A Review of Target Identification Strategies for Drug Discovery: from Database to Machine-Based Methods

Zehua Shangguan1, *
1Faculty of Science, York University, Toronto, Ontario, M3J 0L1
thomash@my.yorku.ca

Abstract. In recent years, target identification has become more efficient than before, and it helped to discover huge amounts of drugs for various diseases. The mystery buried behind was the methods that developed in recent years utilized in the target identification. The advances and research status of database, biological assay and machine-based method in recent years for target identification would be integrated by this review. The various databases help scientists to find information about target property, chemical property or on genome level. The biological assay, such as RNAi, RNA sequencing, DNA microarray, and Gal4/UAS system, is commonly used to identify the target in recent years. The machine-based strategies, such as random forest algorithm and Support Vector Machine (SVM) algorithm, could help scientists identify the target and find compound activity more efficiently. Among the three methods mentioned above, the machine-based methods could have higher efficiency and lower cost while maintaining higher accuracy. Despite the promising properties of machine-based methods, the combination use of biological assay would still be necessary. With wider application of more efficient strategies, target identification, as well as drug discovery process, would gain more rapid development.

1. Introduction
The target identification step is one of the most important steps in the drug development process.[1] Suitable strategies to identify the target would help to reduce the time and efforts that scientists have to pay in this stage. However, the efficiency of finding targets would always be slow and inefficient. Through 2006, the targets that the scientists found, which would desirable for drug development, were only 324.[2] Currently, the methods to do the target identification have been expanded enormously and the amounts of desirable drug targets have increased to a thousand.[3] In this article, the advanced methods that developed recently will be summarized in categories of target database, biological assay and machine-based methods. These methods are important for the target identification of the drug, which is the first step of the drug development and can largely improve the number of the targets identified and reduce the cost during the drug development process. There are two major strategies for drug development; one is the deconvolution method and the other is the bottom-up strategy. The bottom-up development strategy is a traditional way, which would be through the process of screening thousands of small organic molecules in mammalian cells after the target was identified [4]. After the screening, the phenotypic change of the mammalian cells and the effect of those organic molecules brought to the target would be recorded. The target deconvolution strategy, on the other hand, would begin with a drug compound and then identified the targets of it [5]. The database, biological assays, and machine-based methods are thus extremely important for target identification both in the two major development methods. The database could help researchers find their desired drug target and understand the property of small organic molecules. The biological assays could test the effects the
small molecules possess on the target. The machine-based method could predict the relationship between drugs and targets, which makes a lower cost than traditional experimental assays [6]. In this scenario, the current review summarized the advances and research status of the applied databases, biological assays and machine-based methods for target identification, in the hope of providing certain help to researchers in understanding these methods more quickly, especially for using the computational methods for target identification as well as drug development.

2. Drug target database development in recent years

There are huge amounts of database constructed for the drug target development, and these databases would help the scientists find their desired drug targets easier.

The therapeutic target database was first released in 2019. It provides a direct pathway to find the information about target function, 3D structure and ligands binding sequence. The target information provided is large, as it includes 2954 human targets and 465 infectious species targets [7]. It is helpful for the researchers to find the information they needed from this database. Another database that collected large amounts of protein, drug, target sequence is called DrugBank. DrugBank was first released in 2006, and it had over 80000 entries, which consisted of 77529 proteins, 2420 nucleic acids and 3795 DNA complex [8]. The Drugbank contains information about the drug target and drug action. In addition, Drugbank database also includes information about drug-food effects and drug-drug effects. PharmGKB, a database including the genotype, molecular and clinical data, would help the researchers to study how the genetic variation in patients affected the efficacy of some specific drugs [9]. PharmGKB focused on the genetic variation effect on the drug response. It was extremely important for the researchers to understand the drug effects on different races. BindingDB is a database to study the binding affinities, and it shows the interaction between the drug molecules and target proteins [10]. The BindingDB database involves around 20000 protein-ligand complex with determined binding affinity, and involves information such as protein sequence, ligand and protein names [11]. It is useful to understand the drug-protein interaction by searching in the BindingDB database. The Comparative Toxic Genomic Database was released in 2004 and includes the data about the relationship between the gene and proteins, chemicals and drugs. The comparative toxic genomic database involves 45 million toxic-genomic relationships for 51300 genes and 16300 chemicals [12]. The toxin and toxin target database shows 3673 toxins with 41733 synonyms [13]. This toxin and toxin target database is useful when studying the relationship between the toxin and symptoms brought by the toxin [14]. The canSAR database is a database showing over one million bioactive drug molecules and their biochemical activity [15]. The chemical property information and biochemical activity involved in the canSAR database could help the researchers to do the target validation in the drug development process. The canSAR database focus more on the cancer drug discovery compared with the other database. The ChEMBL database collected large amounts of the 2D structures of drug-like molecules [16]. Furthermore, information about the bioactivity of those drug-like molecules were also collected in this database, and the target, protein sequences could be searched by keywords in the ChEMBL database [17]. The ChEMBL database could help the researchers to understand the different molecules’ bioactivity and help to do the drug development. Those databases and their URL address were summarized in Table 1.

