Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2015, Article ID 246374, 5 pages
http://dx.doi.org/10.1155/2015/246374

Research Article

Identification of Chemical Toxicity Using Ontology Information of Chemicals

Zhanpeng Jiang, Rui Xu, and Changchun Dong

School of Software, Harbin University of Science and Technology, Harbin 150080, China

Correspondence should be addressed to Zhanpeng Jiang; jzpp@vip.qq.com

Received 26 January 2015; Revised 20 March 2015; Accepted 22 March 2015

Academic Editor: Tao Huang

With the advance of the combinatorial chemistry, a large number of synthetic compounds have surged. However, we have limited knowledge about them. On the other hand, the speed of designing new drugs is very slow. One of the key causes is the unacceptable toxicities of chemicals. If one can correctly identify the toxicity of chemicals, the unsuitable chemicals can be discarded in early stage, thereby accelerating the study of new drugs and reducing the R&D costs. In this study, a new prediction method was built for identification of chemical toxicities, which was based on ontology information of chemicals. By comparing to a previous method, our method is quite effective. We hope that the proposed method may give new insights to study chemical toxicity and other attributes of chemicals.

1. Introduction

In drug discovery, detecting the toxicity of candidate drugs is a very important procedure. Some approved drugs such as phenacetin [1] and troglitazone [2], which have passed Phase III clinical trials, have to be withdrawn from the market, because their unexpected toxicities were detected. Pharmaceutical companies thus lost millions of dollars. In view of this, it is necessary to detect the toxicity of chemicals before they are selected as candidate drugs. However, evaluating the toxicity of a certain chemical requires comprehensive experimental testing, which costs millions of dollars and takes many years. On the other hand, with the advance of the combinatorial chemistry, a large number of synthetic compounds have surged, inducing that detecting chemical toxicities through traditional methods is an impossible task. Thus, quick, effective, and non-animal-involved prediction methods are urgently necessary.

In recent years, some prediction methods have been built for detecting chemical toxicities. Most of them can only deal with a single toxicity at the same time [3, 4], that is, predict a certain chemical to be toxic or nontoxic for a single toxicity. To detect all toxicities of a chemical, these methods have to be executed many times. Recently, Chen et al. built a multiclass prediction method using chemical-chemical interaction information [5], which can provide a candidate toxicity sequence ranging from the most likely toxicity to the least likely one. Their method was applied to detect the toxicities of chemicals listed in Accelrys Toxicity Database [6], in which six types of toxicity are reported: (1) acute toxicity; (2) mutagenicity; (3) tumorigenicity; (4) skin and eye irritation; (5) reproductive effects; (6) multiple dose effects. In this study, we employed the data in Chen et al.'s study [5] and adopted a new kind of information of chemicals to identify chemical toxicities. ChEBI ontology, integrated in a well-known database ChEBI (Chemical Entities of Biological Interest) [7], reports the ontology information of chemicals and is composed of the following subontologies: (1) molecular structure; (2) biological role; (3) application; (4) subatomic particle. Since gene ontology [8], the ontology information for proteins has been deemed to be a useful tool to investigate protein-related problems [9–12]. It is believed that ChEBI ontology is also a useful tool for studying chemicals and building effective prediction methods to identify chemical attributes. Here, we established a prediction method based on this information and compared to the method reported in [5]. The results indicate that this information is suitable to identify chemical toxicity. And we hope that the proposed method may stimulate extensive investigation based on this information, thereby promoting the study of chemicals and drug discovery.

2. Materials and Methods

2.1. Dataset. The toxicity information of chemicals was retrieved from a previous study [5], which was collected from the Accelrys Toxicity Database [6]. Six types of toxicity are reported in this database; there are (1) acute toxicity; (2) mutagenicity; (3) tumorigenicity; (4) skin and eye irritation; (5) reproductive effects; (6) multiple dose effects. Thus, the toxic chemicals in Accelrys Toxicity Database can be assigned to six classes. To investigate the problem of predicting chemical toxicity more throughout, we also employed the nontoxic chemicals, which were also retrieved from Chen et al.'s study [5]. These chemicals were collected from DrugBank (http://www.drugbank.ca/) [13] and Human Metabolome database (HMDB) (http://www.hmdb.ca/) [14]. Totally, 174,137 chemicals were collected and each of them was nontoxic or had at least one type of toxicity.

