REVIEW ARTICLE

A Review of Drug Repositioning Based Chemical-induced Cell Line Expression Data

Fei Wang[a], Xiujuan Lei[b], and Fang-Xiang Wu[b,c,*]

[a]Division of Biomedical Engineering, University of Saskatchewan, Saskatoon, Canada; [b]School of Computer Science, Shaanxi Normal University, Xi'an, Shaanxi, China; [c]Department of Mechanical Engineering, Department of Computer Science, University of Saskatchewan, Saskatoon, Canada

ARTICLE HISTORY

Received: October 01, 2017
Revised: August 10, 2018
Accepted: October 21, 2018

DOI:
10.2174/0929867325666181101115801

Abstract: Drug repositioning is an important area of biomedical research. The drug repositioning studies have shifted to computational approaches. Large-scale perturbation databases, such as the Connectivity Map and the Library of Integrated Network-Based Cellular Signatures, contain a number of chemical-induced gene expression profiles and provide great opportunities for computational biology and drug repositioning. One reason is that the profiles provided by the Connectivity Map and the Library of Integrated Network-Based Cellular Signatures databases show an overall view of biological mechanism in drugs, diseases and genes. In this article, we provide a review of the two databases and their recent applications in drug repositioning.

Keywords: Drug repositioning, computational biology, bioinformatics, drug candidate, connectivity map, the library of integrated network-based cellular signatures.

1. INTRODUCTION

In recent years, researchers are showing more and more interest in drug repositioning. Drug repositioning is to find an existing drug for healing a disease other than its originally healed disease. In the past decades, *de novo* drug development often takes 10 to 15 years and 0.8 to 1.5 billion dollars to invent a new drug for patients [1]. However, finding an existing drug's new application can reduce time and cost greatly. All compounds that no matter which disease they are invented to heal can be used as a drug candidate in drug repositioning research [2].

There are also some famous drug repositioning examples. Sildenafil is proposed for healing hypertension and angina pectoris at first, but it shows erectile effects in the clinic [3]. Because of the lower cost and shorter period, many institutions, including governments and universities, support drug repositioning largely. For example, the FDA has created several public databases for drug repositioning [4].

In the past years, the discovery of new uses of old drugs is mostly through serendipity [5] or compound screening. The molecular dynamics analysis is a useful approach in drug repositioning. Chen *et al*. created a model of molecular dynamics analysis and proposed three inhibitors for NSCLC patients [6].

In other studies, with the growth of drug-disease association databases, several new repositioning strategies had emerged with various databases. TTD and DrugBank are two databases which provide drug-target interaction information. TTD database contains the known therapeutic proteins and nucleic acid targets, the path way information and the drugs directed at the targets [7]. During the subsequent updates, more information about drug resistance mutations, target gene expressions, target combinations of multi-target drugs and drug combinations are provided [8]. On the other hand, a comprehensive summary of drugs and targets are provided in DrugBank database [9]. During the updates of the DrugBank database, the pharmacological, pharmacogenomic and molecular biological data, drug-

*Address correspondence to this author at the School of Computer Science, Shaanxi Normal University, Xi'an, Shaanxi, China and Department of Mechanical Engineering, Department of Computer Science, University of Saskatchewan, Saskatoon, Canada;
Tel: 1(306)9665280; E-mail: faw341@mail.usask.ca

0929-8673/19 $58.00+.00

drug and drug-food interactions, drug metabolism data and many other informations are added to the database [10]. The drug-drug and drug-target interactions from these databases are widely used in drug repositioning [11, 12]. Additionally, these databases are also used in some studies of gene expression data and drug repositioning approaches [13].

Besides the drug-drug and drug-target interactions, the miRNAs and miRNA-disease interactions are used in drug repositioning. Tang *et al*. proposed several strategies of miRNA classifiers to predict the primary site of tumor [14]. Zeng *et al*. proposed a bilayer network of three sub-networks, including the miRNA-disease network, the miRNA-miRNA similarity network and the disease similarity network [15]. The proposed network can be used to predict potential connections between miRNAs and diseases, which is a useful part of drug repositioning.

To focus on the connections of drugs and diseases in gene expression level, the Connectivity Map (CMap) and the Library of Integrated Network-Based Cellular Signatures (LINCS) databases store the gene expression data of different cell lines induced by many chemicals and thus are widely used in drug repositioning.

In 2006, Lamb *et al.* created a large public reference database of signatures of drugs and genes, and developed pattern-matching tools to detect similarities among these signatures [16]. The reference database contains a collection of chemical-induced gene expression profiles, describing chemical-induced reference states [17]. The reference database of CMap build 01 contains 564 individual reference profiles for 453 small molecule compounds, and the build 02 contains 6,100 individual reference profiles for 1,309 small molecule compounds. A query signature is a list of genes investigating a particular biological condition. Differentially expressed genes between disease and normal conditions are often used to build a signature of this disease. A matching algorithm is to compare the gene signature and the reference profiles for computing their similarity.

After the completion of the CMap database, researchers want to reduce the cost of producing expression profiles. In order to solve it, many institutes participate into a LINCS project, which aims to create a network-based understanding of biology by cataloging changes in gene expression and other cellular processes that occur when cells are exposed to a variety of perturbing agents [18]. The LINCS database Phase I was completed in 2013 and published in the National Center for Biotechnology Information (NCBI) [19]. The Phase II is still going on. The statistics of the CMAP and LINCS databases are shown in Table **1**.

Table 1. The statistics of CMAP and LINCS L1000 data.

Databases	Perturbations	Reference profiles	Cell lines
CMAP data build 2	1306	6100	5
LINCS L1000 data phase I	42080	1319138	82

Since the CMap database has been completed for about 10 years and the LINCS project has not been finished yet and is relatively new, many researchers pay their attention to the applications of the CMap database and ignore the LINCS database. However, in recent years, more and more researchers use the LINCS database as well as the CMap database in drug repositioning.