Those databases showed in the Table 1 gave the researchers a chance to overview the historical data they needed for drug development. Those databases involved the information of toxicity, bioactivity, binding affinity of small drug molecules and the DNA sequences, 3D and 2D structure of the target proteins.
Table 1. The summary of the database and their URL address for the research to do drug development processes

| Database                  | URL                        |
|---------------------------|----------------------------|
| Drugbank database         | www.drugbank.ca            |
| PharmGKB                  | www.pharmgkb.org           |
| BindingDB                 | https://www.bindingdb.org/bind/index.jsp |
| Comparative toxic genomic database | http://ctdbase.org          |
| toxin and Toxin target database | http://www.t3db.ca/        |
| ChEMBL database           | https://www.ebi.ac.uk/chembl/ |
| canSAR database           | http://cansar.icr.ac.uk    |
| Therapeutic target database | http://bidd.nus.edu.sg/group/cjtttd/ |

3. Biological assays used by the researchers to discover drug targets

There are several biological assays used by researchers in recent years to find new drug targets. The gene transfection strategy allows the DNA molecules to transfer from one cell to another. The RNAi method allows reverse genetic screening to find the difference between the phenotypes of targets. The enzyme-linked immunosorbent assay, which has a reaction between antigen and antibody, and the color generated by the reaction, will be detected. This method is usually used in quantitative determination. The real-time PCR assay added the fluorescent group to the PCR system, and then the researchers could observe the PCR process in real-time monitoring. The gene knockout method is widely used, as it could silence a gene function and change the phenotype of the target organism. Researchers could thus observe the effects brought by the gene knockout and predict the function of the gene.

The RNAi with Gal4/UAS system is commonly applied in the snr1 messenger RNA silencing research. Researchers recognized that the silencing of snr1 is essential for neuron development [18]. Based on the biological assay, the Phosphatidylinositol 3-kinase was studied and results showed it could affect the tumor cell growth and survival [19].

From the Food and Drug Administration data, the cost of target identification was around 1.8 billion dollars in the last 13 years [20]. That illustrated the high expensive cost of drug development. Although the target identification process cost huge amounts of money, they provided critical internet databases for future drug development.

Table 2. The description of commonly used biological assays

| Biological Technology | Description |
|-----------------------|-------------|
| Gene transfection     | Transportation of nucleic acid in different cells |
| RNAi                  | allowed the reverse genetic screening to find the difference between the phenotypes of target |
| DNA microarray        | Method used to detect the expression level of different genes by make the unknown DNA sequence hybridized to the spots of microarray. |
| RT-PCR Analysis       | monitoring PCR process in real time |
| gene knockout method  | silence a gene function and change the phenotype of the target organism, understand the function of the gene |
| MARCM technique       | A technique achieved by the flippase in order to create different genotype cells in an individual |
| RNA sequencing        | A way to measure RNA levels in cells and know the gene expression |
As Table 2 showed, there are different types of biological technology used in recent years. The RNAi, MARCM technique and gene knockout method are commonly used in the drosophila genetics with Gal4/UAS system. The Gal4/UAS system helps to do the dominant modifier and know the Pax-FOXO1 pathology in the drosophila [21]. The MARCM and gene knockout methods are used to generate different cells genotype in order to know the relationship between csk, Ras type and the tumor growth [22]. These methods have been developed in recent years and help scientists understand human diseases by making the homolog gene copies in model organism drosophila.

