Journal of Theoretical Biology 417 (2017) 1–7

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

PDC-SGB: Prediction of effective drug combinations using a stochastic gradient boosting algorithm

Qian Xu, Yi Xiong*, Hao Dai, Kotni Meena Kumari, Qin Xu, Hong-Yu Ou, Dong-Qing Wei

State Key Laboratory of Microbial Metabolism, and College of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, China

ARTICLE INFO

Keywords:
Drug combinations
Feature selection
Stochastic gradient boosting
Feature patterns

ABSTRACT

Combinatorial therapy is a promising strategy for combating complex diseases by improving the efficacy and reducing the side effects. To facilitate the identification of drug combinations in pharmacology, we proposed a new computational model, termed PDC-SGB, to predict effective drug combinations by integrating biological, chemical and pharmacological information based on a stochastic gradient boosting algorithm. To begin with, a set of 352 golden positive samples were collected from the public drug combination database. Then, a set of 732 dimensional feature vector involving biological, chemical and pharmaceutical information was constructed for each drug combination to describe its properties. To avoid overfitting, the maximum relevance & minimum redundancy (mRMR) method was performed to extract useful ones by removing redundant subsets. Based on the selected features, the three different type of classification algorithms were employed to build the drug combination prediction models. Our results demonstrated that the model based on the stochastic gradient boosting algorithm yield out the best performance. Furthermore, it is indicated that the feature patterns of therapy had powerful ability to discriminate effective drug combinations from non-effective ones. By analyzing various features, it is shown that the enriched features occurred frequently in golden positive samples can help predict novel drug combinations.

1. Introduction

In traditional drug design, the paradigm of the one-drug-one-target had been the dominating approach in the drug discovery for a long time. However, the old strategy is unlikely to efficaciously deal with certain complex diseases, such as cancer and diabetes, which are regulated by multiple signaling pathways or molecular networks (Csermely et al., 2013; Jia et al., 2009). It is becoming increasingly apparent that such one-drug-one-target paradigm shows limited efficacy, which is often due to factors such as network robustness, redundancy, compensatory actions, and counter-target activities (Jia et al., 2009). Instead, systems-oriented drug design (such as multi-target drug or drug combination) becomes a more productive and promising strategy with higher efficacy but less side effects to overcome those limitations (Chen et al., 2016d; Fan et al., 2014; Min et al., 2013; Sun et al., 2015; Xiao et al., 2013a, 2013b, 2015).

Generally, multiple agents or drugs are simultaneously administrated to form effective drug combinations for disease treatment with significant improvements on drug efficacy and safety. In clinical practice, the effective drug combinations always consist of Food and Drug Administration (FDA)-approved drugs or existing bioactive compounds that have entered clinical trials and passed safety tests. The effective drug combinations could be used by patients without toxic side effects. It is quite time and resource consuming to identify all the effective drug combinations using experimental techniques, since the number of possible drug combinations will expand exponentially with the increasing number of single drugs available in the market. Therefore, computational prediction of drug combinations becomes a significant and challenging task.

During the last decade, a wide variety of computational models have already been developed to aid in the discovery of effective drug combinations. In the first type of methods, the Loewe additivity and Bliss independence were proposed for quantifying synergy between a pair of drugs (Ryall and Tan, 2015). The Loewe additivity model is based on the assumption that the two drugs act through a similar mechanism while the Bliss independence criterion assumes that they act by an independent mechanism. Experimental identification of drug combinations typically involves generating dose response curves with a pair of drugs in separate or in combination. The experimental dose response curve data can then be compared to the predictions of Loewe additivity or Bliss independence to determine if the drugs are acting synergistically. The second type of methods are system approaches,

* Corresponding author.
E-mail address: xiongyi@sjtu.edu.cn (Y. Xiong).

http://dx.doi.org/10.1016/j.jtbi.2017.01.019
Received 26 September 2016; Received in revised form 6 January 2017; Accepted 14 January 2017
Available online 16 January 2017

including large-scale modeling of cell signaling networks, network motif analysis, statistics-based models, correlation identification in gene signatures, and functional genomics. Network-based methods were constructed on the genetic variations, protein-protein interaction (PPI), functional modules, and signaling pathways (Csermely et al., 2013; Ryall and Tan, 2015; Wang et al., 2012; Wu et al., 2010). Among the second type of computational methods, many studies were conducted based on feature-based methods. Sun et al. used biological features from gene expression data of multiple drugs (Sun et al., 2014). Zhao et al. integrated the biological (drug targets) and pharmacological (drug therapy) features (Zhao et al., 2011). Chen et al. combined biological (PPI and target enrichment pathways) and chemical (chemical-chemical interaction, CCI) features (Chen et al., 2013a). However, to our best knowledge, there is still no method which integrates biological, chemical, and pharmaceutical information altogether to predict drug combinations.

In the present study, we developed a computational method for Prediction of effective Drug Combinations using a Stochastic Gradient Boosting algorithm, termed PDC-SGB, which integrates biological (the PPI information of targets and disease pathways), chemical (2-Dimensional substructures) and pharmacological (therapy information). In order to reduce the redundancy among the features, the Minimum Redundancy & Maximum Relevance (mRMR) approach was employed for feature selection (Peng et al., 2005). Then, an advanced machine learning algorithm, that is, stochastic gradient boosting algorithm was applied to construct drug combination prediction models based on the selected features.

As demonstrated by a series of recent publications (Chen et al., 2016c; Jia et al., 2015; Jia et al., 2016b; Lin et al., 2014; Qiu et al., 2016a, 2016b), to establish a really useful feature-based statistical predictor for a biological system and also to make the presentation logically crystal clear, we should follow the five-step guidelines (Chou, 2011): (a) construct or select a valid benchmark dataset to train and test the predictor; (b) formulate the drug combinations samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted; (c) develop a powerful algorithm (or engine) to perform the prediction; (d) properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the predictor; (e) establish a user-friendly web-server for the predictor that is accessible to the public. Below, let us elaborate how to deal with these steps one-by one.

