ELDTIP: An Ensemble Learning-based method for DTI Prediction

Author: Xiangyu Zou*

Institute of Problem Solving: Beijing University of Aeronautics and Astronautic,China

Abstract: Exploring drug-target interactions has always been an important step in drug development. However, exploring drug-target interaction is time-consuming and laborious. A large number of studies try to use artificial intelligence methods to predict possible drug-target interactions to reduce the workload of the wet-lab identification experiments. However, the accuracy of existing methods is still limited.
This paper proposes an ensemble learning-based drug-target interaction prediction method (ELDTIP in short). First, the multiple similarity matrices of drugs or proteins are integrated by singular value decomposition (SVD) to obtain their low-dimensional feature vectors. After that, by concatenating the low-dimensional feature vectors of specific drugs and targets, the feature vector of a drug-target pair are obtained. An ensemble learning model based on gradient boosting decision tree (GBDT) was constructed to predict whether this pair of drug-target can interact with each other. The main contributions of ELDTIP are as follows:(1): ELDTIP uses SVD to integrate multiple similarity matrices, which can retain more valuable information of the original feature. (2): ELDTIP uses the ensemble learning-based model, GBDT, which can make full use of the unknown DTIs in the dataset and mitigate the influence of class imbalance. Experimental results show that the performance of ELDTIP is higher than that of several state-of-the-art DTI prediction methods.

1. Introduction

Drug discovery and development has always been an important topic in the medical field and has great market prospects. However, its development processes are time-consuming and expensive. Drug works by binding to various molecular-target vias drug-target interactions (DTIs). Protein is an important group of target. Drugs can enhance or inhibit the activity of the proteins by binding with them. With the development of artificial intelligence technology in recent years, a large number of studies try to use machine learning methods to predict possible DTIs. The predict results are provide to biologists for subsequent DTI identification wet-lab experiment. In this way, the workload of drug discovery can be reduce.

In recent years, many efficient machine learning methods have been proposed. For example, DDR uses graph mining and machine learning algorithms to predict drug-target interactions. It first integrates multi-source drug similarity and protein similarity. After that DDR constructs graph-based feature vectors for each drug-target pair and classifies them using a random forest model (Rawan S. Olayan, Haitham Ashoor and Vladimir B. Bajic, 2017).The GRMF algorithm focuses on the "isolated nodes" in the network, and adds preprocessing steps to cope with the sparse DTI network. DTINet focuses on integrating the multi-sources drug features and protein features. Then it makes predictions based on these features through matrix decomposition algorithms. DTINet achieves significant performance improvements in drug-target interaction prediction compared to other state-of-the-art methods(Yunan Luo , Xinbin Zhao , 2017). All of these methods effectively predict drug-target interactions from different perspectives, but they also have some problems. DDR effectively solves the problem of high false positive prediction rate of existing DTI prediction methods, but its similarity fusion algorithm can only apply in matrix level when retaining or deleting the similarity matrix, which easily leads to the deletion of some useful information after data processing. The GRMF takes into account missed interactions and is more reasonable in data processing, but does not employ multi-source data(Ali Ezzat, Peilin Zhao, Min Wu, 2017). The general steps of this paper are as follows: First, the multiple similarity matrices of drugs or proteins are integrated by singular value decomposition (SVD) techniques, and the low dimensional feature of drugs and proteins are obtained. After obtaining the feature vector of a drug-target pairwise by concatenating the feature of this two molecules, an ensemble learning-based prediction module is established to predict whether the drug and target can interact with each other. Compared with the previous method, ELDTIP uses SVD singular value decomposition to integrate multi-source data, which effectively retains the more valuable information in the original feature. The ensemble learning algorithm based on GBDT makes full use of the unknown DTIs in the dataset and alleviates the negative impact of class

* E-mail: 23000230084@qq.com

imbalance on the prediction results. The experimental results also show that ELDTIP has a higher accuracy in predicting the targets for drugs.

