Methods xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Methods

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

Boosting compound-protein interaction prediction by deep learning

Kai Tian [a,1], Mingyu Shao [a,1], Yang Wang [c], Jihong Guan [b], Shuigeng Zhou [a,*]

[a] Shanghai Key Lab of Intelligent Information Processing, School of Computer Science, Fudan University, Shanghai 200433, China
[b] Department of Computer Science and Technology, Tongji University, Shanghai 201804, China
[c] School of Software, Jiangxi Normal University, Nanchang 330022, China

ARTICLE INFO

Article history:
Received 12 April 2016
Received in revised form 20 June 2016
Accepted 28 June 2016
Available online xxxx

Keywords:
Compound-protein interaction
Deep learning
Deep neural network (DNN)

ABSTRACT

The identification of interactions between compounds and proteins plays an important role in network pharmacology and drug discovery. However, experimentally identifying compound-protein interactions (CPIs) is generally expensive and time-consuming, computational approaches are thus introduced. Among these, machine-learning based methods have achieved a considerable success. However, due to the nonlinear and imbalanced nature of biological data, many machine learning approaches have their own limitations. Recently, deep learning techniques show advantages over many state-of-the-art machine learning methods in some applications. In this study, we aim at improving the performance of CPI prediction based on deep learning, and propose a method called DL-CPI (the abbreviation of *D*eep *L*earning for *C*ompound-*P*rotein *I*nteractions prediction), which employs *deep neural network* (DNN) to effectively learn the representations of compound-protein pairs. Extensive experiments show that DL-CPI can learn useful features of compound-protein pairs by a layerwise abstraction, and thus achieves better prediction performance than existing methods on both balanced and imbalanced datasets.

1. Introduction

In network pharmacology [1], the assumption of "one drug for one target for one disease" (on which traditional drug discovery is based) is challenged and the relationships between drugs and targets become complicated. One drug may act on multiple targets while there are also proteins that are targeted by two or more compounds. Therefore, identification of the interactions between chemical compounds and proteins plays a critical role in network pharmacology, drug discovery, drug target identification, elucidation of protein functions, and drug repositioning [1,2].

Since experimentally identifying compound-protein interactions (CPIs) is generally expensive and time-consuming [3,4] and has undesirable coverage and throughput [5], various *in silico* approaches have been developed to speed up the experimental process, and meanwhile to cut down the cost.

Up to now, most computational approaches for prediction CPIs are based on the structures of compounds and proteins and/or the interactions among them. For example, Cheng et al. [6] developed multitarget-quantitative structure-activity relationships (QSAR) and chemogenomic methods for CPI prediction, in which substructure patterns and sequence descriptors are calculated for each molecule and protein respectively. Li et al. [7] proposed to predict CPIs by molecular docking, which docks a molecule into the possible binding sites of proteins and ranks the interactions by calculating their interaction energy. Liu et al. [8] detected potential CPIs using pharmacophore mapping approach. In this approach, a given molecule is mapped onto each pharmacophore model (a set of spatial arrangement features essential for a molecule to interact with proteins) of proteins, and the fit value between the molecule and the pharmacophore is calculated. The top ranked hits are then considered CPI candidates. Cobanoglu et al. [9] presented a probabilistic matrix factorization (PMF) method to predict drug-target interactions, where the connectivity matrix of the bipartite graph is decomposed into two matrices of latent variables. Cheng et al. [10,2] developed a network-based inference method that uses the known CPI bipartite network topology similarity to predict novel CPIs, which employs a mass diffusion-like process across the CPI network.

In addition, machine-learning based methods, which have been successfully applied to various prediction problems in biology [11,12], have the potential to effectively learn the relationships among compounds and target proteins to predict new drug-target

* Corresponding author.
E-mail addresses: ktian14@fudan.edu.cn (K. Tian), shaomy@fudan.edu.cn (M. Shao), yang1995t@163.com (Y. Wang), jhguan@tongji.edu.cn (J. Guan), sgzhou@fudan.edu.cn (S. Zhou).
[1] These two authors contributed equally to this work.

http://dx.doi.org/10.1016/j.ymeth.2016.06.024
1046-2023/

interactions [5] from the viewpoint of chemogenomics [13]. Actually, a number of machine learning methods have been proposed to predict CPIs. For instance, in Wang et al.'s [14] work, the substructure descriptors of ligands and sequence descriptors of proteins are extracted and concatenated to form an ligand-protein interaction (LPI) vector and support vector machine (SVM) is used to predict LPIs. Cheng et al. [6] adopted feature selection techniques to reduce the high dimensionality of the chemogenomics space before training SVM and achieved a high AUC (the area under the receiver operating characteristics), while Tabei et al. [15] enhanced the prediction performance of linear SVM by applying an improved minwise hashing algorithm to construct new compact fingerprints for compound-protein pairs. Kim et al. [16] applied both SVM and logistic regression to CPI prediction and found that drug-drug interaction is a promising feature for drug target interaction prediction. Yu et al. [17] employed random forests (RF) by integrating chemical, genomic, and pharmacological information to predict CPIs, and obtained comparable performance to SVM with approximately half time cost.

