Briefings in Bioinformatics, 22(4), 2021, 1–11

https://doi.org/10.1093/bib/bbaa348
Problem Solving Protocol

Computational drug repositioning based on the relationships between substructure–indication

Jingbo Yang[†], Denan Zhang[†], Lei Liu[†], Guoqi Li, Yiyang Cai, Yan Zhang, Hongbo Jin and Xiujie Chen

Corresponding authors: Xiujie Chen, College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150001, China. Tel.: +86-451-86352422. E-mail: chenxiujie@ems.hrbmu.edu.cn; Hongbo Jin, Department of Physiology, Harbin Medical University, Harbin 150001, China. E-mail: kinga@126.com.

[†]These authors contribute equally to this work.

Abstract

At present, computational methods for drug repositioning are mainly based on the whole structures of drugs, which limits the discovery of new functions due to the similarities between local structures of drugs. In this article, we, for the first time, integrated the features of chemical-genomics (substructure–domain) and pharmaco-genomics (domain–indication) based on the assumption that drug–target interactions are mediated by the substructures of drugs and the domains of proteins to identify the relationships between substructure–indication and establish a drug–substructure–indication network for predicting all therapeutic effects of tested drugs through only information on the substructures of drugs. In total, 83 205 drug–indication relationships with different correlation scores were obtained. We used three different verification methods to indicate the accuracy of the method and the reliability of the scoring system. We predicted all indications of olaparib using our method, including the known antitumor effect and unknown antiviral effect verified by literature, and we also discovered the inhibitory mechanism of olaparib toward DNA repair through its specific sub494 (o = C–C: C), as it participates in the low synthesis of the poly subfunction of the apoptosis pathway (hsa04210) by inhibiting the Inositol 1,4,5-trisphosphate receptor(s) (ITPRs) and hydrolyzing poly (ADP ribose) polymerases. ElectroCardioGrams of four drugs (quinidine, amiodarone, milrinone and fosinopril) demonstrated the effect of anti-arrhythmia. Unlike previous studies focusing on the overall structures of drugs, our research has great potential in the search for more therapeutic effects of drugs and in predicting all potential effects and mechanisms of a drug from the local structural similarity.

Key words: drug reposition; substructure–domain associations; local structural similarity

Introduction

Drug repositioning or repurposing is a process of finding new indications for existing drugs, and it has become a promising approach for drug research and development (R&D), because it can reduce the risk of failure in phases II and III of clinical investigation due to drug ineffectiveness, severe adverse drug reactions and poor dynamic characteristics [1–6]. At present, the screening of drug repositioning is mainly performed by means of experimental *in vivo* or *in vitro* models, but because of the limits of experimental methods such as the high cost, lengthy duration

Jingbo Yang is a Ph.D. student in College of Bioinformatics Science and Technology, Harbin Medical University.
Denan Zhang is a lecturer in College of Bioinformatics Science and Technology, Harbin Medical University.
Lei Liu is an associate professor in College of Bioinformatics Science and Technology, Harbin Medical University.
Guoqi Li is a research student in College of Bioinformatics Science and Technology, Harbin Medical University.
Yiyang Cai is a student in College of Bioinformatics Science and Technology, Harbin Medical University.
Yan Zhang is a research student in College of Bioinformatics Science and Technology, Harbin Medical University.
Hongbo Jin is a professor in Department of Physiology, Harbin Medical University.
Xiujie Chen is a professor in College of Bioinformatics Science and Technology, Harbin Medical University.

Submitted: 21 August 2020; **Received (in revised form):** 13 October 2020

and low efficiency [4]. Computational approaches to drug repositioning have become more and more important in the age of big data [7, 8]. Association analysis and similarity searching as common computational strategies including machine learning, network analysis and text mining, etc. [9] have been used to identify new indications for approved drugs [10, 11]. They are based on the assumption that the structures of drugs determine their functions and phenotypes. For example, the research of drug action and reposition by Lorio *et al.* [12] was based on the hypothesis that drugs with similar gene expression patterns will show similar pharmacological effects. The study exploring drug repurposing by Lee *et al.* [13] is based on the suggestion that if drugs have the same or similar side effect profiles, the drugs may have the potential to treat the same diseases profile. Similarly, methods based on phenotypic data (e.g. drug–disease relationships), or 'guilt by association', assumed that if the diseases are similar, then the drugs used for treating these diseases can be shared [14].

Although researchers have made many contributions to finding novel drugs and new indications, there still remains room for improvement because the whole structure of a drug does not always coincide with its function and cannot reflect the specific binding properties mediated by the functional groups of the drug and the active pockets of the protein. For example, Yildirim *et al.* [15] showed that most of the drugs that target the same proteins have distinct chemical structures. Keiser *et al.* [16] demonstrated that structurally similar drugs might bind proteins with dissimilar functions. Therefore, it is suggested that a strategy based on the binding between the functional group of a drug and the active pocket of a protein can better express the local structural similarity of drugs and discover drugs with novel structures. Some researchers have made some associated attempts. For example, Wang *et al.* developed a new method called predicting drug targets with domains to predict potential target proteins for drug repositioning based on the interaction between a drug and the active pockets of a protein, and the results show that it is superior to methods using the interaction between the whole structures of proteins and drugs to discover new drug targets [17]. Song Peng Zu *et al.* established the model of global optimization-based inference of chemogenomic features to identify drug–target interactions (GIFT) from the perspective of the interactions between the substructures (representing active groups) of drugs and domains (representing active pockets) of proteins and found that GIFT had a better interpretive performance and a higher drug–target verification rate compared to previous methods [18]. Yoshihiro Yamagishi's team did a domain–adverse reaction association study to predict new adverse reactions and their mechanisms [19].

The above information suggests that methods based on interactions between the functional group of a drug and the active pocket of a protein can better express the local structural similarity of the drug. However, current calculation methods using functional groups and active pockets have two limits. First, both the potential drug–target and drug-therapeutic effects relationships are predicted from a single level of chemical-genomics (substructure–domain) or pharmaco-genomics (domain–indication) [19]. Second, the purposes of these studies is to discover novel drug–target or indication–targets relationships rather than substructures–indication associations, which are of greater value for drug repositioning. Associations of substructure–indication are core to predicting the therapeutic effects of candidate drugs by requiring only the substructure information of the drugs.