To obtain a well-defined dataset, the chemicals with no ontology information were excluded, resulting in 4,177 chemicals. Thus, we obtained a dataset **S** consisting of 4,177 chemicals, in which 3,769 chemicals were toxic and 408 chemicals were nontoxic. As mentioned in the above paragraph, each toxic chemical has at least one type of toxicity. For convenience, let us tag the six types of toxicity using $t_1, t_2, \ldots, t_6$ and nontoxicity using t_7. Accordingly, the dataset **S** can be separated into seven subsets formulated by

$$\mathbf{S} = \mathbf{S}_1 \cup \mathbf{S}_2 \cup \mathbf{S}_3 \cup \mathbf{S}_4 \cup \mathbf{S}_5 \cup \mathbf{S}_6 \cup \mathbf{S}_7, \tag{1}$$

where $\mathbf{S}_i$ consisted of chemicals having toxicity t_i. The number of chemicals in each subset (i.e., number of chemicals having each type of toxicity) is listed in Table 1, column 3, from which we can see that the acute toxicity was a greatest type of toxicity containing most chemicals, followed by mutagenicity, multiple dose effects, and so forth, while the number of nontoxic chemicals was least. Since some chemicals may have more than one type of toxicity, that is, they may occur in more than one set of $\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_6$, the sum of numbers in seven subsets was larger than the total number of chemicals in **S**. Thus, it is a multilabel classification problem. Figure 1 gives the number of chemicals having 1–7 types of toxicity. Like many previous studies dealing with multilabel classification problem [5, 15, 16], the proposed method would give a series of candidate toxicities for each query chemical with the sequence from most likely toxicity to the least likely one.

Table 1: Distribution of chemicals in **S** and $\mathbf{S_c}$.

Tag of toxicity	Type of toxicity	Number of chemicals in $\mathbf{S}^a$	Number of chemicals in $\mathbf{S_c}^b$
t_1	Acute toxicity	3144	2993
t_2	Mutagenicity	1850	1814
t_3	Tumorigenicity	881	871
t_4	Skin and eye irritation	954	935
t_5	Reproductive effects	1099	1080
t_6	Multiple dose effects	1600	1570
t_7	Nontoxic	408	374
Total	—	9936	9637

[a] **S** is a chemical set consisting of 4,177 chemicals, which was used to examine our method.

[b] $\mathbf{S_c}$ is another chemical set consisting of 3,955 chemicals, which was used to compare our method with a previous method.

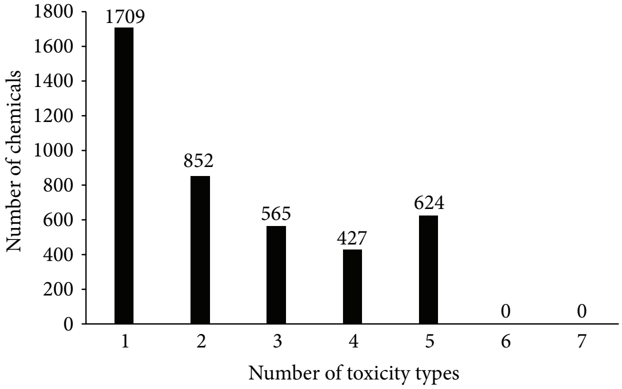

Figure 1: A histogram illustrating the number of chemicals having 1–7 types of toxicity.

2.2. Construction of a Graph by Ontology Information of Compound. The ontology information of compound was retrieved from ChEBI (http://www.ebi.ac.uk/chebi/init.do) [7]. We downloaded a file named as "chebi.obo" (accessed November 2014) from its ftp website: ftp://ftp.ebi.ac.uk/pub/databases/chebi/ontology/, which contains larger number of ontology terms and their descriptions. Since the ontology terms can be conceived as graph-theoretical structures, a graph can be constructed according to the information of all ontology terms, in which nodes represent ontology terms and edges denote the relationship between two terms. By using the entries "is a" and "relationship" in the obtained file to indicate the relationship between two terms, we constructed a large graph G with 45,206 nodes and 113,549 edges.