2. Materials and methods

2.1. Data sets

The data set of drug combinations were collected from the Drug Combination Database (DCDB, version 2.0, October 2014 release) (http://www.cls.zju.edu.cn/dcdb/) (Liu et al., 2014b), which includes 1363 effective drug combinations retrieved from the published clinical studies and FDA orange book. Among them, we collected 946 pairwise drug combinations from 817 single drugs. A drug combination would be further removed if the feature information of either of its components was not available. Finally, the set of 352 pairwise drug combinations from 337 distinctive single drugs were remained as our Golden Positive Samples (GPS) (Table S1). The non-effective drug combinations were generated by randomly pairing drugs that appeared in the data set of positive samples. Among the 56,264 possible non-effective drug combinations, we randomly selected 1760, 1056, and 352 combinations as the Golden Negative Samples (GNS), which were the synthesized data sets where the positive-to-negative ratio is 1:5, 1:3, 1:1, respectively. In the subsequent sections, we used the balanced data set that the number of negative pairs was equal to that of positive pairs if there was no statement (the other two negative data sets gave the similar level classification performances and hence only the results on balanced data set were shown in the rest of the article if not specified). In order to remove the sample bias and achieve a robust result, the process for randomly selecting negative samples was repeated by ten times to generate ten groups of negative samples for model training. The final performance was reported by averaging the performance of the ten runs.

2.2. Feature construction and selection

In this study, we integrated six types of features to describe the drug combinations, which include the molecular 2D structures, structural similarity, anatomical therapeutic similarity, protein-protein interaction, chemical-chemical interaction, and disease pathways. The target proteins, molecular 2D structures, and the Anatomical Therapeutic Chemical code information of the drugs were extracted from DrugBank (http://www.drugbank.ca/) (Law et al., 2014).

2.2.1. Molecular 2D structures

The software package Molecular Operating Environment (MOE, http://www.chemcomp.com/) was used to calculate the 2D MACCS (Molecular ACCess System) fingerprints of drug molecules (Vilar et al., 2008). The fingerprint of a drug is represented as a feature vector of 166 elements, in which each element of the vector represents the existence or nonexistence of a specific substructure (Vilar et al., 2014). For each drug combination, if both drugs in a pair have the same substructure, it is encoded as 3; if only one component has a given substructure, it is encoded as 2; if neither of them has it, it is encoded as 1. The 166th bit representing “fragment” was not included since all drugs had only one fragment. In total, the feature vectors of 165 elements were used to represent the molecular 2D structural information of a drug combination.

2.2.2. Structural similarity between drugs

Tanimoto coefficient (TC) was used to measure the similarity of 2D structural fingerprints between two drug molecules. Given the two drugs d_i and d_j, the TC between them was calculated as follows:

$$\mathrm{TC}(\mathrm{T_i}, \mathrm{T_j}) = \frac{\mathrm{T}_i \cap \mathrm{T_j}}{\mathrm{T}_i \cup \mathrm{T}_j} \quad (1)$$

in which, T_i and T_j were the fingerprints of d_i and d_j, respectively. TC ranges from 0 (minimum similarity) to 1 (maximum similarity).

2.2.3. Protein-protein interaction similarity

The target proteins of drugs have been demonstrated to play an important role in the prediction of effective drug combinations (Chen et al., 2013a; Xu et al., 2012). In this work, we included the PPI information of drug target proteins to infer the potential effective drug combinations. The PPI information of target proteins was obtained from the Biological General Repository for Interaction Datasets (BioGrid Version 3.2.116) (http://thebiogrid.org/) (Chatr-Aryamontri et al., 2015). The protein interaction information of a drug includes a set of the proteins which interact at least one target of the drug. The calculation of PPI similarity of two drugs was similar to the definition of structural similarity as described in Eq. (1).

2.2.4. Anatomical therapeutic similarity

The Anatomical Therapeutic Chemical (ATC) coding system is used for the classification of drugs according to the organ or system on which they act and their therapeutic, pharmacological and chemical properties. A drug has one code or more than one codes. In this system, drugs are catalogued into different groups at 5 levels. We considered the first 3 levels since the most of drugs in our data sets have no similarity in the fourth and fifth level. The kth level drug ATC similarity (S_k) between d_i and d_j was defined as:

$$\mathrm{S_k}(\mathrm{d_i}, \mathrm{d_j}) = \frac{\mathrm{ATC_k}(\mathrm{d}_i) \bigcap \mathrm{ATC_k}(\mathrm{d_j})}{\mathrm{ATC_k}(\mathrm{d_i}) \bigcup \mathrm{ATC_k}(\mathrm{d_j})} \quad (2)$$

where $ATC_k(d_i)$ represents all the ATC codes of drug d_i at the kth level, k ranges from 1 to 3.

In addition, the mean, maximum and minimum values of the three scores were also taken as the features of therapeutic similarity in the study.

$$\mathrm{Mean}(\mathrm{d_i}, \mathrm{d_j}) = \frac{\sum_{k=1}^{n} \mathrm{S_k}(\mathrm{d_i}, \mathrm{d_j})}{\mathrm{n}} \quad (n=3) \tag{3}$$

$$\mathrm{Max}(\mathrm{d_i}, \mathrm{d_j}) = \max S_k(d_i, d_j)(k=1, 2, 3) \tag{4}$$

$$\mathrm{Min}(\mathrm{d_i}, \mathrm{d_j}) = \min S_k(d_i, d_j)(k=1, 2, 3) \tag{5}$$

2.2.5. Disease pathway influence

Since the target proteins of different drugs could be involved in a same disease pathway, we mapped the drugs to the corresponding disease pathways in MSigDB (http://www.broadinstitute.org/gsea/msigdb/) (Chen et al., 2013a; Subramanian et al., 2005), and constructed two scores to evaluate their influence to that pathway. Similar to the feature of molecular 2D structures, the first score was defined as follows. For each drug combination, if the target proteins of drugs in a pair were involved in the same disease pathway it is encoded as 3; if only the target protein of one drug was involved in the given disease pathway, it is encoded as 2; if neither of them was, it is encoded as 1.

The second score was the -$\log_{10}$ of the hypergeometric test p value of gene set G_i, which includes all target proteins of the drug d_i and its direct neighbors in the PPI network extracted from BioGrid. The influence of drug d_i to disease pathway DP_j could be measured by the disease pathway enrichment score, which was calculated in the similar way as (Chen et al., 2013a; Huang et al., 2012).

$$\mathrm{Score_i^j} = -\log_{10}\left(\sum_{k=m}^{n} \frac{(\mathrm{M/k})(\mathrm{N} - \mathrm{M/n} - \mathrm{k})}{(\mathrm{N/n})}\right) \tag{6}$$

where N is the number of genes in human, M is the number of genes annotated to the disease pathway DP_j, n is the number of genes in G_i, and m is the number of genes both in G_i and DP_j. The higher enrichment score indicates that this drug is more likely to have influence on the given disease pathway.