Therefore, in this article, we integrated the known relationships of substructure–drug, indications–drug, drug–target and protein–domain to perform the following tasks based on the correlation analysis and network analysis. First, supervised machine learning was used to identify the substructure–domain and indication–domain relationships. Then, a substructure–domain–indication complex network was constructed, and significant correlations between substructure and indication were identified and quantified. Finally, according to the substructure information of the drugs and based on our scoring system, we ranked all indications of every drug and analyzed the mechanisms of indications from the perspective of local structures.

Method

Data

Drug–target

The interaction data of drugs and target proteins were obtained from the DrugBank database (5.1.0 version) [20]. Only approved drugs and Uniprot-linked target proteins were selected, which included 2021 drugs, 2670 targets and 9797 drug–target pairs.

Drug–substructures

The chemical substructures of the drugs were obtained from the PubChem database (2018_05version) [21]. In total, 1781 out of 2021 drugs had substructure information. Finally, 653 substructures were mapped to 1781 drugs.

Target–domains

The domains of target proteins were obtained from the Uniprot database (2018_03version) [22], in which domain information for 2431 target proteins was found. A total of 4430 target–domain relationships was obtained, including 2431 targets and 1555 domains.

Drug indications

In this article, drug indications are represented by the third level of the ATC codes of the drugs. Drugs with ATC–encoded indications in the DrugBank database were extracted by the advanced search. In total, 4025 drug–indication relationships were obtained, including 1726 drugs and 216 indications.

In a word, a total of 8969 drug–target interactions of substructure–drug–target–domain relationships were obtained, including 1781 drugs with 653 substructures and 2431 target proteins with 1555 domains. A total of 8053 drug–target interactions of indication–drug–target–domain relationships was obtained, including 1726 drugs with 216 indications and 1991 target proteins with 1356 domains.

Binary vector representation of data

With reference to the approaches of Jacob *et al.* [23] and Yasuo *et al.* [24], we represented the collected features in the form of binary vectors and tensor products. Assume that there is a set of drug–target relationships $(D\ {}^*T)$, and a drug–target relationship represented by (d, t) where $d = 1, \ldots D, t = 1, \ldots T$. In the set of D drugs, if there are S substructures, a drug d can be represented by a S-dimensional binary vector, $\phi(d) = (a_1, a_2, \ldots, a_S)^{\mathrm{T}}$ where $a_i 0, 1, i = 1, \ldots, S$. Similarly, in the set of T targets, if there are U domains, a target protein t can be represented by a U-dimensional binary vector, $\phi(t) = (b_1, b_2, \ldots, b_U)^{\mathrm{T}}$ where

$b_i 0, 1, i = 1, \dots, U$. By combining the two binary vectors, any drug–target relationships(d, t) can be represented by the substructure–domain associations

$$\phi\left(d, t\right) = \phi(d) \otimes \phi(t) = \left(a_1 b_1, \dots, a_1 b_U, \dots, a_S b_1, \dots, a_S b_U\right)^{\mathrm{T}} \quad (1)$$

Additionally, a drug–target relationship (d, t) can be represented by the indication–domain associations in a similarly way.

Selection of binary classifiers

The L1 regularized logistic regression (L1LOG) and L1 regularization support vector machine (L1SVM) are suitable for large-scale and sparse data matrixes due to overcoming the polynomial complexity of their kernel [25]. In this article, we utilized the two methods of the jar package LIBLINEAR, which have a higher accuracy performance in the case of our data set [24]. The process of the optimization model is shown in the parameter selection of the machine learning model in Supplementary Materials.

Identifying and weighting the features of substructure–domain and indication–domain

For each feature, we obtained the weighted score $W_{f_i^1}$ for the method of L1LOG and $W_{f_i^2}$ for the method of L1SVM, respectively. To increase the reliability of the results, we have selected the overlapping set of each feature between the two methods and recalculated the weight value according to the rank of the two-weight score of each feature for the two methods. The calculation method of the weight value is as follows:

$$W\left(f_i\right) = \frac{r_1 + r_2}{2N} \quad (2)$$

where r_1 and r_2 are the rank number of the two weighted scores $W_{f_i^1}$ and $W_{f_i^2}$ of each feature f_{SD} or f_{ID}, respectively, and N is the number of common features identified by the two methods. Therefore, each feature has a weight value between [0,1] to evaluate the credibility of this feature. The higher the weight value$W(f_i)$, the more reliable is f_{SD} or f_{ID}.

Extracting and quantifying the feature of substructure–indication

We hypothesized that if there is at least one shared domain between f_{SD} and f_{ID}, then there is a correlation between the substructure in f_{SD} and the indication in f_{ID}. We obtained the substructure–indication associations using the domain as a bridge according to the correlation scores S_{ij}, for which the formula is shown below, to evaluate the strength of each substructure–indication association.

$$S_{ij} = \sum\nolimits_{i=1}^{n} W\left(f_{SD}\right) W\left(f_{ID}\right) \quad (3)$$

Here, n represents the number of domains shared by the substructure i and the indication j. $W(f_{SD})$ and $W(f_{ID})$ represent the weight value of the substructure–domain f_{SD} and indication–domain f_{ID}.

Prediction indications of drugs

This section includes four processes.

Firstly, the specificity of the substructures was identified. It was found that rare substructures (herein, also called the specific substructures distributed in fewer drugs) are often more informative compared to substructures that are widely present in a large number of drugs [3]. To highlight the contributions of specific substructures of drugs to the therapeutic effect and avoid the influence of these substructures (which exist widely in most drugs) on the specificity of drug indications, we supposed that the stronger the specificity of the substructure in the drugs, the greater is the contribution of the substructure to the substructure–indication relationships based on the below formula for calculating the contribution of each substructure to the indication

$$C_k = \exp\left(-{f_k}^2/\sigma^2 h^2\right) \quad (4)$$

where f_k represents frequency of occurrence of substructure k in the drug set. σ represents the standard deviation of f_k for all substructures. In addition, h is a control parameter, which, here, we take as 5.

Secondly, effective therapeutic effects of drugs were predicted. For each drug, we obtained the indication list with different correlation scores through calculating all the correlation scores of the corresponding substructure–indication relationships. To extract significant relationships (called effective relationships) between drugs and indications, we downloaded known drug–indication interactions from DrugBank, and for any drug–indication relationship, we counted all correlation scores of the corresponding substructure–indication relationships according to the following formula:

$$R_{ij} = \sum_{k=1}^{k=Q} S_{kj} * C_k \quad (5)$$

where R_{ij} represents the ith drug's predictive association score for the jth indication. k represents the kth substructure of the drug that has Q shared substructures with the jth indication. S_{kj} is the correlation score between substructure k and indication j (according to the method of extracting and quantifying the feature of substructure–indication). C_k represents the score for calculating substructure k's specific weight value.

