Analytica Chimica Acta 909 (2016) 41–50

Contents lists available at ScienceDirect

Analytica Chimica Acta

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

Improved prediction of drug-target interactions using regularized least squares integrating with kernel fusion technique

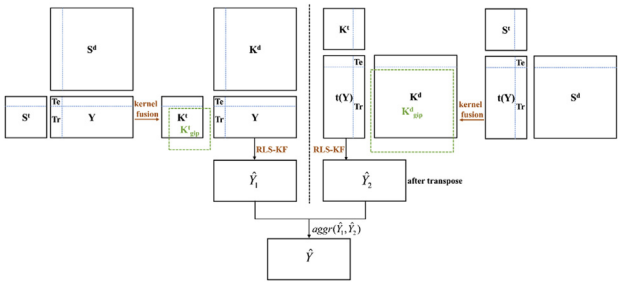
Ming Hao, Yanli Wang*, Stephen H. Bryant**

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

HIGHLIGHTS

- A nonlinear kernel fusion algorithm is proposed to perform drug-target interaction predictions.
- Performance can further be improved by using the recalculated kernel.
- Top predictions can be validated by experimental data.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 24 September 2015
Received in revised form
6 January 2016
Accepted 7 January 2016
Available online 14 January 2016

Keywords:
Drug-target interactions
Regularized least squares
Kernel fusion
PubChem BioAssay
Drug repositioning

ABSTRACT

Identification of drug-target interactions (DTI) is a central task in drug discovery processes. In this work, a simple but effective regularized least squares integrating with nonlinear kernel fusion (RLS-KF) algorithm is proposed to perform DTI predictions. Using benchmark DTI datasets, our proposed algorithm achieves the state-of-the-art results with area under precision–recall curve (AUPR) of 0.915, 0.925, 0.853 and 0.909 for enzymes, ion channels (IC), G protein-coupled receptors (GPCR) and nuclear receptors (NR) based on 10 fold cross-validation. The performance can further be improved by using a recalculated kernel matrix, especially for the small set of nuclear receptors with AUPR of 0.945. Importantly, most of the top ranked interaction predictions can be validated by experimental data reported in the literature, bioassay results in the PubChem BioAssay database, as well as other previous studies. Our analysis suggests that the proposed RLS-KF is helpful for studying DTI, drug repositioning as well as polypharmacology, and may help to accelerate drug discovery by identifying novel drug targets.

Published by Elsevier B.V.

1. Introduction

Identifying interactions between chemical compounds and target proteins plays a fundamental role in drug discovery processes. Pharmaceutical companies, on the one hand, would like, as soon as possible, to detect hidden adverse events (such as adverse drug reactions), which has been a major global health concern, causing side effects, hospitalizations, even deaths [1]. On the other hand, they also would like to explore adverse events to find new applications [2] (drug repositioning or drug repurposing). Both of the purposes can be attributed to accurately identify the potential drug-target interactions (DTI). It is well known that experimental validation of interactions is costly and laborious. Therefore, application of in silico methods for this challenge is needed.

Several traditional methods [3,4], such as ligand-based QSAR

* Corresponding author.
** Corresponding author.
E-mail addresses: ywang@ncbi.nlm.nih.gov (Y. Wang), bryant@ncbi.nlm.nih.gov (S.H. Bryant).

http://dx.doi.org/10.1016/j.aca.2016.01.014
0003-2670/Published by Elsevier B.V.

(quantitative structure–activity relationship) and receptor-based docking, are often used to predict DTI. However, they often have limitations. For QSAR, its performance might be decreased when the training samples are not enough. For docking, it largely depends on the 3D crystal structures of protein targets. Therefore, it is difficult to study DTI for membrane proteins due to the limited number of known 3D structures. In addition, docking-based methods are not computationally efficient and previous studies mostly focused on one single target. With the advent of chemogenomics research accelerated by high-throughput screening (HTS) campaigns of large-scale chemical libraries and the completion of human genome project, more chemical and genomic data are now publicly available, which enables researchers to study DTI at a large scale, such as studying interactions among multiple drugs and multiple targets using computational approaches.