In this review, we try to provide a survey on the CMap and LINCS databases and their applications in drug repositioning in recent years. In Section 2, the principles of the CMap and LINCS databases are briefly described. Section 3 reviews the main application of the CMap and LINCS databases in three categories of drug repositioning. Section 4 concludes this review with some discussions.

2. METHODS

The CMap database build 01 was proposed by Lamb *et al*. in 2006 [16]. The main concept of the CMap database is to use a reference database containing drug-specific gene expression profiles and compare it with a gene signature [20]. This gene signature is a list of genes, which is supposed to characterize a disease or other biological condition (As shown in Fig. **1.A**). The method is simply performed by submitting a gene signature and returned a list of compounds. These compounds are presumptive efficacy for the disease. At least, these compounds enhance the understanding of the disease, such as the metabolic mechanism of the disease and possible chemical structure of drugs. The final goal of the CMap database is to predict potentially drug candidates for a disease.

After building a gene signature of a disease, it is used to query the CMap gene expression profiles. The CMap database is a collection of gene expression profiles representing a set of microarray experiments. All experiments are conducted with a microarray platform (Affymetrix HT_HG_U133A array with 22,283 probes). The experiments are carried out in several cell lines to compounds (drugs and small molecules) at different concentrations and time points against vehicle controls. Although researchers can generate profiles in a wide diversity of established and primary cells, practicality limits them to only a few cell lines that are stably grown over long periods of time [16]. The CMap database includes the breast cancer epithelial cell line MCF7, prostate cancer cell line PC3, nonepithelial lines HL60 (leukemia) and SKMEL5 (melanoma). The MCF7 cell line is treated in a different environment and named ssMCF7. The CMap gene profiles are based on the five cell lines in total.

The initial database contains 564 gene expression profiles, representing 453 individual instances (*i.e.,* one treatment and vehicle pair). The updated version (Build 2) contains 6,100 expression profiles, representing 1,309 chemical compounds. These two versions are based on the same platform and cell lines, in order to use the same analysis methods for researchers. The instance is the basic unit of data in the CMap database. An instance is uniquely identified by an instance identifier. For each instance, the fold change of treatment to control values is calculated for each probe, sorted in decreasing order and converted to a rank vector. The probe that is most up-regulated is given a value 1 and the most down-regulated probe is given a value 22,283, which is the total number of genes expressed in each cell line. Then the CMap instance matrix is a 22,283*6,100 matrix (for Build 2). The rankings of instances are used in the query process.

Since a gene signature contains up-regulated and down-regulated genes, it is divided into two sets: an *s_u* set that contains up-regulated genes and an *s_d* set that contains down-regulated genes. The *s_u* set is on the top of *s_d* set. Then the gene signature is compared to each rank-ordered profile to determine whether up-regulated genes tend to appear near the top of the list and down-regulated genes tend to appear near the bottom (“positive connectivity”) or vice versa (“negative connectivity”), yielding a “connectivity score” ranging from +1 to -1 [16]. The so-called connectivity score is estimated based on the Kolmogorov-Smirnov statistic [21]. All instances in the database are ranked according to their connectivity scores. Those at the top are most strongly positive correlated to the gene signature while those at the bottom are most strongly negative correlated.

After the first introduction of the CMap method and database, there have been numerous tools and extensions of this approach. In order to have an easy or accuracy result, researchers propose some automatic tools to calculate connectivity score, for instance, the statistically significant connections' map (sscMap) [22]. The sscMap combines the matching algorithm and the CMap dataset in it. After users just input a gene signature they can receive a list of drug candidates.

After the completion of the CMap database, the LINCS project based on the L1000 technology [18] is proposed in order to reduce the cost for obtaining gene expression profiles. To make a balance between reflecting more information and reducing more cost, it is estimated that 978 genes are sufficient to recover 82% information on the CMap database with a low cost. To select 978 landmark genes, more than 12,000 samples from the NCBI’s Gene Expression Omnibus (GEO) [24] are assembled. In order to minimize the bias toward a particular lineage or cellular state, the principal component analysis (PCA) is used as a dimensionality reduction procedure. Then the *k*-means (*k* is the number of landmark genes) clustering is performed to identify tight clusters of co-regulated transcripts. For each tight cluster, the landmark gene is identified as the transcript closest to the centroid. Besides the landmark genes, all the other genes are termed as inferred genes. Since the probes in the platform Affymetrix HG U133A are mapped to 13,210 unique genes, the number of inferred genes is 12,232. Based on the assembled dataset, each inferred gene is predicted from landmark genes via linear regression. The LINCS project records all the data of both landmark genes and inferred genes, researchers can use any kind of genes.

The process of generating LINCS data contains five stages, as shown in Fig. (**2**). The data of each stage are available from the LINCS database. The L1000 platform contains 1,058 probes for 978 landmark genes and 80 control genes chosen for their invariant expression across cell states [23]. The data of the first stage (Level 1) are unprocessed image data which are collected from the scanners. It contains fluorescence intensity values of the 1058 probes on a L1000 microarray. The Level 2 data are a collection of gene expression profiles (numerical values) of landmark genes derived from Level 1 data.

The Level 3 data are collections of expression profiles of both landmark genes and inferred genes. To reduce the biases due to the different dyeing efficiency of the fluorescent dyes and different microarrays, expression values of the landmark genes obtained in Level 2 are normalized to invariant gene set curves and quantile normalized across each microarray. The expression values of inferred genes are calculated from landmark genes via linear regression.

To obtain a measure of relative gene expression, a robust z-scoring procedure is adopted to transform gene expression values in Level 3 into the robust z-scores, which consists of Level 4 data of the LINCS database. Since the LINCS experiments are typically done in 3 replicates, all the replicates are aggregated into a single consensus gene signature in Level 5. Each of the 3 replicates has a weight in the aggregation, which is the sum of its Spearman correlations to the other 2 replicates.

