT18 reveal potential drug targets. *Microb Pathog* 142:104096. <https://doi.org/10.1016/j.micpath.2020.104096>
 141. Miryala SK, Anbarasu A, Ramaiah S (2020) Role of SHV-11, a class A β -Lactamase, gene in multidrug resistance among *Klebsiella pneumoniae* strains and understanding its mechanism by gene network analysis. *Microb Drug Resist* 26:900–908. <https://doi.org/10.1089/mdr.2019.0430>
 142. Liu B, Pop M (2009) ARDB—antibiotic resistance genes database. *Nucleic Acids Res* 37:D443–447. <https://doi.org/10.1093/nar/gkn656>
 143. Tomczak K, Czerwińska P, Wiznerowicz M (2015) The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp Oncol (Pozn)* 19:A68–77. <https://doi.org/10.5114/wo.2014.47136>
 144. Xie B, Ding Q, Han H, Wu D (2013) miRCancer: a microRNA-cancer association database constructed by text mining on literature. *Bioinformatics* 29:638–644. <https://doi.org/10.1093/bioinformatics/btt014>
 145. Jiang Q, Wang Y, Hao Y, Juan L, Teng M, Zhang X, Li M, Wang G, Liu Y (2009) miR2Disease: a manually curated database for microRNA deregulation in human disease. *Nucleic Acids Res* 37:D98–104. <https://doi.org/10.1093/nar/gkn714>
 146. Li Y, Qiu C, Tu J, Geng B, Yang J, Jiang T, Cui Q (2014) HMDD v2.0: a database for experimentally supported human microRNA and disease associations. *Nucleic Acids Res* 42:D1070–1074. <https://doi.org/10.1093/nar/gkt1023>
 147. Ponten F, Jirstrom K, Uhlen M (2008) The Human Protein Atlas—a tool for pathology. *J Pathol* 216:387–393. <https://doi.org/10.1002/path.2440>
 148. Gupta R, Verma R, Pradhan D, Jain AK, Umamaheswari A, Rai CS (2019) An in silico approach towards identification of novel drug targets in pathogenic species of *Leptospira*. *PLoS ONE* 14:e0221446. <https://doi.org/10.1371/journal.pone.0221446>
 149. Ye Y, Hua Z, Huang J, Rao N, Guo F (2013) CEG: a database of essential gene clusters. *BMC Genom* 14:769. <https://doi.org/10.1186/1471-2164-14-769>
 150. Sarangi AN, Lohani M, Aggarwal R (2015) Proteome mining for drug target identification in *Listeria monocytogenes* strain EGD-e

