Contents lists available at ScienceDirect

Neurocomputing

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

Predicting drug–target interaction using positive-unlabeled learning

Wei Lan [a], Jianxin Wang [a,*], Min Li [a,*], Jin Liu [a], Yaohang Li [d], Fang-Xiang Wu [b], Yi Pan [a,c]

[a] School of Information Science and Engineering, Central South University, Changsha 410083, China
[b] Division of Biomedical Engineering and Department of Mechanical Engineering, University of Saskatchewan, Saskatoon, SKS7N5A9, Canada
[c] Department of Computer Science, Georgia State University, Atlanta, GA 30303, USA
[d] Department of Computer Science, Old Dominion University, VA 23507, USA

ARTICLE INFO

Article history:
Received 1 September 2015
Received in revised form
15 March 2016
Accepted 27 March 2016
Available online 31 May 2016

Keywords:
Drug–target interaction
Sequence similarity
Structure similarity
Random walk
Positive-unlabeled learning

ABSTRACT

Identifying interactions between drug compounds and target proteins is an important process in drug discovery. It is time-consuming and expensive to determine interactions between drug compounds and target proteins with experimental methods. The computational methods provide an effective strategy to address this issue. The difficulties of drug–target interaction identification include the lack of known drug–target association and no experimentally verified negative samples. In this work, we present a method, called PUDT, to predict drug–target interactions. Instead of treating unknown interactions as negative samples, we set it as unlabeled samples. We use three strategies (Random walk with restarts, *KNN* and heat kernel diffusion) to part unlabeled samples into two groups: reliable negative samples (*RN*) and likely negative samples (*LN*) based on target similarity information. Then, majority voting method is used to aggregate these strategies to decide the final label of unlabeled samples. Finally, weighted support vector machine is employed to build a classifier. Four datasets (enzyme, ion channel, GPCR and nuclear receptor) are used to evaluate the performance of our method. The results demonstrate that the performance of our method is comparable or better than recent state-of-the-art approaches.

1. Introduction

The development of a new drug is a cost- and time-consuming process. According to the US Food and Drug Administrations (FDA) statistical data, the cost of new molecular entity discovery is approximately $1.8 billion and it takes averagely 13 years [1]. In addition, only about 20 new molecular entities are approved by FDA each year. Therefore, it is an important issue in reducing these expenses in drug discovery. The computational methods provide an effective strategy to address this issue [2].

With the development of high-throughput techniques, a great deal of drug–target interaction data has been generated [3–5]. Several databases have been established to store interaction information and provide relevant retrieval servers. For example, DrugBank [6] database is a popular web resource containing information on drugs and drug targets which contains 7740 drug entries in the present version. ChEMBL [7] maintained by the European Bioinformatics Institute (EBI) is a manually curated chemical database of bioactive molecules with drug-like properties. In version 19, it contains 10,579 targets and 1,637,862 compound records and 2,843,338 bioactivity evidences. Supertarget [8] is an online and freely accessible database which contains over 6000 target proteins.

The computational methods have been boosted to predict drug–target interactions on account of the availability of interaction data. The traditional computational methods for drug–target interaction identification can be classified into three categories: ligand-based methods [9,10], docking-based methods [11,12] and literature text mining methods [13]. These approaches have achieved great successful in drug target interaction prediction. However, these methods have some limitations: the ligand-based methods rely on the number of known ligands, the docking-based methods need the information of protein structure, and literature text mining based methods are unable to find unknown and interesting interactions.

Recently, more and more statistical methods have been proposed to predict drug target interactions by integrating biological knowledge such as drug chemical structures, target protein sequence, gene expression and known drug–target interactions [14–17]. The assumption of these approaches is that similar drugs show similar patterns of interactions with targets in drug–target interaction network [18,19]. Chen et al. [15] presented network-based random walk with restart method, called NRWRH, to predict relationships between drugs and targets by integrating drug–drug chemical structure similarity network, protein–protein sequence

[*] Corresponding authors.
E-mail addresses: jxwang@mail.csu.edu.cn (J. Wang), limin@mail.csu.edu.cn (M. Li).

http://dx.doi.org/10.1016/j.neucom.2016.03.080
0925-2312/© 2016 Elsevier B.V. All rights reserved.

similarity network and known drug–target interaction network into a heterogeneous network. Cheng et al. [14] proposed three inferring methods including drug-based similarity inference (DBSI), target-based similarity inference (TBSI) and network-based inference (NBI) to predict drug–target interactions. Similar work has been accomplished by Alaimo et al. [20], they presented DT-hybrid approach which extends network-based inference method by domain-based knowledge to detect drug–target interactions. Emig et al. [21] integrated different network-based methods to predict drug targets of a specific disease. These methods are easy to be implemented. However, these methods are unable to apply to drugs without any targets information. In addition, Bleakley and Yamanishi [17] employed bipartite local models to predict relationships between drugs and targets. Further work has been completed by Mei et al. [22], they integrated neighbour information into bipartite local models for drug target interaction identification. The Gaussian interaction profile kernel and weighted nearest neighbour were integrated for drug–target interaction prediction [23]. The Bayesian matrix factorization and binary classification [24] and probabilistic matrix factorization [25] were proposed to detect drug–target interactions. The common limitation of these supervised learning approaches is to treat unknown drug–target interactions as negative samples, which may affect predictive accuracy. Xia et al. [16] developed a semi-supervised method (NetLapRLS) for drug–target interaction identification by using positive and unlabeled samples. Chen and Zhang [26] presented NetCBP method by maximizing the rank coherence with respect to known knowledge to identify associations between drugs and targets. These semi-supervised methods can make use of unlabeled information. But they need to combine two different classifiers in the final.