All the data of the five stages constitute the LINCS database. Among them, the Level 2 and 3 data are the most commonly used data for drug repositioning. Since the LINCS database Phase I have finished and the Phase II is still on-going, researchers prefer to use the Phase I data in their study.

The principles of the CMap and LINCS databases are shown in Fig. (**1**) [16]. Fig. (**1.A**) represents a gene signature of a biological condition, which is used to query the expression profiles. Figs. (**1.B1**, **1.B2**) and so on represent the matches of a gene signature and expression profiles. Fig. (**1.C**) is the rank of all these matches in which the strongly positive correlations are on the top while the strongly negative correlations are on the bottom.

3. APPLICATIONS

After the CMap and LINCS databases have been published, many researchers use the two databases to study drug repositioning. Many of them use these two databases as an essential step of their drug repositioning approaches, some referred the databases to verify the experimental results of their approaches while others use partial content of the two databases in their experiments. In this paper, we show the three usages of the CMap and LINCS databases in the rest of this section. Because the CMap database is completed earlier than the LINCS database, there are more researchers used the CMap database than the LINCS database.

3.1. As an Essential Step

This is the main application of the CMap and LINCS databases. When researchers are studying drug repositioning, they think the two databases are useful and use them as one step of their approaches. Because of the properties of the two databases, researchers prefer to generate a gene signature of a disease and query it to the databases. This is a widely used type of drug repositioning approaches.

In 2014, in order to deal with the huge amount of data about biomedical information, Temesi *et al.* proposed a knowledge-recycling strategy of compound set enrichment analysis [25]. In their method, the CMap database plays an important role. First, they collect 1941 FDA approved drugs from public databases and rank them. Second, the chemical substances are mapped from the CMap database to their compounds.

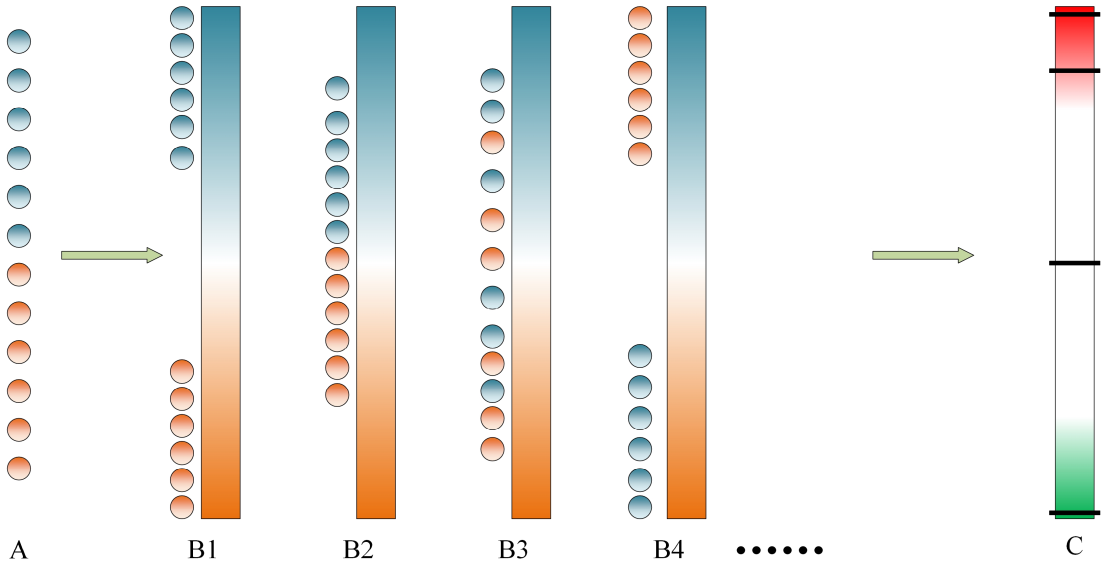

Fig. (1). The principles of CMAP and LINCS.

Third, a candidate compound (marked as *c*) is selected for repositioning and the similarity between *c* and their compounds is calculated based on the structure, target, expression profile and side effects. Finally, a ranked list of compounds is generated and used for drug repositioning.

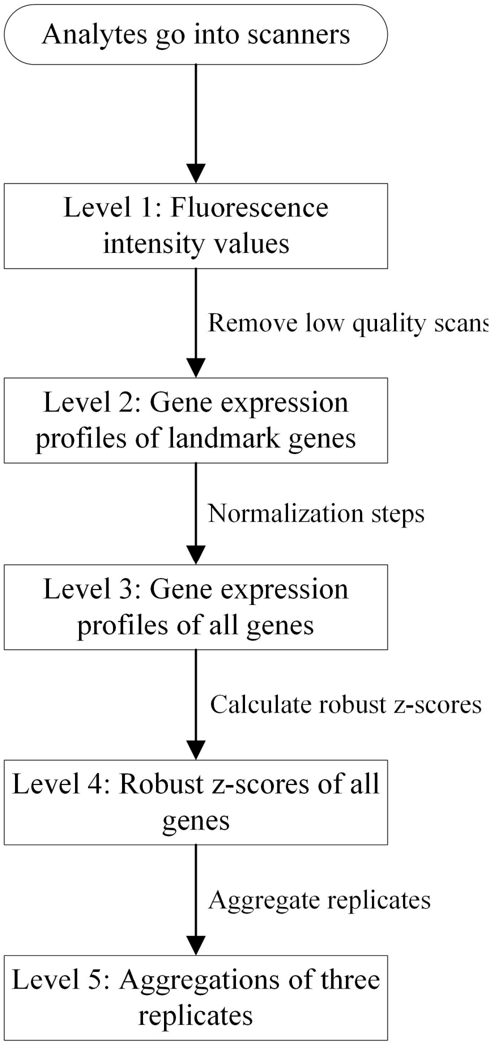
Analytes go into scanners
Level 1: Fluorescence intensity values
Remove low quality scans
Level 2: Gene expression profiles of landmark genes
Normalization steps
Level 3: Gene expression profiles of all genes
Calculate robust z-scores
Level 4: Robust z-scores of all genes
Aggregate replicates
Level 5: Aggregations of three replicates

Fig. (2). The 5 stages of LINCS L1000 data.

