Toxicology 304 (2013) 173–184

Contents lists available at SciVerse ScienceDirect

Toxicology

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

A systematic identification of multiple toxin–target interactions based on chemical, genomic and toxicological data

Wei Zhou[a,1], Chao Huang[a,1], Yan Li[b], Jinyou Duan[c], Yonghua Wang[a,*], Ling Yang[d]

[a] College of Life Science, Northwest A&F University, Yangling, Shaanxi 712100, China
[b] School of Chemical Engineering, Dalian University of Technology, Dalian, Liaoning 116024, China
[c] College of Science, Northwest A&F University, Yangling, Shaanxi 712100, China
[d] Lab of Pharmaceutical Resource Discovery, Dalian Institute of Chemical Physics, Chinese Academy of Sciences, Dalian, Liaoning 116023, China

ARTICLE INFO

Article history:
Received 2 September 2012
Received in revised form
19 December 2012
Accepted 20 December 2012
Available online 11 January 2013

Keywords:
Multiple toxin–target interactions
In silico prediction
SVM
RF
Network toxicology

ABSTRACT

Although the assessment of toxicity of various agents, -omics (genomic, proteomic, metabolomic, etc.) data has been accumulated largely, the acquirement of toxicity information of variety of molecules through experimental methods still remains a difficult task. Presently, a systems toxicology approach that integrates massive diverse chemical, genomic and toxicological information was developed for prediction of the toxin targets and their related networks. The procedures are: (1) by use of two powerful statistical methods, i.e., support vector machine (SVM) and random forest (RF), a systemic model for prediction of multiple toxin–target interactions using the extracted chemical and genomic features has been developed with its reliability and robustness estimated. And the qualitative classification of targets according to the phenotypic diseases has been taken into account to further uncover the biological meaning of the targets, as well as to validate the robustness of the in silico models. (2) Based on the predicted toxin–target interactions, a genome-scale toxin–target-disease network exampled by cardiovascular disease is generated. (3) A topological analysis of the network is carried out to identify those targets that are most susceptible in human to topical agents including the most critical toxins, as well as to uncover both the toxin-specific mechanisms and pathways. The methodologies presented herein for systems toxicology will make drug development, toxin environmental risk assessment more efficient, acceptable and cost-effective.

1. Introduction

With thousands of new chemicals being synthesized year by year, increased efforts are being devoted to evaluating their toxicity properties. Undoubtedly, the toxicity evaluation task of such high volume of compounds is of fundamental importance to both the ecosystems and human health. Normally, in silico methods are effective ways for the job of virtual screening of unknown molecules even before their synthesis (Pritchard et al., 2003; Wang et al., 2008; Zhang et al., 2012), which clearly is important to complement the experimental approaches for reducing time and cost, and thus accelerating the prioritization of those compounds of interest. However, all these techniques have their inherent limitations in either the predictivity, application domain or even algorithms themselves (Butina et al., 2002). More severely, most available toxic data involve diverse kinds of compounds, but are evaluated by a same or similar toxicological endpoint (lethal doses, macroscopic toxicity) (Huang et al., 2009). This makes the precise prediction of a toxin mechanism from a molecular level is often impossible, let alone to consider the multiple toxin–targets interactions.

Due to both the vastness of chemical space (toxins) and the diversity of biological systems (targets), the prediction and characterization of the two domains' interface is difficult. In addition, the interaction patterns of toxins and targets are usually complicated by the fact that they are not simple one-to-one events, as one toxin may bind to multiple target proteins, and different toxins may also bind to the same protein target with similar biological activities (Yabuuchi et al., 2011). Thus it is compelling for considering multitarget strategies over single-target approaches to study the complex interactions, which strategies, however, are seldom studied at present.

Recently, several novel attempts have been made to fulfill this goal. For instance, a chemical genomics approach whose salient motivation is that similar ligands may interact with similar proteins has been used to explore novel bioactive molecules of a target (Klabunde, 2007; Yamanishi et al., 2010). The network

[*] Corresponding author. Tel.: +86 029 87092262.
E-mail address: yh_wang@nwsuaf.edu.cn (Y. Wang).
[1] These authors contributed equally.

0300-483X/$ – see front matter. Crown Copyright © 2012 Published by Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.tox.2012.12.012

approaches may also provide a chance to explore complex biosystem interactions, which in biology have been proven useful for organizing and/or extracting meaningful information from high-dimensional biological data (Yu et al., 2012). And advances in this direction should be helpful to uncover the biological significance of ligand–target interactions. Despite of these efforts, to our knowledge, little is known of the underlying complex interactions between the toxins and targets, and a systems-level characterization of multiple toxin–target associations has not yet been reported up to date.

Generally, the quantitative prediction of biological activities (IC50, EC50, Ki values, etc.) of chemicals should be valuable for precise charactering these candidates. But in many cases, it is not easy to comprehensively retrieve enough reliable biological information for ligands, particularly for large datasets. This is also true for the present compound–protein interactions of this work, which are consisted of heterogeneous data of various resources with different bioassay systems. In addition, it is also difficult to construct an accurate model for predicting activity values of ligands due to the unavailability of reliable and consistent activity information from the present available data. However, a qualitative prediction system that identifies the potential toxin–target relationships may eventually overcome the above limitations. For example, the classification methods usually do not need accurate biological data but a qualitative description of biological groupings of chemicals is enough to build reasonable models. For those widely applied mathematical tools, such as the support vector machine (SVM) and random forest (RF), generally speaking, they are similar to the multiple linear regression (MLR) method. The main difference is that MLR is mainly involved in solving linear fitting problems whereas SVM and RF is nonlinear, which thus in most cases are more appropriate to biological problems due to the inherent nonlinear property in biology.

In this work, a series of computational models were established to identify the complex toxin–target interactions. The procedures are: firstly, by employing two powerful statistical methods, i.e., SVM and RF, the models were constructed with their predictive capacity evaluated by both the internal cross-validation and external tests, which ended up with good performance in both the reliability and robustness. Subsequently, according to the applicability domain (AD) and feature analysis of the models, those compounds predicted with high or poor accuracies were individually identified. Finally, as an example, a genome-scale toxin–target network for cardiovascular diseases was generated, and the topology analysis of which may provide us further insights into the toxin–target interaction mechanism and specific action pathways.

2. Materials and methods

2.1. Building of dataset

Data for toxins and targets with their interaction information were extracted from the Toxin and Toxin–Target Database (T3DB, http://www.t3db.org), which currently contain over 2900 small molecules and peptide toxins, 1300 targets and more than 33,800 toxin–target associations. The original database was manually built from numerous sources, including the electronic databases, government documents, textbooks and scientific journals following such criteria: (i) these compounds can be found in the home, environment or workplace with medical consequence records like acute reaction, injury or death; (ii) they are routinely identified as hazardous resources in relatively low concentrations (<1 mM for some, <1 μM for others); (iii) they appear on multiple toxin/poison lists provided by the government agencies or the toxicological and medical literature; (iv) these substances must be identified as specific toxic components with known chemical structures.

Since some molecular descriptors of chemicals and peptides cannot be calculated, two kinds of toxic substances, i.e., arsenic, lead, mercury, phosphorus, restrictocin, etc., were omitted in this study. Additionally, those compounds including sodium, potassium salts were calculated for their water-dissolved products to obtain the molecular descriptors. Finally, a data set of 26,277 toxin–target pairs composed of 2257 toxins and corresponding 949 targets was compiled. The names and ID codes of the toxins and proteins were provided in Table S1.

Supplementary material related to this article found, in the online version, at http://dx.doi.org/10.1016/j.tox.2012.12.012.

2.2. Calculation of chemical and protein descriptors

Chemical descriptors were calculated using DRAGON 5.4 program (http://www.talete.mi.it/index.htm), which has been proven successful in evaluation of molecular structure–activity or structure–property relationships (Wang et al., 2010). As a result, 1664 descriptors were calculated from 20 molecular descriptor blocks: constitutional descriptors, topological descriptors, two-dimensional (2D) autocorrelations, molecular properties et al. (with details referred to DRAGON manual). After eliminating those descriptors that were not available for each molecule or were constant values for all molecules, 1547 molecular descriptors were finally adopted in the subsequent processing (Table S2).

Supplementary material related to this article found, in the online version, at http://dx.doi.org/10.1016/j.tox.2012.12.012.

