Journal of Biomedical Informatics 54 (2015) 132–140

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Characterizing and optimizing human anticancer drug targets based on topological properties in the context of biological pathways

Jian Zhang [a,1], Yan Wang [b,1], Desi Shang [c,1], Fulong Yu [c], Wei Liu [d], Yan Zhang [a], Chenchen Feng [a], Qiuyu Wang [e], Yanjun Xu [c], Yuejuan Liu [a], Xuefeng Bai [a], Xuecang Li [a], Chunquan Li [a,*]

[a] School of Medical Informatics, Daqing Campus, Harbin Medical University, Daqing 163319, PR China
[b] Majorbio Bio-Pharm Technology Co., Ltd., Shanghai 201203, PR China
[c] College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150081, PR China
[d] Department of Mathematics, Heilongjiang Institute of Technology, Harbin 150050, PR China
[e] School of Nursing, Daqing Campus, Harbin Medical University, Daqing 163319, PR China

ARTICLE INFO

Article history:
Received 20 September 2014
Accepted 17 February 2015
Available online 24 February 2015

Keywords:
Optimization
Drug targets
Topology
Pathways

ABSTRACT

One of the challenging problems in drug discovery is to identify the novel targets for drugs. Most of the traditional methods for drug targets optimization focused on identifying the particular families of "druggable targets", but ignored their topological properties based on the biological pathways. In this study, we characterized the topological properties of human anticancer drug targets (ADTs) in the context of biological pathways. We found that the ADTs tended to present the following seven topological properties: influence the number of the pathways related to cancer, be localized at the start or end of the pathways, interact with cancer related genes, exhibit higher connectivity, vulnerability, betweenness, and closeness than other genes. We first ranked ADTs based on their topological property values respectively, then fused them into one global-rank using the joint cumulative distribution of an N-dimensional order statistic to optimize human ADTs. We applied the optimization method to 13 anticancer drugs, respectively. Results demonstrated that over 70% of known ADTs were ranked in the top 20%. Furthermore, the performance for mercaptopurine was significant: 6 known targets (ADSL, GMPR2, GMPR, HPRT1, AMPD3, AMPD2) were ranked in the top 15 and other four out of the top 15 (MAT2A, CDKN1A, AREG, JUN) have the potentialities to become new targets for cancer therapy.

1. Introduction

The identification of novel drug targets is a major challenge in medicine and biology. Traditionally, most studies have focused on the identification of particular families of "druggable targets" and achieved significant success [1,2]. However, genes rarely function in isolation in a complex biological system, especially when a patient undergoes drug treatment (such as the anticancer therapy). A growing body of evidence indicate that drug design should focus on all the drug-affected genes simultaneously from the genome-wide perspective [3]. Recently, with the development of high-throughput biological experimental technologies, a large collection of gene expression profiles with drug treatments is available, such as the Connectivity Map (CMap) [4]. Many expression-based approaches have emerged for further understanding of drug mechanisms in the whole genome [5–7]. These methods have primarily focused on single drug target without considering the interactions among them. Studies have demonstrated that the drugs affect not only their intended targets but also other genes that interact with them or trigger the downstream molecular events [8]. The increasing use of protein–protein interactions (PPIs) allowed for network-based approach to predict novel drug targets [9–11]. These studies showed that some topological properties, including betweenness, closeness and connectivity, within protein–protein network could distinguish known ADTs from other genes significantly. Thus, the topological properties within protein–protein network can be used to better assess the potentiality of a node as a novel drug target [9–11].

Biological pathways that belong to more accurate network-based data have superior properties that naturally suited to discover novel drug targets [12,13]. Firstly, the pathways were reliable because they were curated manually from scientific

* Corresponding author.
E-mail address: lcqbio@163.com (C. Li).
[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.jbi.2015.02.007
1532-0464/

literature, such as the KEGG PATHWAY Database [14]. Secondly, it was convenient to consider a pathway in isolation as a particular function module [15,16]. Thus, it was feasible to map the drug targets into certain pathways to elucidate the mechanisms of action of the drug. Finally, the biological pathways were directional. So, it was practical to assess the importance of the potential drug targets according to its position or interaction [16,17]. For example, the insulin receptor (INSR), which was involved in the insulin pathway and the adherens junction pathway, plays a more important role in insulin pathway due to its terminal position and high connectivity compared to its role in adherens junction pathway [18]. Taken together, the biological pathways are highly valuable and powerful for optimization of the drug targets.

In this study, we carefully analyzed the topological properties of human ADTs in the context of biological pathways and found seven topological properties that could distinguish known ADTs from other genes significantly. We proposed an optimization approach for human ADTs through integrating all seven pathway-level topological properties to rank the candidate ADTs. Our pathway-based method can help solve some limitations of PPI-based methods, including (1) the lack of consideration regarding the functional module in which several genes carry out a specific function together, (2) the difficulty in understanding the global importance of any gene in one module or across all of these various functional modules, (3) the directionless assessment of information transfer across the genes in functional module. We applied this method to 13 anticancer drugs respectively, and achieved good optimization, particularly for the drugs mercaptopurine and methotrexate. In conclusion, the optimization strategy we developed, which was based on pathway-level topological properties, offers a new sight and could aid in the discovery of novel anticancer drug targets.