The Ewing sarcoma (EWS) is a serious cancer for which about 70% of patients die. In 2016, Goss *et al.* proposed a method to target EWS-FLI1, which was specific for EWS cancer cells and required for tumorigenesis [26]. In their study, gene expression data from a EWS model is used to identify EWS-FLI1 target genes. These genes are then used to query the CMap database. As a result, the iron chelator is identified as a drug. In their experiments, the CMap analysis also predicts another candidate, etoposide. Currently, etoposide is a useful compound for EWS. In order to identify new therapeutic options, in 2017, Pessetto *et al.* proposed a multi-pronged approach, including *in silico* predictions and an *in vitro* screen [27]. In the *in silico* predictions section, the CMap database is employed to select 20 drug candidates for *in vitro* validation (also including etoposide).

In 2016, Lee *et al.* proposed that withaferin A was an antidiabetic drug candidate and reported that it showed great properties in mice experiments [28]. In their analysis, endoplasmic reticulum (ER) stress is one of the mechanisms that can reduce leptin signaling and then induce leptin resistance during the progression of obesity. In addition, the spliced form of X-box binding protein-1 (XBP1s) is one of the cardinal regulators of ER homeostasis. In order to find molecules that are similar with XBP1s in gene expression profiles, they search the CMap database. After inputting six different query signatures, researchers found that celastrol is a drug candidate. Meanwhile, withaferin A has a large similarity with celastrol. The experiments in mice also show that withaferin A is effective.

Valosin-containing protein (VCP/p97) ATPase is an important participator in a pathway of ER. The specific pathway is an ER-associated protein degradation (ERAD) pathway, which is implicated in cancers and many other diseases. In 2017, Segura-Cabrera *et al.* proposed a new drug repositioning approach to detect drugs that repress (VCP/p97) ATPase [29]. A gene signature of READ is generated and used to query the CMap database. As a result, three drug candidates are proposed as potential inhibitors of VCP/p97 and ERAD.

Because of the different regenerative capacity between the central nervous system (CNS) and peripheral nervous system (PNS), Chandran *et al.* proposed an approach to generate drug candidates that can improve the dorsal root ganglia (DRGs) neurite outgrowth and optic nerve outgrowth [30]. Similarly, two gene signatures are built and used to query the CMap database. There are three drug candidates which appeared in both two drug candidate lists.

Glioblastoma multiforme (GBM) is a serious brain cancer that many patients can't survive for more than 14 months. In 2017, Xiao *et al.* proposed that a diabetes mellitus drug can be used to repress GBM [31]. The GBM patient data from GEO database are downloaded. After differential expression analysis, a gene signature is obtained which contains 20 up-regulated and down-regulated genes. Then the signature is used to query the CMap database and a candidate drug list is generated. The diabetes mellitus drug, repaglinide, is the top 6 drug candidate in the list. Repaglinide also shows anti cancer effect in the experiments.

Chronic allograft damage is a leading cause of allograft failure. It is defined by the interstitial fibrosis and tubular atrophy (IF/TA). In order to find a solution to reduce it, Li *et al.* generated an IF/TA dataset in GEO and used a meta-analysis approach to process it [32]. They finally obtain a gene signature of 85 genes. Then the gene signature is used to query the CMap database and a list of candidate drugs is generated. Kaempferol and esculetin are at the top of the list. The two drug candidates also show an effect in experiments.

Recently, traditional Chinese medicine attracted many researchers' attention. Many researchers want to find a succedaneum of a specific traditional Chinese medicine. Justicidin A (JA) is a natural ligand that is isolated from Justicia procumbens. In 2017, Won *et al.* generated gene signatures from HT-29 cells and Hep cells which were treated by Justicidin A [33]. Then those signatures are used to query the CMap database and a drug candidate list with 30 compounds is obtained. After analysis, they proposed that 15-delta prostaglandin J2 is a similar compound with JA.

In recent years, drug repositioning approaches are being more and more sophisticated than before [34]. Some researchers change their thoughts in studying drug repositioning. In 2017, Iorio *et al.* talked about methods for constructing drug networks [34]. Two kinds of networks are built and the transcriptional networks are based on the CMap and LINCS databases. Mode of Action by Network Analysis (MANTRA) and Drug-set Enrichment Analysis (DSEA) are tools based on the transcriptional networks. They also show the performance of these approaches in drug repositioning in their examples.

3.2. As an Evidence

Besides using the CMap and LINCS databases in drug repositioning approaches, researchers also use them in other fields. Since the CMap database has been proposed for more than ten years and the LINCS database Phase I also has been completed for several years, their accuracies have been proved in many aspects. When doing their studies, researchers prefer to verify their results by comparing with the CMap and LINCS databases.

In 2016, Chen *et al.* presented an algorithm, named Mpath, which derives multi-branching developmental trajectories using neighborhood-based cell state transitions [35]. In the experiments, Mpath is applied to two datasets, including murine conventional dendritic cell (cDC) and human myoblasts [36]. Murine cDCs contain two main lineages, which are cDC1 and cDC2. The Mpath-generated trajectories detect a branching event at the pre-dendritic cells (preDC) stage revealing that preDC subsets are exclusively committed to cDC1 or cDC2 lineages. In order to further describe the Mpath-derived state transition network, the CMap analysis is performed in the experiments. The cDC1 and cDC2 signature genes are used to identify cDC subset-primed cells. The CMap analysis shows that their method can detect the branching events successfully.

