Computers in Biology and Medicine 68 (2016) 101–108

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

ARWAR: A network approach for predicting Adverse Drug Reactions

Hossein Rahmani [a,b,*], Gerhard Weiss [b], Oscar Méndez-Lucio [c], Andreas Bender [c]

[a] School of Computer Engineering, Iran University of Science and Technology, Tehran, Iran
[b] Maastricht University, PO Box 616, Maastricht 6200 MD, The Netherlands
[c] Centre for Molecular Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, United Kingdom

ARTICLE INFO

Article history:
Received 19 June 2015
Accepted 10 November 2015

Keywords:
Side effects
Adverse Drug Reactions
Network approach
Human Drug Network
Multi-label classification
Graph augmentation
Random Walk with Restarts

ABSTRACT

Predicting novel drug side-effects, or Adverse Drug Reactions (ADRs), plays an important role in the drug discovery process. Existing methods consider mainly the chemical and biological characteristics of each drug individually, thereby neglecting information hidden in the relationships among drugs. Complementary to the existing individual methods, in this paper, we propose a novel network approach for ADR prediction that is called Augmented Random-WAlk with Restarts (ARWAR). ARWAR, first, applies an existing method to build a network of highly related drugs. Then, it augments the original drug network by adding new nodes and new edges to the network and finally, it applies Random Walks with Restarts to predict novel ADRs. Empirical results show that the ARWAR method presented here outperforms the existing network approach by 20% with respect to average Fmeasure. Furthermore, ARWAR is capable of generating novel hypotheses about drugs with respect to novel and biologically meaningful ADR.

1. Introduction

A typical drug discovery cycle, from target identification to clinical use, can take approximately 14 years [1] with an associated cost of 800 million US dollars [2]. One of the main causes of failure in the process of the drug development is the existence of Adverse Drug Reactions (ADRs). ADRs are known as a serious clinical problems and are estimated to result in more than 2 million hospitalizations [3] and more than 100,000 deaths in the United States per year [4]. Additionally, in case of serious ADR pharmaceutical companies are forced to withdraw their drugs from the market, which involves significant danger for patients, as well as major financial implications to the companies involved. Therefore, predicting the ADRs prior to market introduction of the drug is necessary and has been considered as a very challenging issue in drug development.

Laboratory-based approaches for predicting and evaluating the potential ADRs are very costly and time consuming. Therefore, using computational approaches for early identification of potential ADRs in the drug discovery process gained much attention in the recent years.

The general pattern for computational methods is as follows: First, they consider different chemical and biological properties of the drugs. Second, they transform the considered properties into numerical features. Third, they develop a systematic way of understanding, predicting and interpreting the desired and undesired effects of drugs [5–16]. The main difference among these methods lies in the type of properties they consider for the ADRs analysis.

In the most domains, more interesting knowledge can be mined from the relationships among entities [17]. For example, several studies [18–21] showed that considering the relationships among different diseases reveals informative patterns and is indeed useful for different prediction tasks. To the best of our knowledge, most of the existing methods focus on each drug individually neglecting the informative knowledge that could be gained from the hidden relationships among different drugs. However, there are some more recent approaches that follow "similar drugs have similar ADRs" pattern and consider relationships among drugs for predicting ADRs [22,23]. For example, Vilar et al. [22] and Luo et al. [23] calculated similarity between two drugs according to drugs' 3D molecular structure and their docking profiles, respectively, to predict potential ADR for new drugs.

Extending upon previous approaches, we now in this work consider also information hidden in the relationships among drugs. We apply the previous method [24] (as described in Section 6.2) to discover hidden relationships among drugs. Rahmani et al. [24] consider target proteins of drugs, Protein–Protein Interaction (PPI) networks, functional and structural information of PPI networks to discover the most informative relationships among drugs and accordingly, build a network among highly related drugs. Network representation of relationships among drugs provides the unique

* Corresponding author.
E-mail address: h.rahmani@maastrichtuniversity.nl (H. Rahmani).

http://dx.doi.org/10.1016/j.compbiomed.2015.11.005
0010-4825/

opportunity to apply the successful off-the-shelf network-based classifiers in other domains for predicting novel ADRs. The effectiveness of the network approach for generating novel hypotheses about drugs has been shown in previous studies [24–29].

In this paper, we explore methods implementing the network approach for predicting the novel ADRs. We examine two classifiers for this purpose. The first classifier is called the Majority Rule Method (MRM) [30] and considers the ADR of neighboring drugs in the network for the prediction. MRM has been used as a baseline method in different domains [31–33]. Considering the limitations of MRM, we propose a second classifier that is called Augmented Random WAlk with Restarts (ARWAR). Our empirical results show that ARWAR outperforms MRM significantly with respect to Fmeasure and is capable of generating novel hypotheses about ADR.

This paper is structured as follows: Section 2 discusses the previous methods on ADR prediction. We model, formally, the task of ADR prediction as a multi-label classification problem in Section 3. Section 4 discusses MRM and its limitations as one of the prominent methods for the multi-label classification problem. We describe our proposed ARWAR approach in details in Section 5. In Section 6, a drug network is constructed and then evaluated in terms of interpretability and novel ADR predictions. Section 7 concludes.

2. Background

Considering different types of input data, we categorize the existing computational methods for the task of novel ADR prediction into three categories.