We calculated the scores of two components d_i (i=1, 2) in a drug combination in each of the 186 disease pathways, generating $dp_i^1, dp_i^2, \ldots, dp_i^{186}$($i$=1, 2), where dp_i^j was the score defined in Eq. (6). For each drug combination (d_1, d_2), the 372 features can be derived from these enrichment scores as follows.

$$\mathrm{dp_1^1} + \quad \mathrm{dp_2^1}, \quad \mathrm{dp_1^2} + \mathrm{dp_2^2}, \quad \ldots, \mathrm{dp_1^{186}} + \mathrm{dp_2^{186}} \tag{7}$$

$$|d\mathrm{p_1^1} - \mathrm{dp_2^1}|, \quad |\mathrm{dp_1^2} - \mathrm{dp_2^2}| \quad , \ldots, |d\mathrm{p_1^{186}} - \mathrm{dp_2^{186}}|. \tag{8}$$

2.2.6. Chemical-chemical interaction confidence score

The information of chemical-chemical interactions was retrieved from STITCH (http://stitch.embl.de/) (Kuhn et al., 2014). Each interaction has a combined confidence score which describes the overall similarity between each pair of chemicals. The larger confidence score indicates the higher probability of two chemicals to interact with each other. These confidence scores were scaled to the range [0,1] as the input feature values.

In total, each drug combination was represented by a 732 dimensional feature vector to describe its properties, where the 165 dimensions from the molecular 2D substructure, 1 dimension from structural similarity between drugs, 1 from PPI similarity, 6 from anatomical therapeutic similarity information, 558 (186+186+186) from the disease pathway information, and 1 from CCI similarity. To avoid overfitting, the feature selection procedure was performed to analyze these features and extract useful ones by removing redundant subsets. mRMR was applied here which is a powerful method for ranking the features through maximizing the dependency between the selected features and the classification variables, in the meanwhile, minimizing the correlation of the inner features (He et al., 2010).

2.3. Model construction

Machine learning algorithms have been widely applied in the bioinformatics field (Liu et al., 2016a, 2016b, 2015; Shen and Chou, 2009, 2010). Three different classification algorithms were implemented to build prediction models in this work. They were support vector machine (SVM), naïve bayes (NB), and stochastic gradient boosting (SGB) (Friedman, 2002). Prediction models based on different machine learning algorithms were implemented by applying classification and regression training (Caret) package in R, including parameter optimizing, model training, and evaluating (Kuhn, 2008). In SVM algorithm, Gaussian kernel was employed as the kernel function. The cost factor c for outlier samples and gamma γ in kernel function were optimized by grid search. In NB algorithm, Laplace smoothing was used in the case that the posterior probability was zero. In SGB algorithm, gradient boosting is a machine learning technique for regression and classification problems, which constructs a prediction model using an ensemble of weak classifiers, typically decision trees. It builds the model in a stage-wise fashion, and constructs additive regression models by sequentially fitting a simple parameterized function (base learner) to current pseudo-residuals by least squares at each iteration. The pseudo-residuals are the gradient of the loss functional being minimized. It is worth noting that both the approximation accuracy and execution speed of gradient boosting can be substantially improved by incorporating randomization into the procedure (Friedman, 2002). The parameters of SGB (depth of interacting, the number of tress, and shrinkage) were optimized for prediction models.

In statistical prediction, the following three cross-validation methods are often used to examine a predictor for its effectiveness in practical application: independent test, subsampling test, and jackknife test. However, of the three test methods, the jackknife test is deemed the least arbitrary that can always yield a unique result for a given benchmark dataset as elaborated in (Chou, 2011) and demonstrated by Eqs.28–30 in (Chou, 2011). Accordingly, the jackknife test has been widely recognized and increasingly used by investigators to examine the quality of various predictors (Chen et al., 2016a; Dehzangi et al., 2015; Hajisharifi et al., 2014; Khan et al., 2015; Nanni et al., 2014; Tahir and Hayat, 2016). However, to reduce the computational time, we adopted the 10-fold cross-validation in this study as done by many investigators with some classification algorithm as the prediction engine to train the prediction model (i.e. train the parameters). Then, the independent test was applied to evaluate the trained prediction models, and used for comparison among different methods.

2.4. Model evaluation

The performance of prediction models was evaluated by:

$$\mathrm{Recall} = \frac{TP}{TP + FN} \tag{9}$$

$$\mathrm{Precision} = \frac{TP}{TP + FP} \tag{10}$$

$$F_1 = \frac{2*Recall*Precision}{Recall + Precision} \tag{11}$$

$$\mathrm{MCC} = \frac{TP*TN - FP*FN}{\sqrt{(TP + FP)*(TN + FN)*(TP + FN)*(TN + FP)}} \tag{12}$$

where TP is the number of correctly predicted effective drug combinations, TN is the number of correctly predicted non-effective drug combinations, FP is the number of non-effective drug combinations predicted as effective ones, and FN is the number of effective drug

combinations wrongly predicted as non-effective ones. (Baldi et al., 2000; Xiong et al., 2011a, 2011b, 2012). Since the four metrics, particularly the MCC, are lacking intuitiveness and not easy to understand for most biologists. The readers could take the advantage of using Eq.14 of (Chen et al., 2013b) and Eq.11 of (Lin et al., 2014) for more intuitive and easier-to-understand metrics formulations. The advantage of using this kind of intuitive merits has been concurred by a series of studies published very recently (Guo et al., 2014; Jia et al., 2016b, 2016c; Liu et al., 2016a, 2016b, 2014a; Qiu et al., 2014). The set of metrics is valid only for the single-label systems. For the multi-label systems whose existence has become more frequent in system biology (Chou et al., 2011, 2012; Lin et al., 2013; Wu et al., 2011) and system medicine (Qiu et al., 2016b; Xiao et al., 2013c), a completely different set of metrics as defined in (Chou, 2013) is used. Moreover, a receiver operating characteristic (ROC) curve is plotted by the sensitivity versus (1-specificity) for a binary classifier at dynamic thresholds ranging from 0 to 1. The area under the curve (AUC) was used as a measure to evaluate the predictive performance.