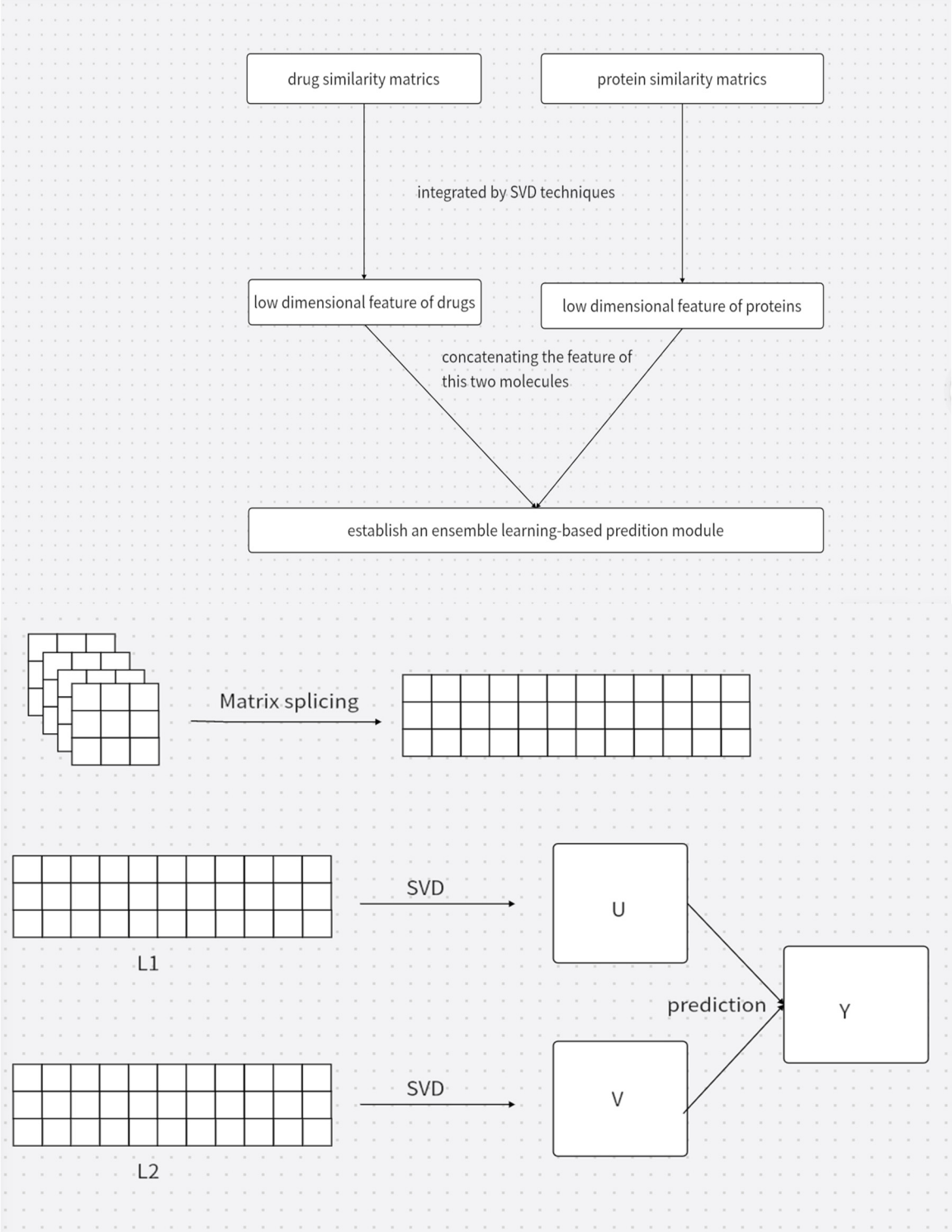

Figure 1 The flow chart of ELDTIP

2. Related work

We know that the principle of general random forest (RF) is to build n decision trees, classify the feature vectors of the original sample respectively, and finally use voting, and the more people think that the sample belongs to which category, the sample is divided into this category, that is, the minority obeys the majority; The principle of GBDT is different, first, the feature vector of drugs and proteins is spliced horizontally to obtain a feature vector with a dimension of 2k, and then a decision tree is built to predict, the prediction number at this time is not a