- and structure-based virtual screening of a candidate drug target penicillin binding protein 4. *J Microbiol Methods* 111:9–18. <https://doi.org/10.1016/j.mimet.2015.01.011>
151. Melak T, Gakkhar S (2015) Maximum flow approach to prioritize potential drug targets of *Mycobacterium tuberculosis* H37Rv from protein-protein interaction network. *Clin Transl Med* 4:61. <https://doi.org/10.1186/s40169-015-0061-6>
 152. Lohani M, Dhasmana A, Haque S, Wahid M, Jawed A, Dar SA, Mandal RK, Areeshi MY, Khan S (2017) Proteome mining for the identification and in-silico characterization of putative drug targets of multi-drug resistant *Clostridium difficile* strain 630. *J Microbiol Methods* 136:6–10. <https://doi.org/10.1016/j.mimet.2017.02.008>
 153. Agamah FE, Mazandu GK, Hassan R, Bope CD, Thomford NE, Ghansah A, Chimusa ER (2020) Computational/in silico methods in drug target and lead prediction. *Brief Bioinform* 21:1663–1675. <https://doi.org/10.1093/bib/bbz103>
 154. Raman K, Yeturu K, Chandra N (2008) targetTB: a target identification pipeline for *Mycobacterium tuberculosis* through an interactome, reactome and genome-scale structural analysis. *BMC Syst Biol* 2:109. <https://doi.org/10.1186/1752-0509-2-109>
 155. Wu Z, Li W, Liu G, Tang Y (2018) Network-based methods for prediction of drug-target interactions. *Front Pharmacol* 9:1134. <https://doi.org/10.3389/fphar.2018.01134>
 156. Wong YH, Lin CL, Chen TS, Chen CA, Jiang PS, Lai YH, Chu L, Li CW, Chen JJ, Chen BS (2015) Multiple target drug cocktail design for attacking the core network markers of four cancers using ligand-based and structure-based virtual screening methods. *BMC Med Genom* 8(Suppl 4):S4. <https://doi.org/10.1186/1755-8794-8-s4-s4>
 157. Coates AR, Hu Y (2007) Novel approaches to developing new antibiotics for bacterial infections. *Br J Pharmacol* 152:1147–1154. <https://doi.org/10.1038/sj.bjp.0707432>
 158. Chung BK, Dick T, Lee DY (2013) In silico analyses for the discovery of tuberculosis drug targets. *J Antimicrob Chemother* 68:2701–2709. <https://doi.org/10.1093/jac/dkt273>
 159. Bloomingdale P, Nguyen VA, Niu J, Mager DE (2018) Boolean network modeling in systems pharmacology. *J Pharmacokinet Pharmacodyn* 45:159–180. <https://doi.org/10.1007/s10928-017-9567-4>
 160. Potapov AP, Goemann B, Wingender E (2008) The pairwise disconnectivity index as a new metric for the topological analysis of regulatory networks. *BMC Bioinform* 9:227. <https://doi.org/10.1186/1471-2105-9-227>
 161. Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D (1999) Detecting protein function and protein-protein interactions from genome sequences. *Science* 285:751–753. <https://doi.org/10.1126/science.285.5428.751>
 162. Saha S, Sengupta K, Chatterjee P, Basu S, Nasipuri M (2018) Analysis of protein targets in pathogen-host interaction in infectious diseases: a case study on *Plasmodium falciparum* and *Homo sapiens* interaction network. *Brief Funct Genomics* 17:441–450. <https://doi.org/10.1093/bfpg/elx024>
 163. Reisdorf WC, Chhugani N, Sanseau P, Agarwal P (2017) Harnessing public domain data to discover and validate therapeutic targets. *Expert Opin Drug Discov* 12:687–693. <https://doi.org/10.1080/17460441.2017.1329296>
 164. Barh D, Tiwari S, Jain N, Ali A, Santos AR, Misra AN, Azevedo V, Kumar A (2011) In silico subtractive genomics for target identification in human bacterial pathogens. *Drug Dev Res* 72:162–177. <https://doi.org/10.1002/ddr.20413>
 165. Penrod NM, Moore JH (2014) Influence networks based on co-expression improve drug target discovery for the development of novel cancer therapeutics. *BMC Syst Biol* 8:12. <https://doi.org/10.1186/1752-0509-8-12>
 166. Vinayagam A, Gibson TE, Lee HJ, Yilmazel B, Roesel C, Hu Y, Kwon Y, Sharma A, Liu YY, Perrimon N, Barabasi AL (2016) Controllability analysis of the directed human protein interaction network identifies disease genes and drug targets. *Proc Natl Acad Sci U S A* 113:4976–4981. <https://doi.org/10.1073/pnas.1603992113>
 167. Zong N, Kim H, Ngo V, Harismendy O (2017) Deep mining heterogeneous networks of biomedical linked data to predict novel drug-target associations. *Bioinformatics* 33:2337–2344. <https://doi.org/10.1093/bioinformatics/btx160>
 168. Peng C, Lin Y, Luo H, Gao F (2017) A comprehensive overview of online resources to identify and predict bacterial essential genes. *Front Microbiol* 8:2331. <https://doi.org/10.3389/fmicb.2017.02331>
 169. Li K, Du Y, Li L, Wei DQ (2020) Bioinformatics approaches for anti-cancer drug discovery. *Curr Drug Targets* 21:3–17. <https://doi.org/10.2174/1389450120666190923162203>
 170. Chen H, Engkvist O, Wang Y, Olivecrona M, Blaschke T (2018) The rise of deep learning in drug discovery. *Drug Discov Today* 23:1241–1250. <https://doi.org/10.1016/j.drudis.2018.01.039>
 171. Gao D, Chen Q, Zeng Y, Jiang M, Zhang Y (2020) Applications of machine learning in drug target discovery. *Curr Drug Metab* 21:790–803. <https://doi.org/10.2174/1567201817999200728142023>
 172. Mayr A, Klambauer G, Unterthiner T, Steijaert M, Wegner JK, Ceulemans H, Clevert DA, Hochreiter S (2018) Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. *Chem Sci* 9:5441–5451. <https://doi.org/10.1039/c8sc00148k>
 173. Lee H, Grosse R, Ranganath R, Ng AY (2011) Unsupervised learning of hierarchical representations with convolutional deep belief networks. *Commun ACM* 54:95–103. <https://doi.org/10.1145/2001269.2001295>
 174. Bengio Y (2009) Learning deep architectures for AI. *Found Trends Mach Learn* 2:1–127. <https://doi.org/10.1561/2200000006>
 175. Sturm N, Mayr A, Le Van T, Chupakhin V, Ceulemans H, Wegner J, Golib-Dzib JF, Jeliakova N, Vandriessche Y, Böhm S, Cima V, Martinovic J, Greene N, Vander Aa T, Ashby TJ, Hochreiter S, Engkvist O, Klambauer G, Chen H (2020) Industry-scale application and evaluation of deep learning for drug target prediction. *J Cheminform* 12:26. <https://doi.org/10.1186/s13321-020-00428-5>
 176. Ferrero E, Dunham I, Sanseau P (2017) In silico prediction of novel therapeutic targets using gene-disease association data. *J Transl Med* 15:182. <https://doi.org/10.1186/s12967-017-1285-6>
 177. Gao D, Morini E, Salani M, Krauson AJ, Chekuri A, Sharma N, Ragavendran A, Erdin S, Logan EM, Li W, Dakka A, Narasimhan J, Zhao X, Naryshkin N, Trotta CR, Effenberger KA, Woll MG, Gabbeta V, Karp G, Yu Y, Johnson G, Paquette WD, Cutting GR, Talkowski ME, Slaugenhaupt SA (2021) A deep learning approach to identify gene targets of a therapeutic for human splicing disorders. *Nat Commun* 12:3332. <https://doi.org/10.1038/s41467-021-23663-2>
 178. Wang Q, Feng Y, Huang J, Wang T, Cheng G (2017) A novel framework for the identification of drug target proteins: combining stacked auto-encoders with a biased support vector machine. *PLoS ONE* 12:e0176486. <https://doi.org/10.1371/journal.pone.0176486>
 179. Zeng X, Zhu S, Lu W, Liu Z, Huang J, Zhou Y, Fang J, Huang Y, Guo H, Li L, Trapp BD, Nussinov R, Eng C, Loscalzo J, Cheng F (2020) Target identification among known drugs by deep learning from heterogeneous networks. *Chem Sci* 11:1775–1797. <https://doi.org/10.1039/c9sc04336e>
 180. Lee I, Keum J, Nam H (2019) DeepConv-DTI: Prediction of drug-target interactions via deep learning with convolution on