The dipeptide composition was used to transform the variable length of proteins to the fixed length feature vectors, which has already been used in the protein structural classifications, compound–protein interaction predictions and protein subcellular localizations fields (Yabuuchi et al., 2011). In our previous work, we also adopted the dipeptide composition-based descriptors to predict the drug–target interactions (Yu et al., 2012). Dipeptide composition encapsulates information about the fraction of amino acids and their local order, which gives a fixed pattern length of 400 (20 × 20). The fraction of each dipeptide was calculated using the following equation:

$$\text{Fraction of dep}(i) = \frac{\text{total number of dep}(i)}{\text{total number of all possible dipeptides}} \quad (1)$$

where dep(i) is one dipeptide i of 400 dipeptides.

2.3. Construction of training and test sets

To distinguish the interacted toxin–target pairs from the non-interaction ones, an experimental dataset including both positive and negative samples which were represented by concatenating chemical descriptors and protein descriptors (1547 + 400 dimensions) was firstly established. This dataset was then split into two subsets, i.e., a training set used to build the model and an independent test set to validate the model's accuracy. According to whether the toxin and/or the target in the test set were in the training set or not, we designed four models: Model I for "general" prediction (all toxins versus all targets); Model II for new-toxins versus known-targets; Model III for known-toxins versus new-targets; Model IV for new-toxins versus new-targets. Toxins and targets in the training set are called 'known' whereas those not in the training set are called 'new'.

In details, the training and test sets of the four models were produced as follows: (1) creating the positive training and test sets. Firstly, an initial positive test set and an initial positive training set were obtained by randomly splitting the whole positive samples. Then, for Model I, the initial positive training and test sets were directly used as final positive training and test sets, respectively. For Models II and III, the final subdata sets were generated by removing the samples of known toxins/new targets (or the new toxins/known targets) in the initial positive test and training sets. And deleting the samples containing the known toxins and targets from the initial positive test set generated the final subsets of Model IV. (2) Creating the negative training and test sets. As information about non-interaction pairs was unavailable, we randomly generated the negative samples from the unknown interaction pairs not overlapping with those interaction pairs. To ensure the balance of positive and negative data, an equal number of negative samples were added to each positive training and test sets by randomly choosing the unknown interactions in the corresponding positive training or test sets. As a result, for Model I, II, III and IV, their training sets contained 42,044, 42,250, 41,942, 39,816 samples respectively, and the test sets contained 10,510, 10,304, 10,612 and 290 samples respectively. To avoid the attributes in greater numeric ranges dominating those in smaller numeric ranges, these descriptor vectors were separately scaled to the range of −1 to 1.

2.4. Support vector machine

The support vector machine developed by Vapnik (1998) is a well-known large margin classifier. Due to its remarkable generalization performance, it has been used in bioinformatics and cheminformatics (Yu et al., 2012). The notable feature of SVM is that it explicitly relies on the structure risk minimization (SRM) principle from statistical learning theory (Cristianini and Shawe-Taylor, 2000), which is superior to the traditional empirical risk minimization (ERM) principle employed in conventional neural networks (Jiang et al., 2006). SVM classification is based on constructing a maximal margin hyperplane in the high multidimensional space that optimally separates two different groups. The maximal margin is defined as the closest distance from any point to the separating hyperplane.

To describe an SVM precisely, suppose our data are given as a set of labeled training vectors (x_i, y_i), $i = 1, \ldots, m$) that are classified to two classes ($y_i \in \{-1, 1\}$) (1 and −1, in our case, representing the interaction and non-interaction toxin–target pairs, respectively) and each of vector is an m-dimensional feature vector ($x_i = (x_i^1, x_i^2, \ldots, x_i^m)$). Using this notation an SVM classifier is produced as follows:

$$f(x) = \text{sign}\left(\sum_{i=1}^{n} \alpha_i y_i k(x_i, x) + b_0\right) \tag{2}$$

where x is the new object to be classified, n is the number of the training samples, $f(x)$ is a decision function and $k(x_i, x)$ is a kernel function that is used to measure the similarity between two samples. A popular radial basis function (RBF) was used in this research. The constants b_0 and α_i are obtained by solving a quadratic programming problem. A new toxin–target pair is then classified as positive (negative) if $f(x)$ is positive (negative). In this study, the SVM classification was conducted by using the LIBSVM suite of program (http://www.csie.ntu.edu.tw/~cjlin/libsvm). The parameters of the SVM with the radial basis function (RBF) kernel were optimized using a grid search (Hsu et al., 2003).

2.5. Random forests

Random forests method is a non-parametric machine-learning algorithm based on model aggregation ideas (Breiman, 2001), which is effective for tracking the classification and regression tasks in many scientific areas (Svetnik et al., 2003). It is a combination of randomized decision trees, which ensemble produces a corresponding number of outputs aggregated to obtain one final prediction. The training algorithm of the RF for classification can be summarized as follows: (i) Select N bootstrap samples $\{B_1, B_2, \ldots, B_N\}$ from the initial samples. (ii) Grow an unpruned tree T_p ($p = 1, \ldots, N$) with each training set B_p. At each node, randomly sample the subset of input variables rather than all of the predictors to determine the best split. The tree is grown to the maximum size and not pruned back. (iii) Predict new data by majority voting of the N trees. During the training process, RF ensures its own reliable statistical characteristics via the use of Out-Of-Bag (OOB) samples. The samples in the original data set that do not occur in a bootstrap sample are called OOB samples. For each OOB sample, the predicted values of the trees that have not been built using this OOB sample are calculated. Then, aggregate the OOB predictions and calculate the error rate, namely, the OOB estimate of error rate.

Additionally, one of the most importance features of RF is the outputs of the variable importance. To estimate the variable importance for a special variable j, the values of the jth variable are randomly permuted for the OOB samples. Then the measure for the jth variable is simply $M - M_j$, where M is the average margin based on the OOB prediction and M_j is the average margin based on the OOB prediction with the jth variable permuted. For classification problem, the margin is replaced by the prediction accuracy. If substantially decreased prediction accuracy is produced, it indicates that the variable j has strong association with the response.

In this work, the Random Forest soft package developed by Leo Breiman et al. was used to build the RF prediction models (available at http://www.stat.berkeley.edu/users/breiman/). Default settings were used for the parameters: 500 for the number of trees and the square root of the total number of variables for the number of randomly selected variables, respectively.

2.6. Performance evaluation

With the purpose of deriving reliable in silico models, both internal and external validations methods were applied. Furthermore, all predictive models were evaluated and verified with 5-fold cross-validation. By using the internal validation, the training set was firstly split into five approximately equal-sized subsets randomly, where four subsets were selected as the training set to develop a model and the remaining samples as test set. This process was repeated five times to ensure every subset can be predicted as a validation set once. Meanwhile, external validations were performed by using different test sets for all models. Finally, the performance of the models built by RF and SVM methods were compared.

The prediction performance in the classification system was evaluated by several parameters. The accuracy (ACC), sensitivity (SEN), specificity (SPE) and precision (PRE) were used to measure the accuracy of overall, positive prediction, negative prediction and the positive predictive value of the model, respectively. The ACC, SEN, SPE and PRE were calculated according to the following equations:

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{3}$$

$$\text{SEN} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4}$$

$$\text{SPE} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{5}$$

$$\text{PRE} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{6}$$

here, the TP, TN, FP and FN represent the number of true-positives, true-negatives, false-positives and false-negatives, respectively. Meanwhile, the performance was evaluated by using a receiver operating curve (ROC), that is, the plot of false-positive rate (1 − SPE) versus the true-positive rate (SEN) based on the various thresholds. In addition to a simple output of a yes/no decision, RF and SVM predicted score was used to estimate the SVM and RF confidence of the predicted outputs. The scoring method of RF was defined as the percentage of trees voting for "yes" (interaction). The SVM score is based on the idea that samples lying closer to the hyperplane have a larger probability of being misclassified than examples lying far away (Rüping, 2004). The flowchart of the modeling procedure is shown in Fig. 1.

