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.
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