2.3. Prediction Method. As mentioned in Section 2.2, a graph was constructed according to the ontology information of compounds. It can be observed that the corresponding ontology terms of two adjacent nodes in G have some special relationship. And it can be further inferred that if two nodes are with small distance in G, the corresponding ontology terms have close linkage. In view of this, using the distance in G to quantitatively measure the relationship between two ontology terms is reasonable. For two terms a_1 and a_2, let us denote the distance of the corresponding nodes in G by $d(a_1, a_2)$.

For two chemicals c_1 and c_2, let $a_{11}, a_{12}, \ldots, a_{1k}$ be the ontology terms of c_1 and let $a_{21}, a_{22}, \ldots, a_{2l}$ be the ontology terms of c_2. It is obvious that if $d(a_{1i}, a_{2j})$ $(1 \le i \le k, 1 \le j \le l)$ is small, c_1 and c_2 are highly related and have high probability to share same structures, functions, and so on. Thus, we gave the following formulation to measure the common features of chemicals c_1 and c_2:

$$S(c_1, c_2) = \min \left\{ d\left(a_{1i}, a_{2j}\right) : 1 \le i \le k,\ 1 \le j \le l \right\}, \tag{2}$$

where $d(a_{1i}, a_{2j})$ denote the distance of terms a_{1i} and a_{2j} in the graph constructed in Section 2.2, which can be obtained by Dijkstra's algorithm [17]. The smaller the $S(c_1, c_2)$ is, the closer the relationship c_1 and c_2 have.

The proposed prediction method highly relied on the result of (2). To introduce the method clearly, it is necessary to employ some notations. Let $\mathbf{S}'$ be a training set consisting of n chemicals, say $c_1, c_2, \ldots, c_n$; that is, $\mathbf{S}' = \{c_1, c_2, \ldots, c_n\}$. The toxicity information of each c_i $(1 \le i \le n)$ can be represented by

$$T(c_i) = [b_{i1}, b_{i2}, \ldots, b_{i7}]^T \quad (1 \le i \le n), \tag{3}$$

where b_{ij} $(1 \le j \le 7)$ was defined by

$$b_{ij} = \begin{cases} 1 & \text{if } c_i \text{ has toxicity } t_j \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

For a query chemical c, its score of having toxicity t_j was calculated as follows.

(1) For each chemical c_i in the training set $\mathbf{S}'$, calculate $S(c, c_i)$ according to (2). Then, find all nearest neighbors, say $c_1, c_2, \ldots, c_k$, without generalization, such that $S(c, c_i) = \min\{S(c, c') : c' \in \mathbf{S}'\}$ $(1 \le i \le k)$.

(2) For each t_j, the score of c having toxicity t_j was calculated by

$$P\left(c \triangleright t_j\right) = \sum_{i=1}^{k} b_{ij}. \tag{5}$$

It is easy to observe that the score of c having toxicity t_j is the number of chemicals among $c_1, c_2, \ldots, c_k$ which have toxicity t_j. Since $c_1, c_2, \ldots, c_k$ are highly related to c, larger $P(c \triangleright t_j)$ indicates that many closely related training chemicals of c have toxicity t_j, inducing that the probability of c having toxicity t_j is high. In particular, $P(c \triangleright t_j) = 0$ suggests that the score of c having toxicity t_j is zero, inducing that the possibility of c having this toxicity is zero.