Also, the DNA microarray, RT-PCR and RNA sequencing are critically important in genetic analysis. The use of DNA microarray is a revolutionized approach that has been discovered in recent years. This method is designed for large scale DNA mapping and sequencing. More recently, the DNA microarray is gradually being substituted by RNA sequencing due to the extremely high efficiency of the later one. RNA sequencing is a genome-wide analysis for different gene expression, and it would be highly efficient when combined with next generation sequencing strategies [23]. Apart from sequencing technology, PCR strategies have also gained development. RT-PCR is considered quantitative PCR, which allowed real-time PCR reaction monitoring and being widely used in drug development.

4. Machine-based methods applied in target identification

The machine-based methods are widely used in the drug development process and have lots of advantages compared with the traditional experimental assay. The use of machine-based methods, such as machine learning algorithms, could simulate the condition that the target would bound by the desirable molecules. Based on the prediction model, the scientists could understand the availability of the target. To perform a drug target identification project in silico, there are three steps needed. Scientists would firstly select the database, then select and apply the suitable learning algorithms and finally evaluate the performance of the simulation process [24].

By utilizing the machine-based methods, great contributions would be generated in scientific field. A good example of using machine-based method is the application of random forest (RF) algorithm in the prediction of drug activity against cancer cells based on chemical properties and genomic information in Anderson and Lind’s experiment [25]. The RF algorithm helped classify and designate the data from huge amounts of data set in this project. The classification model established here was built by integrating the data from high throughput screening and machine-learned algorithm to predict the drug activity against the different cancer cells. In addition, Support Vector Machine (SVM) is another machine learning algorithm that has been widely and successfully applied in drug discovery. The SVM algorithm could classify the compound from a different database and make a prediction for the possibility of the studied compound to be active. In Zararsiz’s experiment, the dataset consisted of 631 compounds and SVM algorithm was applied to classify the active drugs and inactive drugs from the dataset [26]. Furthermore, in Doig’s experiment, the authors used the SVM algorithm as a prediction model to predict the potential drug target [27]. The consequence of their experiment was compelling and it showed SVM based model had high accuracy on predictions of the drug targets.

The machine-based learning methods are popular in recent years and help scientists to do drug discovery projects much easier than before. Although it had high efficiency and low cost with high accuracy, evaluation, and verification made by biological assay are still needed. The combination of machine-based methods and biological assay could be a great way for scientists to explore.

5. Conclusion

This review summarized the using of database, biological assays and machine-based methods in drug discovery in recent years. These methods are becoming more and more popular and gradually replace some inefficient methods. The using of RNA sequencing has been more frequent in recent years instead of traditional DNA microarray to measure specific gene expression. Also, machine-based methods, such as SVM algorithm and RF algorithm, have contributed to drug discovery, especially the target identification step. The SVM algorithm helped classify the active compound from the dataset and the RF algorithm could predict the drug activity to the cancer cells. The database integrated
information about the compound property, target property, and small organic molecule property, all of which are important for drug discovery. The combination of biological assay and machine-based method could be a way for scientists to explore in order to find a more efficient and less costly way to identify the drug targets. Therefore, the development of the machine-based method and biological assay is critically important for drug discovery.

6. Reference

[1]. Horien, C. and P. Yuan, Drug Development. The Yale Journal of Biology and Medicine, 2017. 90(1): p. 1-3.

[2]. Overington, J.P., B. Al-Lazikani, and A.L. Hopkins, How many drug targets are there? Nature Reviews Drug Discovery, 2006. 5(12): p. 993-996.

[3]. Santos, R., et al., A comprehensive map of molecular drug targets. Nature reviews. Drug discovery, 2017. 16(1): p. 19-34.

[4]. Lindsay, M.A., Target discovery. Nature Reviews Drug Discovery, 2003. 2(10): p. 831-838.