In 2016, in order to identify candidate target genes and molecules, Zickenrott *et al.* proposed a differential network-based methodology [37]. Many datasets are combined in the proposed method to construct gene regulatory networks (GRNs). In order to show the performance of the proposed method in identifying candidate target genes and molecules, six samples from the CMap database are chosen, including celastrol+androgen, gedunin+androgen, celastrol, cobalt chloride, estradiol and genistein. The results show that the proposed method can identify genes and drugs accurately.

In 2017, Liu *et al.* conducted a research to reveal which gene was more likely to be associated with disease-related functions [38]. In the CMap matrix, all elements are the logarithmic transformed fold change (lnFC) of probe sets. In their study, lnFC>0.69 or lnFC<-0.69 cutoffs are used for the determination of significantly up-regulated genes and down-regulated genes in each cell line, respectively. Then the numbers of these up-regulated and down-regulated genes are counted as the differential expression number (DEN). After that, a lot of comparisons and analysis between the high DEN genes and other genes are performed. All the CMap cell lines are applied in the experiments and the results are unsurprising as the higher DEN genes are more likely to be disease-related.

Besides the CMap database, researchers also use the LINCS database to confirm their results from other approaches. In 2016, Grammer *et al.* used a Lupus Treatment List (LRxL)- SLE Treatment Acceleration Trials (STAT) approach to rank potential treatments for Systemic Lupus Erythematosus (SLE) patients by using Combined Lupus Treatment Scoring (CoLTs) [39]. The CoLTs can give scores to all candidates based on five fields: scientific rationale, experience in lupus mice/human cells (pre-clinical), previous clinical experience in autoimmunity, drug properties and safety profiles including adverse events. The scoring contents suggest that the LRxL-STAT is a literature-search approach. In order to examine its performance and accuracy, the LINCS database is generated to conduct a comparison between the observed changes in experiments and the abnormalities of meta-analysis. The lupus B cell lines are applied in the experiments. Both the experimentally based the LINCS database and other scoring systems suggest that ustekinumab is a treatment for SLE.

In 2015, Siavelis *et al.* combined five widely used databases, including the CMap and LINCS databases, to generate a merged drug candidate of Alzheimer's disease [40]. The CMap tool ssCMap is also used at the same time. In the method, the compound which appears in at least two approaches is collected. The final result contains 27 unique drugs. After then, all the drug candidates are analyzed based on their chemical structure, pathway and ontology enrichment, and network analysis. Among the candidates, histone deacetylase and other three drugs are suggested that they may play important roles in healing Alzheimer's disease. Their results show a potential application of current databases that can reduce weaknesses and bias of them.

3.3. As a Material

Since an excellent design of the experiment is very helpful in research, when researchers are dealing with troublesome experiments, one of the useful solutions is to learn from others. Meanwhile, the CMap and LINCS databases contain well designed experiments that can be followed by other researchers. Besides, the cancer cell lines of the CMap and LINCS databases are also important resources that researchers use the specific cancer cell lines in their work.

In 2012, Sirota *et al.* collected drug-exposure gene expression microarray data of 164 distinct small molecules from the CMap database [41]. All the data are measured on four cell lines, including MCF7, PC3, HL60, and SKMEL5. The microarray data are used as a reference database. It is queried by all individual disease signatures that apply the CMap matching strategy. This study is a simple usage of the CMap database.

In 2015, in order to predict the interactions between drugs and immune cells, Kidd *et al.* presented an integrative computational approach in a system-wide manner [42]. After generating an immune cell state data of 304 transitions from 211 immune cell types, 1,309 drug perturbation profiles in the CMap database are collected and matched to the 304 immune cell state changes. These matches make up a system-wide interaction map that contains 397,936 potential connections. In their study, the cancer cell lines are regarded as an important part of the CMap database and used in the research.

In 2016, Ryan *et al.* proposed a gene expression biomarker that predicts Estrogen Receptor α (ER α) modulation [43]. This ERα biomarker is a list of differentially expressed genes. It is generated after exposure to ERα modulators. In the experiments, biosets from MCF7 cell line in the CMap database are collected to identify the biomarker. In short, the results are greatly based on the CMap database.

There is a traditional Chinese medicine Shexiang Baoxin pill that is widely used in healing cardiovascular disease (CVD) in China. Fang *et al.* proposed a network-based method for mechanistic investigation of a drug's treatment of cardiovascular diseases [44]. In the experiments, their study is also based on MCF7 cell line of the CMap database.

Since the LINCS database contains more cancer cell lines, many researchers use the LINCS database in their experiments. In 2017, Chen *et al.* proposed a repositioned drug candidate for reducing the growth of hepatocellular carcinoma (HCC) cells [45]. In the experiments, two cell lines from the LINCS database are used, including HepG2 and Huh 7. Based on the two cell lines, the authors use 978 landmark genes of the LINCS database in their prediction.

In 2017, Zhang *et al.* proposed an approach based on another kinds of networks, the drug similarity network and target protein similarity network [46]. The networks are used to predict the drug-target interactions, which are an important part of drug repositioning. Many datasets are collected to construct the networks, including drug category data from DrugBank [47], LINCS database and some other databases. All the datasets are necessary for their approach.

From the above we can see that some researchers prefer to refer the CMap and LINCS databases as a material of their experiments instead of generating a gene signature and querying the databases.

CONCLUSION

In this review, we have provided a survey about the CMap and LINCS databases methods and applications in recent years. The CMap and LINCS databases are two kinds of perturbation databases that offer an opportunity for drug repositioning. The applications of these two databases are focusing on identifying new therapeutic targets, drug repositioning opportunities and finding new mechanism of action for existing compounds. Most of the potentials of the CMap and LINCS databases are undoubtedly beneficial in research and useful in drug industries. However, there are also some limitations. One of the existing problem is that most compounds have only one replicate per cell line for each experiment. This may cause some statistical challenges. Then cell lines of the CMap and LINCS databases are also limited, and do not cover all possible disease cell lines. Moreover, not all compounds are applied to all the cell lines. It may cause a bias in research. These weaknesses of the CMap and LINCS databases must be addressed in the future. One possible solution is to integrate more public databases when using the CMap and LINCS databases, such as RNA-seq database.