As mentioned in Section 2.1, the investigated problem is a multilabel classification problem. Only giving the most likely candidate toxicity is not enough. Fortunately, we can output a series of candidate toxicities according to the scores of the query chemical having 7 types of toxicity. The toxicity which receives the highest score is the most likely toxicity, while the toxicity receiving the second highest score is the second likely toxicity and so forth. For example, if the rank of seven scores for a certain query chemical c is

$$\begin{aligned} P\left(c \triangleright t_1\right) &\ge P\left(c \triangleright t_4\right) \ge P\left(c \triangleright t_2\right) > P\left(c \triangleright t_3\right) = P\left(c \triangleright t_5\right) \\ &= P\left(c \triangleright t_6\right) = P\left(c \triangleright t_7\right) = 0, \end{aligned} \tag{6}$$

it suggests t_1 (i.e., acute toxicity) is the most likely toxicity for c, followed by t_4 (i.e., skin and eye irritation) and t_2 (i.e., mutagenicity), while the other types of toxicity are not predicted to be candidate toxicities for c. Furthermore, t_1 is called the first prediction, t_4 the second prediction, and so forth.

Table 2: Performance of the methods on $\mathbf{S}$ and $\mathbf{S_c}$.

Prediction order	Our method on $\mathbf{S}^{a}$	Our method on $\mathbf{S_c}^{b}$	Chen et al.'s method on $\mathbf{S_c}^{b}$
1st	75.17%	75.40%	75.14%
2nd	43.52%	45.18%	49.87%
3rd	28.47%	29.76%	34.11%
4th	23.34%	24.15%	29.94%
5th	16.78%	17.98%	27.00%
6th	9.74%	10.24%	19.97%
7th	3.16%	3.16%	5.54%

[a] $\mathbf{S}$ is a chemical set consisting of 4,177 chemicals, which was used to examine our method.
[b] $\mathbf{S_c}$ is another chemical set consisting of 3,955 chemicals, which was used to compare our method with a previous method.

2.4. Accuracy Measurements. For a query chemical, the proposed method can provide a series of candidate toxicities. In view of this, we should calculate the accuracy for each order prediction. The kth prediction accuracy can be computed by [5, 15]

$$\mathrm{ACC}_k = \frac{CP_k}{N} \quad k = 1, 2, \ldots, 7, \tag{7}$$

where CP_k is the number of chemicals whose kth prediction is correct and N is the total number of chemicals that are predicted by the method. Since it is difficult to know the number of toxicities for a query chemical, the first prediction accuracy is the most important measure to evaluate the performance of the method. In addition, an effective prediction method for a multilabel classification problem should rank the candidate toxicities well; that is, prediction accuracies should follow a decreasing trend with the increasing of the prediction order.

Besides, to evaluate the performance of prediction method on the whole, another measurement was also adopted [5, 15]. It measures the proportion of the true toxicities covered by the first m predictions of chemicals, which can be calculated by

$$W_m = \frac{\sum_{i=1}^{N} \Psi_i^m}{N_i}, \tag{8}$$

where Ψ_i^m is the number of true toxicities of the ith chemical which are listed among its first m predictions and N_i is the total number of true toxicities of the ith chemical. Generally, m is always taken as the smallest integer bigger than or equal to the average number of toxicities of chemicals processed by the method; that is, $m = \lceil \sum_{i=1}^{N} N_i / N \rceil$. It is obvious that larger W_m indicates the true toxicities are arranged in the front of candidate toxicities.

3. Results and Discussion

3.1. Performance of the Method. For the 4,177 chemicals in $\mathbf{S}$, the prediction method was executed to identify their toxicities evaluated by jackknife test [15]. The seven prediction accuracies thus obtained by (7) are listed in Table 2,

Table 3: Chemicals with closest relationship of CID104975.

Compound ID	Tag of toxicity	Ontology information	Shortest path to CHEBI25957
CID995	t_1, t_2, t_3, and t_6	CHEBI:28851	CHEBI:25957, CHEBI:25959, CHEBI:25961, and CHEBI:28851
CID2236	t_1, t_2, t_3, and t_6	CHEBI:2825	CHEBI:25957, CHEBI:25959, CHEBI:25961, and CHEBI:2825
CID6763	t_1, t_2, and t_3	CHEBI:37454	CHEBI:25957, CHEBI:25959, CHEBI:25961, and CHEBI:37454
CID13257	t_2	CHEBI:35860	CHEBI:25957, CHEBI:25959, CHEBI:25961, and CHEBI:35860

column 2. It can be observed that the first prediction accuracy was 75.17%, the second one was 43.52%, and the third one was 28.47%. Furthermore, seven prediction accuracies always followed a decreasing trend with the increasing of the prediction order, indicating the proposed method arranged the candidate toxicities of all tested chemicals quite well. In addition, the average number of toxicities of chemicals in **S** was about 2.38. Thus, the first three predictions of all chemicals in **S** were collected, obtaining the accuracy of 61.87% by (8), which means the proportion of the true toxicities of chemicals in **S** covered by their first three predictions. All of these indicate that the proposed method is quite effective for identification of chemical toxicities.