2. Materials and methods

2.1. Datasets

2.1.1. Known anticancer drug targets (ADTs)

The drug targets dataset was downloaded from the KEGG and DrugBank database [14,19]. We extracted all the anticancer drugs and their targets according to the disease and the Anatomical Therapeutic Chemical (ATC) information. Finally, we obtained 573 ''anticancer drug–target'' relationships and 155 targets.

2.1.2. Four types of special neighbor genes

To test whether the known ADTs interacted with the cancer related genes, cancer genes (CGs), cancer hallmark genes (CHMGs), known anticancer drug target genes (KADTGs) and genes encoding the nuclear membrane proteins (NMPGs) were used as four types of special neighbor genes, which were important for cancer initiation, progression and therapy in the following analysis.

766 CGs were obtained from the report by Li et al. [20]. CHMGs in the study were defined as genes functionally involved in the six cancer hallmarks, including self-sufficiency in growth signals, insensitivity to antigrowth signals, evasion of apoptosis, limitless replicative potential, sustained angiogenesis, tissue invasion and metastasis [21]. We obtained 1320 genes as CHMGs, which were functionally annotated with ''DNA repair'', ''cell growth'', ''cell proliferation'', ''angiogenesis'', ''cell migration'', and ''locomotion'' from Gene Ontology (GO) [22]. KADTGs mean genes known as anticancer drug targets, we obtained 155 KADTGs from KEGG and DrugBank. NMPGs mean genes that encode the nuclear membrane proteins. In particular, we selected 149 NMPGs with a nuclear membrane subcellular location from the Uniprot Knowlegebase (UniprotKB) [23].

2.1.3. Biological pathways

Information regarding the biological pathways was obtained from the KEGG PATHWAY database [14], including metabolic and non-metabolic pathways. We used the R-based software package SubpathwayMiner to reconstruct all pathways graphically [15]. This type of reconstruction retains the raw information of the pathways, particularly the structures, and provides detailed and reliable information for analyzing the ADT topological properties based on these biological pathways.

2.1.4. Anticancer drug gene expression data

In order to optimize the targets for anticancer drugs from the whole genome, we utilized transcriptional data for cultured human cancer cells treated with anticancer drugs obtained from CMap (http://www.broadinstitute.org/cmap/). The library contains 6100 instances of 4 cancer cell lines treated with 1309 distinct small molecules [4]. We downloaded all the instances, gene expression profiles and their associated annotation file ''cmap_instances_02.xls'' from the CMap website. According to the annotation information, 42 among the 1309 bioactive small molecules were anticancer drugs. The instance information that these drugs corresponded to were used to extract the anticancer drug gene expression data.

2.1.5. Gene expression profiles of cancers with the survival time information

To validate that the candidate targets with prior rank have closely correlation with cancer patients' survival time, we downloaded 9 gene expression profiles of cancers with the survival time information from GEO database (http://www.ncbi.nlm.nih.gov/geo/): 3 lung adenocarcinoma profiles (GSE13213, GSE3141, GSE8894), 3 breast cancer profiles (GSE2990, GSE4922, GSE1456), and 3 colon cancer profiles (GSE12945, GSE17536, GSE14333).

2.2. Methods

2.2.1. Identifying candidate targets for anticancer drugs

We identified the candidate targets for each anticancer drug according to its expression profiles from the CMap database [24]. For each instance of drug, we matched perturbation and control pairs of expression profiles according to descriptions of the instances in the file ''cmap_instances_02.xls''. Then we used fold-change analysis to identify differentially expressed genes (DEGs) for each instance with $|\log_2\text{fold-change}|>\log_2 1.5$ (gene expression up-regulated or down-regulated 1.5 folds) between the corresponding treatments and control gene expression profiles. The DEGs were merged if the corresponding instances belonged to the same drug, these genes were considered to be anticancer drug affected genes. As a result, of 42 anticancer drugs in CMap, we obtained the corresponding 13,082 DEGs which were significantly affected by at least two anticancer drugs. An anticancer drug can correspond to 1345 genes on average. They were considered as the candidate targets for anticancer drug and were used to identify the cancer related crucial pathways.

2.2.2. Identifying cancer related crucial pathways (CRCPs)

The pathways that satisfied the following three rules were considered to be CRCPs: associated with cancer initiation and progression; important for cancer therapy; and prone to be affected by the anticancer drugs. Subsequently, we performed pathway annotated analysis for three gene sets: cancer gene set (766 genes), known ADT gene set (155 genes), and anticancer drug affected gene set (13082 genes). The concurrently annotated pathways for these three gene sets were regarded as the CRCPs and used for analyzing the topological properties of the known ADTs.