classification tree in a random forest, for example, assuming that the probability of giving a value of 1 is λ, and we want its value to be 1, then the error at this time is 1-λ, which can also be called loss, or residual, that is, how much is the difference from the real label; For the second tree, we want to make up for the error of the first tree, that is, to make the output of the second tree 1-λ, assuming that its true output value is β, then the error is 1-λ-β; In the same way, to build a third tree, if the output value at this time is 1-λ-β, then the loss is zeroed at this time, and we fit the true label of the sample with three trees, which is the general idea of GBDT. In addition, as mentioned earlier, when building GBDT, it is necessary to consider the problem of cross-validation, there are trees X and Y, which only contain the training set, use them to train GBDT, after training, add the part of the test set to the trained GBDT, you can get the prediction data, and compare with the data in the test set, you can estimate the loss.

This algorithm is realized by calculating the ROC curve, and the area AOC under the ROC curve is calculated. Compared with the previously mentioned methods, the random forest algorithm is used in DDR, and n trees are constructed at the same time, n trees are independent of each other, do not affect each other, and they are equal relations, and finally n trees are used to predict the result of voting, and whoever votes more is considered to be the final result; The gradient boosting decision tree we constructed is not an equal relationship, and the decision trees constructed later are all to make up for the remaining problems of the previous decision tree and fit the residuals. It can be seen that random forest and gradient boosting decision trees are fundamentally different in algorithm thinking.

In order to be able to easily evaluate the effectiveness of the algorithm, it is necessary to divide on the original data set to divide who is the training set and who is the test set. For the original drug-protein interaction matrix, we divide all 1s and 0s into five parts, 4 for training and 1 for testing, and these five data are to be used in the final prediction. Then the data is preprocessed, set the number of drugs to m, then the drug similarity matrix is a square matrix of m*m, set the number of matrices to n, and splice all drug similarity matrices horizontally to obtain a comprehensive matrix L with a dimension of m*nm, and then learn the processed data representation, L=UΣVT, where Σ is the singular value, for the matrix of m*nm, there are at most m singular values, through the singular value decomposition, the dimensionality reduction of the original matrix can be realized, and the SVD singular value decomposition of L is set, and k singular values are retained, L can be reduced to a matrix R of m*k dimension. The k-dimensional feature vector representing m drugs, the similarity matrix of proteins, can be processed in the same way, the last step is to make decisions, this algorithm uses GBDT (gradient boosting decision tree) processing. We know that the principle of general random forest (RF) is to build n decision trees, classify the feature vectors of the original sample respectively, and finally use voting, and the more people think that the sample belongs to which category, the sample is divided into this category, that is, the minority obeys the majority; The principle of GBDT is different, first, the feature vector of drugs and proteins is spliced horizontally to obtain a feature vector with a dimension of 2k, and then a decision tree is built to predict, the prediction number at this time is not a classification tree in a random forest, for example, assuming that the probability of giving a value of 1 is λ, and we want its value to be 1, then the error at this time is 1-λ, which can also be called loss, or residual, that is, how much is the difference from the real label; For the second tree, we want to make up for the error of the first tree, that is, to make the output of the second tree 1-λ, assuming that its true output value is β, then the error is 1-λ-β; In the same way, to build a third tree, if the output value at this time is 1-λ-β, then the loss is zeroed at this time, and we fit the true label of the sample with three trees, which is the general idea of GBDT. In addition, as mentioned earlier when constructing GBDT, it is necessary to consider the problem of cross-validation, there are trees X and Y, they only contain the training set, use them to train GBDT, after training, add the part of the test set to the trained GBDT, you can get the prediction data, and the test set data comparison, you can estimate the loss, this algorithm is realized by calculating the ROC curve, calculating the area AOC under the ROC curve. Compared with the previously mentioned methods, the random forest algorithm is used in DDR, and n trees are constructed at the same time, n trees are independent of each other, do not affect each other, and they are equal relations, and finally n trees are used to predict the result of voting, and whoever votes more is considered to be the final result; The gradient boosting decision tree we constructed is not an equal relationship, and the decision trees constructed later are all to make up for the remaining problems of the previous decision tree and fit the residuals. It can be seen that random forest and gradient boosting decision trees are fundamentally different in algorithm thinking.