3.2. Understanding the Method by Listing an Example. To better understand our method, this section listed an example. CID104975 is a chemical with toxicity t_2 (mutagenicity) and t_3 (tumorigenicity). Its ontology term is CHEBI:25957. According to the method, we computed the distance between CHEBI:25957 and ontology terms of other chemicals in **S**, thereby calculating the relationship between CID104975 and other chemicals by (2). Four chemicals, listed in Table 3, were found to be closely related to CID104975; they are CID995, CID2236, CID6763, and CID13257. Their toxicities and ontology terms are listed in Table 3, column 2 and column 3, respectively. By the method, the toxicity t_1 received 3 votes, t_2 4 votes, t_3 3 votes, t_6 2 votes, and other toxicities no votes. Accordingly, we obtained that the candidate toxicities for CID104975 were t_2, t_1, t_3, and t_6. It is obvious that the first and third predictions were correct, while the second prediction was incorrect.

3.3. Comparison of Other Methods. In this section, we employed another kind of chemical information, which has been applied for identification of chemical toxicities in Chen et al.'s study [5]. Their method used chemical-chemical interaction information, which has been deemed to be useful information for study of chemical-related problems [5, 15, 18, 19], to build the prediction method, and gave good performance.

To compare our method and Chen et al.'s method in a fair circumstance, a chemical set, consisting of 3,955 chemicals, was extracted from **S**, called $\mathbf{S_c}$, such that each chemical in $\mathbf{S_c}$ has both ontology information and interaction information; that is, each chemical can be predicted by these two methods. The number of chemicals in $\mathbf{S_c}$ on each type of toxicity is listed in Table 1, column 4, from which we can see that the distribution of 3,955 chemicals on seven types of toxicity is similar to chemicals in **S**. Also some chemicals have two or more toxicities. Our method and Chen et al.'s method were all executed on $\mathbf{S_c}$ with their performance being evaluated by jackknife test. Listed in Table 2, columns 3 and 4, are seven prediction accuracies. It can be seen that the first prediction accuracy of our method was 75.40%, which is little higher than 75.14% of Chen et al.'s method. However, with the increasing of prediction order, the prediction accuracies of Chen et al.'s method were higher than those obtained by our method. It is reasonable because the ontology information of chemicals is not very complete at present, which induces that many relations of ontology terms have not been detected. Furthermore, we also calculated the measurement defined in (8). Since the average number of toxicities of chemical in $\mathbf{S_c}$ was about 2.44, the first three predictions of chemicals in $\mathbf{S_c}$, which were obtained by two methods, were collected, thereby obtaining the accuracy of 61.70% for our method and 65.31% for Chen et al.'s method. It is also caused by the aforementioned reason. Although, if one considers more than one toxicity for a certain chemical, our method is not better than Chen et al.'s method, the first prediction accuracy of our method is higher than that of Chen et al.'s method, which is the most important one because one always pays more attention to the most likely toxicity for a chemical. In view of this, we believe that our method has superiority for identification of chemical toxicities.

4. Conclusions

This study gave a new prediction method to identify chemical toxicities. By utilizing the ontology information of chemicals reported in ChEBI, one can predict the toxicities of a certain chemical with quite high quality. It is hopeful that this method may promote the study of chemicals.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] U. C. Dubach, B. Rosner, and T. Stürmer, "An epidemiologic study of abuse of analgesic drugs. Effects of phenacetin and salicylate on mortality and cardiovascular morbidity (1968 to

1987)," *The New England Journal of Medicine*, vol. 324, no. 3, pp. 155–160, 1991.