- protein sequences. *PLoS Comput Biol* 15:e1007129. <https://doi.org/10.1371/journal.pcbi.1007129>
181. Mamoshina P, Vieira A, Putin E, Zhavoronkov A (2016) Applications of deep learning in biomedicine. *Mol Pharm* 13:1445–1454. <https://doi.org/10.1021/acs.molpharmaceut.5b00982>
182. Hu Y, Zhao T, Zhang N, Zhang Y, Cheng L (2019) A Review of recent advances and research on drug target identification methods. *Curr Drug Metab* 20:209–216. <https://doi.org/10.2174/1389200219666180925091851>

Authors and Affiliations

Xuting Zhang¹ · Fengxu Wu² · Nan Yang¹ · Xiaohui Zhan³ · Jianbo Liao³ · Shang kang Mai³ · Zunnan Huang^{1,4} 

¹ Key Laboratory of Big Data Mining and Precision Drug Design of Guangdong Medical University, Key Laboratory for Research and Development of Natural Drugs of Guangdong Province, School of Pharmacy, Guangdong Medical University, No. 1 Xincheng Road, Songshan Lake District, Dongguan 523808, China

² Hubei Key Laboratory of Wudang Local Chinese Medicine Research, School of Pharmaceutical Sciences, Hubei University of Medicine, Shiyan 442000, China

³ The Second School of Clinical Medicine, Guangdong Medical University, Dongguan 523808, China

⁴ Southern Marine Science and Engineering Guangdong Laboratory (Zhanjiang), Zhanjiang 524023, China