3. Materials and methods

3.1 Dataset

The data set used in this paper is obtained from a published work. The dataset involved 708 drugs, 1512 proteins and 1923 known DTIs. It also contains four drug similarity matrices, which denote as M1∈708*708, M2 ∈708*708, M3∈708*708, M4∈708*708. The three protein similarity matrices in the data set are denoted as N1 ∈ 1512*1512, N2 ∈ 1512*1512, N3 ∈ 1512*1512. The drug-target interaction matrix, denoted as Y ∈ 708*1512. The elements in Y is ∈{0, 1}, for example, if Y(4,2) = 1, then drug d4 and protein t2 have an interaction, if Y(4,2) = 0, then there is no interaction between them.

3.2 Representation learning

We treat the four types of similarities between a drug and other drugs as the features of it. In this way , we concatenate the four drug similarity matrices to obtain the

drug feature matrix L1 ∈ R708* （708*4） , where L1=M1⊕M2⊕M3⊕M4. The feature matrix of target , L2, can obtain by the same way. L1 and L2 are decomposed by SVD, respectively,and let the reduced drug identity matrix be Fr∈^{x*x}, then Fr can be calculated according to the following equation:

$$L1 = UsigmaVT \quad (1)$$

$$Fr = Usigma1/2 \quad (2)$$

where U is the left singular value matrix and V is the right singular value matrix. sigma is a diagonal array, and the singular values of L1 are arranged on the diagonal of sigma in descending order. Similarly, the reduced dimensional protein feature martix Fp ∈ ^{x*x} can be computed by the same operation.

3.3 Prediction

In our dataset, there are serious class imbalance between the known DTIs (positive samples) and unknown DTIs (negative samples). Selecting an equal number of positives and negatives when training the dataset may loss some valuable information contains in the negatives samples. As a result, we chose GBDT as the classifier, which can build multiple decision trees and use different negative samples to train different decision trees, ensuring that negative samples are fully utilized. If the feature vector of drug di is denoted by A[di], the feature vector of target tj is denoted by B[tj], and the feature vector of the drug-target pair denoted by $V[i,j] = A\left[d_i\right] \oplus B\left[t_j\right]$, the score for the existence of interaction between drug di and target tj can be calculated according to the following equation.

$$\text{score}(i,j) = \sum_{k=1}^{g} T_k\left(V_f(i,j)\right) \quad (3)$$

where Tk () denotes the prediction score of the kth decision tree in GBDT for the feature vector Vf(i,j). The higher the score(i , j) , the more probability that di is interacted with tj. Define Yij=score(i , j) to denote the interaction score matrix, which can be evaluated for loss by negative log-likelihood. To improve the prediction accuracy, to make the prediction result as high as possible for positive samples and converge to 0 for negative samples, the prediction model can be optimized by the following equation.

$$\text{loss} = \sum_{i,j} \log\left(1+\exp\left(-2Y_{ij}\hat{Y}_{ij}\right)\right) \quad (4)$$

where Yij denotes the actual interaction between di and tj.

4. Experimental results

We use five-fold cross-validation to evaluate the performance of the prediction method.

For all known and unknown DTIs in the dataset, we divide them into five groups randomly. In each fold of cross-validation, we use four sets of positive samples and four sets of negative samples to train the model, leaving one set of positive and one set of negative samples as the test set.

4.1 Evaluation metrics

AUC is a commonly used metric to verify the correctness of a method, which refers to the areas under the Receiver operating characteristic (ROC) curve. In this paper, we evaluate the performance of each DTI prediction method by calculating the AUC of them.