3. Results

3.1. Analysis of the features of drug combinations

As described above, a feature vector of 732 dimensions were constructed to represent a pairwise drug combination. However, the resultant high-dimensional feature vector increased the possibility of being relevant or redundant among its feature elements. In this section, we aimed to produce an mRMR feature list. The mRMR method ranked each feature according to both its relevance to the target classification variable and the redundancy between the features (Huang et al., 2010). The features with scores larger than zero were selected for model construction in the next step. The top 50 features ranked by mRMR were investigated to analyze the importance of different types of features. As shown in Fig. 1, the features of CCI confidence score and ATC codes accounted for more proportion than other types of the features in the selected ones. The percentages of different features in Fig. 1 were normalized by the dimensions of the features since the number of elements or dimensions from different types of features had a wide range.

Next, to examine the distinguishing power of different features, the statistical analysis was performed, such as Kolmogorov-Smirnov test, which is a nonparametric test for determining whether two samples of observations come from the same distribution. Fig. 2 shows that the distributions of chemical-chemical interaction confidence scores were different in golden positive samples and golden negative samples. The p value was 9.17178E-59, confirming the statistical significance of the difference of CCI confidence scores between two groups of samples.

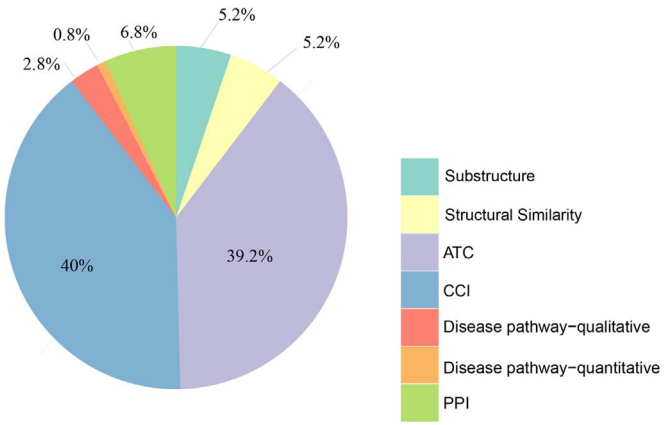

Fig. 1. Distribution of the top ranked features by mRMR.

3.2. Performance comparison of different classification algorithms

The aim of this section is to compare the classification performance of different machine learning algorithms (SVM, Naïve Bayes, and SGB) for prediction of effective drug combinations. Fig. 3 shows ROC curves of models of the three type of classification algorithms by 10-fold cross validation on the training data set. As shown in Fig. 3, prediction model based on SGB performed the best, with its *AUC* up to 0.9775, followed by SVM, which was inferior to that of SGB, while the one using Naïve Bayes was the worst.

Table 1 shows the classification performances of the three types of algorithms on the independent tests. Among the three methods, all performance metrics of SGB and SVM were higher than that of Naïve Bayes. Moreover, not only the *AUC* score of SGB was higher than that of SVM, but SGB gave a higher F_1 of 0.8979 and *MCC* of 0.8046 compared with SVM (0.8182 for F_1, 0.6598 for *MCC*) as well.

Overall, the prediction model based on SGB achieved best performance. We chose the SGB as the classification engine to construct the prediction models in the following parts.

3.3. Predictive power of features using SGB

In previous sections, we calculated the scores of features ranked by the mRMR, which implied that their different levels of distinguishing power to identify effective drug combinations. In this section, we further validated the prediction power of features by SGB algorithm.

Fig. 4 presents the ROC curves of the models using different types of features by SGB. The area under ROC of the model based on ATC was 0.8103, which implied the feature of ATC had strong potential to predict whether a drug combination is effective or not. The model based on the feature of CCI was also performed well, while the distinguishing power of 2D substructure was interior.

Furthermore, we combined the different types of features to validate their predictive power. As we mentioned above, the six types of features were categorized into three group: biological information (PPI and disease pathways), chemical information (2D substructure, structural similarity and CCI), and pharmacological information (therapeutic similarity). The performance was significantly improved (the AUC was increased from 0.6562 to 0.8428) by adding therapeutic information to biological features. When we added chemical features to biological ones, the performance was also greatly improved (AUC from 0.6562 to 0.8459). The combination of the therapeutic and chemical features yielded the better AUC value (0.9033). When combined with the biological features, the prediction model gave the best performance (AUC was 0.9519). These results demonstrate that these three groups of features were capable of providing complementary information for discriminating effective drug combination from non-effective ones. KS test was applied here to analyze the significance of the addition of chemical features to the biological features (p=1.083e-05), and the improvement of the addition of therapeutic features was also significant, as shown in Table 2. Therefore, the SGB model integrating all three groups of features was selected as the final model in our study.

The pharmacological features (ATC part) had most promising ability to prediction (*AUC*=0.8280). To test whether ATC was a dominant feature of the classification model, we re-evaluated the feature importance by leaving out the feature of ATC. It was shown that the prediction performance (*AUC*=0.8459) was declined in comparison to that of the classifier using all features.

3.4. Case study

In order to validate whether the PDC-SGB we proposed here was effective in real application, we then applied our final model in any pairs of drugs which have not yet been known effective or not in the 65 marketed hypertension drugs (Table S4).

