Journal of Biomedical Informatics 59 (2016) 248–257

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Inferring new drug indications using the complementarity between clinical disease signatures and drug effects

Dongjin Jang [a,b,1], Sejoon Lee [b,1], Jaehyun Lee [a,b], Kiseong Kim [a,b], Doheon Lee [a,b,*]

[a] Department of Bio and Brain Engineering, KAIST, 291 Daehak-ro, Yuseong-gu, Daejeon 305-701, Republic of Korea
[b] Bio-Synergy Research Center, 291 Daehak-ro, Yuseong-gu, Daejeon 305-701, Republic of Korea

ARTICLE INFO

Article history:
Received 5 June 2015
Revised 31 October 2015
Accepted 9 December 2015
Available online 17 December 2015

Keywords:
Drug repositioning
Electronic clinical information
Clinical disease signatures
Clinical drug effects

ABSTRACT

Background: Drug repositioning is the process of finding new indications for existing drugs. Its importance has been dramatically increasing recently due to the enormous increase in new drug discovery cost. However, most of the previous molecular-centered drug repositioning work is not able to reflect the end-point physiological activities of drugs because of the inherent complexity of human physiological systems.

Methods: Here, we suggest a novel computational framework to make inferences for alternative indications of marketed drugs by using electronic clinical information which reflects the end-point physiological results of drug's effects on the biological activities of humans. In this work, we use the concept of complementarity between clinical disease signatures and clinical drug effects. With this framework, we establish disease-related clinical variable vectors (clinical disease signature vectors) and drug-related clinical variable vectors (clinical drug effect vectors) by applying two methodologies (i.e., statistical analysis and literature mining). Finally, we assign a repositioning possibility score to each disease–drug pair by the calculation of complementarity (anti-correlation) and association between clinical states ("up" or "down") of disease signatures and clinical effects ("up", "down" or "association") of drugs. A total of 717 clinical variables in the electronic clinical dataset (NHANES), are considered in this study.

Results: The statistical significance of our prediction results is supported through two benchmark datasets (Comparative Toxicogenomics Database and Clinical Trials). We discovered not only lots of known relationships between diseases and drugs, but also many hidden disease–drug relationships. For example, glutathione and edetic-acid may be investigated as candidate drugs for asthma treatment. We examined prediction results by using statistical experiments (enrichment verification, hyper-geometric and permutation test $P < 0.009$ in Comparative Toxicogenomics Database and Clinical Trials) and presented evidences for those with already published literature.

Conclusion: The results show that electronic clinical information is a feasible data resource and utilizing the complementarity (anti-correlated relationships) between clinical signatures of disease and clinical effects of drugs is a potentially predictive concept in drug repositioning research. It makes the proposed approach useful to identity novel relationships between diseases and drugs that have a high probability of being biologically valid.

1. Introduction

Drug repositioning is defined as the process of finding new uses outside the scope of the original medical indications for existing drugs. Its importance has been dramatically increasing recently due to the enormous increase in new drug discovery cost [1,2]. Because 90% of new drug candidates fail in early tests of safety and efficacy in *de novo* drug discovery, many researchers are applying repositioning strategies to discover novel therapeutics of known drugs.

Drug and target protein characteristics based on chemical structures and the properties of ligands and receptors have been used to identify new targets for existing drugs. Keiser et al. make use of chemical similarities between drugs [3]. The Keiser hypothesis was that structurally similar chemicals tend to have similar properties. In other words, similar molecules exhibit similar biological

* Corresponding author at: Department of Bio and Brain Engineering, KAIST, 291 Daehak-ro, Yuseong-gu, Daejeon 305-701, Republic of Korea.
E-mail addresses: djjang@kaist.ac.kr (D. Jang), sejuning@kaist.ac.kr (S. Lee), jaeh@kaist.ac.kr (J. Lee), iames@kaist.ac.kr (K. Kim), dhlee@kaist.ac.kr (D. Lee).
[1] These authors equally contributed to this work.

http://dx.doi.org/10.1016/j.jbi.2015.12.003
1532-0464/

activities. Chang et al., Kinnings et al., and Zahler et al. focused on physical interactions between drugs and targets, which is called docking method [4–6]. Their methods for drug repositioning assume that a chemical would be a feasible candidate for a treatment of a disease when the chemical physically binds to the target protein of interest, which has already been reported to play an important role in development or treatment of the disease. Bleakley et al., Mei et al., and van Laarhoveb et al. take into account unified resources including chemical structures, amino-acid sequences of target proteins, and chemical–protein interaction network [7–9]. Even though these approaches can be considered as systematic and comprehensive means to find new molecular targets for existing drugs, there are some drawbacks for finding new uses of known drugs. Their primary limitation is an inconsistency between the results from these methods and clinical therapeutic effects. Additionally, in physical binding simulations between drugs and its targets, three-dimensional-structure libraries of both chemical compounds and target proteins are required. Unfortunately, 3D libraries of both chemical compounds and proteins have not been completely identified so far. Although information about 3D structure can be inferred from 2D structure, this may also introduce errors that may occur in the transformation.