For each indication I_j of a drug, we supposed that when its R_{ij} is greater than or equal to the median of the R_{ij} scores corresponding to the real drug, indication I_j is significant and is also an effective indication of the drug.

Thirdly, the contribution ranks of drugs for an indication were identified. We used the formula Score_{ATC} to evaluate the contribution of drug i to indication j. The higher the value of Score_{ATC}, the stronger is the likelihood that the drug will produce the indication j:

$$\mathrm{Score}_{ATC} = \frac{R_{ij}}{\sum_{k=1}^{k=P} S_{kj} * C_k} \quad (6)$$

where P is the number of substructures which are associated with indication j by predicted in method of extracting and quantifying the feature of substructure–indication. S_{kj} represents the correlation score between substructure k and indication j. C_k represents the score for calculating substructure k's specific weight value.

Fourthly, the ranks of indications for a drug were identified. We used the formula Score_{Drug} to evaluate contribution of

Table 1. The number of two kinds of features extracted by two methods

Features	L1LOG	L1SVM	Overlap	Significantly correlated associations
Substructure–domain	142 006	125 547	111 560	1131
Indication–domain	5829	5703	5652	2788

indication j to drug i.

$$\mathrm{Score}_{\mathrm{Drug}} = \frac{R_{ij}}{\sum_{j=1}^{j=T} R_{ij}} \tag{7}$$

Here, T is the total number of indications for a drug.

Finally, in order to obtain the significant drug–indication pairs, we use the product value ($\mathrm{Score}_{\mathrm{rank}}$) of $\mathrm{Score}_{\mathrm{ATC}}$ and $\mathrm{Score}_{\mathrm{Drug}}$ to measure the strength of the relationship between an indication and a drug:

$$\mathrm{Score}_{\mathrm{rank}} = \mathrm{Score}_{\mathrm{ATC}}\,{}^{*}\mathrm{Score}_{\mathrm{Drug}} \tag{8}$$

Results

Feature of substructure–domain and indication–domain

We used two methods (L1LOG and L1SVM) to extract the features of substructure–domain associations based on the known drug–target interactions. According to the same rule, the indication–domain associations were obtained. The overall extraction of the two features is shown in Table 1.

We only selected the substructures with the highest weight value $W(f_i)$ corresponding to each domain as the effective correlated domain–substructure associations because we considered that the effect of each domain is mainly induced by the substructure that is most closely related to it and defined these effective domain–substructure and domain–indication as significantly correlated when their weight values $W(f_i)$ were higher than their respective average values. Finally, we obtained 1131 substructure–domain associations with significantly correlated weight values, including 241 substructures and 1131 domains (Supplementary File 1). Additionally, we obtained 2788 indication–domain associations with significantly correlated weight values, composed of 1087 domains and 205 indications (Supplementary File 2).

Substructure–indication features

We assigned the specific weight values C_k of substructures to association scores between a substructure and an indication as the final association score of each substructure–indication association. Finally, 1748 substructure–indication associations were obtained, including 227 substructures and 205 indications (Supplementary File 3). The higher the association score, the more we believe that the substructure is related to the indication.

We found that different substructures provide different contributions to an indication. We selected substructures with higher-than-average scores associated with the indication, indicating significant contribution substructures, which were also defined as the dominant substructures of the indication. In the same way, it was also defined that the substructure was associated with only one indication as a specific substructure of the indication. The same definition applies to the dominant indications or the specific indications of a substructure. For example, it is obvious that P01B is the dominant indication of SUB189 (shown in Figure 1B). In addition, SUB62, SUB368 and SUB189 are the obvious dominant substructures of P01B (shown in Figure 1A).

The relationships between drugs and indications

We mapped drug–substructure relationships to these predicted indication–substructure associations. For each indication, we designed a relational score $\mathrm{Score}_{\mathrm{rank}}$ to assess the contribution of each drug to the indication to find repositioning drugs or the dominant drugs for the indication in this article (Supplementary File 4). We obtained 83 205 significant drug–indication pairs including 1479 drugs and 178 indications. Based on these significant interactions, a drug–substructure–indication network was used to analyze the complex relationships of drug–indication. As shown in Table 2, for P01B (antimalarial drugs), the six dominant drugs are artenimol, pivmecillinam, oxygen, eribulin, stiripentol and glucosamine, and specific substructures related to P01B are also shown.

Verification of drug–indication relationships

The predicted drug–indication relationships were verified in four aspects:

Validation based on the DrugBank database

We first used 2156 known drug–ATC relationships recorded in the DrugBank to verify our predicted 83 205 drug–ATC associations. Of the 2156 drug–ATC relationships of the 1479 drugs in DrugBank database, 1464 pairs were predicted by this method, and the coverage rate of the DrugBank databases was 67.90%.

Meanwhile, we checked the coverage rate of the relationship of the top 10 of our predicted results in Drugbank. In total, 610 drugs (accounting for 41.19% of the total predicted drugs) and their top 10 ATCs (at least one ATC in the top 10) were confirmed in DrugBank.

Literature validation based on PubMed

To extend the hit rate in the literatures of PubMed, we manually replaced the ATC in each pair of drug–ATC with one or several related keywords, such as C03E (combination of diuretic and potassium-sparing drugs) corresponding to potassium and diuretic. All corresponding ATC keywords are shown in the Supplementary File 5. If a drug and the corresponding keywords could be matched in the abstract of a document at the same time, this meant that the drug–ATC relationship was verified by the literature. The overall literature verification rate for the 83 205 drug–ATC relationships was 71.44%.

The verification rates of the top 1 to top 100 ATCs corresponding each drug are shown in Figure 2A. As shown in Figure 2, the higher the prediction position (those with larger scores), the higher is the verification rate, which shows the credibility of our scoring system. Similarly, the verification rates of the top 1 to top 100 drugs corresponding to each ATC are shown in Figure 2B to have the similar trend.