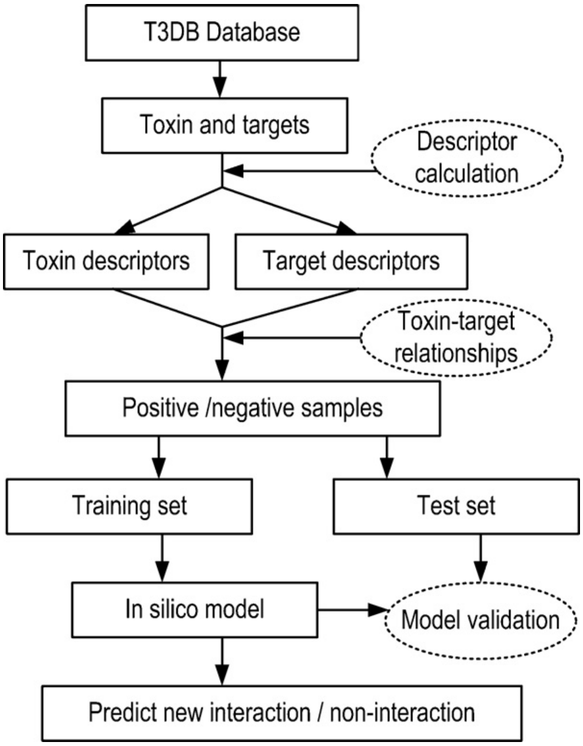


Fig. 1. A framework for toxin–target interaction prediction.

2.7. Network construction

Proteins rarely function in isolation and outside the cell; instead, they operate as part of highly interconnected cellular networks referred to as interactome networks. With the recent explosion of publicly available high throughput biological data, the analysis of networks has gained significant attentions in biological and even toxicological fields, due to the fact that such analysis can provide a unifying language to describe relations within complex systems and to understand the physiological functions. In this work, we have combined a set of systematic tools to (i) analyze the properties of toxin–target networks, (ii) assess retrospectively and prospectively the network-based relationships between the toxins and their targets, quantifying ongoing trends and shifts in the discovery of toxic mechanisms, and (iii) quantify the interrelationships between targets and disease-related gene products.

In our work, the toxins and target proteins were used to build the toxin–target network by Cytoscape 2.8.1, a standard tool for biological network visualization and data integration (Smoot et al., 2011). In the visualized network, the toxin–target network was produced by linking all toxins in the dataset and the disease-related proteins, which were represented as nodes and intermolecular interactions. The heterogeneous nodes corresponded to either toxins or target proteins, and edges for the interactions between them. The edge is placed between a toxin node and a target node if the protein is a known target of the toxin. Finally, the quantitative properties of these networks were analyzed by the NetworkAnalyzer (Assenov et al., 2008) and CentiScaPe 1.2 (Scardoni et al., 2009).

3. Results and discussion

The inherent complexity of interactions between toxins and targets has proposed a huge challenge in the area of predictive toxicology. Actually, the combination of many potential interactions and various endpoints contributing to an overall effect has always determined that the prediction of such toxicity is a Herculean task. In the present work, firstly, we built and evaluated four in silico models developed by SVM and RF approaches. Then based on the combinational assessments of the predicted results, a comprehensive toxin–target network was constructed and analyzed

Table 1
Statistics of the prediction performances.

		Model I	Model II	Model III	Model IV	Average
SEN (SVM/RF)	Training	94.62%	94.84%	94.87%	94.81%	94.79%
	Test	94.45%	94.50%	94.87%	95.04%	94.72%
SPE (SVM/RF)	Training	95.62%	78.38%	60.38%	26.90%	65.32%
	Test	95.01%	82.90%	41.54%	11.72%	57.79%
PRE (SVM/RF)	Training	92.65%	92.98%	92.71%	92.61%	92.74%
	Test	90.15%	90.34%	90.80%	90.04%	90.33%
ACC (SVM/RF)	Training	87.97%	92.29%	92.80%	91.72%	91.20%
	Test	82.23%	86.53%	95.78%	98.62%	90.79%
AUC (SVM/RF)	Training	92.79%	93.11%	92.87%	92.77%	92.89%
	Test	90.56%	90.73%	91.16%	90.52%	90.74%

to predict more potential interactions between toxins and targets.

3.1. Model evaluation and comparison

The statistical parameters SEN, SPE, PRE, ACC and the AUC (area under the ROC curve) were used to estimate the performance of the derived models, as shown in Table 1. In order to obtain accurate comparisons, the SVM and RF models were built with the same training and test sets.

As seen from Table 1, all the models evaluated by the internal five-fold cross-validation show significantly consistent prediction performances: an average SEN of 94.79% and 94.72% for the binding patterns, an average SPE of 92.74% and 90.33% for the non-binding interactions, an average PRE of 92.89% and 90.74%, an average ACC of 93.76% and 92.53%, as well as an average AUC of 97.79 and 97.98, respectively. Further comparison reveals that RF models are relatively worse in SEN, SPE, PRE and ACC than SVM models except RF Model IV, which is slightly better in SEN (95.04%) than SVM Model IV (94.81%). For purpose of evaluating the performance, the relevant ROC curves for SVM and RF models were calculated and drawn in Fig. 2. The results demonstrate that the SVM method possesses quite good power on detecting toxin–target interactions with high true-positive rates versus low false-positive rates based on the prediction score for various threshold values. For instance, in the SVM Model I, when the true positive rate possesses 40% or 80%, the corresponding false positive rate is as low as ∼2% or ∼3%. In conclusion, the obtained models are satisfactory for both the training and test sets, with no evident overfitting or over training phenomenon, exhibiting strong robustness and capability to predict the multiple toxin–target interactions.

Even with excellent fitness and predictions in the training process, the models may still lack a generalization ability for novel data. Therefore, a reliable validation procedure, i.e. an external testing of the models should be carried out to evaluate the real predictive power of the models and confirm the inexistence of chance correlations. Here, these models were validated by four independent external validations in order to guarantee all models using different test sets.

As a result, the predictability performance of the general dataset (test set I, Model I) by SVM method is the best, as reflected by the statistical values (the SEN of 95.62%, the SPE of 87.97%, the PRE of 88.83% and the ACC of 91.80%). This result is similar to that of the internal five-fold cross-validation, unveiling that the obtained models are unlikely to be over-fitted. For test set II (new toxins–known targets dataset, Model II) and test set III (new targets–known toxins dataset, Model III), the prediction accuracies of the SVM models are 78.38% and 60.38% for SEN, 92.29% and

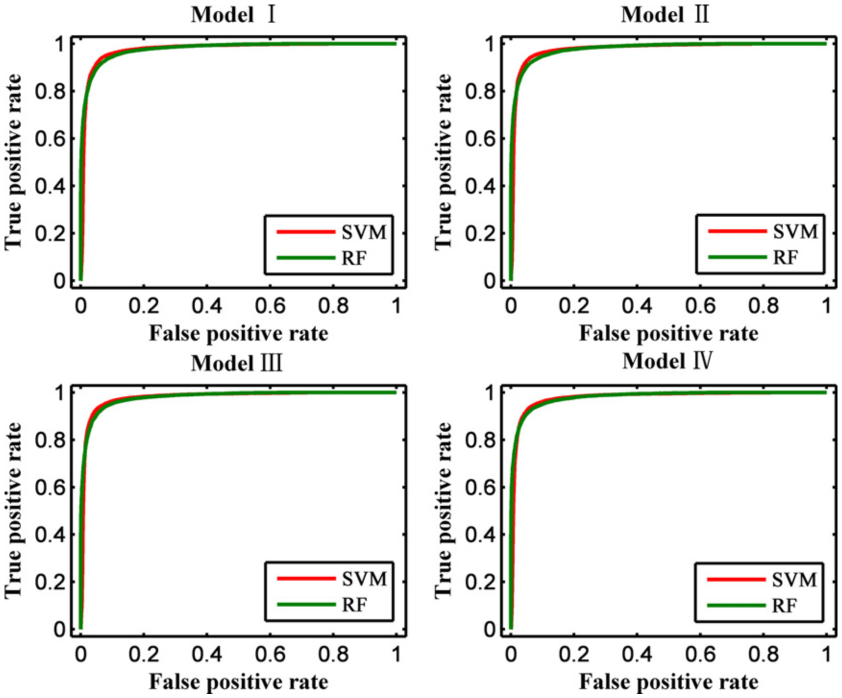


Fig. 2. ROC curves obtained by five-fold cross-validation for the SVM (red) and RF methods (green). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

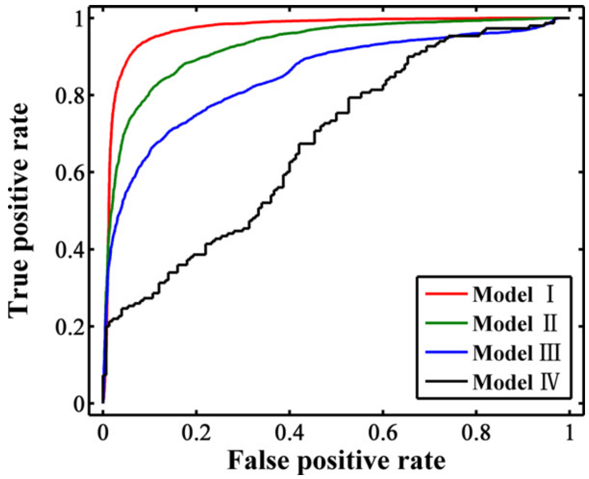


Fig. 3. ROC curves obtained by external validation for the four models.