As a consequence, our PDC-SGB model predicted the 17 potential

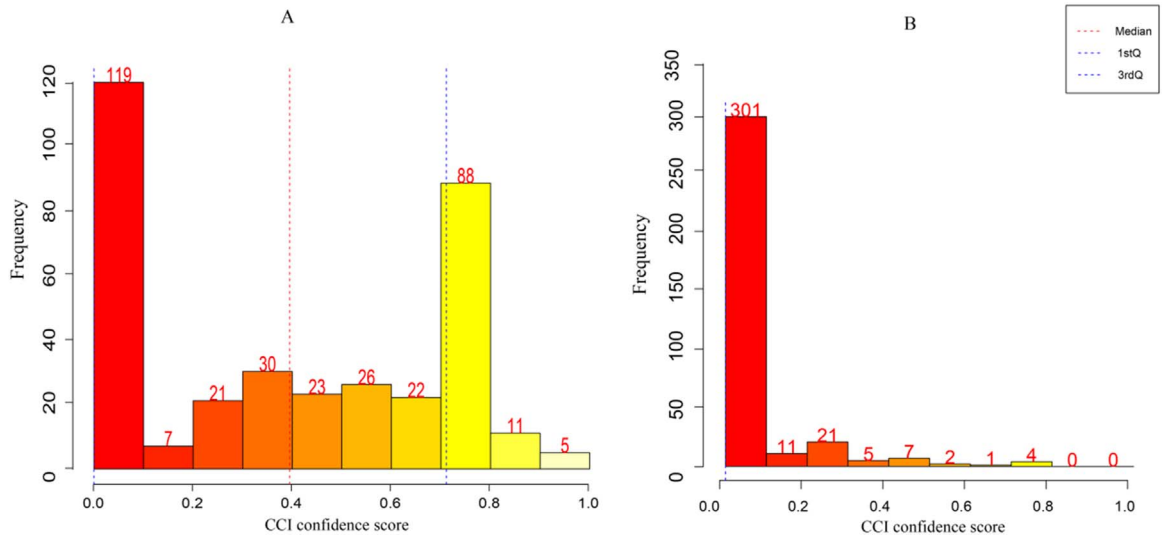

Fig. 2. The distribution of CCI confidence score in GPS and GNS. (A) Histogram of CCI confidence score distribution in GPS; (B) Histogram of CCI confidence score distribution in GNS.

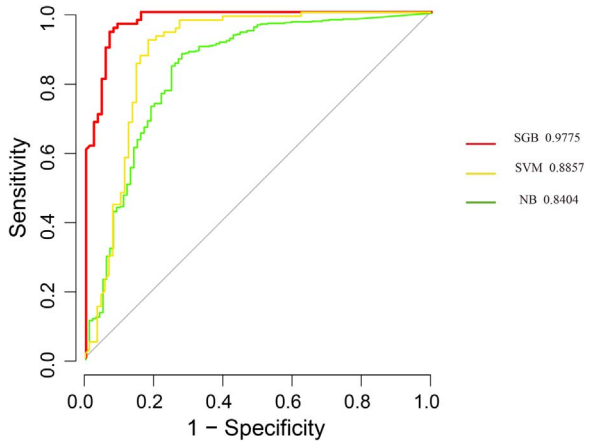

Fig. 3. ROC curves for models based on three types of machine learning methods by 10-fold cross validation on the training data set.

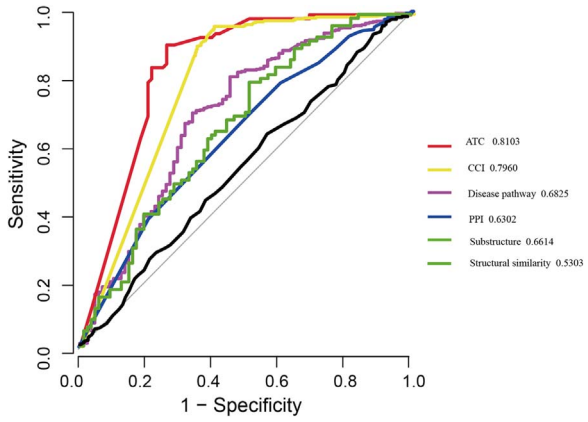

Fig. 4. Performance comparison of different features by SGB using 10-fold cross-validation test.

Table 1
The performance of the prediction models based on different classification algorithms on the independent test.

Algorithms	*AUC*	F_1	*MCC*	*Recall*	*Precision*
Naïve bayes	0.8476	0.6222	0.5442	0.6214	0.6243
SVM	0.8815	0.8182	0.6598	0.7761	0.8662
SGB	0.9519	0.8979	0.8046	0.8693	0.9292

effective drug combinations with high confidence scores larger than 0.6. A literature search in PubMed showed that 6 out of our 17 predictions have already been reported to be effective drug combinations in the literature, although they have not yet been approved by the FDA (Table S5). They may become novel effective drug combinations for further study.

For example, the predicted drug combination of Amlodipine and Dipyridamole which played complementary roles in the treatment of hypertension and angina would reduce the rate of neuronal cell death compared with the amlodipine alone in the treatment of cerebrovascular stroke and in hypertensive patients (Yamagata et al., 2004). When the propranolol was combined with the indapamide for the treatment of hypertension, the patients had a progressive and significant improvement in the controlled process of blood pressure. It showed that indapamide alone controlled blood pressure in 82% of the patients, while the combined therapy controlled blood pressure in 85% of patients (Athanassiadis et al., 1990).

Table 2
Performance comparison of prediction models using various combinations of features.

Feature Type	AUC	F_1	MCC	Recall	Precision
Biological	0.6562	0.6296	0.2444	0.6443	0.6177
Pharmacological	0.8280	0.5260	0.4826	0.4227	0.7119
Chemical	0.7929	0.6029	0.5599	0.4972	0.7660
Biological-Therapeutic	0.8428	0.7812	0.5836	0.7466	0.8200
Biological-Chemical	0.8459	0.7573	0.5895	0.6716	0.8740
Chemical-Therapeutic	0.9033	0.6674	0.6145	0.6045	0.7471
Biological-Chemical-Therapeutic	0.9519	0.8979	0.8046	0.8693	0.9292

Table 3
Performance comparison of prediction models with Zhao et al.'s method.

Methods	*AUC*	F_1	MCC	Recall	Precision
Zhao et al.'s method	–	0.6263	–	0.8769	0.4871
Our method	0.9519	0.8979	0.8046	0.8693	0.9292

3.5. Comparison with other methods

We employed our dataset to the available method proposed by (Zhao et al., 2011), and the performances comparison were shown in

Table 3. Zhao et.al presented a screening method based on frequent pattern of system molecular, biological and pharmacological data. The features of target protein, therapy, side effect, indications and pathways were constructed to represent drug combinations. The training set was used to calculate the enrichment score for each feature pair while the remaining group was used as the validation set to evaluate the performance of the feature pairs. Besides, Zhao et al. predicted a drug pair as an effective combination if its confidence score is above the threshold instead of using machine learning algorithms. It is obvious that the performance of our method is better than the Zhao et al.'s method.