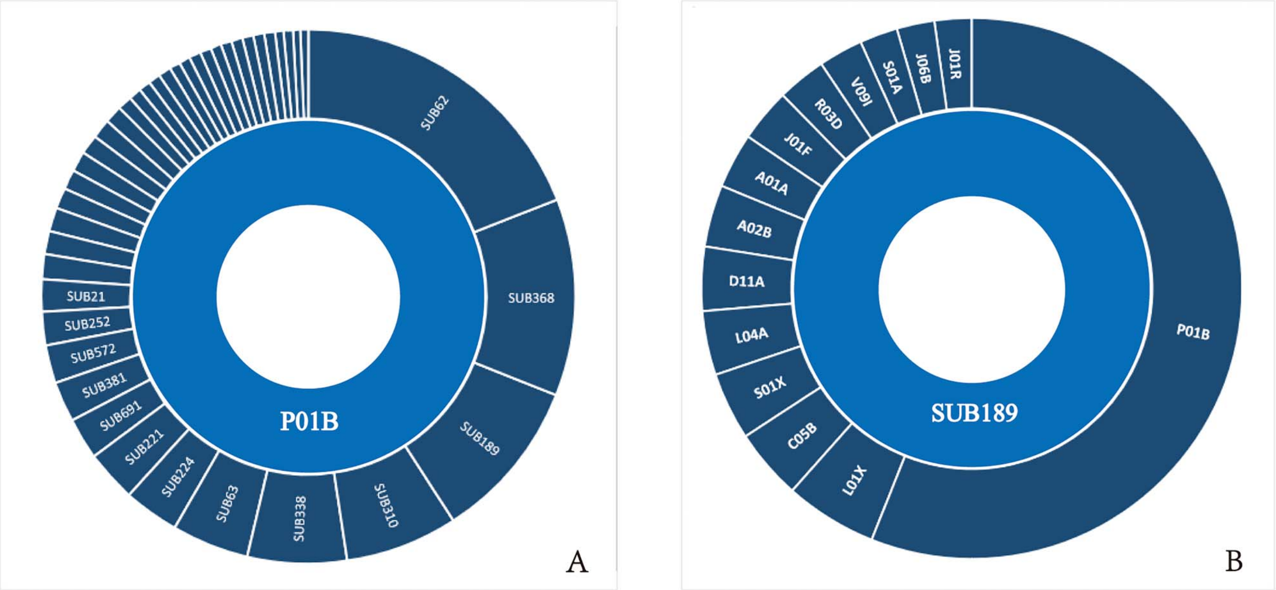

Figure 1. (**A**) The distribution of P01B-substructure correlation scores. (**B**) The distribution of Sub189-indication correlation scores.

Table 2. The predicted drugs and their dominant substructures associated with P01B

Predicting drugs related to P01B	Score ranks of drug-P01B	Dominant substructures related to P01B	Correlation scores S_{ij} of substructure-P01B
Artenimol	0.2400	SUB310(O–O)	8.8250
		SUB368(C(~H)(~O)(~O))	5.8623
		SUB224(≥2 saturated or aromatic heteroatom-containing ring size 7)	4.3880
		SUB221(≥2 any ring size 7)	4.0188
Pivmecillinam	0.1273	SUB368	5.8623
		SUB224	4.3880
		SUB221	4.0188
Oxygen	0.0904	SUB310	8.8249
Eribulin	0.0674	SUB224	4.3880
		SUB221	4.0188
		SUB217(≥1 saturated or aromatic heteroatom-containing ring size 7)	0.8659
Stiripentol	0.0635	SUB368	5.8623
Glucosamine	0.0634	SUB368	5.8623
...			
Simeprevir	1.73E-06	SUB252(≥1 saturated or aromatic heteroatom-containing ring size 10)	0.0327
		SUB686(O = C–C–C–C–N)	0.0083
		SUB701(O–C–C–C–C–C–O–C)	0.0074
Ketotifen	1.46E-06	SUB252	0.03268
		SUB533(S–C:C–[#1])	0.0105

Verification based on the CTD database

The drug–disease relationships recorded in the indications of our predicted 1479 drugs were obtained by the crawler. We combined the drug–ATC relationships recorded in DrugBank database with the drug–disease relationships obtained by the crawler to form the associations of ATC–disease.

We transformed the drug–disease relationships of the curated tags in the CTD database into 44 254 drug–ATC relationships, and compared them with the 83 205 drug–ATC relationships of the 973 drugs that we predicted. There were 13 541 (approximately 30.60%) coinciding drug–ATC relationships between them.

After removing the common drug–ATC pairs of the above three verification methods, 60 766 drug–ATC associations out of the 83 205 that we predicted were verified, with a total verification rate of 73.03%.

Comparison with three other methods

Furthermore, we compared this method with three other methods, which are the predictions of the first level ATC code, the second level ATC code and the third level ATC code of drugs. We took the coverage rate of the predicted result to the drug–ATC relationships in DrugBank database as the benchmark for

