Biochimica et Biophysica Acta 1844 (2014) 207–213

Contents lists available at ScienceDirect

Biochimica et Biophysica Acta

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

Prediction of drug target groups based on chemical–chemical similarities and chemical–chemical/protein connections ☆

Lei Chen [a,1], Jing Lu [b,1], Xiaomin Luo [b], Kai-Yan Feng [c,*]

[a] College of Information Engineering, Shanghai Maritime University, Shanghai 201306, People's Republic of China
[b] Drug Discovery and Design Center (DDDC), Shanghai Institute of Materia Medica, Shanghai 201203, People's Republic of China
[c] BGI-Shenzhen, Beishan Industrial Zone, Yantian District, Shenzhen 518083, People's Republic of China

ARTICLE INFO

Article history:
Received 29 November 2012
Received in revised form 20 May 2013
Accepted 22 May 2013
Available online 1 June 2013

Keywords:
Drug-target interaction network
Chemical–chemical similarity
Chemical–chemical connection
Chemical–protein connection
Jackknife test

ABSTRACT

Drug-target interaction is a key research topic in drug discovery since correct identification of target proteins of drug candidates can help screen out those with unacceptable toxicities, thereby saving expense. In this study, we developed a novel computational approach to predict drug target groups that may reduce the number of candidate target proteins associated with a query drug. A benchmark dataset, consisting of 3028 drugs assigned within nine categories, was constructed by collecting data from KEGG. The nine categories are (1) G protein-coupled receptors, (2) cytokine receptors, (3) nuclear receptors, (4) ion channels, (5) transporters, (6) enzymes, (7) protein kinases, (8) cellular antigens and (9) pathogens. The proposed method combines the data gleaned from chemical–chemical similarities, chemical–chemical connections and chemical–protein connections to allocate drugs to each of the nine target groups. A jackknife test applied to the training dataset that was constructed from the benchmark dataset, provided an overall correct prediction rate of 87.45%, as compared to 87.79% for the test dataset that was constructed by randomly selecting 10% of samples from the benchmark dataset. These prediction rates are much higher than the 11.11% achieved by random guesswork. These promising results suggest that the proposed method can become a useful tool in identifying drug target groups. This article is part of a Special Issue entitled: Computational Proteomics, Systems Biology & Clinical Implications. Guest Editor: Yudong Cai.

1. Introduction

Drug-target interaction is an important consideration in the drug discovery pipeline [1]. It is very well known that an essential problem in drug discovery and deployment is unacceptable toxicity, an issue that may cause the withdrawal of drugs even after they are brought into the market, thus threatening pharmaceutical companies and their consumers. Therefore, it is necessary to develop an effective method to identify the sensitivity and toxicity of drug candidates before they are synthesized and launched [2]. The toxicity of a drug candidate is exerted by interaction with target proteins in human tissues. Therefore, identification of drug target proteins is beneficial for the analysis of drug toxicity and other related problems. However, it is expensive and time-consuming to identify drug target proteins experimentally. A computational approach is an alternative way to tackle this problem as it is very efficient and may provide some additional information.

Many efforts have been devoted to identify drug target proteins in the past few years. Zhu et al. used literature text mining to search for co-occurrences of drugs and genes [3]. Cheng et al. identified drug-target interactions based on docking simulation [4,5]. However, these two methods have their own limitations: literature text mining is prone to redundancy due to multiple gene and compound names, and docking is only available for the proteins with known 3D structures. Campillos et al. used phenotypic side-effect similarity to predict drug-target interactions, an approach suitable for marketed drugs [6]. Prado-Prado et al. developed some multi-target QSAR models to carry out drug-target prediction [7,8]. 3D structural parameters for targets and 3D molecular descriptors for drugs were used as input for an artificial neural network. Cheng et al. proposed a network-based inference method that used drug-target bipartite network topology similarity to suggest new targets for known drugs [9]. Chen et al. developed a simple but novel computational approach combining chemical structure information and protein functional domain information for identification of drug-target interactions [10].

In recent years, some previous studies showed that connected compounds may also have similar biological functions [11–13]. Since target proteins can be regarded as the properties of drug

☆ This article is part of a Special Issue entitled: Computational Proteomics, Systems Biology & Clinical Implications. Guest Editor: Yudong Cai.

* Corresponding author. Tel.: +86 13332956293.
E-mail addresses: chen_lei1@163.com (L. Chen), lujing@mail.shcnc.ac.cn (J. Lu), xmluo@mail.shcnc.ac.cn (X. Luo), kaiyan.feng@gmail.com, fengkaiyan@genomics.org.cn (K.-Y. Feng).

[1] These authors contributed equally to this work.

1570-9639/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.bbapap.2013.05.021

compounds, in this paper we attempt to use the information gleaned from chemical–chemical similarities, chemical–chemical connections and chemical–protein connections to identify drug target proteins. The sheer numbers of candidate pairs of drug and target proteins are too large to be handled easily. Further, the search space is very wide if one directly predicts drug target proteins. Thus, it is necessary to reduce the number of candidate target proteins for each query drug. According to the data of KEGG (Kyoto Encyclopedia of Genes and Genomes, http://www.genome.jp/kegg/) [14], the target proteins can be divided into the following ten groups: (1) G protein-coupled receptors, (2) cytokine receptors, (3) nuclear receptors, (4) ion channels, (5) transporters, (6) enzymes, (7) protein kinases, (8) cellular antigens, (9) cytokines and (10) pathogens. If a computational method can correctly predict the target groups of a query drug, the number of its possible target proteins can be reduced rendering the result useful for some further analyses.

