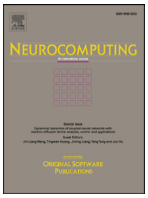
Neurocomputing 260 (2017) 284–293

Contents lists available at ScienceDirect

Neurocomputing

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



Drug–target interaction prediction with Bipartite Local Models and hubness-aware regression

Krisztian Buza[a,b,*], Ladislav Peška[c]

[a] Brain Imaging Center, Research Center for Natural Sciences, Hungarian Academy of Sciences; Magyar tudósok körútja 2., 1117 Budapest, Hungary
[b] Knowledge Discovery and Machine Learning, Institut für Informatik III, Rheinische Friedrich-Wilhelms-Universität Bonn; Römerstr. 164, 53117 Bonn, Germany
[c] Department of Software Engineering, Faculty of Mathematics and Physics, Charles University; Malostranske nam. 25, 11800 Prague, Czech Republic

ARTICLE INFO

Article history:
Received 21 September 2016
Revised 24 February 2017
Accepted 21 April 2017
Available online 11 May 2017

Communicated by Prof. S. Mitra

Keywords:
Drug-target interaction prediction
Bipartite local models
Hubness-aware machine learning
Regression

ABSTRACT

Computational prediction of drug–target interactions is an essential task with various applications in the pharmaceutical industry, such as adverse effect prediction or drug repositioning. Recently, expert systems based on machine learning have been applied to drug–target interaction prediction. Although hubness-aware machine learning techniques are among the most promising approaches, their potential to enhance drug–target interaction prediction methods has not been exploited yet. In this paper, we extend the Bipartite Local Model (BLM), one of the most prominent interaction prediction methods. In particular, we use BLM with a hubness-aware regression technique, EC*k*NN. We represent drugs and targets in the similarity space with rich set of features (i.e., chemical, genomic and interaction features), and build a projection-based ensemble of BLMs. In order to assist reproducibility of our work as well as comparison to published results, we perform experiments on widely used publicly available drug–target interaction datasets. The results show that our approach outperforms state-of-the-art drug–target prediction techniques. Additionally, we demonstrate the feasibility of predictions from the point of view of applications.

1. Introduction

Due to the large number of drug compounds and pharmacological targets, many of the interactions between these entities are unknown. More complete knowledge about drug–target interactions will not only contribute to better understanding the pharmacology of drugs, but it is also relevant for the prediction of adverse effects and drug repositioning, i.e., use of an existing medicine to treat a disease that has not been treated with that drug yet. The relevance of the later application is also underlined by the fact that only a few dozens of new drugs are approved by FDA each year. Moreover, the average costs related to discovery of a new drug are approximately $1.8 billion, and the process takes more than 10 years [18].

In addition, the incomplete knowledge about the interactions between drugs and pharmaceutical targets in case of drugs affecting the central nervous system (CNS) further emphasizes the need for computational prediction approaches: while CNS plays an essential role, the costs associated with disorders affecting CNS are enormous: solely in Europe, the total annual costs associated with brain disorders is estimated to be approximately 800 billion EUR [25].

The biochemical validation of hypothesized drug–target interactions is laborious, time-consuming and expensive [31,49]. Therefore, computational methods have been proposed for the prediction of drug–target interactions [5,22,23,34]. Traditional techniques include approaches based on molecular docking [10,15,29], ligand chemistry [21], [26] and text mining [53].

A serious limitation of docking-based approaches is that they require information about the three-dimensional structure of candidate drugs and targets which is often not available, especially for G-protein coupled receptors (GPCRs) and Ion Channels. Additionally, the performance of ligand-based approaches decrease in case if only few ligands are known.

In response to the above limitations of classic approaches, expert systems based on machine learning techniques have been proposed for the prediction of drug–target interactions [14,50,51]. Recent approaches are based on matrix factorization [12,14,52], restricted Boltzmann machines [48], network-based inference [9,11,35,43], positive-unlabeled learning [22] and the integration of

[*] Corresponding author at: Knowledge Discovery and Machine Learning, Institut für Informatik III , Rheinische Friedrich-Wilhelms-Universität Bonn; Römerstr. 164, 53117 Bonn, Germany

E-mail addresses: buza@biointelligence.hu, chrisbuza@yahoo.com (K. Buza), peska@ksi.mff.cuni.cz (L. Peška).

http://dx.doi.org/10.1016/j.neucom.2017.04.055
0925-2312/© 2017 Elsevier B.V. All rights reserved.

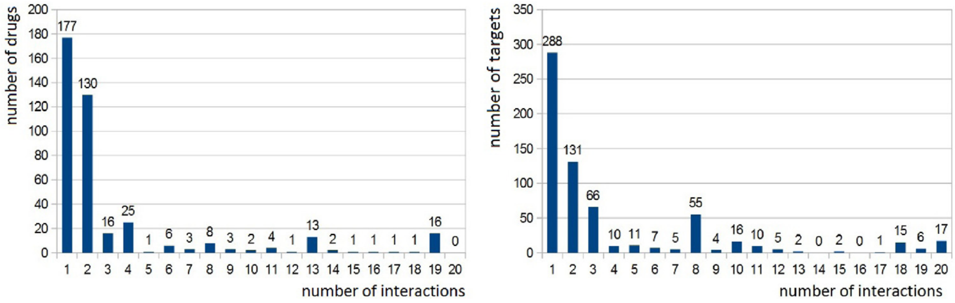


Fig. 1. The degree distribution in the *Enzyme* drug–target network. The horizontal axis corresponds to the number of interactions, whereas the vertical axis corresponds to the number of drugs (left) or targets (right). For example, the first column in the left diagram shows that 177 drugs interact with only one (but not necessarily the same) target. In contrast, some drugs (and targets, resp.) participate in surprisingly many interactions, e.g., there are 16 drugs, each of which interacts with 19 targets.

multiple sources of information [33,42]. See also [32,37] for excellent surveys.