92.80% for SPE, 91.05% and 89.35% for PRE, 85.34% and 76.59% for ACC, as well as 93.79 and 84.91 for AUC, respectively.

These statistics demonstrate that the SVM models are more suitable to set II than set III, similarly as the RF models (the AUC for set II and set III is 93.09 and 80.60, respectively). This may be due to the smaller information space of the targets compared with the toxins in the training set. Compared with the first three test sets, the prediction capability of the test set IV (new toxins–new targets dataset, Model IV) is relatively weak, as demonstrated by the SEN of 26.90%, the SPE of 91.72%, the PRE of 76.47% and the ACC of 59.31%. The lack of sample size of dataset IV to represent the general cases may lead to its poor prediction performance when compared with the first three models (Yamanishi et al., 2008).

The aforementioned results suggest that all models could span almost the entire performance space and their high predictive power may depend heavily on the composition of the test set. The ROC curves of each model based on the external validation were plotted in Fig. 3. Parameters in each model are chosen by the AUC score as an objective function. According to these scores (from 97.05 to 67.75), SVM models exhibit the most potent prediction ability for Set I, followed by Set II, Set III and Set IV.

For RF models, the external set I is relatively worse than that of SVM models in SEN, SPE and ACC. In contrast, the set II of RF model presents better statistical results in SEN, while the sets of III and IV of RF models all slightly outperform the SVM models both in SPE and PRE. From these results we can see that the predictive abilities of SVM and RF models are quite similar to each other despite of the slight difference, both demonstrating proper reliability and robustness.

In addition, all models accurately indentified those negative samples (non-interaction) with a very high specificity (82.23–98.62%) of all the datasets, though the negative samples were initially randomly produced. However, compared with this high specificity, the sensitivity is low, especially when the toxins or/and targets information is insufficient in the models. One explanation for the low sensitivity is that the actual non-interaction space is very huge compared with the interaction space, making it much more easily to capture the non-interaction pairs than the interaction pairs. This, from a statistical point of view, reveals that a toxin binding to the target is quite specific, thus to find a new interaction toxin–target pair by chance should be extremely difficult. Although the sensitivities are low in both Model III and Model IV, reliable predictions are still possible due to the high precision of 90.77% and 89.47% of RF method. In another word, our method provides an effective way to eliminate as many false positive predictions as possible and to obtain a high enrichment of true positive in the predicted interaction sets. All these outcomes demonstrate that our models exhibit proper performance and universality for multiple toxin–target interactions prediction.

The results presented above illustrate that the predictive power of SVMs and RFs is slightly different based on the same training and test sets. Although both SVMs and RFs are effective resources for building accurate classifiers, SVMs show superiority to RFs in predicting the toxin–target interactions. Firstly, SVMs show a better generalization ability to build models. The reason may be that SVMs method embodies the structural risk minimization principle, which minimizes an upper bound of the generalization error rather than minimizes the training error. Secondly, SVMs allow us to use a tensor product space, with no extra calculation time with respect to the joint space, and provide a versatile choice of similarity measures for targets and toxins (Yu et al., 2012). In addition, the SVMs algorithm appears to be marginally more accurate, and especially it can be applied when some experimental data are missing. In Models II, III and IV with the toxins or/and targets data not included in the models, SVMs performed obviously better with accuracy of 85.34%, 76.59%, 59.31% than RFs with corresponding accuracy of 84.71%, 68.66% and 55.17%, respectively. Typically, SVM algorithm uses a portion of training set as support vectors for classifications. If the missing experimental data are non-support vectors, they won't affect the model performance (Cheng et al., 2011).

Although SVMs apparently outperform RFs, RFs are still an effective method with some advantages over SVMs. The algorithm is robust against overfitting since each tree in the ensemble grows on an independently bootstrapped subsample of the data. As a large number of low-correlated decision trees are averaged, RFs can achieve both low bias and low variance. Actually, RFs provide a reliable error estimate by using the so-called OOB data. The preselection of variables is not required because the RF algorithm is quite robust to noise in predictors. As only a limited random number of predictors are used to seek for the best split at each node, the diversity of the forest is produced and the cost of the computational load is reduced. Pruning the trees is not necessary which results in low bias and high variance trees and also reduced computation time (Grimm et al., 2008). As we investigate the computational cost for SVMs and RFs methods, RF classifier also cost less time when achieving similar good performance. In this work, all the programs were implemented on a Dell computer (Redhat Linux Operating System) with 2.8 GHz AMD Phenom (tm) II X6 1055T processer and 12 GB RAM. The total execution time of the cross-validation experiment of SVM (24 h) is much longer than that of the RF (9 h) approach. Finally, a *T* test was performed to estimate whether our methods to evaluate the difference of our models' prediction ability are good or not. With this test, the significance can be analyzed and used to distinguish between any two models. We take the overall accuracy ACC as the test parameter duo to that it is the general statistical parameter to evaluate the predictive power of a model. All the obtained results show that there exists extremely big difference between the prediction results (ACC) of the four models with P-values <0.01. This demonstrates that the predictive capacities of the four models we built are extremely different, and subject to such an order of Model I > Model II > Model III > Model IV.

3.2. Applicability domain and feature analysis

The applicability domain is of crucial importance to provide good accuracy estimation of the classification models, which is obtained by visualizing the samples in a multidimensional space. It provides additional information to identify which samples are classified with the best accuracy or unreliable predictions. The selection of the most reliable prediction can dramatically improve the performance of the methods while decreasing the coverage of the predictions (Tetko et al., 2006). For this purpose it is important

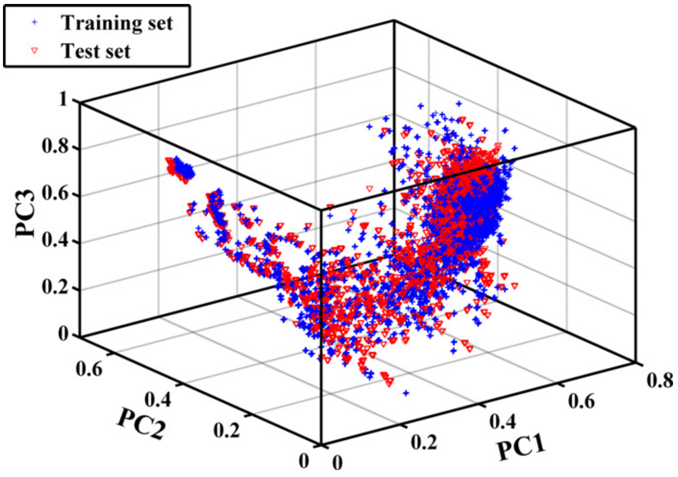


Fig. 4. Distribution of the Model I over the first three PCs.

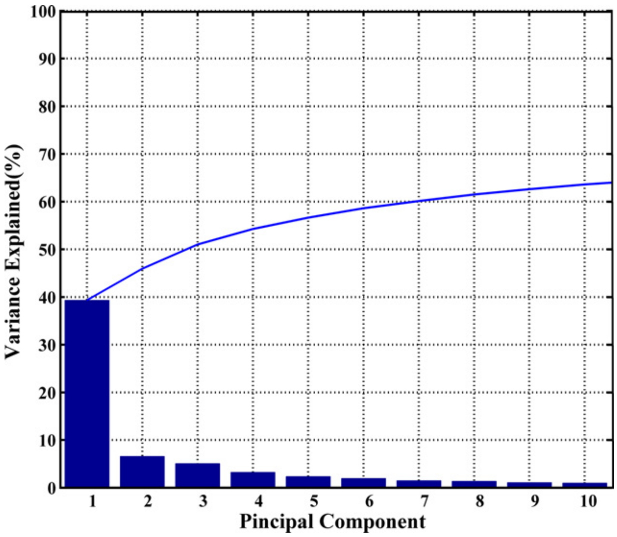


Fig. 5. The pareto chart of the variance explained by the first 10 principal components for the experimental dataset.

to know the proportion of samples that fall within the AD of a certain model. Therefore, in order to reduce the redundant information and ensure that useful information can be processed in a low-dimensional space, the principle component analysis (PCA) (Wold et al., 1987) was applied to analyze the ADs of these obtained models. This process is achieved by transforming the original matrix to a smaller data set with uncorrelated variables, i.e., principal components (PCs). Fig. 4 displays the training and test samples visual distribution of Model I over the first three PCs. Large overlap can be found between the two kinds of samples in “chemical–biological” space, indicating good structure diversity and versatility of chemical and biological properties among them. This provides good foundation for screening toxins and target proteins with interactions. Moreover, the original 1947 components have been compressed and analyzed by PCA resulting in 10 PCs (explaining 63.58% of the total variance), in which the obtained first three PCs account for about 51.0% of the total variance. The descriptions for these PCs are shown in Fig. 5. All these results demonstrate that the applicability domain of these models is wide enough to overlap most of the whole “chemical–biological” space.