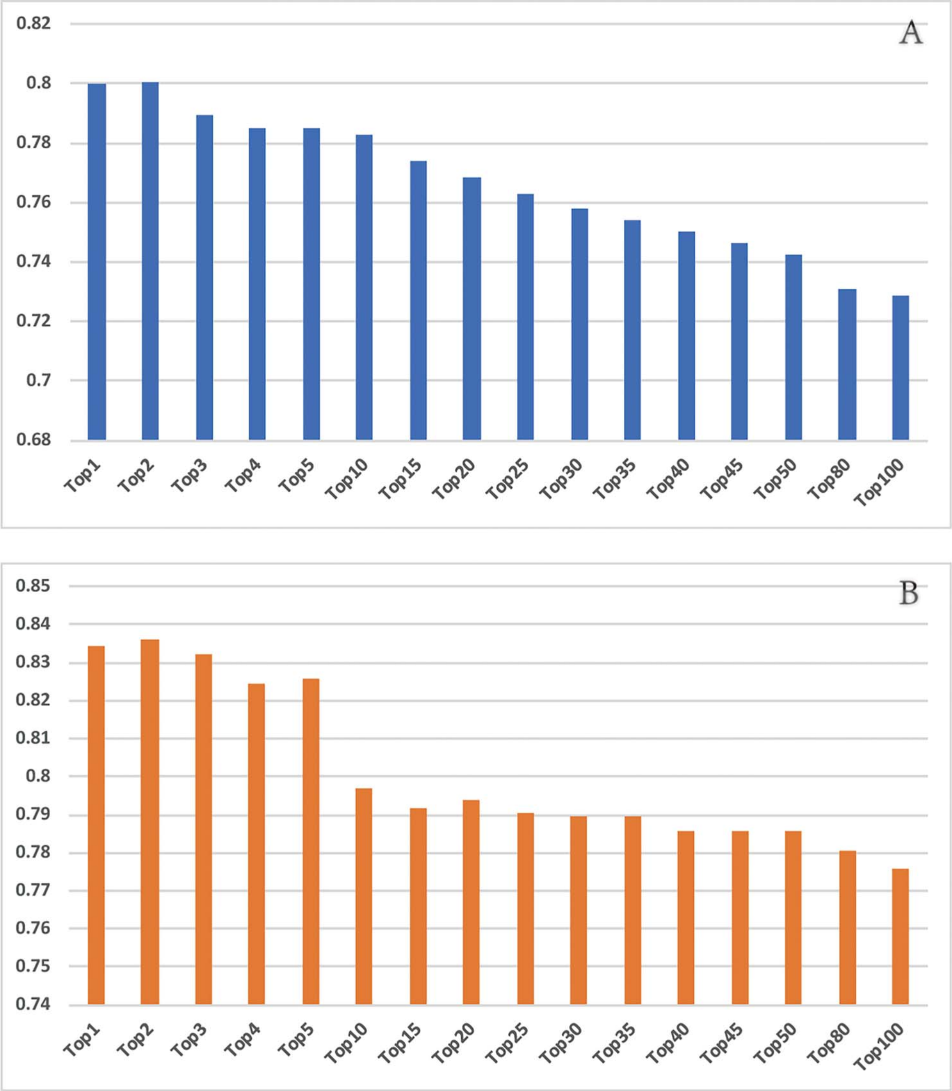

Figure 2. **(A)** The literature verification rates of the Top100 ATCs corresponding each drug. **(B)** Verification rates of the Top100 drugs corresponding each indication.

the performance of the method. Our method has obvious advantages in the prediction accuracy of the first level ATC and the second level ATC, while our method is inferior to the SPACE method in the third level ATC prediction. Although our method is lower than SPACE method in coverage, our method can not only predict the main action spectrum by only require the substructure information of a compound, but also know the specific structures of therapeutic effects, and can quantitatively the strength of the drugs with one indication, so that we can find the most suitable drug for the indication, as well as its influence function. See the detailed information of the comparison of our method with other three methods [26–28] in comparison to other drug prediction methods in Supplementary Materials.

Case analysis

Predicting indications of olaparib

Olaparib is a known oral inhibitor of poly (ADP ribose) polymerase (PARP) protein, which plays a key role in DNA repair and genome stability [29]. In this article, based on drug–substructure–indication associations and drug–substructure–domain–protein extended associations, we predicted 61 indications and 420 targets of olaparib. The known antineoplastic and immunomodulatory effect recorded in DrugBank database are ranked 29 in our predicted list of Olaparib. The lowest interaction intensity score is 0.92, and the highest interaction intensity score is 3.90. A total of 11 targets (shown in Table 3) was enriched in the apoptotic pathway of KEGG (hsa04210) [30], which is considered the mechanisms of the antitumor effect.

Here, we mainly focus on the top 6 targets out of the 11 targets (ITPR1, ITPR2, ITPR3, PARP1, PARP2 and PARP3) for further analysis, and these targets participate in the low synthesis of poly subfunctions in the pathway hsa04210, as shown in Figure 3A. In our prediction, SUB494 (o = C–C: C) of olaparib binds to the receptors of inositol 1,4,5-triphosphate on the calcium channel of the endoplasmic reticulum. This may accelerate the release of calcium ions, leading to the activation of the caspase cascades [31, 32]. Additionally, SUB494 (o = C–C: C) of olaparib targets PARPs, which inhibits repairing the damaged DNA strands to cause apoptosis by hydrolyzing of PARPs (as shown in Figure 3B).

Table 3. Information of 11 targets of olaparib enriched in the apoptotic pathway

Target	Gene name	Strength score	Related substructure
Q14643	ITPR1	3.90	SUB494
Q14571	ITPR2	3.90	SUB494
Q14573	ITPR3	3.90	SUB494
Q9UGN5	PARP2	2.95	SUB494
P09874	PARP1	2.95	SUB494
Q9Y6F1	PARP3	2.95	SUB494
Q04206	RELA	1.96	SUB673
P05412	JUN	1.00	SUB19
P31749	AKT1	1.00	SUB19
P19838	NFKB1	1.00	SUB673, SUB624
P25963	NFKBIA	1.00	SUB624

This is consistent with the fact that olaparib destroys DNA repair and replication pathways by inhibiting PARP1, PARP2 and PARP3 [33].

Meanwhile, the dominant ATC corresponding to olaparib was J05A (antiviral agent). There is no record of the indications of olaparib in the DrugBank database. About 9 of the 420 predicted targets of olaparib were enriched in the B cell receptor signaling pathway (hsa04662), and 12 targets were enriched in the T cell receptor signaling pathway (hsa04660). The locations of these targets in the two pathways are shown in Figure 4A and B. These two pathways have been proven to be involved in antiviral activity [34–37] through regulating B cell ontogenetic autoimmunity and proliferation in hsa04662 and the differentiation immune response in hsa04660.

Prediction of antiarrhythmic drugs

Cardiac arrhythmias are a major health problem associated with reduced quality of life and substantial morbidity and mortality. Therefore, researchers would like to know which substructures of drugs can antagonize this effect. Based on the predicted drug–substructure–domain–indication relationships, we obtained 675 drugs used against cardiac arrhythmias with different relational scores (selected from Supplementary File 4). We selected amiodarone, quinidine, milrinone and fosinopril from the predicted antiarrhythmic drugs (C01B) for verification using ElectroCardioGrams (ECG; see the experimental validation of antiarrhythmic drugs in the Supplementary Materials for detailed information). Among them, amiodarone and quinidine are drugs that have been proven to have a C01B effect in the DrugBank database, while milrinone is a positive inotropic cardiotonic agent with vasodilator properties. Fosinopril is used to treat mild to moderate hypertension, and as an adjunct treatment for congestive heart failure.

In the study, through excessive administration of ouabain, an abnormal ECG was induced in rats, mainly involving ventricular arrhythmia, including ventricular premature beat, continuous ventricular premature beat, ventricular premature bigeminy, short array ventricular tachycardia, bidirectional ventricular tachycardia and ventricular fusion of short ventricular tachycardia. After observing and recording the abnormal ECG, quinidine, amiodarone, fosinopril and milrinone were given intraperitoneally. The ECG of the rats gradually returned to normal, showing a sinus heart rate and regular rhythm. The results showed that these drugs had antiarrhythmic effects (Supplementary Figure 2 in Supplementary Materials).

