Journal of Biomedical Informatics xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Screening drug target proteins based on sequence information

Jiao T. Wang [a,b,*], Wei Liu [a], Hailin Tang [a,c], Hongwei Xie [a]

[a] College of Mechatronic Engineering and Automation, National University of Defense Technology, Changsha, China
[b] Statistics Department, Harvard University, Cambridge 20138, USA
[c] Second Military Medical University, Shanghai, China

ARTICLE INFO

Article history:
Received 24 October 2013
Accepted 14 March 2014
Available online xxxx

Keywords:
Machine learning
Drug target
SVM

ABSTRACT

Identifying new drug target (DT) proteins is important in pharmaceutical and biomedical research. General machine learning method (GMLM) classifiers perform fairly well at prediction if the training dataset is well prepared. However, a common problem in preparing the training dataset is the lack of a negative dataset. To address this problem, we proposed two methods that can help GMLM better select the negative training dataset from the test dataset. The prediction accuracy was improved with the training dataset from the proposed strategies. The classifier identified 1797 and 227 potential DT proteins, some of which were mentioned in previous research, which added correlative weight to the new method. Practically, these two sets of potential DT proteins or their homologues are worth considering.

1. Introduction

At the molecular level, the main targets for drugs are proteins (mainly enzymes, receptors and transport proteins) and nucleic acids (DNA and RNA). In recent years, novel drug identification research has been widely conducted. For example, Hughes et al. [1] summarized the key preclinical stages of the drug discovery process and demonstrated that data mining of available biomedical data has led to a significant increase in target identification. Yang et al. [2] introduced tentative strategies of integrating different sources for target discovery. Imming et al. [3] considered the natural properties of DT proteins and conducted a comprehensive analysis of the DT proteins to estimate the number of known targets. Many papers have been published to address the screening drug target protein based on sequence information (see, e.g., [4–8]).

Because of the uncertainty of the drug action mechanism for each specific drug-protein target, there is no wide consensus on the optimal computational identification method. Several prediction methods have recently been developed. They performed fairly well given their specific biological hypothesis (i.e., based on side-effect similarity [9], based on ligands [10], or based on chemical structure and genomic sequence information [11]) and were aimed at integrating as much information as possible. Current knowledge is limited about what makes a protein a drug target, and there are also limits on the reliability of the biological assumptions. The latter can restrict the corresponding biological hypothesis. Therefore, it is important to develop a method that does not depend on the DT protein properties and can balance the training dataset to avoid the overfit of the prediction results. To address this problem, Li and Lai [12] developed a drug target prediction method based solely on protein sequence information without the knowledge of family/domain annotation or the protein 3D structure.

Once the training dataset is prepared, traditional machine learning strategies perform fairly well to predict new DT proteins. However, the process inherently suffers from an overfit problem [13]. This is because the training datasets have two classes. One is called the positive dataset (proteins that are known as DT proteins), and the other is called the negative dataset (proteins that are not DT proteins). The accuracy and completeness of the predictions are limited by our inability to be certain that proteins in the negative dataset are not DT proteins. To provide negative training datasets to the classifier for training, it is necessary to use some of the test data (proteins that will be predicted) for the negative dataset. The commonly used strategy is to randomly extract the negative training dataset from the putative non-DT dataset [12,14].

In this study, two strategies for selecting the negative dataset for a classifier are presented. The diagram of our workflow is shown in Fig. 1. The first strategy is aimed at increasing prediction accuracy in cross-validation, and the second strategy is aimed at filtering out as many non-DT proteins as possible. The proposed training dataset helps enhance the classifier's cross-validation accuracy. Based on the combination of SVM and different kernel functions, the classifier was trained by two datasets. The SVM with

* Corresponding author. Address: 1 Oxford Street, Cambridge, MA 02138, USA.
E-mail addresses: wangtengjiao619@gmail.com, tengjiaowang@fas.harvard.edu (J.T. Wang).