In 2008, Yamanishi and colleagues [5] proposed a bipartite network method for the integration of chemical and genomic spaces to predict DTI of four classes of protein targets, i.e., enzymes, ion channels (IC), G protein-coupled receptors (GPCR) and nuclear receptors (NR). Their models suggested many potential interaction pairs between drugs and targets. As a following study, Bleakley et al. [6] proposed a novel supervised inference method to predict unknown drug-target interactions from the same datasets used by Yamanishi and co-workers. Their kernel-based models using support vector machine (SVM) transformed the edge-prediction problem into the binary classification problem of points with label. Results from their models gave high performance in terms of AUC (area under receiver operating characteristic curve) and AUPR (area under precision–recall curve).

van Laarhoven et al. [7] used a simple machine learning method called (kernel) regularized least squares (RLS) to predict DTI by using only the topological information from the adjacency matrix of drug-target network. Then they defined a kernel on the topology profiles, called Gaussian interaction profile (GIP) kernel. Using the only defined kernel, results from their models exhibited a significant improvement for AUPR over results of the state-of-the-art methods at that time. Furthermore, they found that by combining the topological information with others (such as chemical and genomic information), the performance could further be improved. However, their method was focusing on the setting where both drugs and targets are known, which means that they used known interactions for predicting novel ones. Thus, for the situation where both drugs and targets are new (meaning that there are not interactions between them), these models are not feasible. In order to overcome such limitation, Mei and co-workers [8] introduced a neighbor-based interaction-profile inferring (NII) method and integrated it into the existing bipartite local model (called BLM-NII). By incorporating NII algorithm, the performance of DTI predictions for the four benchmark datasets presented a significant improvement, which turned out to be the best results.

Apart from the aforementioned popular methods for predicting DTI, various novel statistical methods were also proposed, such as restricted Boltzmann machines [9], Bayesian matrix factorization [10], even ranking-based method [11]. All these methods exhibited good performance but those kernel-based methods have been the most popular ones.

It is noted that the previous kernel-based methods [7,8] for DTI predictions used only a simple linear combination of different kernels as input to form final kernel matrix. However, that approach may not be appropriate when linear relationship is not evident among kernels. Thus, in this work, we explored a nonlinear kernel fusion (KF) technique, which was originally applied successfully in patient similarity network by Wang et al. [12], to combine different kernels for predicting DTI. The kernel fusion algorithm can derive both shared and complementary information from various kernel matrices, even those from a small number of samples. In order to validate the effectiveness of our proposed algorithm, we integrated a simple but effective regularized least squares (incorporating NII) with novel nonlinear kernel fusing (RLS-KF) technique, and compared the results of DTI predictions for the four benchmark DTI datasets [5] with those from previously reported methods. Moreover, we recalculated the kernel matrices of drug compounds and target proteins, and results based on this exhibited a further improvement especially for the small NR dataset. Importantly, most of the top predicted interaction pairs have been successfully validated by either experimental data reported in the literature, confirmatory assay results in the PubChem BioAssay database, as well as by results in other previous studies.

2. Material and experimental methods

2.1. Dataset

Four drug-target interaction networks, including enzymes, ion channels, G protein-coupled receptors and nuclear receptors in human, originally studied by Yamanishi et al. [5], were used as the benchmark datasets in the current work. These interaction information was retrieved from KEGG BRITE [13], BRENDA [14], SuperTarget [15] and DrugBank [16] databases. Protein sequences of the target proteins were obtained from the KEGG GENES database [13]. Target sequence similarity matrices (denoted by S^t, which is an M by M square matrix, where M denotes the number of targets) between proteins were computed using a normalized version of Smith-Waterman score [17]. Chemical compounds were derived from the KEGG DRUG and COMPOUND databases [13]. Chemical structure similarity matrices (denoted by S^d, which is an N by N square matrix, where N denotes the number of drugs) between compounds were computed using the SIMCOMP tool [18]. The M by N adjacency matrix, Y, where $Y_{ij} = 1$ if drug i interacts with target j, and $Y_{ij} = 0$ otherwise, was the same to that used in the previous study [5]. Table 1 lists the summary of all four datasets.

2.2. Problem formalization

Given three matrices, S^t, S^d and Y, the task is how to make use of them to predict interactions between drug compounds and target proteins, which includes four scenarios of interactions between existing/new drugs and targets as described in the literature [8]. A brief diagram (Fig. 1) is given to explain the notation of existing/new drugs and targets, which assumes there are 4 targets (T_1 through T_4) and 5 drugs (D_1 through D_5) in total. Taking the first drug, D_1, as a query drug, the purpose of current work is to predict if D_1 interacts with T_1 (in the test set) by using the related information from the training set (labeled in red). If there is at least one interaction known between D_1 and any target from T_2 through T_4, then the current query drug is denoted as an existing drug (Fig. 1A and B), or a new drug otherwise (Fig. 1C and D). Similarly for the definition of existing targets and new targets, if there is at least one interaction between T_1 (in the test set) and any drug from D_1 through D_5, the current target is denoted as an existing target (Fig. 1A and C), or a new target otherwise (Fig. 1B and D). Thus, four

Table 1
Summary of the four benchmark datasets.

Data	Enzymes	IC	GPCR	NR
Number of targets	664	204	95	26
Number of drugs	445	210	223	54
Number of interactions	2926	1476	635	90