The accuracy and quality of these models are greatly affected by a very large number and diverse types of molecular descriptors (1547 dimensions) and protein descriptors (400 dimensions) in the context of toxin–target pairs. On the basis of the variable importance outputs of RF Model I, the top 30 chemical and protein descriptors are picked out and shown in Fig. 6. The top 30 chemical descriptors as shown in Fig. 6A are mainly from five blocks including the Burden eigenvalues, 3D-MoRSE descriptors, eigenvalue-based indices, constitutional descriptors and GETAWAY descriptors, which are usually applied in toxicological analysis (Zhu et al., 2008). The chemical descriptors employed in these statistical models may give some insights into the toxicological behavior to bind to the specific protein target. For examples, the Burden eigenvalues BEHe7, BEHm1, BEHp1, BEHv6, BELe3, BELe6,

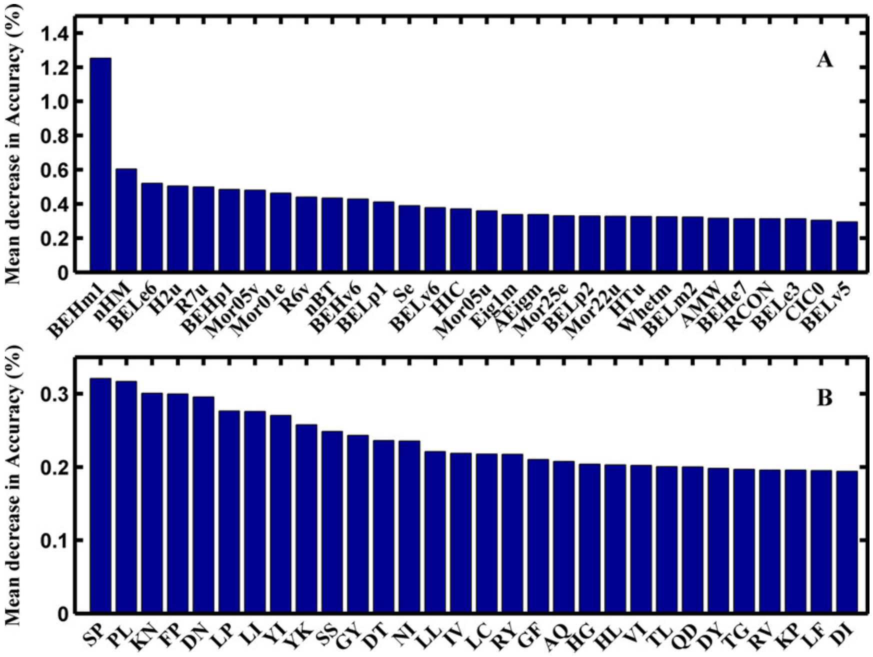


Fig. 6. The relative importance of descriptors: (A) the top 30 chemical descriptors and (B) the top 30 protein descriptors.

Table 2
The statistical parameters of each kind of disease.

Disease	SEN	SPE	PRE	ACC	AUC
Cardiovascular diseases	95.02%	84.52%	85.77%	89.72%	96.33%
Congenital, hereditary, neonatal diseases abnormalities	80.77%	89.47%	63.64%	87.86%	92.68%
Digestive system diseases	90.68%	85.60%	75.36%	87.26%	93.52%
Endocrine system diseases	91.76%	85.46%	79.60%	87.87%	95.40%
Female urogenital diseases and pregnancy complications	90.48%	84.00%	78.08%	86.50%	93.27%
Hemic and lymphatic diseases	96.77%	85.19%	83.33%	90.21%	95.64%
Immune system diseases	97.91%	84.38%	89.13%	92.05%	96.18%
Male urogenital diseases	93.47%	86.30%	87.07%	89.86%	96.37%
Mental disorders	93.50%	89.34%	88.21%	91.25%	96.55%
Musculoskeletal diseases	96.44%	84.52%	91.04%	91.91%	96.00%
Neoplasms	96.27%	84.93%	87.20%	90.79%	96.36%
Nervous system diseases	94.53%	86.33%	87.12%	90.38%	96.11%
Nutritional and metabolic diseases	93.52%	86.83%	86.19%	89.96%	97.30%
Pathological conditions, signs and symptoms	94.77%	87.35%	85.29%	90.59%	96.27%
Respiratory tract diseases	92.54%	88.13%	78.98%	89.56%	95.61%
Skin and connective tissue diseases	98.12%	82.56%	92.48%	93.24%	96.34%
Substance-related disorders	92.36%	87.20%	84.30%	89.40%	95.73%

BELm2, BELp1, BELp2, BELv5 and BELv6 are descriptors characterizing the molecular size, polarizability and electronegativity. The 3D-MoRSE descriptors Mor01e, Mor25e, Mor05u, Mor22u and Mor05v mainly reflect the molecular size and 3D information. Clearly molecular size, shape, charge and polarity are important for the ligand to bind with its targets (Quillin et al., 2000). Besides, many metals, particularly those heavy ones are toxic, thus the nHM (number of heavy atoms) makes significant contributions to the classification models. Fig. 6B shows the relative contribution of the 30 most important protein descriptors to the classification model. The most important descriptor for our model is SP, describing the combination of Ser and Pro. Except for SP, the PL, KN and FP are also crucial for the classification, which can be explained by the combination of Pro and Leu, Lys and Asn, Phe and Pro, respectively. This result shows that dipeptide composition, which provides information about amino acid composition as well as the local order of amino acids, is also useful index for the classifications (Yu et al., 2012), and is a better feature as compared with the amino acid composition alone for constructing the feature of a protein sequence. Meanwhile, the average of mean decrease in accuracy of all proetin (0.20%) is higher than that of the chemicals (0.05%), confirming the previous conjecture that the protein descriptors are more relevant than chemical indices.

To further expand the application of our models in real application cases, we have adopted a classification scheme and separated targets into 17 categories according to the phenotypic diseases (Table S3). Here, the known 319 disease-associated targets were selected to perform the analysis by the general SVM Model I. All the statistical parameters for each phenotype are acceptable as shown in Table 2. We can see that the resulted ACC scores for these diseases range from 86.50% to 93.24%, all exhibiting good performance. Interestingly, it seems that the best result with ACC of 93.24% is the targets that are related to the skin and connective tissue diseases, while the worst is the female urogenital diseases and pregnancy complications-related proteins (ACC = 86.50%). This means that cutaneous reactions upon contacting with a substance can be relatively accurately predicted, but long term effects such as the congenital, hereditary, neonatal diseases abnormalities (ACC = 87.86%) and the female urogenital diseases and pregnancy-related proteins are relatively poor predicted. In addition, the immune system diseases (ACC = 92.05%); hemic and lymphatic diseases (ACC = 90.21%); cardiovascular diseases (ACC = 89.72%) closely related to blood system have also been properly assessed, as blood is responsible for transporting and directly contacting with toxins in body. Last but not the least, it is found that diseases with more sufficient target information have also been much better predicted, such as the neoplasms (ACC = 90.79%); pathological conditions, signs and symptoms (ACC = 90.59%); as well as nervous system diseases (ACC = 90.38%). In contrast, the diseases with less target information such as the substance-related disorders were predicted with relatively lower accuracy (ACC = 89.40%). All this indicates that the improvement in both the quality and quantity of diseases-targets information could enhance the predictability of our models as well. In conclusion, these results illustrate that the obtained models might be further applied to predict toxin-deduced diseases based on the relationships of protein targets with their phenotypic diseases.

Supplementary material related to this article found, in the online version, at http://dx.doi.org/10.1016/j.tox.2012.12.012.