Gene expression information is also widely used for drug repositioning [10–12]. Several studies have tried to identify novel uses of existing drugs by analyzing patterns of gene expression-signatures of both drug-associated and disease-associated gene sets. Although those studies have suggested computational approaches to discover new drug indications by taking experimental outcomes of molecular activity into account, there are still problems that make some improvements necessary. First is the gap of chemical responses between molecular and phenotypic levels of the entire human system. For instance, in complex physiological systems, it may be hard to represent the overall molecular responses of chemicals with only gene expression data because gene expression profiles are derived from separated-cell lines treated with each chemical. Secondly, the experimental process requires enormous time and cost to retain enough expression data to use in research.

Clinical information may provide new opportunities to directly connect chemicals to clinical therapeutic effects in complex physiological systems because clinical information not only indicates the phenotypic states of disease-conditions, but also reflects the end-physiological results of chemical impacts on human biological activity [13,14].

One of the strategies for drug repositioning with clinical information is the side-effects-based approach [15,16]. In these studies, an underlying assumption was that if many side-effects are shared between drugs having different indications, then the drugs could be repositioned.

Recently, many researchers in biomedical fields have focused on another resource of the electronic clinical data [17–19]. Electronic clinical information encompasses a variety of medical histories, such as diagnoses, prescriptions, and laboratory test results, and they are accumulated when medical services are provided to patients in medical institutions. With the rapid increase in the adoption of the electronic clinical information systems, there is now plenty of clinical data, which provides a promising opportunity to investigate hidden connections between diseases and clinical variables [20]. In addition, most doctors prescribe medicines based on the clinical status or symptoms of patients. Therefore, electronic clinical information can directly help identify alternative indications of already approved drugs.

One potential source in electronic clinical information for drug repositioning is the free text in the clinical notes that has explicit information about disease–drug relationships [21–23]. The basic idea of the study is to count the number of times that a disease and a drug co-occur in the same free text and compare that count to the number of expected co-occurrences by chance. This approach can be used easily, but it may lead to many false-positives. Moreover, it does not consider other useful structured data for drug repositioning included in the electronic clinical data, such as laboratory test and medical survey results that can capture clinical disease signatures.

Here, we suggest a systematic framework (Fig. 1) to make inferences for new uses of known drugs using the structured data in the electronic clinical information. The electronic clinical information was from National Health and Nutrition Examination Survey (NHANES) which was provide by National Center for Health Statistics (NCHS) in the USA [24]. We adopted the clinical-level complementarity between clinical disease signatures and drug effects. In this framework, we established clinical variable vectors for diseases and drugs by applying two methodologies (statistical analysis on electronic clinical information and literature mining on PubMed). Then, we assigned a score to each disease–drug pair based on both complementarity and association between clinical variable vectors for diseases and drugs. We validated our prediction results by making use of two independent datasets, Comparative Toxicogenomics Database (CTD) [25] and Clinical Trials [26]. CTD is a database containing chemical-disease, chemical-gene and gene-disease interactions manually curated from literature, and Clinical Trials is a database for clinical studies of human participants in 190 countries (http://clinicaltrials.gov/).

The statistical significance of our prediction results is supported through two benchmark datasets (enrichment verification, hypergeometric, and permutation test $P < 0.009$ in CTD and Clinical Trials). Through our prediction results, we discovered that glutathione and edetic-acid may be investigated as candidate drugs for asthma treatment. These results show that using electronic clinical information and the concept of clinical-level complementarity can offer promising insights into drug repositioning research.

2. Materials and preprocessing

2.1. Data sources

We obtained electronic clinical information from the cross-sectional epidemiological database called the National Health and Nutrition Examination Survey (NHANES). It is a major program of the Center for Disease Control and Prevention (CDC), and its aim is to assess the health and nutritional status of adults and children in the USA. NHANES includes demographics, dietary habits, health-related questions, and results of laboratory tests. These datasets are available at http://www.cdc.gov/nchs/nhanes.htm.

Drug information was retrieved from the DrugBank database [27]. Of the 6811 drug entities in the database, 1578 FDA-approved drugs, and their 22,143 synonyms were included in our research.

Literature information came from the biomedical literature resources of PubMed. We downloaded and aggregated all publicly available abstracts and Medical Subject Headings (MeSH) terms of each published work, ranging from 1950 to 2011. We then retrieved a total of 11,563,353 PubMed abstracts and their MeSH information. Because MeSH terms indicate essential keyword annotations of the literature, we utilized MeSH information in this study.

The validation in this work used two independent datasets (CTD [25] and Clinical Trials database [26]). A total of 469,609 chemical-disease relationships manually curated from literature, and 197,234 studies that are currently under Clinical Trials were gathered from the CTD and Clinical Trials database, respectively.