[5]. Terstappen, G.C., et al., Target deconvolution strategies in drug discovery. Nature Reviews Drug Discovery, 2007. 6(11): p. 891-903.

[6]. Kim, E., A.S. Choi, and H. Nam, Drug repositioning of herbal compounds via a machine-learning approach. BMC Bioinformatics, 2019. 20(10): p. 247.

[7]. Wang, Y., et al., Therapeutic target database 2020: enriched resource for facilitating research and early development of targeted therapeutics. Nucleic Acids Research, 2020. 48(D1): p. D1031-D1041.

[8]. Wishart, D.S., et al., DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic acids research, 2008. 36(Database issue): p. D901-D906.

[9]. Gong, L., et al., PharmGKB: an integrated resource of pharmacogenomic data and knowledge. Current protocols in bioinformatics, 2008. Chapter 14: p. Unit14.7-Unit14.7.

[10]. Gilson, M.K., et al., BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. Nucleic acids research, 2016. 44(D1): p. D1045-D1053.

[11]. Liu, T., et al., BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. Nucleic Acids Research, 2007. 35(suppl_1): p. D198-D201.

[12]. Davis, A.P., et al., Comparative Toxicogenomics Database (CTD): update 2021. Nucleic Acids Research, 2020.

[13]. Wishart, D., et al., T3DB: the toxic exposome database. Nucleic acids research, 2015. 43(Database issue): p. D929-D934.

[14]. Lim, E., et al., T3DB: a comprehensively annotated database of common toxins and their targets. Nucleic acids research, 2010. 38(Database issue): p. D781-D786.

[15]. Halling-Brown, M.D., et al., canSAR: an integrated cancer public translational research and drug discovery resource. Nucleic acids research, 2012. 40(Database issue): p. D947-D956.

[16]. Gaulton, A., et al., ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Research, 2012. 40(D1): p. D1100-D1107.

[17]. Gaulton, A., et al., The ChEMBL database in 2017. Nucleic acids research, 2017. 45(D1): p. D945-D954.

[18]. Jeibmann, A., et al., Identification of genes involved in the biology of atypical teratoid/rhabdoid tumours using Drosophila melanogaster. Nature Communications, 2014. 5(1): p. 4005.

[19]. Courtney, K.D., R.B. Corcoran, and J.A. Engelman, The PI3K Pathway As Drug Target in Human Cancer. Journal of Clinical Oncology, 2010. 28(6): p. 1075-1083.

[20]. Zhang, Y.-N., et al., Oleanolic acid and its derivatives: New inhibitor of protein tyrosine phosphatase 1B with cellular activities. Bioorganic & Medicinal Chemistry, 2008. 16(18): p. 8697-8705.

[21]. Galindo, K.A., et al., A rapid one-generation genetic screen in a Drosophila model to capture rhabdomyosarcoma effectors and therapeutic targets. G3 (Bethesda, Md.), 2014. 5(2): p. 205-217.

[22]. Newton, H., et al., Systemic muscle wasting and coordinated tumour response drive tumourigenesis. Nature Communications, 2020. 11(1): p. 4653.
[23]. Stark, R., M. Grzelak, and J. Hadfield, RNA sequencing: the teenage years. Nature Reviews Genetics, 2019. 20(11): p. 631-656.
[24]. Yang, H., et al., A Review of Recent Advances and Research on Drug Target Identification Methods. Current Drug Metabolism, 2019. 20(3): p. 209-216.
[25]. Lind, A.P. and P.C. Anderson, Predicting drug activity against cancer cells by random forest models based on minimal genomic information and chemical properties. PLOS ONE, 2019. 14(7): p. e0219774.
[26]. Korkmaz, S., G. Zararsiz, and D. Goksuluk, Drug/nondrug classification using Support Vector Machines with various feature selection strategies. Computer Methods and Programs in Biomedicine, 2014. 117(2): p. 51-60.
[27]. Bakheet, T.M. and A.J. Doig, Properties and identification of human protein drug targets. Bioinformatics, 2009. 25(4): p. 451-457.