The features about the pharmacological, target and biological pathways were employed in both Zhao et al.'s and our study. There are three main differences between the two studies: first of all, the confidence score of chemical interaction which was one of the key factors for the discrimination of drug combination in our study was not used in Zhao et al.'s study. Secondly, as for the feature vectors, Zhao et al. used 3 features of target, indication and therapy information which had promising performance when each of them was applied to build the model alone, while in our method, the final feature vectors were chosen by mRMR, one popular feature selection method. Last but not the least, in Zhao et al.'s study, they chose a simple method (maximization of the F_1 score), whereas we employed the sophisticated machine learning method (SGB) to build the classification model for prediction of drug combinations.

4. Discussion

In this study, the six types of features, which were categorized into three main groups of features (biological, pharmacological, and chemical properties were integrated to build PDC-SGB model for prediction of effective drug combinations.

The CCI confidence score of d_1 and d_2 in drug combination (d_1, d_2) was the individual feature show strong predictive power, which indicated that interaction of drugs as chemicals was a key element for classification of drug combinations. The other predictive features with strong power were the similarity confidence scores calculated from the first level of ATC code and the mean score of first three levels, which means that the therapy information of drugs plays an important role in identification of drug combinations. However, it was shown that the performance of biological part was relatively lower, which may be attributed by the two main factors. Firstly, the incompleteness of molecular networks or biological pathways led to the poor quality of biological features with noise. Secondly, the representation method we used for constructing the biological features were too simple, and could not be able to represent the change of biological mechanism after a drug is taken in.

Moreover, our result showed that models based on all features outperformed better than the model based on a random predictor, which suggests that our features could contribute to distinguishing whether a drug combination is effective or not. Among the features, the feature based on Tanimoto coefficient of molecular 2D structure was not well predictive, because the simple similarity of the chemical drug could not reflect the complex mechanism of drug combinations.

5. Conclusions

Drug combination therapy is a promising strategy for combating the complex diseases. In this study, we proposed a new computational method to predict effective drug combinations by integrating biological, chemical and pharmacological information. Among the selected features, the CCI and therapy information of drugs showed powerful potential to discriminate effective drug combinations from the non-effective ones. Three different classification algorithms were applied to build prediction models. The prediction model based on SGB gave the best performance among the three kind of models which was selected as our final model.

Some of the potential combinations we predicted had literature validation, which proved the effectiveness of our model. We believed that our method will help narrow the search space of possible drug combinations in future. As demonstrated in a series of recent publications (Chen et al., 2013b; Chen et al., 2016b; Jia et al., 2016a; Qiu et al., 2016c; Xiao et al., 2016) in developing new prediction methods, user-friendly and publicly accessible web-servers will significantly enhance their impacts (Chou, 2015), we shall make efforts in our future work to provide a web-server for the prediction method presented in this paper.

Acknowledgment

The authors would like to thank the two anonymous reviewers for their insightful suggestions on strengthening the presentation, quality and rigorous of this article. This work was supported by the funding from National Key Research Program (Contract No. 2016YFA0501703), the grant from National Natural Science Foundation of China (NSFC, Grant No. 31371261), and the grants from NSFC for Young Scholars (Grant No. 31601074 and 31400704).

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.jtbi.2017.01.019.

References

Athanassiadis, D.I., Dimopoulos, C.G., Tsakiris, A.K., Cokkinos, D.F., Tourkantonis, A.A., Toutouzas, P.K., Boutin, B., Guez, D., 1990. Clinical efficacy and quality of life with indapamide alone or in combination with beta blockers or angiotensin-converting enzyme inhibitors. Am. J. Cardiol. 65, 62H–66H.

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C.A., Nielsen, H., 2000. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 16, 412–424.

Chatr-Aryamontri, A., Breitkreutz, B.J., Oughtred, R., Boucher, L., Heinicke, S., Chen, D., Stark, C., Breitkreutz, A., Kolas, N., O'Donnell, L., Reguly, T., Nixon, J., Ramage, L., Winter, A., Sellam, A., Chang, C., Hirschman, J., Theesfeld, C., Rust, J., Livstone, M.S., Dolinski, K., Tyers, M., 2015. The BioGRID interaction database: 2015 update. Nucleic Acids Res 43, D470–D478.

Chen, J., Long, R., Wang, X.L., Liu, B., Chou, K.C., 2016a. dRHP-PseRA: detecting remote homology proteins using profile-based pseudo protein sequence and rank aggregation. Sci. Rep. 6, 32333.

Chen, L., Li, B.Q., Zheng, M.Y., Zhang, J., Feng, K.Y., Cai, Y.D., 2013a. Prediction of effective drug combinations by chemical interaction, protein interaction and target enrichment of KEGG pathways. Biomed. Res Int 2013, 723780.

Chen, W., Feng, P.M., Lin, H., Chou, K.C., 2013b. iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. Nucleic Acids Res 41, e68.

Chen, W., Ding, H., Feng, P., Lin, H., Chou, K.C., 2016a. iACP: a sequence-based tool for identifying anticancer peptides. Oncotarget 7, 16895–16909.

Chen, W., Tang, H., Ye, J., Lin, H., Chou, K.-C., 2016b. iRNA-PseU: Identifying RNA pseudouridine sites. Mol. Ther. Nucleic Acids 5, e332.

Chen, X., Ren, B., Chen, M., Wang, Q., Zhang, L., Yan, G., 2016d. NLLSS: predicting synergistic drug combinations based on Semi-supervised learning. PLoS Comput. Biol. 12, e1004975.

Chou, K.C., 2011. Some remarks on protein attribute prediction and pseudo amino acid composition. J. Theor. Biol. 273, 236–247.

Chou, K.C., 2013. Some remarks on predicting multi-label attributes in molecular biosystems. Mol. Biosyst. 9, 1092–1100.

Chou, K.C., 2015. Impacts of bioinformatics to medicinal chemistry. Med. Chem. 11, 218–234.

Chou, K.C., Wu, Z.C., Xiao, X., 2011. iLoc-Euk: a multi-label classifier for predicting the subcellular localization of singleplex and multiplex eukaryotic proteins. PLoS One 6, e18258.