LIST OF ABBREVIATIONS

cDC	=	Conventional Dendritic Cell
CMap	=	Connectivity Map
CNS	=	Central Nervous System
CoLTs	=	Combined Lupus Treatment Scoring
CVD	=	Cardiovascular Disease
DEN	=	Differential Expression Number
DSEA	=	Drug-set Enrichment Analysis
ER	=	Endoplasmic Reticulum
ERα	=	Estrogen Receptor α
ERAD	=	ER-associated Protein Degradation
EWS	=	Ewing Sarcoma
FDA	=	US Food and Drug Administration
GBM	=	Glioblastoma Multiforme
GEO	=	Gene Expression Omnibus
GRN	=	Gene Regulatory Network
HCC	=	Hepatocellular Carcinoma
IF/TA	=	Interstitial Fibrosis and Tubular Atrophy
JA	=	Justicidin A
LINCS	=	Library of Integrated Network-Based Cellular Signatures
LRxL	=	Lupus Treatment List
MANTRA	=	Mode of Action by Network Analysis
NCBI	=	National Center for Biotechnology Information
NSCLC	=	Non-Small Cell Lung Cancer
PCA	=	Principle Component Analysis
PNS	=	Peripheral Nervous System
preDC	=	Pre-Dendritic Cell
SLE	=	Systemic Lupus Erythematosus
sscMAP	=	Statistically Significant Connection's Map
STAT	=	SLE Treatment Acceleration Trials
TTD	=	Therapeutic Target Database
VCP/p97	=	Valosin-Containing Protein
XBP1	=	X-box Binding Protein-1

CONSENT FOR PUBLICATION

Not applicable.

FUNDING

This work is supported in part by Natural Science and Engineering Research Council of Canada (NSERC), China Scholarship Council (CSC) and by the National Natural Science Foundation of China under Grant No. 61772552 and No. 61428209.

CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

Declared none.

REFERENCES

[1] Emmert-Streib, F.; Tripathi, S.; Simoes, R. D. M.; Hawwa, A. F.; Dehmer, M. The human disease network: Opportunities for classification, diagnosis, and prediction of disorders and disease genes. *Systems Biomedicine*, **2013**, 1(1), 20-28.

[2] Shameer, K.; Readhead, B.; T Dudley, J. Computational and experimental advances in drug repositioning for accelerated therapeutic stratification. *Current topics in medicinal chemistry*, **2015**, 15(1), 5-20.

[3] Boolell, M.; Allen, M.J.; Ballard, S.A.; Gepi-Attee, S.; Muirhead, G.J.; Naylor, A.M.; Osterloh, I.H.; Gingell, C. Sildenafil: an orally active type 5 cyclic GMP-specific phosphodiesterase inhibitor for the treatment of penile erectile dysfunction. *International journal of impotence research*, **1996**, 8(2), 47-52.

[4] Li, J.; Zheng, S.; Chen, B.; Butte, A. J.; Swamidass, S. J.; Lu, Z. A survey of current trends in computational drug repositioning. *Briefings in bioinformatics*, **2015**, 17(1), 2-12.

[5] Bolgar, B.; Arany, A.; Temesi, G.; Balogh, B.; Antal, P.; Matyus, P. Drug repositioning for treatment of movement disorders: from serendipity to rational discovery strategies. *Current topics in medicinal chemistry*, **2013**, 13(18), 2337-2363.

[6] Chen, L.; Zou, B.; Lee, V. H.; Yan, H. Analysis of the Relative Movements Between EGFR and Drug Inhibitors Based on Molecular Dynamics Simulation. Current Bioinformatics, 2018, 13(3), 299-309.

[7] Chen, X.; Ji, Z. L.; Chen, Y. Z. TTD: therapeutic target database. Nucleic acids research, 2002, 30(1), 412-415.

[8] Li, Y.H.; Yu, C.Y.; Li, X.X.; Zhang, P.; Tang, J.; Yang, Q.; Fu, T.; Zhang, X.;Cui, X.;Tu, G.; Zhang, Y.; Li, S.; Yang, F.; Sun, Q.; Qin, C.; Zeng, X.; Chen, Z.; Chen, Y.Z.; Zhu, F. Therapeutic target database update 2018: enriched resource for facilitating bench-to-clinic research of targeted therapeutics. Nucleic acids research, 2018, 46(D1), D1121-D1127.

[9] Wishart, D. S.; Knox, C.; Guo, A. C.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z.; Woolsey, J. DrugBank: a comprehensive resource for *in silico* drug discovery and exploration. Nucleic acids research, 2006, 34(suppl_1), D668-D672.

[10] Wishart, D. S.; Feunang, Y. D.; Guo, A. C.; Lo, E. J.; Marcu, A.; Grant, J. R.; 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. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic acids research, 2018, 46(D1), D1074-D1082.

[11] Cheng, F.; Liu, C.; Jiang, J.; Lu, W.; Li, W.; Liu, G.; Zhou, W.; Huang, J.; Tang, Y. Prediction of drug-target interactions and drug repositioning via network-based inference. PLoS computational biology, 2012, 8(5), e1002503.

[12] Chen, L.; Chu, C.; Zhang, Y. H.; Zheng, M.; Zhu, L.; Kong, X.; Huang, T. Identification of drug-drug interactions using chemical interactions. Current Bioinformatics, 2017, 12(6), 526-534.

