Journal of Biomedical Informatics xxx (2013) xxx–xxx

Contents lists available at SciVerse ScienceDirect

Journal of Biomedical Informatics

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

A semi-supervised approach to extract pharmacogenomics-specific drug–gene pairs from biomedical literature for personalized medicine

Rong Xu [a,*], QuanQiu Wang [b]

[a] Medical Informatics Division, Case Western Reserve University, OH, USA
[b] ThinTek LLC, Palo Alto, CA, USA

ARTICLE INFO

Article history:
Received 16 August 2012
Accepted 1 April 2013
Available online xxxx

Keywords:
Pharmacogenomics
Text mining
Information extraction
Personalized medicine

ABSTRACT

Personalized medicine is to deliver the right drug to the right patient in the right dose. Pharmacogenomics (PGx) is to identify genetic variants that may affect drug efficacy and toxicity. The availability of a comprehensive and accurate PGx-specific drug–gene relationship knowledge base is important for personalized medicine. However, building a large-scale PGx-specific drug–gene knowledge base is a difficult task. In this study, we developed a bootstrapping, semi-supervised learning approach to iteratively extract and rank drug–gene pairs according to their relevance to drug pharmacogenomics. Starting with a single PGx-specific seed pair and 20 million MEDLINE abstracts, the extraction algorithm achieved a precision of 0.219, recall of 0.368 and *F*1 of 0.274 after two iterations, a significant improvement over the results of using non-PGx-specific seeds (precision: 0.011, recall: 0.018, and *F*1: 0.014) or co-occurrence (precision: 0.015, recall: 1.000, and *F*1: 0.030). After the extraction step, the ranking algorithm further improved the precision from 0.219 to 0.561 for top ranked pairs. By comparing to a dictionary-based approach with PGx-specific gene lexicon as input, we showed that the bootstrapping approach has better performance in terms of both precision and *F*1 (precision: 0.251 vs. 0.152, recall: 0.396 vs. 0.856 and *F*1: 0.292 vs. 0.254). By integrative analysis using a large drug adverse event database, we have shown that the extracted drug–gene pairs strongly correlate with drug adverse events. In conclusion, we developed a novel semi-supervised bootstrapping approach for effective PGx-specific drug–gene pair extraction from large number of MEDLINE articles with minimal human input.

1. Background

1.1. Pharmacogenomics and personalized medicine

Pharmacogenomics (PGx) is important for personalized medicine. Different patients respond differently to the same drug, with genetics accounting for 20–95% of the variability [1]. Pharmacogenomics is the study of how human genetic variations affect an individual's response to drugs, with foci on drug metabolism, absorption, and distribution [2]. Pharmacogenomics plays an important role in identifying drug responders and non-responders, avoiding adverse events, and optimizing drug dose [3,4]. Recently, the U.S. Food and Drug Administration (FDA) has become a strong pharmacogenomics advocate in an effort to make drugs safer and more effective [5,6]. In order to improve the quality of already-marketed drugs, the FDA has updated certain drug labels to include PGx information. Currently, over one hundred FDA-approved drugs have PGx information on their labels that describe genes responsible for drug exposure, clinical response variability, and risk for adverse events.[1] One of the well-known PGx-specific drug–gene associations is warfarin-CYP2C9. Gene CYP2C9 encodes an important cytochrome P450 (CYP) enzyme that plays a major role in the metabolizing of more than 100 therapeutic drugs, one of which is warfarin. The genetic polymorphisms of CYP2C9 are associated with altered enzyme activity leading to toxicity at normal therapeutic doses of warfarin. Understanding how the genetic variants contribute to various drug responses is an essential step of personalized medicine [1,7,8]. The success of personalized drug treatment largely depends on the availability of accurate and comprehensive knowledge bases of PGx-specific drug–gene relationships, such as warfarin-CYP2C9 and irinotecan-UGT1A.

1.2. Automatic methods in extracting PGx-specific drug–gene pairs from literature

There are substantial research efforts in constructing PGx knowledge bases using both manual and automatic approaches. The Pharmacogenomics Knowledge Base (PharmGKB) is the largest

* Corresponding author. Fax: +1 216 368 0207.
E-mail address: rxx@case.edu (R. Xu).

[1] http://www.fda.gov/Drugs/ScienceResearch/ResearchAreas/Pharmacogenetics/ucm083378.htm.

1532-0464/$ - see front matter © 2013 Published by Elsevier Inc.
http://dx.doi.org/10.1016/j.jbi.2013.04.001

manually created resource of information on how variations in human genetics lead to variations in drug response (http://www.pharmgkb.org) [9]. The PharmGKB project involves a large number of curators who read the literature and manually extract relationships among genes, drugs, and diseases from the pharmacogenetic literature. However, manually extracting PGx knowledge and other biomedical information in general from published literature and transforming it into machine-understandable knowledge is a difficult task because biomedical knowledge and terminology comprise huge, dynamic, and highly complicated fields. In addition, human curators are liable to error and subjective bias.