[2] "AstraZeneca Decides to Withdraw Exanta," 2006, http://www.astrazeneca.com/Media/Press-releases/Article/20060214–AstraZeneca-Decides-to-Withdraw-Exanta.

[3] M. Zheng, Z. Liu, C. Xue et al., "Mutagenic probability estimation of chemical compounds by a novel molecular electrophilicity vector and support vector machine," *Bioinformatics*, vol. 22, no. 17, pp. 2099–2106, 2006.

[4] Y. Wang, J. Lu, F. Wang et al., "Estimation of carcinogenicity using molecular fragments tree," *Journal of Chemical Information and Modeling*, vol. 52, no. 8, pp. 1994–2003, 2012.

[5] L. Chen, J. Lu, J. Zhang, K.-R. Feng, M.-Y. Zheng, and Y.-D. Cai, "Predicting chemical toxicity effects based on chemical-chemical interactions," *PLoS ONE*, vol. 8, no. 2, Article ID e56517, 2013.

[6] Accelrys Software Inc, *Accelrys Toxicity Database 2011.4*, Accelrys Software Inc., San Diego, Calif, USA, 2011.

[7] K. Degtyarenko, P. De matos, M. Ennis et al., "ChEBI: a database and ontology for chemical entities of biological interest," *Nucleic Acids Research*, vol. 36, no. 1, pp. D344–D350, 2008.

[8] M. Ashburner, C. A. Ball, J. A. Blake et al., "Gene ontology: tool for the unification of biology," *Nature Genetics*, vol. 25, no. 1, pp. 25–29, 2000.

[9] M. A. Mahdavi and Y.-H. Lin, "False positive reduction in protein-protein interaction predictions using gene ontology annotations," *BMC Bioinformatics*, vol. 8, article 262, 2007.

[10] C.-S. Yu, C.-W. Cheng, W.-C. Su et al., "CELLO2GO: a web server for protein subcellular localization prediction with functional gene ontology annotation," *PLoS ONE*, vol. 9, no. 6, Article ID e99368, 2014.

[11] C. Bettembourg, C. Diot, and O. Dameron, "Semantic particularity measure for functional characterization of gene sets using gene ontology," *PLoS ONE*, vol. 9, no. 1, Article ID e86525, 2014.

[12] K.-C. Chou and Y.-D. Cai, "A new hybrid approach to predict subcellular localization of proteins by incorporating gene ontology," *Biochemical and Biophysical Research Communications*, vol. 311, no. 3, pp. 743–747, 2003.

[13] D. S. Wishart, C. Knox, A. C. Guo et al., "DrugBank: a knowledgebase for drugs, drug actions and drug targets," *Nucleic Acids Research*, vol. 36, no. 1, pp. D901–D906, 2008.

[14] D. S. Wishart, D. Tzur, C. Knox et al., "HMDB: the human metabolome database," *Nucleic Acids Research*, vol. 35, no. 1, pp. D521–D526, 2007.

[15] L. Chen, W.-M. Zeng, Y.-D. Cai, K.-Y. Feng, and K.-C. Chou, "Predicting anatomical therapeutic chemical (ATC) classification of drugs by integrating chemical-chemical interactions and similarities," *PLoS ONE*, vol. 7, no. 4, Article ID e35254, 2012.

[16] P. Du, T. Li, and X. Wang, "Recent progress in predicting protein sub-subcellular locations," *Expert Review of Proteomics*, vol. 8, no. 3, pp. 391–404, 2011.

[17] T. H. Gormen, C. E. Leiserson, R. L. Rivest, and C. Stein, Eds., *Introduction to Algorithms*, MIT Press, Cambridge, Mass, USA, 1990.

[18] L.-L. Hu, C. Chen, T. Huang, Y.-D. Cai, and K.-C. Chou, "Predicting biological functions of compounds based on chemical-chemical interactions," *PLoS ONE*, vol. 6, no. 12, Article ID e29491, 2011.

[19] L. Chen, J. Lu, T. Huang et al., "Finding candidate drugs for hepatitis C based on chemical-chemical and chemical-protein interactions," *PLoS ONE*, vol. 9, no. 9, Article ID e107767, 2014.