For a given threshold δ , a positive sample is considered as a true positive (TP) sample if its prediction score is greater than δ. A positive sample is defined as a false negative sample (FN) if its score is lower than δ. If the score of the negative sample is greater than δ, it is considered as a false positive sample (FP). If it is not, it is considered as a true negative sample (TN). Receiver operating characteristic (ROC) curves can be constructed by calculating the true positive rate (TPR) and false positive rate (FPR) at different δ. TPR and FPR can be defined as follows.

$$TPR = \frac{TP}{TP+FN}$$
$$FPR = \frac{FP}{TN+FP} \quad (5)$$

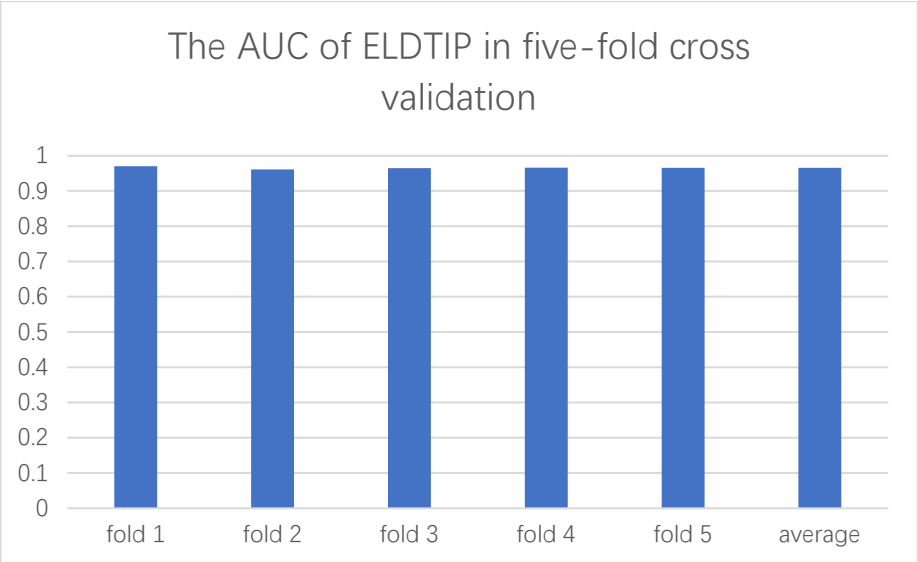

Figure 2 The AUC of ELDTIP in each fold of cross-validation

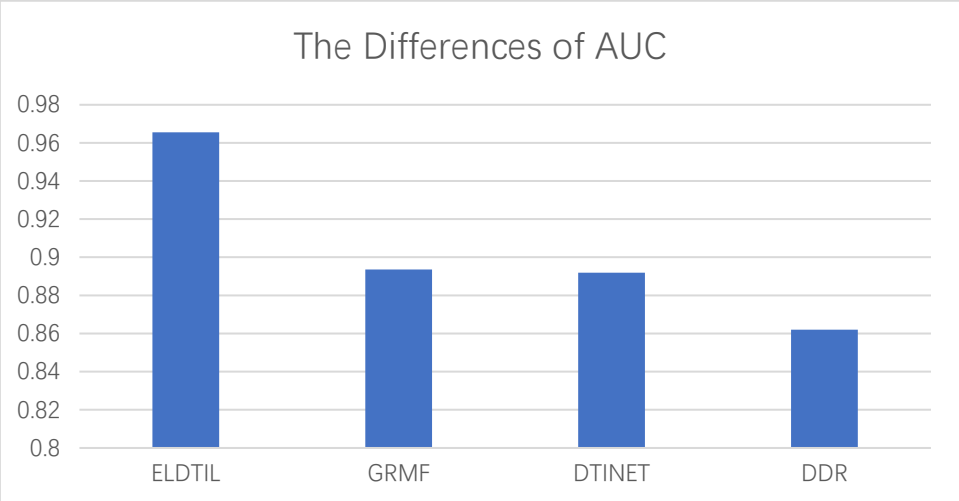

Figure 3 The AUC of ELDTIP and other state-of-the-art DTI perdiction method

4.2 Performance Comparison

The AUC values of ELDTIP in the five-fold cross-validation are shown in the figure, and it can be seen that our algorithm has relatively stable performance on different test sets.