3.3. *Comprehensive prediction for potential toxin–target interactions*

Network analysis has become a cornerstone of fields as diverse as systems biology, which is helpful for revealing the known/unknown interactions of a given system in global view. Recently, the emerging tools of network medicine have offered a platform to explore systematically not only the molecular complexity of a particular disease, leading to the identification of disease modules and pathways, but also the molecular relationships among apparently distinct (patho) phenotypes. Given the complexity of biological system, it is important to generate networks to uncover multiple potential interactions. Therefore, we built a comprehensive network for identifying multiple potential toxin–target interactions, which, in turn, can address some fundamental properties of the proteins toward the understanding of the toxins. Here we have selected a set of heart disease-related proteins to test the reliability of our models, since heart disease has been ranked as one of the major causes of mortality posing a serious threat to human health (Gu et al., 2009). Accordingly, 51 typical drug targets, which are related to heart disease and 2257 toxins from the T3DB database, were used to construct the comprehensive toxin–target interactions network by the optimal SVM Model I.

As a result, by using the top 500 scoring toxin–target interactions, a network has been generated where a compound and a protein are connected to each other if the protein is a known target of the compound (toxin–target network). Fig. 7 shows a global view of toxin–target interactions network with color-coded nodes (toxin: orange, protein target: blue), which contains 201 nodes and 500 edges, with 150 toxins and 51 targets. Most toxins target only a few protein targets, but some have many protein targets. Likewise, the protein targets also display rich landscape of interacting toxins. This indicates that the availability of toxin–target interactions

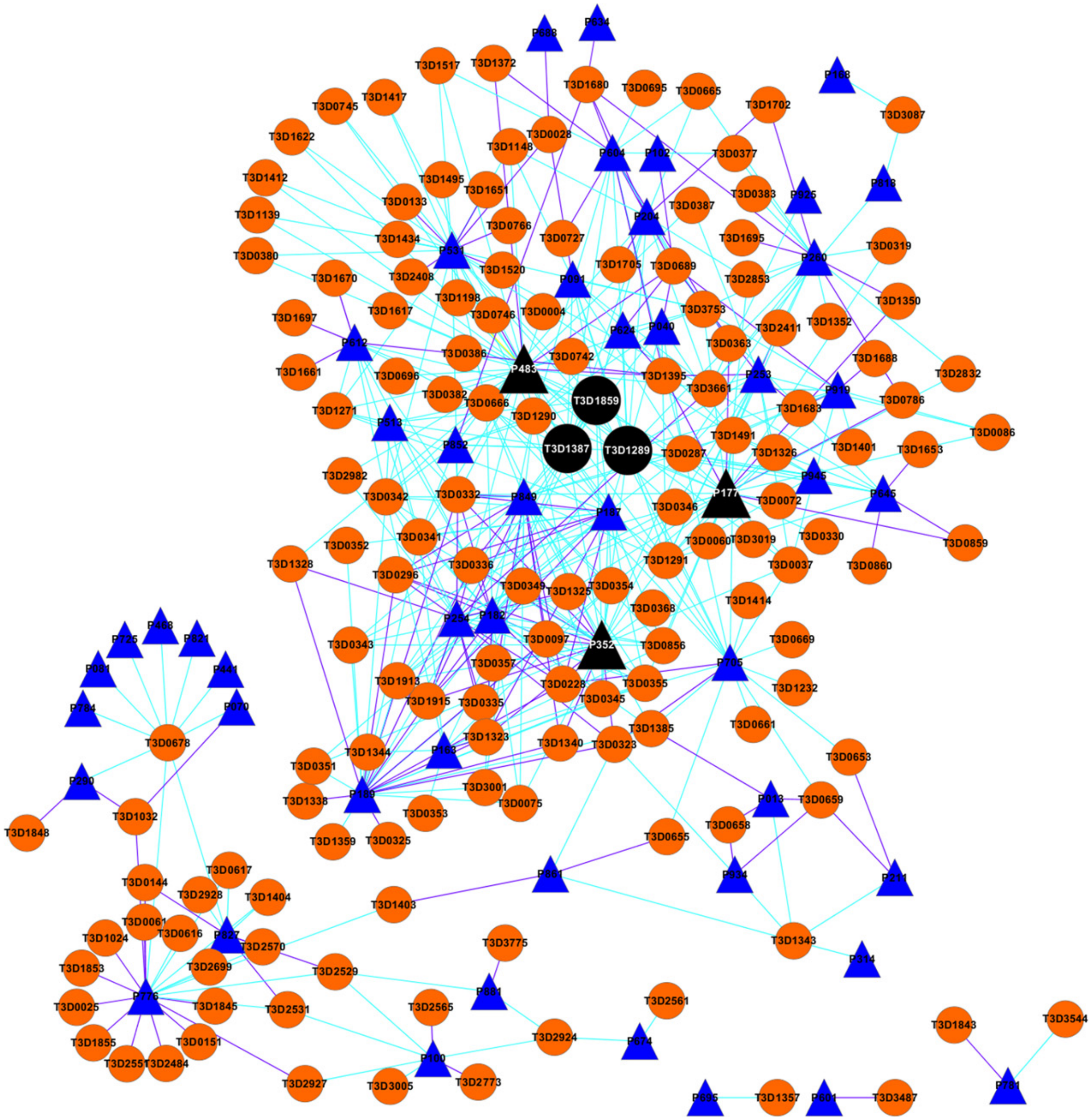


Fig. 7. Toxin–target interactions network (purple line: validated, cyan line: predicted, red ball: toxin, blue triangle: heart disease-related protein, black ball: important toxins, black triangle: important targets). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

network is by itself a useful compendium that reflects current and potential multiple interactions.

Network data structures are amenable to many sophisticated forms of computational analysis, which can uncover important, nonobvious properties of nodes and the relationships between them (Lee et al., 2009). Here the centralization, density, heterogeneity, node degree distribution and betweenness were analyzed to investigate both the global and topological properties of this toxin–target network (Dong and Horvath, 2007). A first general overview of the global topological properties of the network came from the min, max and average values of all computed centralities along with the diameter (13.0) and the average distance of the network (4.48) suggests a relatively decentralized network, in which proteins are not strongly functionally interconnected. And the centralization of 0.172 and density of 0.025 also indicate that the network structure would be rather decentralized than centralized. Moreover, the heterogeneity value of 1.324 shows that a few nodes are more central compared to other nodes in this network, which reveals that the toxin–target space is partial to certain toxins and proteins.

Topological analysis of network may offer insights into biologically relevant connectivity patterns, which may pinpoint highly influential toxins or targets. The most essential characteristic of a node is its degree (the number of connections or edges the node has to other nodes), which tells us how many direct links the node holds. In our network, of all the 51 protein targets, 19 have considerable strong interactions with ≥10 toxins, and 13 protein targets are linked to more than 15 toxins. Protein P483 (Cytochrome c oxidase subunit 5A, mitochondrial) exhibits the highest number of interactions with 39 toxins, which can be susceptibly attacked by toxins, possibly inducing heart diseases (Wikipedia, 2009). Following on are protein P352 (sodium/potassium-transporting ATPase subunit alpha-2) and P177 (Cytochrome c oxidase subunit 1) with 31 and 30 toxins, respectively (black triangle in Fig. 7). Actually,

Table 3
The representative prediction results of toxin–target interactions.

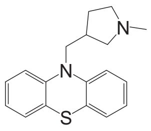
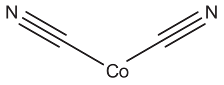
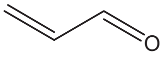
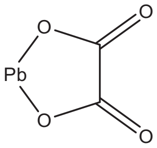
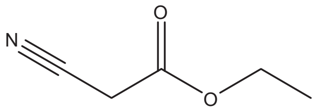
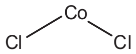
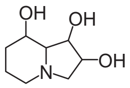
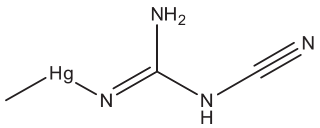
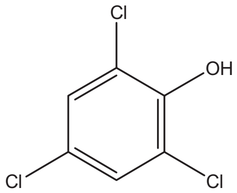
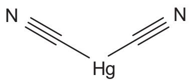
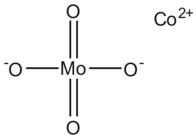
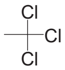
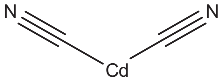
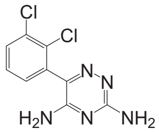
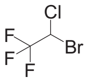
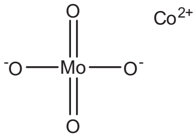
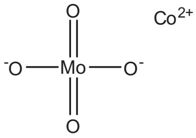
No.	Toxin ID	Toxin name	Chemical structure	Protein ID	Binding score
1	T3D0028	Cyanide[a]		P531	0.9998
2	T3D1491	Aluminum antimonide		P253	0.9998
3	T3D2924	Methdilazine		P100	0.9996
4	T3D2408	Boron phosphide[a]		P531	0.9985
5	T3D1395	Cobalt(II) cyanide[a]		P604	0.9985
6	T3D0332	Lead tetroxide[a]		P187	0.9983
7	T3D0037	Acrolein		P177	0.9983
8	T3D1271	Tin(II) oxide		P531	0.9983
9	T3D1323	Lead oxalate[a]		P182	0.9971
10	T3D1680	Ethyl cyanoacetate[a]		P260	0.9970
11	T3D0659	Cobalt(II) chloride[a]		P013	0.9970
12	T3D3087	Swainsonine		P818	0.9968
13	T3D0341	Mercury(II) sulfide		P254	0.9959
14	T3D1350	Methylmercuric dicyanamide[a]		P177	0.9958
15	T3D0086	2,4,6-Trichlorophenol		P919	0.9958
16	T3D0330	Lead oxide		P177	0.9957
17	T3D0363	Mercury(II) cyanide[a]		P604	0.9957