In this article, we further explored the structural basis of the antiarrhythmic effects of the four drugs. According to the correlation scores of the substructure–indication associations (selected from Supplementary File 3), we determined that SUB410 (O(:C)(:C)), SUB182 (≥1 saturated or aromatic heteroatom-containing ring size 6), SUB673(O = C–C = C–[#1]) and SUB4(≥32 H) are the dominant substructures of amiodarone, quinidine, milrinone and fosinopril for antiarrhythmic therapy, respectively. These specific substructures have strong practical significance. First, they represent that drugs containing these specific substructures may have antiarrhythmic effects, which are not limited to the predicted drugs, further expanding the space for drug discovery. Second, in future drug rationalization design, the specificity of drug substructures can be used to retain the therapeutic substructures, reduce the substructures that produce drug side effects, and optimize the function of the drug.

Discussion and conclusion

In studies of computational drug repositioning, few researchers have integrated chemical-genomics features (substructure–domain) with the pharmaco-genomic features (domain–indication) to identify the substructure–indication and final drug–indication relationships to predict all indications of a drug through only the substructure information of the drug.

In this article, we took the lead in establishing drug–indication relationships by combining the substructures of drugs, the domains of proteins, the indications of drugs and drug–target interactions, providing a novel strategy for computational drug repositioning from the substructural level to highlight the local structural similarity.

The strategy not only can illustrate the relationships between substructures and indications but also provide a scoring system to evaluate the strength of each association. According to the substructure information of a test drug and our scoring system, we can find all indications and the dominant indications of the drug and provide a mechanistic explanation of the predicted indications. The traditional methods for predicting a drug phenotype are generally involve finding the target proteins of the drug, and the indications are inferred by the corresponding bioactivities derived from the functions of the target proteins that interact with the drug. However, these methods have the limitation that they cannot explain the specific linking mechanisms between drugs and proteins because of using their whole structure. For example, it is impossible to explain the phenomena in which a protein can interact with multiple drugs that are not similar in terms of whole structures or one drug can interact with multiple proteins that are not evolutionarily similar because this is a manifestation of the specific connections

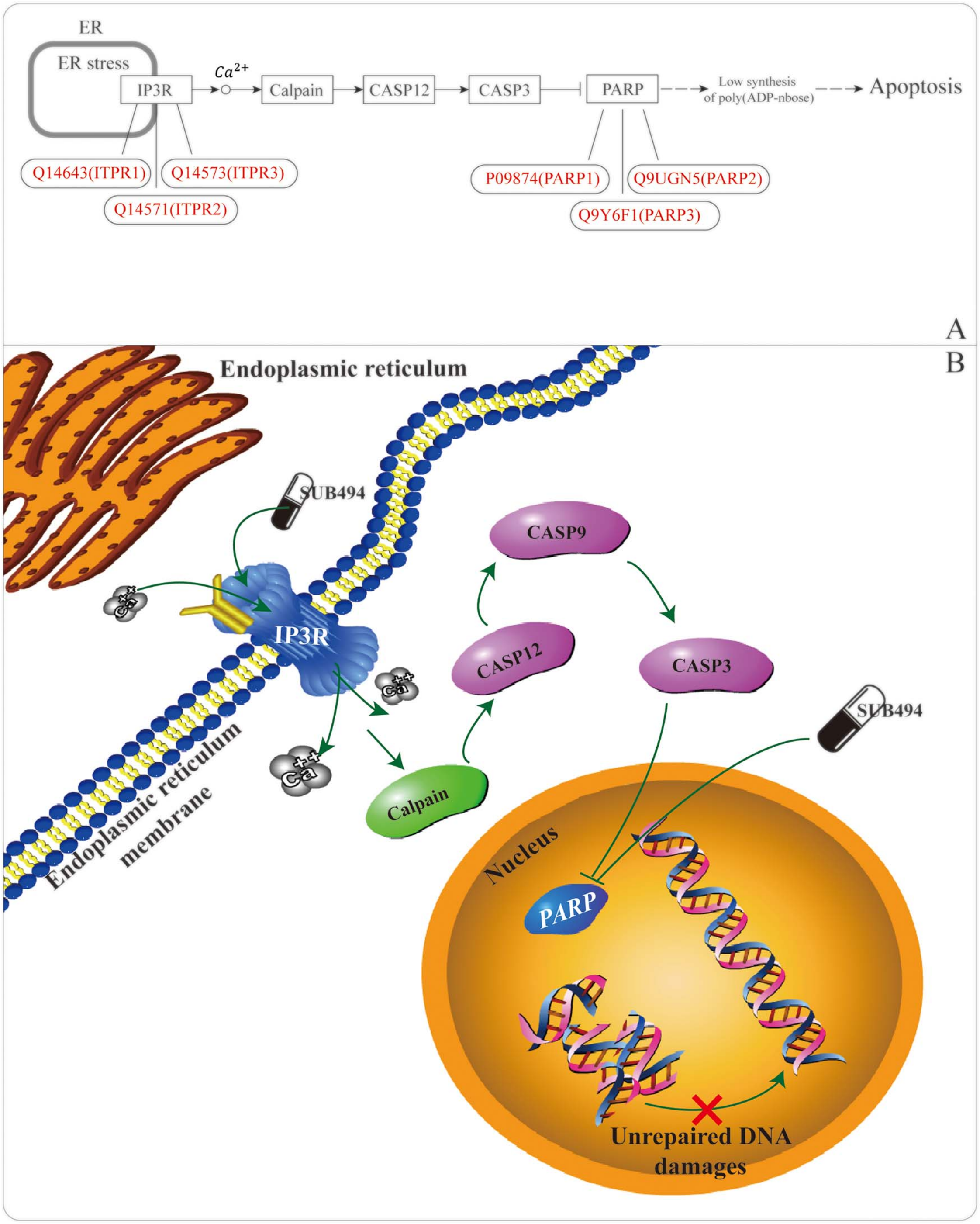

Figure 3. (**A**) The location of the Top6 targets of olaparib which enriched in the apoptotic pathway of KEGG (hsa04210). (**B**) The process of SUB494 (o = C-C: C) of olaparib inhibiting repairing the damaged DNA strand and causing apoptosis by hydrolyzing of PARPs.

between drug substructures and protein domains. The relationship between the drug's substructure and the protein's domain is the essential reason for the interaction between a drug and a target. It is also the most primitive driving force for the overall response of a drug to the body.