Despite these approaches have achieved good performance, there are some limitations and difficulties for drug–target interactions prediction. Firstly, most of the methods adopt sequence information to measure the similarity of two proteins. More studies demonstrate that the structure information is more conservative than sequence information. Therefore, the structure information of target protein may be better suited for drug–target interaction identification. Secondly, there are no experimentally verified negative samples. Traditional methods treat the non-interaction data as negative sample which is unreasonable as those non-interaction data may contain undetected drug–target interactions. Thirdly, some methods are unable to predict new drugs without any targets, which limits the application in practice.

In this paper, we propose a framework to predict drug–target interaction based on positive-unlabeled learning. Comparing with existing approaches, we integrated multiple target resources including target structure information, target function category information and target function annotation information. In addition, we treat unknown drug target interactions as unlabeled set *U* instead of negative set *N*. Three strategies (Random walk with restarts, *KNN* and heat kernel diffusion) are used to classify unlabeled samples into two groups: reliable negative samples (*RN*) and likely negative samples (*LN*) based on target similarity information and majority voting method is used to aggregate these strategies to decide the final label of unlabeled samples. The weighted support vector machines are employed to build a multi-level classifier to predict drug target interactions based on positive set, reliable negative set and likely negative set. The experiments are conducted on four datasets (including Enzyme, Ion Channel, GPCR and Nuclear Receptor). The experimental results demonstrate that our method outperforms state-of-the-art approaches.

2. Materials and methods

2.1. Data preparation

In this paper, we use four drug–target interaction networks in human involving Enzyme, Ion Channel, GPCR and Nuclear Receptor which are first analysed by Yamanishi et al. [27]. These datasets can be downloaded from http://web.kuicr.kyoto-u.ac.jp/supp/yoshi/drugtarget/. Table 1 show some information of four datasets. The drug–target interaction data are collected from the KEGG BRITE [28], BRENDA [29], SuperTarget [8] and DrugBank [6].

Drug chemical structure information is retrieved from the DRUG AND COMPOUND Sections in the KEGG LIGAND [28]. The chemical structure similarity between compounds is calculated by using SIMCOMP, which gives a score based on the size of common substructures with graph alignment [17,30]. The chemical structure similarity has been widely applied in drug–target interaction prediction [15].

The sequence similarity between targets is calculated by normalized Smith–Waterman algorithm based on the information of amino acid sequence of target proteins extracted from KEGG GENE database [31]. Given two proteins A_i and A_j, the sequence similarity between two proteins is calculated as:

$$Sim\_seq(A_i, A_j) = \frac{SW(A_i, A_j)}{\sqrt{SW(A_i, A_i)}\sqrt{SW(A_j, A_j)}} \tag{1}$$

where $SW(A_i, A_j)$ is the score of Smith–Waterman algorithm.

The 3D structures of target protein are obtained from PDB database [32]. There are over 98,000 3D structures existing in PDB database. For some target proteins without known 3D structure in PDB, the SWISS-MODEL [33], which is a popular method to generate reliable three-dimensional protein structure models based on homology modeling, is employed to predict their 3D structures. The structure similarity between targets is calculated by utilizing TM-align tool [34].

The function category (FC) information of targets is extracted from HUGO Gene Nomenclature committee [35] based on their gene families. It is organized as hierarchically structure. The FC-based similarity is calculated by counting the common sub-codes from top to bottom in their hierarchy. For example, the FC code of target contains *N* entries. If the first *M* entries of two targets are the same, then the similarity of two targets can be calculated as *M*/*N*.

The Gene ontology (GO) information is useful for evaluating gene function similarity [36,37]. In here, we obtain GO data from Gene Ontology Consortium [38]. It contains three kinds of ontology: biological process (BP), cellular component (CC) and molecular function (MF). The GO terms are presented as directed acyclic graphs (DAGs) in which the terms form nodes and the two kinds of semantic relations (is-a and part-of) form edges. Then, we calculated functional similarity of targets based on the semantic similarities [39] of GO terms used for gene annotation.

2.2. Positive-unlabeled learning for drug–target prediction (PUDT)

Our method for drug–target prediction is based on assumption that similar drugs often target on similar proteins. Under this assumption, our method integrates target sequence similarity, target FC-based similarity, target GO-based similarity, target 3D structure similarity, drug chemical structure similarity and drug–target interaction data to predict drug–target interactions. Most methods treat unlabeled samples as negative samples. However, unlabeled samples may contain undetected positive samples which may influence the accuracy of classification. Thus, instead of treating unlabeled samples as negative samples, we use three