Though SVM and logistic regression generally do not perform bad, they cannot capture nonlinear relationships among features, which prevents them from performing perfectly [18,19]. Furthermore, the imbalanced nature of data ubiquitously existing in many bioinformatics problems also degrades the performance of many existing predictors, such as RFs [20]. In the era of big biological data, more effective models are urgently needed to do better predictions with the rapidly amassing data.

Recently, deep learning (DL) techniques have been proved advantageous over traditional state-of-the-art machine learning methods in some applications [21]. In bioinformatics, deep learning has also successfully applied to several problems. For instance, Spencer et al. [22] applied a deep network to *ab initio* protein secondary structure prediction where the position-specific scoring matrix (PSSM), the amino acid residues (RES), and the Atchley factors (FAC) were used as features. Lena et al. [23] introduced a deep spatio-temporal architecture that consists of multidimensional stack of learning modules for contact prediction. Leung et al. [24] developed a deep neural network (DNN) model that can jointly predict the splicing patterns in individual tissues and the differences in splicing patterns across tissues. Moreover, Fakoor et al. [25] applied unsupervised feature learning and deep learning methods to handle cancer diagnosis problems by training a more generalized version of cancer classifier. Chicco et al. [26] proposed a deep AutoEncoder model that achieves better performance than other standard machine learning methods on gene annotation prediction. Wang et al. [27] modeled drug target interaction (DTI) relationships with a two-layer graphical model that is known as restricted Boltzmann machine (RBM). They constructed RBMs for all targets with the same parameters. Unterthiner et al. [28] proposed to use a deep neural network with multiple output units to predict DTIs, they formulated DTI prediction as a multi-task learning problem. Recently, Hamanaka et al. [29] trained a deep belief network to predict compound protein interactions (CPIS) and achieved better performance than SVM. However, training a deep belief network is very time consuming as it is a generative model that is trained by layer-wise pre-training of RBMs.

Among the different DL techniques, deep neural network (DNN) is a feedforward, artificial neural network with multiple hidden layers between inputs and outputs. It can automatically learn complex functions that map inputs to outputs, without hand-crafted features or rules [30,31]. Using techniques such as dropout and momentum training to speed up the training procedure, DNN is shown to be potentially suitable for big data including "omics" datasets [24,18].

In this work, we aim at improving the performance of CPI prediction by deep learning. Concretely, we propose a method DL-CPI (*D*eep Learning for *C*ompound-*P*rotein *I*nteractions prediction), to predict new CPIs by constructing a deep neural network (DNN) model and extracting chemical and protein features from compound-protein pairs as input. By appropriately optimizing the hyperparameters of the model, experimental results show that the DL-CPI method outperforms six existing prediction models on both balanced and imbalanced data. The good performance of our method validates the applicability of the DNN model to the CPI prediction problem.

2. Materials and methods

2.1. The DL-CPI pipeline

Fig. 1 shows the pipeline of our DL-CPI method. In this study, we propose a deep learning approach for predicting compound-protein interactions (DL-CPI, Deep Learning for Compound-Protein Interactions). We first retrieve CPIs from public databases as positive samples and generate negative samples by randomly pairing compounds and proteins and keeping those not appearing in the positive set. Then, we extract the chemical fingerprint of each compound and the domain features for each protein from public databases, respectively. For each example (CPI or compound-protein pair), we concatenate the features of the corresponding compound and protein as the feature vector of the example. Next, we input the feature vectors of both positive and negative examples to the DNN model. After hyperparameter adjustment, we train the DNN model and get the DNN predictor. Finally, we evaluate the prediction performance of the DNN predictor using a set of performance metrics and compare our method with existing prediction approaches. In what follows, we describe the major steps of the pipeline in detail.

2.2. Datasets

2.2.1. Compound-protein interactions

We retrieved CPIs of human from the STITCH database (Version 4.0) [32], a comprehensive resource for both known and predicted interactions of compounds and proteins as positive examples. Eventually, we obtained 612,214 interactions between 51,444 unique proteins and 258,936 unique compounds in total.

2.2.2. Compound data

For each compound, we used its basic substructures as features, and constructed a fingerprint (a binary vector where "1" indicates the presence of a certain feature) of features to represent the compound. The fingerprints were obtained from the PubChem database [33], and each compound is represented as a 881-dimension binary vector.

2.2.3. Protein data

We extracted 5523 domains from the Pfam database [34], and represented each protein as a 5523-dimension vector with binary elements (1 or 0). For each element in the domain feature vector, a value of "1" denotes the presence and "0" denotes the absence of the domain, respectively.

2.2.4. Negative samples

The negative samples were generated by random pairing. We first generated 612,214 negative samples (the same number of positive examples). After removing the negative examples with too few domain features, we got 606,469 negative samples in all. We then built both balanced and imbalanced datasets using randomly paired negative samples. We randomly chose positive examples from the total 612,214 positive examples. The number