The relationships between structure and function can be better reflected from the local perspective. For this reason, the strategy adopted in this article is different from the strategies used in previous studies. For example, compared with the existing computational methods for predicting drug repositioning, we find that our method can help us obtain multiple protein-related drugs with different overall structures by exploring the similarities of the local structures. As seen from our results, we could not only find drugs with similar functions and structures, but more importantly, we could also find drugs with completely different macroscopic structures that act on the same protein spectrum

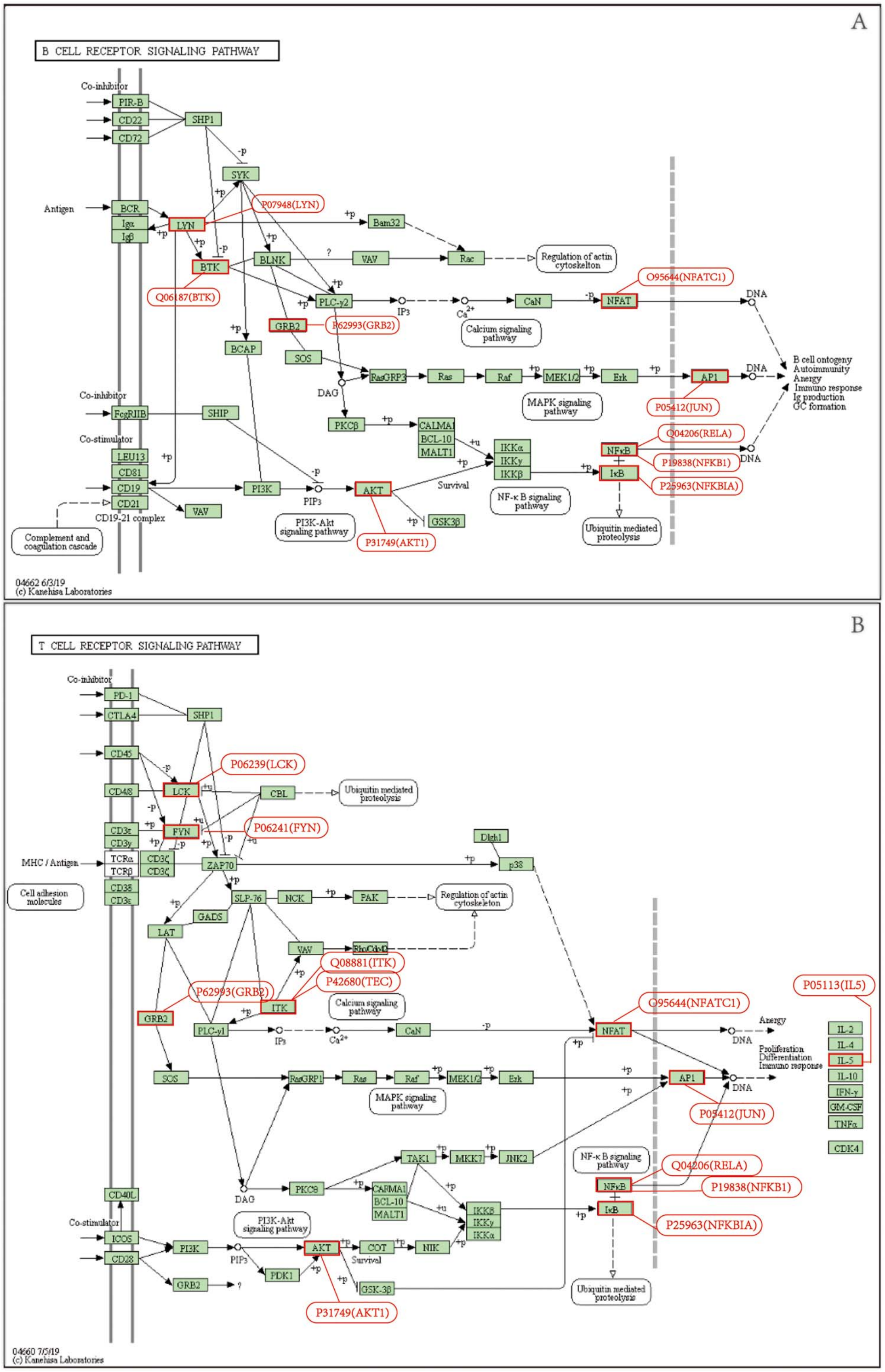

Figure 4. The location of targets of olaparib in B cell receptor signaling pathway (hsa04662) and T cell receptor signaling pathway (hsa04660). The green genes in red boxes are the targets that enriched to these two pathways.

(e.g. milrinone and fosinopril). Substructures and domains are distributed across different compounds and proteins (for example, each substructure is, on average, distributed in 122 drugs, and each domain is distributed, on average, in 1.86 proteins); therefore, our method greatly expands the scope of searching for drugs and indications, overcoming the difficulty of identifying drugs that are not similar in overall structures but have the same functions.

Despite the above advantages, there are many places to improve in the article. First, it is restricted by the number of drug substructures and protein domains. It is not suitable for the prediction of drugs with no substructures and proteins without domains. The number of substructure–domain features also confines the implementation of this research. Second, this method only defines the interaction strength of each substructure–domain associations, but does not outline their effect attributes (e.g. activated effects or suppressed effects). However, these problems will be solved with the development of structural chemistry and structural biology and molecular pharmacology.

Key Points

- A novel strategy for repositioning drugs by highlighting the local similarity.
- The strategy can quantify the strength of the relationships of drug–indication and has the potential to find new drugs with novel structures.
- The strategy can provide specific substructures for an indication.
- The strategy can explain the mechanism of drug–indication relationships from a more fundamental perspective (drug substructures and target domains).

Supplementary data

Supplementary data are available online at https://academic.oup.com/bib.

Funding

This work was supported by the National Natural Science Foundation of China (grant nos. 61671191, 61971166).