Table 3 (*Continued*)

No.	Toxin ID	Toxin name	Chemical structure	Protein ID	Binding score
18	T3D1385	Cobalt(II) molybdenum(VI) oxide		P189	0.9952
19	T3D1290	Boron arsenide		P187	0.9952
20	T3D1325	Lead selenide		P483	0.9947
21	T3D0097	1,1,1-Trichloroethane[a]		P849	0.9946
22	T3D0343	Mercury(I) chloride		P187	0.9942
23	T3D1491	Aluminum antimonide		P040	0.9937
24	T3D1372	Cadmium cyanide[a]		P483	0.9932
25	T3D2570	Lamotrigine[a]		P827	0.9931
26	T3D0342	Mercury(II) oxide		P852	0.9927
27	T3D3001	Halothane		P189	0.9924
28	T3D1385	Cobalt(II) molybdenum(VI) oxide[a]		P013	0.9923
29	T3D1385	Cobalt(II) molybdenum(VI) oxide		P705	0.9920
30	T3D1271	Tin(II) oxide		P612	0.9916

[a] Represents validated.

their crucial roles in heart diseases have already been proven (Schwinger et al., 2003; Wikipedia, 2009). Among the 150 toxins, toxin T3D1289 (aluminum arsenide) possesses the largest number of interacting target proteins (19), followed by the toxins T3D1387 (cobalt sulfide) and T3D1859 (antimony monosulfide) with 15 target proteins (black ball in Fig. 7). These are examples of the highly connected toxins and targets that are closely related to the heart disease (Linna et al., 2004; Ratnaike, 2003; Ross and Adrian, 2009) (Table S4). This result shows that toxins interacting with the target proteins will gain a high probabilistic weight, which helps to prioritize target proteins and interactions on the basis of their potential involvement in the heart disease. The node degree distribution situation is shown in Fig. 8. This result shows that most nodes have low degrees with only a small number of interaction partners (hubs), which further demonstrates that the generation of this network has almost no randomness. Therefore, it proves again that several toxins could affect multiple targets simultaneously, while a target might also have an impact on multiple toxins, which further synergistically influences those pathways related to the disease of interest. In addition, our model and its derived information provide strong theoretical evidence and explanation for multiple toxins–multiple targets interactions phenomenon, which can be further applied to more complex biosystems.

Supplementary material related to this article found, in the online version, at http://dx.doi.org/10.1016/j.tox.2012.12.012.

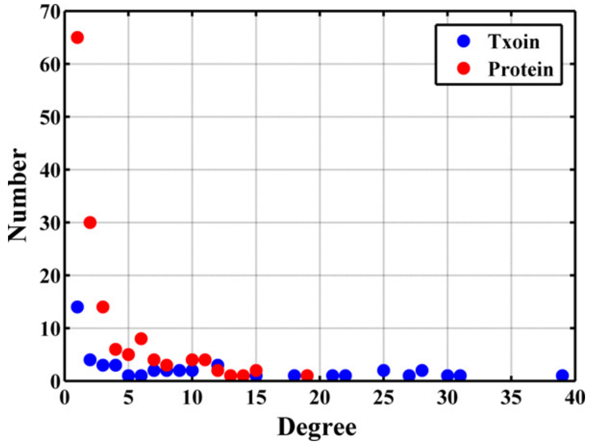


Fig. 8. The node degree distribution of toxin–target network.

"Betweenness" as another elementary property can characterize the importance of the node or edge in the network. It quantitatively reflects the impact of a node exerts on the speading of the information throughout the whole network. Betweenness has the capacity to be located in the shortest communication paths between different pairs of nodes that pass through the node of interest. This property is also defined as traffic, and high traffic nodes are referred to as network bottlenecks. Generally, it is important for the nodes (toxins) to have both higher degree and betweenness (Jeong et al., 2001). This is shown in all the 150 toxin nodes that most of those nodes with higher degree would have larger betweenness, and 45 of the top 55 toxins have both high degree (≥3) and betweenness (≥40); and the number is 22 out of the top 25 toxins. Highly connected nodes are referred to as hubs. This implies that toxin–target interactions network hubs are not generated at random and that they on one hand tend to encode bottlenecks, and on the other hand impact different network regions through both direct and indirect interactions. In this network, many heart disease-related proteins are found to have close relationships to toxins. In addition, it also indicates that if a few heart disease-related proteins are identified, other disease-related components are likely to be found in the network-based vicinity.

In addition, the results show that 500 interactions between toxins and targets are predicted, in which 141 (28%) interactions are validated in T3DB database (purple lines in Fig. 7). Our model also predicted 359 new toxin–target interactions as shown in Fig. 7 (cyan lines). An interesting finding is that about half of the 141 validated interactions (73) and about three quarters (283) of the 359 newly predicted interactions are caused by heavy metals. The representative prediction results are presented in Table 3. It is reported that metal-dependent cell toxicity seems to be closely related to non-specific binding of heavy metal cations to sulfydryl residues in target proteins (Panfoli et al., 2000). This is quite consistent with our findings in the network, which further validates the reasonability of the constructed model.

In conclusion, our toxin–target interactions network reveals a challenge exists in the potential toxin–target interaction predictions as many multiple relationships of toxins and targets still remain unknown or poorly mapped up to date. Furthermore, this framework not only builds a bridge linking the experimental results and theoretical knowledge, but also demonstrates a fact that some drug targets are also susceptible to the attacking of certain toxins. This research will lead to a series of hypotheses that tie the toxicity of compounds to the human health, whose validity and applications will receive more attentions in the near future.

4. Conclusion

In this article, we have developed four in silico models to conduct a system framework of multiple toxin–target interactions from chemical, genomic and toxicological data on a large scale. Our method is based on SVM and RF, which were both evaluated in terms of the sensitivity, specificity, precision and accuracy. All the obtained models were evaluated and verified by both internal and external validations. The results show that all the SVM and RF models exhibit reliable statistical and prediction performance, in which the SVM models are slightly better than the RF ones. Then the applicability domain and feature analysis were carried out to define the area of reliable predictions and to give excellent correlations of our models. In the final part of this study, a comprehensive toxin–target interactions network analysis by using heart disease proteins as an example offered a new framework for prediction of potential multiple toxin–target interactions from a systematic level. Actually, part of the interactions detected by our method has been fully supported by experimental results (Imming et al., 2006; Singh et al., 2006).

The characteristics of our proposed method are the following: Firstly, this approach opens up new opportunities to comprehensively understand the multiple interactions among toxin and target proteins beyond a one-toxin/one-target simply. Secondly, our system could be used as a fast filter in the screening of a huge number of toxins and target proteins on a large scale. Thirdly, it is possible to perform screening of any toxin compound against many target proteins based on our method. Fourthly, we propose a systematic method to predict the toxin–target interactions, even for targets with unknown 3D structure. Fifthly, the toxin–target interactions network can find new toxins and new target proteins simultaneously and infer missing links from the information of known links. Overall, our models are computationally efficient and applicable, which originality lies in the integration of chemical space, genomic space and large-scale toxicological data in a unified framework, as well as in the extraction of correlated sets of chemical substructures and potential multiple toxin–target interactions. To our knowledge, no previous work has possessed all these features. Thus, it is anticipated that our prediction system may become a useful tool to determine new or potential toxins or corresponding targets. From a technical viewpoint, toxins that target RNA or DNA are also critical issues to consider in toxicity prediction development. However, this is a challenging task for the common case of very large molecular targets involving DNA, or RNA due to the insufficiency of toxin targets information so far. Therefore, the prediction of toxins that target RNA or DNA will be an extension of our work and should be carried out lately, through which we expect to bring about more interesting findings.