Upon comparison, our method has the highest AUC value of 0.9655, which is 0.0719 higher than the second-ranked GRMF, 0.0736 higher than the third-ranked DTINET, and 0.1035 higher than the fourth-ranked DDR. It proves that ELDTIP is superior than other prediction method.

This may be due to the fact that (i) we use SVD to integrate data from multiple sources, thus retaining more representative features in the original data in the process of dimensionality reduction. (ii)ELDTIP constructs a GBDT-based classifier, which effectively mitigates the effect of class imbalance while making full use of the negative samples in the dataset.

5. Conclusion

In this work, we proposed a GBDT-based DTI prediction method, named ELDTIP. It integrates multiple data of drugs and targets include their chemical properties, mechanism and clinical functions. ELDTIP has two main innovations, one is the use of SVD to fuse data from multiple sources. This approach can fully retain the valuable information of the original data in the process of dimensionality reduction. Second, it uses an ensemble learning algorithm based on GBDT, which can make full use of the negative samples information in the dataset and alleviate the impact of class imbalance on the prediction results.

Experimental results show that ELDTIP has stable performance on different test sets, and it outperforms several state-of-the-art DTI prediction methods.

Reference

1 Ezzat, A. , Zhao, P. , Min, W. , Li, X. , & Kwoh, C. K. . 2016. Drug-target interaction prediction with graph regularized matrix factorization. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 14(3), 1-1. 10.1109/TCBB.2016.2530062

2 Langville, A. N. , Meyer, C. D. , Albright, R. , Cox, J. , & Duling, D. . 2014. Algorithms, initializations, and convergence for the nonnegative matrix factorization. Eprint Arxiv. doi:http://dx.doi.org/

3 Luo, Y. , Zhao, X. , Zhou, J. , Yang, J. , Zhang, Y. , & Kuang, W. , et al. 2017. A Network Integration Approach for Drug-Target Interaction Prediction and Computational Drug Repositioning from Heterogeneous Information. Research in Computational Molecular Biology. Springer.

4 Mousavian, & Masoudi-Nejad. Drug-target interaction prediction via chemogenomic space: Learning-based methods. 10.1517/17425255.2014.950222

5 Olayan, R. S. , Haitham, A. , & Bajic, V. B. . 2018. Ddr: efficient computational method to predict drug–target interactions using graph mining and machine learning approaches. Bioinformatics(21), 21. 10.1093/bioinformatics/btx731

6 Sun, C. , Cao, Y. , Wei, J. M. , & Liu, J. . 2021. Autoencoder-based drug-target interaction prediction by preserving the consistency of chemical properties and functions of drugs. Bioinformatics. 10.1093/bioinformatics/btab384

7 Sun, C. , Xuan, P. , Zhang, T. , & Ye, Y. . 2020. Graph convolutional autoencoder and generative adversarial network-based method for predicting drug-target interactions. IEEE/ACM Transactions on Computational Biology and Bioinformatics, PP(99), 1-1. 10.1016/j.patcog.2021.108095

8 Tapio, P. , Antti, A. , Sami, P. , Sushil, S. , Agnieszka, S. , & Tang, J. , et al. 2015. Toward more realistic drug–target interaction predictions. Briefings in Bioinformatics(2), 325-337. 10.1093/bib/bbu010

9 Twan, V. L. , Nabuurs, S. B. , & Elena, M. . Gaussian interaction profile kernels for predicting drug–target interaction. Bioinformatics(21), 3036. 10.1093/bioinformatics/btr500

10 Xing, Chen, Clarence, C. , Yan, Xiaotian, & Zhang, et al. 2016. Drug-target interaction prediction: databases, web servers and computational models. Briefings in bioinformatics.10.1093/bib/bbv066