One of the most prominent drug–target interaction prediction techniques is based on Bipartite Local Models (BLMs) [4] in case of which the drug–target interaction prediction task is treated as a link prediction problem in bipartite graphs. Recent works aiming to extend BLMs focused on semi-supervised prediction [50], improved kernels [47] and the incorporation of neighbor-based interaction-profile inferring [24].

However, none of the aforementioned prediction techniques took the presence of hubs into account. With hubs, we mean entities that are connected to surprisingly many other entities of a network. This phenomenon has been observed for various biological, chemical and medical networks, see e.g. [1,17]. Similar observations can be made for drug–target networks as well, e.g., Fig. 1 shows the degree distribution for the drugs and targets in the *Enzyme* drug–target interaction network (we will describe the data in Section 2.1). As one can see, the distributions have long tails, i.e., there are drugs (and targets, resp.) that are connected with surprisingly many targets (drugs, resp.) compared to “average” drugs (targets, respectively).

The presence of hubs has been observed in nearest neighbor graphs, see e.g. [8,28,46], and hubness-aware classifiers have been developed, see [45] for a survey. More recently, hubness-aware regression techniques, including *k*-nearest neighbor with error correction (EC*k*NN), were developed that allow for predictions on a continuous scale [7]. Despite the fact that hubness-aware techniques are among the most promising recent machine learning approaches, their potential to enhance drug–target interaction prediction methods has not been exploited yet: to the best of our knowledge, our initial work [6] is the only one aiming to apply hubness-aware models to the drug–target prediction problem.

In this study, we extend Bipartite Local Models and our previous work [6]. We use EC*k*NN as local model of BLM and propose an enhanced representation of drugs and targets in a multi-modal similarity space (i.e., a representation which incorporates multiple similarity measures). Furthermore, we build a projection-based ensemble and study how the performance depends on the number of base models of the ensemble. As we use hubness-aware local models in the proposed approach, we refer to it as HLM for simplicity. In order to assist reproducibility of our work as well as comparison to published results, we perform experiments on publicly available real-world drug–target interaction datasets. The results show that our approach outperforms other state-of-the-art drug–target prediction techniques.

The rest of this paper is organized as follows: in Section 2 we review the background necessary to understand our work. In particular, we focus on BLM and EC*k*NN. Section 3 presents the proposed approach, followed by its experimental evaluation in Section 4. Finally, conclusions are drawn in Section 5.

Table 1
Number of drugs, targets and interactions in the datasets used in our study.

Dataset	# Drugs	# Targets	# Interactions
Enzyme	445	664	2926
Ion Channels	210	204	1476
G-protein coupled receptors (GPCR)	223	95	635
Nuclear Receptors (NR)	54	26	90
Kinase [36]	68	442	1527

2. Materials and methods

In order to ensure that the paper is self-contained, we begin this section by describing the datasets used in our study and the procedure to obtain drug–drug and target–target similarities. Subsequently, the BLM approach for drug–target interaction prediction is reviewed. This is followed by the description of hubness-aware error correction for nearest neighbor regression.

2.1. Drug–target interaction data

In our study we used five publicly available drug–target interaction datasets from two repositories,[1] namely Enzyme, Ion Channel, G-protein coupled receptors (GPCR), Nuclear Receptors (NR), and Kinase [36]. These datasets have been used in various studies, see e.g. [4,14,39,50,51].

Each of the first four datasets contains a binary interaction matrix between drugs and targets, in which each entry indicates whether the interaction between the corresponding drug and target is known or not. In contrast, Kinase contains continuous values of binding affinity for all drug–target pairs of the data. In order to produce a binary interaction matrix, we used the same cutoff threshold as Pahikkala et al. [39]. Table 1 shows the number of drugs, targets and interactions in the datasets.

In general, drug–drug and target–target similarities may be computed in many ways. Next, we describe the similarities used in our study. In case of the Enzyme, Ion Channel, GPCR and NR datasets, chemical structure similarities were computed using the SIMCOMP [16] graph-alignment algorithm, in order to obtain drug–drug similarities. For Kinase, we used 2D Tanimoto coefficients as drug–drug similarities.

In order to compute the similarity between target proteins of the Enzyme, Ion Channel, GPCR and NR datasets, their amino acid sequences were retrieved from the KEGG GENES [20] database so that similarities between pharmacological targets were determined by sequence alignment methods, such as the Smith–Waterman algorithm. We refer to [51] for more details. For Kinase, we used the normalized Smith-Waterman scores as target–target similarities.

[1] http://web.kuicr.kyoto-u.ac.jp/supp/yoshi/drugtarget/ http://staff.cs.utu.fi/~aatapa/data/DrugTarget/