http://dx.doi.org/10.1016/j.jbi.2014.03.009
1532-0464/

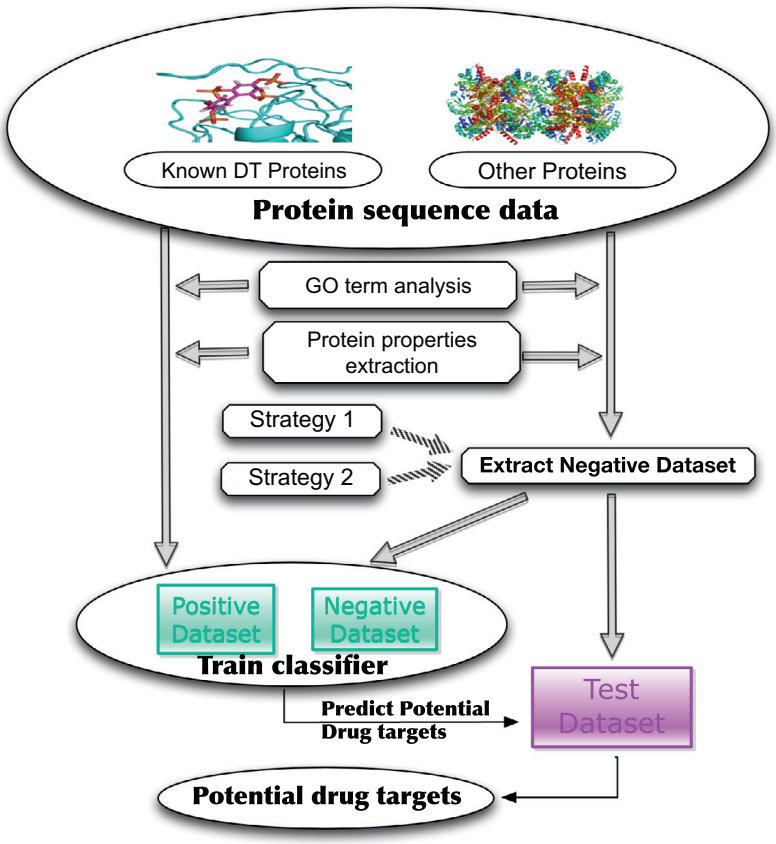

Fig. 1. Diagram of screening potential drug target proteins.

a radial basis function kernel classifier predicted 1797 and 227 DT proteins, respectively.

As demonstrated by a series of recent publications [4,5,15–17] and summarized in a comprehensive review [18], to develop a really useful predictor for a protein or peptide system, one needs to go through the following five steps: (a) select or construct a valid benchmark dataset to train and test the predictor; (b) represent the samples with an effective formulation that can truly reflect their intrinsic correlation with the target to be predicted; (c) introduce or develop a powerful algorithm to conduct the prediction; (d) properly perform cross-validation tests to objectively evaluate the anticipated prediction accuracy; (e) establish a user-friendly web-server for the predictor that is accessible to the public. Below, let us elaborate how to deal with these five steps.

2. Materials and methods

2.1. Data collection

The DT protein information was extracted from the DrugBank database Version 3.0 [13], in which the approved DT proteins set contains approximately 1273 proteins. Moreover, 262 DT proteins were used as approved carriers, 819 DT proteins as approved transporters and 984 DT proteins as approved enzymes. All of the sequence data were extracted from the UniProtKB/SwissProt data file (October, 2010). The non-human proteins were removed from both the DT and non-DT protein datasets. The known DT proteins that we collected from the DrugBank database were removed from the non-DT protein dataset. We applied the redundancy method [14] to obtain our DT and non-DT protein datasets. Then, PISCES [19] was used to remove the sequences with identity larger than 20% for both the DT and the non-DT sequences. We obtained 517 DT proteins for a positive dataset from the 1604 known DT proteins and 3834 proteins as a test dataset for later use by the machine learning algorithm.

2.2. Gene ontology analysis

Since the final goal of drug target prediction is drug development, the classification of drug target should base on the drug property. Therefore, the classification should be standard classification. If the target is membrane or exterior, antibody can be used as drug. Otherwise, the drug must be small molecule, since the antibody cannot penetrate the membrane. If the target is membrane, the required affinity could be weak rather than the drug for cytoplasm or nucleus considering the distribution of drug molecule. If the target is GPCR, the drug should be mono amine and the drug could be agonist and inverse agonist. If the target is enzyme, the drug should be inhibitor. Enhancing the enzyme activity is difficult. There have been many articles and reviews about the drug target [20,21]. The target proteins are classified into membrane, cytoplasm, exterior, nucleus, or, GPCR, nuclear receptor, channel, enzyme, transcription factor, etc.

The DrugBank database categorized DT proteins into three groups: Enzyme, Carrier and Transporter. To determine whether we can consider these three groups of proteins as a whole set of DT proteins for the latter training data of SVM, we analyzed their Gene Ontology (GO) [22] term. GO provides standardized terms to describe the biological properties of gene products. The ontology covers three domains: cellular components (the parts of a cell or its extracellular environment); molecular function (the elemental activities of a gene product at the molecular level, such as binding or catalysis); and biological process (operations or sets of molecular events with a defined beginning and end relevant to the

Please cite this article in press as: Wang JT et al. Screening drug target proteins based on sequence information. J Biomed Inform (2014), http://dx.doi.org/10.1016/j.jbi.2014.03.009