Development of automatic approaches to extract PGx-specific drug–gene relationships from published biomedical literature is an active research area. Both statistical and natural language processing (NLP) methods have been used [10–17]. Recently, we have developed a conditional approach to extract PGx-specific drug–gene pairs from 20 million MEDLINE abstracts using known drug–gene pairs available in PharmGKB as prior knowledge to implicitly classify sentences before relationship extraction. We have demonstrated that the conditional drug–gene relationship extraction approach significantly improves the precision and the *F*1 measure when compared with the unconditioned approach [18]. One common feature among above studies is that these drug–gene relationship extraction algorithms used either PGx-specific gene lexicons as input or PGx-related articles as the text corpus. These gene lexicons were either manually compiled or were derived from PharmGKB drug–gene pairs. PharmGKB is the largest pharmacogenomics knowledge, however the genes in this knowledge are often a mixture of non PGx-specific genes (e.g., IL2, VDR, EGFR, KRAS, ERBB2, and BRCA1) and PGx-specific genes (CYP2C9, VKORC1, ABCB1, UGT1A). Correspondingly, the drug–gene pairs in PharmGKB are also a mixture of non PGX-specific pairs. In addition, the recall of the PharmGKB gene lexicon is also limited. For example, there are total of 60 CYP (cytochrome P450) gene symbols approved by the HUGO Gene Nomenclature Committee (HGNC) (http://www.genenames.org/), but PharmGKB contains only 30 of them. Therefore, in order to increase the recall of extracted drug–gene pairs, we need to either compile a more comprehensive PGx-specific gene lexicon as done in [19], or start from all human genes and develop an algorithm to extract valid drug–gene pairs and classify them by their PGx-relevance.

1.3. Our semi-supervised iterative approach in extracting PGx-specific drug–gene pairs from literature

In this study, instead of using a precompiled PGx-specific gene lexicon, we use all human protein coding genes (total 19,055) as the underlying gene lexicon input to the drug–gene extraction algorithm. Since PGx-specific drug–gene pairs only account for a very small of portion of all drug–gene semantic pairs, using all human genes as the input gene lexicon makes the task of extracting PGx-specific drug–gene pairs more challenging and interesting. Therefore, it is critical to develop a ranking algorithm to rank extracted drug–gene pairs according to their PGx relevance. Another critical difference from our previous knowledge-driven approach [18] is that instead of using a significant portion of PharmGKB drug–gene pairs as prior knowledge, we use only one or a few known PGx-specific drug–gene pairs (e.g. warfarin-CYP2C9, or caffeine-CYP1A2) as seeds to start the whole extraction process. Our previous conditional approach was guided by many known drug–gene pairs and therefore constituted a supervised learning approach. The method we present in this study is a semi-supervised approach since it depends on only a few seeds to start the whole learning process. Our study is based on the assumption that PGx-specific drug–gene pairs are often clustered together in a sentence. If we start with a known PGx-specific pair such as warfarin-CYP2C9, it is likely that sentences containing this pair are also PGx-specific. The other drug–gene pairs extracted from these PGx-related sentences are likely PGx-specific. The likelihood increases as the relatedness of the sentences increases, which depends on the relatedness of other drug–gene pairs in it. For example, using seed pair "warfarin-CYP2C9", we retrieved the following sentence "Genetic factors (VKORC1, **CYP2C9**, EPHX1, and CYP4F2) are predictor variables for **warfarin** response in very elderly, frail inpatients." (PMID19794411). Since this sentence contains a PGx-specific drug–gene pair warfarin-CYP2C9, the sentence itself is highly likely to be related to PGx. The other three drug–gene pairs (warfarin-VKORC1, warfarin-EPHX1, and warfain-CYP4F2) are likely to be PGx-specific pairs.

Recent studies in semi-supervised iterative learning approaches are motivated by the use of a very large collection of texts (web) [20] and the possibility of handling multiple entity types [21]. Semi-supervised pattern learning approaches are advantageous because they require minimal human intervention and no external domain knowledge. Therefore, semi-supervised information extraction systems are able to extract broad types of entities and relationships. Semi-supervised learning approaches have been used to extract information from the web [22–29]. Semi-supervised learning approaches depend on the regularity of language and the data redundancy. A big corpus such as MEDLINE (22 million articles as of the year 2012) is ideal for such tasks. However, the potential for semi-supervised approaches for biomedical information extraction was not fully explored until recently, when we developed semi-supervised pattern learning approaches for disease entity recognition [30] and medical intervention entity recognition [31], *isa* relationship extraction [32], and medical image retrieval from the web [33]. All iterative learning systems suffer from the inevitable problem of spurious patterns and instances introduced in the iterative process. We develop an iterative ranking algorithm to rank extracted drug–gene pairs according to their PGx-relatedness by combining the frequency of drug–gene pairs in MEDLINE with the PGx specificity of other co-occurred drug–gene pairs. The ranking algorithm is similar to the topic sensitive PageRank algorithm developed by Haveliwala [34]. Topic-Sensitive PageRank was based on the PageRank algorithm [35] in order to personalize search rankings using link analysis. Topic-sensitive PageRank computed a set of PageRank vectors, biased using a set of representative topics, in order to capture the importance with respect to a particular topic (details in Section 2).

2. Data and methods

Fig. 1 depicts the iterative process of PGx-specific drug–gene extraction. The system consists of the following components: (1) build a local MEDLINE search engine; (2) iteratively extract drug–gene pairs; (3) rank extracted pairs; and (4) analyze extracted pairs.

2.1. Build local MEDLINE search engine

We have used 20 million MEDLINE abstracts (roughly 100 million sentences) published from 1965 to 2010 as the text corpus for our task of PGx-specific drug–gene relationship extraction. The 2010 MEDLINE/PubMed baseline XML files were downloaded from NLM's anonymous FTP server at ftp://ftp.nlm.nih.gov/nlmdata/.medleasebaseline/. The MEDLINE XML files were then parsed. The abstracts and PMID information from the XML files were extracted. Abstracts were subsequently split into sentences. We used the publicly available information retrieval library Lucene (http://lucene.apache.org) to create a local search engine with

Please cite this article in press as: Xu R, Wang Q. A semi-supervised approach to extract pharmacogenomics-specific drug–gene pairs from biomedical literature for personalized medicine. J Biomed Inform (2013), http://dx.doi.org/10.1016/j.jbi.2013.04.001