References

1. Dickson M, Gagnon JP. The cost of new drug discovery and development. *Discov Med* 2004;**4**(22):172–9.
2. Paul SM, Mytelka DS, Dunwiddie CT, *et al.* How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat Rev Drug Discov* 2010;**9**(3):203–14.
3. Wang Y, Chen S, Deng N, *et al.* Drug repositioning by kernel-based integration of molecular structure, molecular activity, and phenotype data. *PLoS One* 2013;**8**(11):e78518.
4. Shaughnessy AF. Old drugs, new tricks. *BMJ* 2011;**342**:d741.
5. Gupta SC, Sung B, Prasad S, *et al.* Cancer drug discovery by repurposing: teaching new tricks to old dogs. *Trends Pharmacol Sci* 2013;**34**(9):508–17.
6. Chen H, Zhang H, Zhang Z, *et al.* Network-based inference methods for drug repositioning. *Comput Math Methods Med* 2015;**2015**:130620.
7. Zhang P, Wang F, Hu J. Towards drug repositioning: a unified computational framework for integrating multiple aspects of drug similarity and disease similarity. *AMIA Annu Symp Proc* 2014;**2014**:1258–67.
8. Jin G, Wong ST. Toward better drug repositioning: prioritizing and integrating existing methods into efficient pipelines. *Drug Discov Today* 2014;**19**(5):637–44.
9. Li J, Chen B, Butte AJ, *et al.* A survey of current trends in computational drug repositioning. *Brief Bioinform* 2016;**17**(1): 2–12.
10. Dudley JT, Schadt E, Sirota M, *et al.* Drug discovery in a multidimensional world: systems, patterns, and networks. *J Cardiovasc Transl Res* 2010;**3**(5):438–47.
11. Pesquita C, Faria D, Falcão AO, *et al.* Semantic similarity in biomedical ontologies. *PLoS Comput Biol* 2009;**5**(7):e1000443.
12. Iorio F, Bosotti R, Scacheri E, *et al.* Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proc Natl Acad Sci USA* 2010;**107**(33): 14621–6.
13. Lee S, Lee KH, Song M, *et al.* Building the process-drug-side effect network to discover the relationship between biological processes and side effects. *BMC Bioinformatics* 2011;**12**(Suppl 2):S2.
14. Chiang AP, Butte AJ. Systematic evaluation of drug-disease relationships to identify leads for novel drug uses. *Clin Pharmacol Ther* 2009;**86**(5):507–10.
15. Yildirim MA, Goh K-I, Cusick ME, *et al.* Drug-target network. *Nat Biotechnol* 2007;**25**(10):1119–26.
16. Keiser MJ, Roth BL, Armbruster BN, *et al.* Relating protein pharmacology by ligand chemistry. *Nat Biotechnol* 2007;**25**(2):197–206.
17. Wang YY, Nacher JC, Zhao XM. Predicting drug targets based on protein domains. *Mol Biosyst* 2012;**8**(5):1528–34.
18. Zu S, Chen T, Li S. Global optimization-based inference of chemogenomic features from drug-target interactions. *Bioinformatics* 2015;**31**(15):2523–9.
19. Iwata H, Mizutani S, Tabei Y, *et al.* Inferring protein domains associated with drug side effects based on drug-target interaction network. *BMC Syst Biol* 2013;**7**(Suppl 6):S18.
20. Wishart DS, Knox C, Guo AC, *et al.* DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res* 2006;**34**(Database issue):D668–72.
21. Chen B, Wild D, Guha R. PubChem as a source of polypharmacology. *J Chem Inf Model* 2009;**49**(9):2044–55.
22. UniProt, C. The universal protein resource (UniProt) in 2010. *Nucleic Acids Res* 2010;**38**(Database issue):D142–8.
23. Jacob L, Vert JP. Protein-ligand interaction prediction: an improved chemogenomics approach. *Bioinformatics* 2008;**24**(19):2149–56.
24. Tabei Y, Pauwels E, Stoven V, *et al.* Identification of chemogenomic features from drug-target interaction networks using interpretable classifiers. *Bioinformatics* 2012;**28**(18): i487–94.
25. Hinselmann G, Rosenbaum L, Jahn A, *et al.* Large-scale learning of structure-activity relationships using a linear support vector machine and problem-specific metrics. *J Chem Inf Model* 2011;**51**(2):203–13.
26. Hameed PN, Verspoor K, Kusljic S, *et al.* A two-tiered unsupervised clustering approach for drug repositioning through heterogeneous data integration. *BMC Bioinformatics* 2018;**19**(1):129.
27. Chen L, Zeng W-M, Cai Y-D, *et al.* Predicting anatomical therapeutic chemical (ATC) classification of drugs by integrating chemical-chemical interactions and similarities. *PLoS One* 2012;**7**(4):e35254.
28. Liu Z, Guo F, Gu J, *et al.* Similarity-based prediction for anatomical therapeutic chemical classification of drugs by integrating multiple data sources. *Bioinformatics* 2015;**31**(11):1788–95.
29. Munroe M, Kolesar J. Olaparib for the treatment of BRCA-mutated advanced ovarian cancer. *Am J Health Syst Pharm* 2016;**73**(14):1037–41.
30. Kanehisa M, Sato Y, Kawashima M, *et al.* KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res* 2016;**44**(D1):D457–62.
31. Guerra MT, Florentino RM, Franca A, *et al.* Expression of the type 3 InsP3 receptor is a final common event in the

development of hepatocellular carcinoma. *Gut* 2019;**68**(9): 1676–87.

32. Rezuchova I, Hudecova S, Soltysova H, *et al*. Type 3 inositol 1,4,5-trisphosphate receptor has antiapoptotic and proliferative role in cancer cells. *Cell Death Dis* 2019;**10**(3): 186.
33. Lin KY, Kraus WL. PARP inhibitors for cancer therapy. *Cell* 2017;**169**(2):183.
34. Mu Y, Li M, Ding F, *et al*. De novo characterization of the spleen transcriptome of the large yellow croaker (*Pseudosciaena crocea*) and analysis of the immune relevant genes and pathways involved in the antiviral response. *PLoS One* 2014;**9**(5):e97471.
35. Nerreter T, Distler E, Köchel C, *et al*. Combining dasatinib with dexamethasone long-term leads to maintenance of antiviral and antileukemia specific cytotoxic T cell responses in vitro. *Exp Hematol* 2013;**41**(7):604–614 e4.
36. Anikeeva N, Lebedeva T, Clapp AR, *et al*. Quantum dot/peptide-MHC biosensors reveal strong CD8-dependent cooperation between self and viral antigens that augment the T cell response. *Proc Natl Acad Sci USA* 2006;**103**(45): 16846–51.
37. Zhou XC, Dong S-H, Liu Z-S, *et al*. Regulation of gammaherpesvirus lytic replication by endoplasmic reticulum stress-induced transcription factors ATF4 and CHOP. *J Biol Chem* 2018;**293**(8):2801–14.