Chou, K.C., Wu, Z.C., Xiao, X., 2012. iLoc-Hum: using the accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites. Mol. Biosyst. 8, 629–641.

Csermely, P., Korcsmaros, T., Kiss, H.J., London, G., Nussinov, R., 2013. Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review. Pharm. Ther. 138, 333–408.

Dehzangi, A., Heffernan, R., Sharma, A., Lyons, J., Paliwal, K., Sattar, A., 2015. Gram-positive and Gram-negative protein subcellular localization by incorporating evolutionary-based descriptors into Chous general PseAAC. J. Theor. Biol. 364, 284–294.

Fan, Y.N., Xiao, X., Min, J.L., Chou, K.C., 2014. iNR-Drug: predicting the interaction of

drugs with nuclear receptors in cellular networking. Int J. Mol. Sci. 15, 4915–4937.
Friedman, J.H., 2002. Stochastic gradient boosting. Comput. Stat. Data Anal. 38, 367–378.
Guo, S.H., Deng, E.Z., Xu, L.Q., Ding, H., Lin, H., Chen, W., Chou, K.C., 2014. iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition. Bioinformatics 30, 1522–1529.
Hajisharifi, Z., Piryaiee, M., Mohammad Beigi, M., Behbahani, M., Mohabatkar, H., 2014. Predicting anticancer peptides with Chou's pseudo amino acid composition and investigating their mutagenicity via Ames test. J. Theor. Biol. 341, 34–40.
He, Z., Zhang, J., Shi, X.H., Hu, L.L., Kong, X., Cai, Y.D., Chou, K.C., 2010. Predicting drug-target interaction networks based on functional groups and biological features. PLoS One 5, e9603.
Huang, T., Zhang, J., Xu, Z.-P., Hu, L.-L., Chen, L., Shao, J.-L., Zhang, L., Kong, X.-Y., Cai, Y.-D., Chou, K.-C., 2012. Deciphering the effects of gene deletion on yeast longevity using network and machine learning approaches. Biochimie 94, 1017–1025.
Huang, T., Wang, P., Ye, Z.Q., Xu, H., He, Z., Feng, K.Y., Hu, L., Cui, W., Wang, K., Dong, X., Xie, L., Kong, X., Cai, Y.D., Li, Y., 2010. Prediction of deleterious non-synonymous SNPs based on protein interaction network and hybrid properties. PLoS One 5, e11900.
Jia, J., Liu, Z., Xiao, X., Liu, B., Chou, K.C., 2015. iPPI-Esml: An ensemble classifier for identifying the interactions of proteins by incorporating their physicochemical properties and wavelet transforms into PseAAC. J. Theor. Biol. 377, 47–56.
Jia, J., Liu, Z., Xiao, X., Liu, B., Chou, K.C., 2016a. iCar-PseCp: identify carbonylation sites in proteins by Monte Carlo sampling and incorporating sequence coupled effects into general PseAAC. Oncotarget 7, 34558–34570.
Jia, J., Zhang, L., Liu, Z., Xiao, X., Chou, K.C., 2016b. pSumo-CD: predicting sumoylation sites in proteins with covariance discriminant algorithm by incorporating sequence-coupled effects into general PseAAC. Bioinformatics 32, 3133–3141.
Jia, J., Liu, Z., Xiao, X., Liu, B., Chou, K.C., 2016c. pSuc-Lys: Predict lysine succinylation sites in proteins with PseAAC and ensemble random forest approach. J. Theor. Biol. 394, 223–230.
Jia, J., Zhu, F., Ma, X., Cao, Z., Li, Y., Chen, Y.Z., 2009. Mechanisms of drug combinations: interaction and network perspectives. Nat. Rev. Drug Discov. 8, 111–128.
Khan, Z.U., Hayat, M., Khan, M.A., 2015. Discrimination of acidic and alkaline enzyme using Chou's pseudo amino acid composition in conjunction with probabilistic neural network model. J. Theor. Biol. 365, 197–203.
Kuhn, M., 2008. Building Predictive models in R using the caret package. J. Stat. Softw. 28, 1–26.
Kuhn, M., Szklarczyk, D., Pletscher-Frankild, S., Blicher, T.H., von Mering, C., Jensen, L.J., Bork, P., 2014. STITCH 4: integration of protein-chemical interactions with user data. Nucleic Acids Res. 42, D401–D407.
Law, V., Knox, C., Djoumbou, Y., Jewison, T., Guo, A.C., Liu, Y., Maciejewski, A., Arndt, D., Wilson, M., Neveu, V., Tang, A., Gabriel, G., Ly, C., Adamjee, S., Dame, Z.T., Han, B., Zhou, Y., Wishart, D.S., 2014. DrugBank 4.0: shedding new light on drug metabolism. Nucleic Acids Res. 42, D1091–D1097.
Lin, H., Deng, E.Z., Ding, H., Chen, W., Chou, K.C., 2014. iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. Nucleic Acids Res. 42, 12961–12972.
Lin, W.Z., Fang, J.A., Xiao, X., Chou, K.C., 2013. iLoc-Animal: a multi-label learning classifier for predicting subcellular localization of animal proteins. Mol. Biosyst. 9, 634–644.
Liu, B., Long, R., Chou, K.C., 2016a. iDHS-EL: identifying DNase I hypersensitive sites by fusing three different modes of pseudo nucleotide composition into an ensemble learning framework. Bioinformatics 32, 2411–2418.
Liu, B., Fang, L., Long, R., Lan, X., Chou, K.C., 2016b. iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition. Bioinformatics 32, 362–369.
Liu, B., Liu, F., Wang, X., Chen, J., Fang, L., Chou, K.C., 2015. Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. Nucleic Acids Res. 43, W65–W71.
Liu, B., Zhang, D., Xu, R., Xu, J., Wang, X., Chen, Q., Dong, Q., Chou, K.C., 2014a. Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection. Bioinformatics 30, 472–479.
Liu, Y., Wei, Q., Yu, G., Gai, W., Li, Y., Chen, X., 2014b. DCDB 2.0: a major update of the drug combination database. Database (Oxf.), 2014, (bau124).
Min, J.L., Xiao, X., Chou, K.C., 2013. iEzy-drug: a web server for identifying the interaction between enzymes and drugs in cellular networking. Biomed. Res. Int. 2013, 701317.