During the past twenty years, some online compound databases, such as KEGG [14], STITCH (Search Tool for Interactions of Chemicals) [15] and ChEBI (Chemical Entities of Biological Interest) [16] have been established from which users can easily retrieve compound information and properties, providing an opportunity to study some related problems in great detail [10–13,17–19]. Among these databases, KEGG contains an important component, "KEGG LIGAND", providing the chemical substances and reactions while STITCH provides the connection information for chemicals and proteins and ChEBI provides the ontology information for chemicals, Here, we propose a computational method integrating compound information from KEGG (chemical–chemical similarities) and STITCH (chemical–chemical connections and chemical–protein connections) to predict drug target groups. To evaluate the method, a benchmark dataset consisting of 3028 drug compounds was constructed through KEGG and was separated into one training dataset and one test dataset. It was observed, utilizing the jackknife test on the training dataset, that the overall correct rate achieved 87.45%, while the overall correct rate on the test dataset was 87.79%. The high correct prediction rates indicate that the method is likely to facilitate the discovery of new drugs and the screening of drug candidates with unacceptable toxicities.

2. Materials and methods

2.1. Benchmark dataset

The information for 3610 drugs was retrieved from KEGG DRUG (http://www.genome.jp/kegg/drug/). According to the website http://www.genome.jp/kegg-bin/get_htext?br08310.keg, the target proteins of these drugs are classified into ten groups: (1) G protein-coupled receptors, (2) cytokine receptors, (3) nuclear receptors, (4) ion channels, (5) transporters, (6) enzymes, (7) protein kinases, (8) cellular antigens, (9) cytokines and (10) pathogens. Accordingly, these 3610 drugs belong to ten categories based on their target proteins, *i.e.*, drugs belong to one category if their target proteins are in the same group of proteins. However, some drugs' target proteins belong to more than one target group, *i.e.*, these drugs belong to two or more categories. After excluding these drugs, 3537 samples remained. Furthermore, drugs without any information concerning chemical–chemical similarities, chemical–chemical connections and chemical–protein connections were also excluded, resulting in 3030 drugs. At completion of the process, we found that target group "cytokines" only contained two drugs, a number not sufficient for a prediction model and that group was abandoned. Thus, only nine target groups, tagged T_1, T_2, …, T_9 (see Table 1 for the correspondence between target groups and tags) with 3028 samples were investigated, *i.e.*, these samples composing the benchmark dataset S were classified into nine categories, which can be formulated as follows:

$$S = S_1 \cup S_2 \cup \ldots \cup S_9 \quad (1)$$

where S_i contained drugs with tag T_i ($i = 1, 2, \ldots, 9$). The detailed codes of drugs are provided in Online Supporting Information S1.

To sufficiently evaluate the proposed method, we randomly selected 10% (303) of samples in the benchmark dataset S to compose a test dataset S_{te}. The rest 2725 samples in S were used to construct a training dataset S_{tr}. Shown in Table 1 is the number of drugs with different tags in the training and test datasets.

Table 1
Distribution of drugs with different tags in the training and test datasets.

Tag	Target group	Number of drug compounds		
		Training dataset	Test dataset	Total
T_1	G protein-coupled receptors	968	108	1076
T_2	Cytokine receptors	28	4	32
T_3	Nuclear receptors	284	33	317
T_4	Ion channels	343	34	377
T_5	Transporters	33	4	37
T_6	Enzymes	546	54	600
T_7	Protein kinases	26	2	28
T_8	Cellular antigens	7	1	8
T_9	Pathogens	490	63	553
–	Total	2725	303	3028

2.2. Chemical–chemical similarities

It is known that compounds with similar physicochemical properties often share similar biological activities [20]. Thus, using the chemical–chemical similarities data to predict drug target groups may be feasible. Previous studies showed success in identification of drug-target interactions using topology similarity [9,21]. Thus, here, we selected graphical representation, proposed by Hattori et al. [22], to measure the similarity of two drug compounds, a process deemed more effective in reflecting compound structure than other representation methods, such as SMILES [23]. Such representation has been used to investigate different attributes of chemicals and related problems [10,17,18,21]. For two chemicals d_1 and d_2, their 2D (two-dimensional) graphical representations can be readily obtained in which nodes represent atoms and edges represent bonds between corresponding atoms. Subsequently, a maximum common subgraph of these two graphs can be devised to estimate the similarity of d_1 and d_2, denoted by $I_s(d_1,d_2)$, *via* the Jaccard coefficient [24]. For the detailed description of the method, please refer to Hattori et al.'s paper [22]. The similarity scores of the existing compounds are stored in KEGG, which can be obtained *via* the website http://www.genome.jp/ligand-bin/search_compound. In particular, those without similarity scores in the website were set to be zero in this study.

2.3. Chemical–chemical/protein connections

Recently, protein–protein connection data has been used to predict some attributes of proteins [25–27], implying that connected proteins are more likely to share common biological functions. Likewise, connected compounds may also have similar properties, and some previous works have shown that this is true [11–13]. The target proteins can be regarded as properties of drug compounds, an assumption that may also fit with the rule.

Data concerning chemical–chemical connections were downloaded from a well-known database, STITCH (http://stitch.embl.de/) [15]. Each connection in the obtained file was labeled with a confidence score to measure the likelihood that the connection occurs. For two drug compounds d_1 and d_2, the confidence score of their connection was denoted