Conflict of interest statement

Authors declare that there are no conflicts of interest.

Acknowledgement

The work is financially supported by both the Fund of Northwest A & F University, and the National Natural Science Foundations of China (Grant No. 31170796 and No. 11201049).

References

Assenov, Y., Ramírez, F., Schelhorn, S.E., Lengauer, T., Albrecht, M., 2008. Computing topological parameters of biological networks. Bioinformatics 24, 282–284.
Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.
Butina, D., Segall, M.D., Frankcombe, K., 2002. Predicting ADME properties in silico: methods and models. Drug Discov. Today 7, S83–S88.

Cheng, F., Shen, J., Yu, Y., Li, W., Liu, G., Lee, P.W., Tang, Y., 2011. In silico prediction of *Tetrahymena pyriformis* toxicity for diverse industrial chemicals with substructure pattern recognition and machine learning methods. Chemosphere 82, 1636–1643.

Cristianini, N., Shawe-Taylor, J., 2000. An introduction to support Vector Machines: and other kernel-based learning methods. Cambridge University Press, New York, NY, USA, ISBN 0-521-78019-5.

Dong, J., Horvath, S., 2007. Understanding network concepts in modules. BMC Syst. Biol. 1, 24.

Grimm, R., Behrens, T., Märker, M., Elsenbeer, H., 2008. Soil organic carbon concentrations and stocks on Barro Colorado Island-digital soil mapping using Random Forests analysis. Geoderma 146, 102–113.

Gu, J.Y., Yuan, G., Zhu, Y.H., Xu, X.J., 2009. Computational pharmacological studies on cardiovascular disease by Qishen Yiqi Diwan. Sci. China Ser. B 52, 1871–1878.

Hsu, C.W., Chang, C.C., Lin, C.J., 2003. A practical guide to support vector classification. Technical Report. Department of Computer Science and Information Engineering, University of National Taiwan, Taipei, pp. 1–12.

Huang, R., Southall, N., Xia, M., Cho, M.H., Jadhav, A., Nguyen, D.T., Inglese, J., Tice, R.R., Austin, C.P., 2009. Weighted feature significance: a simple, interpretable model of compound toxicity based on the statistical enrichment of structural features. Toxicol. Sci. 112, 385–393.

Imming, P., Sinning, C., Meyer, A., 2006. Drugs, their targets and the nature and number of drug targets. Nat. Rev. Drug Discov. 5, 821–834.

Jeong, H., Mason, S.P., Barabasi, A.L., Oltvai, Z.N., 2001. Lethality and centrality in protein networks. Nature 411, 41–42.

Jiang, Z., Yamauchi, K., Yoshioka, K., Aoki, K., Kuroyanagi, S., Iwata, A., Yang, J., Wang, K., 2006. Support vector machine-based feature selection for classification of liver fibrosis grade in chronic hepatitis C. J. Med. Syst. 30, 389–394.

Klabunde, T., 2007. Chemogenomic approaches to drug discovery: similar receptors bind similar ligands. Br. J. Pharmacol. 152, 5–7.

Lee, S., Park, K., Kim, D., 2009. Building a drug–target network and its applications. Expert Opin. Drug Dis. 4, 1177–1189.

Linna, A., Oksa, P., Groundstroem, K., Halkosaari, M., Palmroos, P., Huikko, S., Uitti, J., 2004. Exposure to cobalt in the production of cobalt and cobalt compounds and its effect on the heart. Occup. Environ. Med. 61, 877–885.

Panfoli, I., Burlando, B., Viarengo, A., 2000. Effects of heavy metals on phospholipase C in gill and digestive gland of the marine mussel *Mytilus galloprovincialis* Lam. Comp. Biochem. Physiol. B 127, 391–397.

Pritchard, J.F., Jurima-Romet, M., Reimer, M.L.J., Mortimer, E., Rolfe, B., Cayen, M.N., 2003. Making better drugs: decision gates in non-clinical drug development. Nat. Rev. Drug Discov. 2, 542–553.

Quillin, M.L., Breyer, W.A., Griswold, I.J., Matthews, B.W., 2000. Size versus polarizability in protein–ligand interactions: binding of noble gases within engineered cavities in phage T4 lysozyme1. J. Mol. Biol. 302, 955–977.

Rüping, S., 2004. A simple method for estimating conditional probabilities for SVMs. Technical Report/Universität Dortmund, SFB 475 Komplexitätsreduktion in Multivariaten Datenstrukturen, No. 2004, 56, http://hdl.handle.net/10419/22569

Ratnaike, R.N., 2003. Acute and chronic arsenic toxicity. Postgrad. Med. J. 79, 391–396.

Ross, C., Adrian, H., 2009. The exposure to and health effects of antimony. Indian J. Occup. Environ. Med. 13, 3–10.

Scardoni, G., Petterlini, M., Laudanna, C., 2009. Analyzing biological network parameters with CentiScaPe. Bioinformatics 25, 2857–2859.

Schwinger, R.H.G., Bundgaard, H., Müller-Ehmsen, J., Kjeldsen, K., 2003. The Na, K-ATPase in the failing human heart. Cardiovasc. Res. 57, 913–920.

Singh, S., Bhalla, A., Verma, S.K., Kaur, A., Gill, K., 2006. Cytochrome-c oxidase inhibition in 26 aluminum phosphide poisoned patients. Clin. Toxicol. 44, 155–158.

Smoot, M.E., Ono, K., Ruscheinski, J., Wang, P.L., Ideker, T., 2011. Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics 27, 431–432.

Svetnik, V., Liaw, A., Tong, C., Culberson, J.C., Sheridan, R.P., Feuston, B.P., 2003. Random forest: a classification and regression tool for compound classification and QSAR modeling. J. Chem. Inf. Comput. Sci. 43, 1947–1958.

Tetko, I.V., Bruneau, P., Mewes, H.W., Rohrer, D.C., Poda, G.I., 2006. Can we estimate the accuracy of ADME-Tox predictions? Drug Discov. Today 11, 700–707.

Vapnik, V., 1998. Statistical Learning Theory. Wiley, New York.

Wang, Y., Li, Y., Ding, J., Jiang, Z., Chang, Y., 2008. Estimation of bioconcentration factors using molecular electro-topological state and flexibility. SAR QSAR Environ. Res. 19, 375–395.

Wang, Z., Li, Y., Ai, C., Wang, Y., 2010. In silico prediction of estrogen receptor subtype binding affinity and selectivity using statistical methods and molecular docking with 2-arylnaphthalenes and 2-arylquinolines. Int. J. Mol. Sci. 11, 3434–3458.

Wikipedia, 2009. Cyanide poisoning. Last Updated 30 March 2009.

Wold, S., Esbensen, K., Geladi, P., 1987. Principal component analysis. Chemometr. Intell. Lab. 2, 37–52.

Yabuuchi, H., Niijima, S., Takematsu, H., Ida, T., Hirokawa, T., Hara, T., Ogawa, T., Minowa, Y., Tsujimoto, G., Okuno, Y., 2011. Analysis of multiple compound–protein interactions reveals novel bioactive molecules. Mol. Syst. Biol. 7, 472.

Yamanishi, Y., Araki, M., Gutteridge, A., Honda, W., Kanehisa, M., 2008. Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. Bioinformatics 24, i232–i240.

Yamanishi, Y., Kotera, M., Kanehisa, M., Goto, S., 2010. Drug–target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. Bioinformatics 26, i246–i254.

Yu, H., Chen, J., Xu, X., Li, Y., Zhao, H., Fang, Y., Li, X., Zhou, W., Wang, W., Wang, Y., 2012. A systematic prediction of multiple drug–target interactions from chemical, genomic and pharmacological data. PLoS One 7 (5), e37608.

Zhang, H., Li, Y., Wang, X., Wang, Y., 2012. Probing the structural requirements of A-type aurora kinase inhibitors using 3D-QSAR and molecular docking analysis. J. Mol. Model. 18 (3), 1107–1122.

Zhu, H., Tropsha, A., Fourches, D., Varnek, A., Papa, E., Gramatica, P., Oberg, T., Dao, P., Cherkasov, A., Tetko, I.V., 2008. Combinatorial QSAR modeling of chemical toxicants tested against *Tetrahymena pyriformis*. J. Chem. Inf. Model. 48, 766–784.