The first category of methods tries to relate drug side-effects to their chemical substructures [11,12,8,5]. Their results indicate that side-effects of drugs are usually associated with the presence of specific chemical substructures. However, their precision is highly dependent on the pre-definition of chemical substructures. This is true for specific toxic features, e.g. nitrogen mustards, but usually toxicity depends on complex combination of substructures that is not captured by these methods.

The second category of methods relates drug side-effects to its protein targets [9,14,34]. Campillos et al. [9] propose a measure for side-effect similarity by considering the relations among terms in the Unified Modeling Language System (UMLS) ontology. Then, they observe a clear correlation between side-effect similarity and the likelihood that two drugs share protein targets. Finally, they exploit this characteristic to predict novel target proteins for drugs. Fukuzaki et al. [14] use cooperative pathways and gene expression profiles to predict ADRs. Brouwers et al. [34] present the contribution of Protein–Protein Interaction (PPI) networks to drug side-effect similarities.

The third category of methods predicts drug side-effects by integrating multiple data sources [10,13,35,36]. Yamanishi et al. [10] describe each drug according to its chemical profile (an 881 dimensional feature vector where each element encodes for the presence/absence of each PubChem chemical structure) and biological profile (an 1368 dimensional feature vector where each element encodes for the presence/absence of each target protein). Then, they apply different machine learning methods to predict potential side-effect profiles for uncharacterized drugs. Huang et al. [13,35] significantly improve the accuracy of ADR prediction by integrating drug target data, PPI networks, drug structure and Gene Ontology (GO) term annotations. Liu et al. [36] apply five different machine learning methods, namely logistic regression (LR), naive Bayes (NB), K-nearest neighbor (KNN), random forest (RF), and support vector machine (SVM) on the integration of chemical, biological and phenotypic (i.e., indications and other known side-effects) properties. Then, they show that SVM outperforms the other methods and phenotypic data are the most informative for the ADR prediction. The latter conclusion can be explained by the existence of high correlation among ADRs.

3. Problem statement

In this section, we model the task of ADR prediction as a network-based multi-label classification problem. Consider an undirected network $G\langle V, E\rangle$ with node set V and edge set E, where each node $v_i \in V$ is annotated with a description $d(v_i) \in D$ and, optionally, a label $l(v_i) \in L$. We assume that there exists a "true" labeling function λ from which l is a sample, i.e., $l(v) = \lambda(v)$ where $l(v)$ is defined. The task of node classification [37] is to predict the labeling set $l(v_i)$ for each unclassified node v_i. If $|L| = 2$ then the classification problem is called binary classification while if $|L| > 2$ then it is called multi-class classification. In case $l(v)$ associated with a set of labels $Y \subseteq L$ then the classification problem is called multi-label classification [38].

In our Human Drug Network (HDN) (as described in details in Section 6.2), each node $v_i \in V$ represents a drug and each edge $e_{ij} \in E$ represents an relationship between two drugs v_i and v_j. Description vector $d(v_i)$ contains the available biological and chemical properties of drug v_i. The labeling function $l(v_i)$ returns a set of ADR for drug v_i ($|l(v_i)| >= 0$). In this context, the task of a multi-label classification is to generate a classifier H that, given an unlabeled drug v_j with description vector $d(v_j)$, is capable of predicting the ADR associated to v_j.

In the following sections, we discuss two network approaches for predicting novel ADR considering relationships among drugs. The first one is called Majority Rule Method (MRM) and has been applied before in several domains [30–33], while the second method is the Random-WAlk with Restarts (ARWAR) method proposed here. Both methods take Human Drug Network, that is partially annotated with ADR, as input and predict new ADR for drugs in HDN as an output. We discuss both methods in the following sections.

4. Majority Rule Method (MRM) and its limitations

K-nearest neighbor (KNN) classifier considers the majority label (s) of k nearest neighbors of unclassified input data in the classification process [39]. KNN is easy to implement, its results are easy to interpret and it has been studied extensively in the literature [40–42]. One specific graph implementation of KNN classifier is called Majority Rule Method (MRM) [30] that assigns to each unclassified node those labels that occur most frequently among its neighbors in the graph. As an example, Fig. 1 shows the simple graph with four nodes $V = \{v_1, v_2, v_3, v_4\}$ and labeling set $l(v_i)$ for each node v_i. MRM predicts $\{l_1\}$ for node v_1 as it occurs most in the neighborhood of v_1.

However, this method suffers from several limitations. First, this method only considers the local neighborhood of the v_i ignoring the remaining information in the network. In Fig. 2, MRM

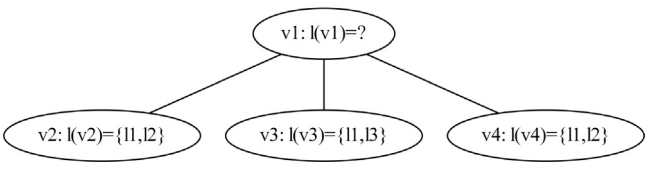

Fig. 1. Simple graph with node set $V = \{v_1, v_2, v_3, v_4\}$ and labeling set $l(v_i)$ for each node v_i. MRM predicts $\{l_1\}$ for node v_1 as it occurs most in the neighborhood of v_1.