[13] Wu, H.; Huang, J.; Zhong, Y.; Huang, Q. DrugSig: A resource for computational drug repositioning utilizing gene expression signatures. PloS one, 2017, 12(5), e0177743.

[14] Tang, W.; Wan, S.; Yang, Z.; Teschendorff, A. E.; Zou, Q. Tumor origin detection with tissue-specific miRNA and DNA methylation markers. Bioinformatics, 2017, 34(3), 398-406.

[15] Zeng, X.; Liu, L.; Lu, L.; Zou, Q. Prediction of potential disease-associated microRNAs using structural perturbation method. Bioinformatics, 2018, 34(14), 2425-2432.

[16] Lamb, J.; Crawford, E.D.; Peck, D.; Modell, J.W.; Blat, I.C.; Wrobel, M.J.; Lerner, J.; Brunet, J.P.; Subramanian, A.; Ross, K.N.; Reich, M.; Brunet, J.P.; Subramanian, A.; Ross K.N.; Reich M.; Hieronymus H.; Wei G.; Armstrong S.A.; Haggarty S.J.; Clemons P. A.; Wei R.; Carr S.A.; Lander E.S.; Golub T.R. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, **2006**, 313(5795), 1929-1935.

[17] Wen, Q.; O'Reilly, P.; Dunne, P.D.; Lawler, M.; Van Schaeybroeck, S.; Salto-Tellez, M.; Hamilton, P.; Zhang, S.D. Connectivity mapping using a combined gene signature from multiple colorectal cancer datasets identified candidate drugs including existing chemotherapies. *BMC systems biology*, **2015**, 9(5), S4.

[18] National Institutes of Health: The Library of Integrated Network-Based Cellular Signatures Project. http://www.lincsproject.org

[19] National Center for Biotechnology Information Search database. https://www.ncbi.nlm.nih.gov

[20] Musa, A.; Ghoraie, L.S.; Zhang, S.D.; Glazko, G.; Yli-Harja, O.; Dehmer, M.; Haibe-Kains, B.; Emmert-Streib, F. A review of connectivity map and computational approaches in pharmacogenomics. *Briefings in bioinformatics*, **2017**, bbw112.

[21] Subramanian, A.; Tamayo, P.; Mootha, V.K.; Mukherjee, S.; Ebert, B.L.; Gillette, M.A.; Paulovich, A.; Pomeroy, S.L.; Golub, T.R.; Lander, E.S.; Mesirov, J.P. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of *Science*, **2005**, 102(43), 15545-15550.

[22] Zhang, S. D.; Gant, T. W. ssCMap: an extensible Java application for connecting small-molecule drugs using gene-expression signatures. *BMC bioinformatics*, **2009**, 10(1), 236.

[23] Subramanian, A.; Narayan, R.; Corsello, S.M.; Peck, D.D.; Natoli, T.E.; Lu, X.; Gould, J.; Davis, J.F.; Tubelli, A.A.; Asiedu, J.K.; Lahr, D.L.; Hirschman, J. E.; Liu, Z.;; Donahue, M.; Julian, B.; Khan, M.; Wadden, D.; Smith, I.; Lam, D.; Liberzon, A.; Toder, C.; Bagul, M.; Orzechowski, M.; Enache, O. M.; Piccioni, F.; Berger, A. H.; Shamji, A.; Brooks, A. N.; Vrcic, A.; Flynn, C.; Rosains, J.; Takeda, D.; Davison, D.; Lamb, J.; Ardlie, K.; Hogstrom, L.; Gray, N. S.; Clemons, P. A.; ; Silver, S.; ; Wu, X.; ; Zhao, W.; ; Read-Button, W.; ; Wu, X.; ; Haggarty, S. J.; ; Ronco, L. V.; ; Boehm, J. S.; ; Schreiber, S. L.; ; Doench, J. G.; ; Bittker, Joshua A.; Root, David E.; Wong, Bang; Golub, Todd R. A Next Generation Connectivity Map: L1000 Platform and The First 1,000,000 Profiles. *bioRxiv*, **2017**, 136168.

[24] Edgar, R.; Domrachev, M.; Lash, A.E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic acids research*, **2002**, 30(1), 207-210.

[25] Temesi, G.; Bolgár, B.; Arany, Á.; Szalai, C.; Antal, P.; Mátyus, P. Early repositioning through compound set enrichment analysis: a knowledge-recycling strategy. *Future medicinal chemistry*, **2014**, 6(5), 563-575.

[26] Goss, K. L.; Gordon, D. J. Gene expression signature based screening identifies ribonucleotide reductase as a candidate therapeutic target in Ewing sarcoma. *Oncotarget*, **2016**, 7(39), 63003.

[27] Pessetto, Z.Y.; Chen, B.; Alturkmani, H.; Hyter, S.; Flynn, C.A.; Baltezor, M.; Ma, Y.; Rosenthal, H.G.; Neville, K.A.; Weir, S.J.; Butte, A.J.; Godwin, A.K. *In silico* and *in vitro* drug screening identifies new therapeutic approaches for Ewing sarcoma. *Oncotarget*, **2017**, 8(3), 4079.

[28] Lee, J.; Liu, J.; Feng, X.; Hernández, M.A.S.; Mucka, P.; Ibi, D.; Choi, J.W.; Ozcan, U. Withaferin A is a leptin sensitizer with strong antidiabetic properties in mice. *Nature medicine*, **2016**, 22(9), 1023-1032.

[29] Segura-Cabrera, A.; Tripathi, R.; Zhang, X.; Gui, L.; Chou, T. F.; Komurov, K. A structure-and chemical genomics-based approach for repositioning of drugs against VCP/p97 ATPase. *Scientific Reports*, **2017**, 7.