Nanni, L., Brahnam, S., Lumini, A., 2014. Prediction of protein structure classes by incorporating different protein descriptors into general Chou's pseudo amino acid composition. J. Theor. Biol. 360, 109–116.
Peng, H., Long, F., Ding, C., 2005. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. Pattern Anal. Mach. Intell. 27, 1226–1238.
Qiu, W.R., Xiao, X., Chou, K.C., 2014. iRSpot-TNCPseAAC: identify recombination spots with trinucleotide composition and pseudo amino acid components. Int J. Mol. Sci. 15, 1746–1766.
Qiu, W.R., Xiao, X., Xu, Z.C., Chou, K.C., 2016a. iPhos-PseEn: Identifying phosphorylation sites in proteins by fusing different pseudo components into an ensemble classifier. Oncotarget.
Qiu, W.R., Sun, B.Q., Xiao, X., Xu, Z.C., Chou, K.C., 2016b. iPTM-mLys: identifying multiple lysine PTM sites and their different types. Bioinformatics 32, 3116–3123.
Qiu, W.R., Sun, B.Q., Xiao, X., Xu, Z.C., Chou, K.C., 2016c. iHyd-PseCp: Identify hydroxyproline and hydroxylysine in proteins by incorporating sequence-coupled effects into general PseAAC. Oncotarget 7, 44310–44321.
Ryall, K.A., Tan, A.C., 2015. Systems biology approaches for advancing the discovery of effective drug combinations. J. Cheminform. 7, 7.
Shen, H.B., Chou, K.C., 2009. Predicting protein fold pattern with functional domain and sequential evolution information. J. Theor. Biol. 256, 441–446.
Shen, H.B., Chou, K.C., 2010. Gneg-mPLoc: a top-down strategy to enhance the quality of predicting subcellular localization of Gram-negative bacterial proteins. J. Theor. Biol. 264, 326–333.
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., 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA 102, 15545–15550.
Sun, Y., Xiong, Y., Xu, Q., Wei, D., 2014. A hadoop-based method to predict potential effective drug combination. Biomed. Res. Int. 2014, 196858.
Sun, Y., Sheng, Z., Ma, C., Tang, K., Zhu, R., Wu, Z., Shen, R., Feng, J., Wu, D., Huang, D., Huang, D., Fei, J., Liu, Q., Cao, Z., 2015. Combining genomic and network characteristics for extended capability in predicting synergistic drugs for cancer. Nat. Commun. 6, 8481.
Tahir, M., Hayat, M., 2016. INuc-STNC: a sequence-based predictor for identification of nucleosome positioning in genomes by extending the concept of SAAC and Chou's PseAAC. Mol. Biosyst. 12, 2587–2593.
Vilar, S., Cozza, G., Moro, S., 2008. Medicinal chemistry and the molecular operating environment (MOE): application of QSAR and molecular docking to drug discovery. Curr. Top. Med. Chem. 8, 1555–1572.
Vilar, S., Uriarte, E., Santana, L., Lorberbaum, T., Hripcsak, G., Friedman, C., Tatonetti, N.P., 2014. Similarity-based modeling in large-scale prediction of drug-drug interactions. Nat. Protoc. 9, 2147–2163.
Wang, Y.Y., Xu, K.J., Song, J., Zhao, X.M., 2012. Exploring drug combinations in genetic interaction network. BMC Bioinform. 13 (Suppl 7), S7.
Wu, Z., Zhao, X.M., Chen, L., 2010. A systems biology approach to identify effective cocktail drugs. BMC Syst. Biol. 4 (Suppl 2), S7.
Wu, Z.C., Xiao, X., Chou, K.C., 2011. iLoc-Plant: a multi-label classifier for predicting the subcellular localization of plant proteins with both single and multiple sites. Mol. Biosyst. 7, 3287–3297.
Xiao, X., Min, J.L., Wang, P., Chou, K.C., 2013a. iGPCR-drug: a web server for predicting interaction between GPCRs and drugs in cellular networking. PLoS One 8, e72234.
Xiao, X., Min, J.L., Wang, P., Chou, K.C., 2013b. iCDI-PseFpt: identify the channel-drug interaction in cellular networking with PseAAC and molecular fingerprints. J. Theor. Biol. 337, 71–79.
Xiao, X., Wang, P., Lin, W.Z., Jia, J.H., Chou, K.C., 2013c. iAMP-2L: a two-level multi-label classifier for identifying antimicrobial peptides and their functional types. Anal. Biochem 436, 168–177.
Xiao, X., Ye, H.X., Liu, Z., Jia, J.H., Chou, K.C., 2016. iROS-gPseKNC: predicting replication origin sites in DNA by incorporating dinucleotide position-specific propensity into general pseudo nucleotide composition. Oncotarget 7, 34180–34189.
Xiao, X., Min, J.L., Lin, W.Z., Liu, Z., Cheng, X., Chou, K.C., 2015. iDrug-Target: predicting the interactions between drug compounds and target proteins in cellular networking via benchmark dataset optimization approachJ Biomol Struct Dyn 33, 2221–2233.
Xiong, Y., Liu, J., Wei, D.Q., 2011a. An accurate feature-based method for identifying DNA-binding residues on protein surfaces. Proteins 79, 509–517.
Xiong, Y., Xia, J., Zhang, W., Liu, J., 2011b. Exploiting a reduced set of weighted average features to improve prediction of DNA-binding residues from 3D structures. PLoS One 6, e28440.
Xiong, Y., Liu, J., Zhang, W., Zeng, T., 2012. Prediction of heme binding residues from protein sequences with integrative sequence profiles. Proteome Sci. 10 (Suppl 1), S20 .
Xu, K.J., Song, J., Zhao, X.M., 2012. The drug cocktail network. BMC Syst. Biol. 6 (Suppl 1), S5.
Yamagata, K., Ichinose, S., Tagami, M., 2004. Amlodipine and carvedilol prevent cytotoxicity in cortical neurons isolated from stroke-prone spontaneously hypertensive rats. Hypertens. Res. 27, 271–282.
Zhao, X.M., Iskar, M., Zeller, G., Kuhn, M., van Noort, V., Bork, P., 2011. Prediction of drug combinations by integrating molecular and pharmacological data. PLoS Comput. Biol. 7, e1002323.