[30] Chandran, V.; Coppola, G.; Nawabi, H.; Omura, T.; Versano, R.; Huebner, E. A.; Zhang, A.; Costigan, M.; Yekkirala, A.; Barrett, L.; Blesch, A.; Michaelevski, I.; Davis-Turak, J.; Gao, F.; Langfelder, P.; Horvath, S.; He, Z.; Benowitz, L.; Fainzilber, M.; Tuszynski, M.; Woolf, C.J.; Geschwind, D.H. A systems-level analysis of the peripheral nerve intrinsic axonal growth program. *Neuron*, **2016**, 89(5), 956-970.

[31] Xiao, Z. X.; Chen, R. Q.; Hu, D. X.; Xie, X. Q.; Yu, S. B.; Chen, X. Q. Identification of repaglinide as a therapeutic drug for glioblastoma multiforme. *Biochemical and Biophysical Research Communications*, **2017**, 488(1), 33-39.

[32] Li, L.; Greene, I.; Readhead, B.; Menon, M.C.; Kidd, B.A.; Uzilov, A.V.; Wei, C.; Philippe, N.; Schroppel, B.; He, J.C.; Chen, R.; Dudley, J.T.; Murphy, B. Novel Therapeutics Identification for Fibrosis in Renal Allograft Using Integrative Informatics Approach. *Scientific Reports*, **2017**, 7.

[33] Won, S. J.; Yen, C. H.; Hsieh, H. W.; Chang, S. W.; Lin, C. N.; Huang, C. Y. F.; Su, C. L. Using connectivity map to identify natural lignan justicidin A as a NF-κB suppressor. *Journal of Functional Foods*, **2017**, 34, 68-76.

[34] Iorio, F.; Saez-Rodriguez, J.; Di Bernardo, D. Network based elucidation of drug response: from modulators to targets. *BMC systems biology*, **2013**, 7(1), p.139.

[35] Chen, J.; Schlitzer, A.; Chakarov, S.; Ginhoux, F.; Poidinger, M. Mpath maps multi-branching single-cell trajectories revealing progenitor cell progression during development. *Nature communications*, **2017**, 7.

[36] Trapnell, C.; Cacchiarelli, D.; Grimsby, J.; Pokharel, P.; Li, S.; Morse, M.; Lennon, N.J.; Livak, K.J.; Mikkelsen, T.S.; Rinn, J.L. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology*, **2014**, 32(4), 381-386.

[37] Zickenrott, S.; Angarica, V. E.; Upadhyaya, B. B.; Del Sol, A. Prediction of disease–gene–drug relationships following a differential network analysis. *Cell death & disease*, **2016**, 7(1), e2040.

[38] Liu, X.; Zeng, P.; Cui, Q.; Zhou, Y. Comparative analysis of genes frequently regulated by drugs based on connectivity map transcriptome data. *PloS one*, **2017**, 12(6), e0179037.

[39] Grammer, A.C.; Ryals, M.M.; Heuer, S.E.; Robl, R.D.; Madamanchi, S.; Davis, L.S.; Lauwerys, B.; Catalina, M.D.; Lipsky, P.E. Drug repositioning in SLE: crowd-sourcing, literature-mining and Big Data analysis. *Lupus*, **2016**, 25(10), 1150-1170.

[40] Siavelis, J. C.; Bourdakou, M. M.; Athanasiadis, E. I.; Spyrou, G. M.; Nikita, K. S. Bioinformatics methods in drug repurposing for Alzheimer's disease. *Briefings in bioinformatics*, **2015**, 17(2), 322-335.

[41] Sirota, M.; Dudley, J.T.; Kim, J.; Chiang, A.P.; Morgan, A.A.; Sweet-Cordero, A.; Sage, J.; Butte, A.J. Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Science translational medicine*, **2011**, 3(96), 96ra77.

[42] Kidd, B.A.; Wroblewska, A.; Boland, M.R.; Agudo, J.; Merad, M.; Tatonetti, N.P.; Brown, B.D.; Dudley, J.T. Mapping the effects of drugs on the immune system. *Nature biotechnology*, **2016**, 34(1), 47-56.

[43] Ryan, N.; Chorley, B.; Tice, R. R.; Judson, R.; Corton, J. C. Moving Toward Integrating Gene Expression Profiling into High-throughput Testing: A Gene Expression Biomarker Accurately Predicts Estrogen Receptor α Modulation in a Microarray Compendium. *Toxicological Sciences*, **2016**, 151(1), 88-103.

[44] Fang, H.Y.; Zeng, H.W.; Lin, L.M.; Chen, X.; Shen, X.N.; Fu, P.; Lv, C.; Liu, Q.; Liu, R.H.; Zhang, W.D.; Zhao, J. A network-based method for mechanistic investigation of Shexiang Baoxin Pill's treatment of cardiovascular diseases. *Scientific Reports*, **2017**, 7.

[45] Chen, B.; Wei, W.; Ma, L.; Yang, B.; Gill, R.M.; Chua, M.S.; Butte, A.J.; So, S. Computational Discovery of Niclosamide Ethanolamine, a Repurposed Drug Candidate That Reduces Growth of Hepatocellular Carcinoma Cells *in Vitro* and in Mice by Inhibiting Cell Division Cycle 37 Signaling. *Gastroenterology*, **2017**, 152(8), 2022-2036.

[46] Zhang, X.; Li, L.; Ng, M. K.; Zhang, S. Drug-target Interaction Prediction by Integrating Multiview Network Data. Computational *Biology and Chemistry*. **2017**.

[47] Knox, C.; Law, V.; Jewison, T.; Liu, P.; Ly, S.; Frolkis, A.; Pon, A.; Banco, K.; Mak, C.; Neveu, V.; Djoumbou, Y.; Eisner, R.; Guo, A.C.; Wishart, D.S. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic acids research*, **2010**, 39(suppl_1), D1035-D1041.

DISCLAIMER: The above article has been published in Epub (ahead of print) on the basis of the materials provided by the author. The Editorial Department reserves the right to make minor modifications for further improvement of the manuscript.

PMID: 30381060