Journal of
Pharmacogenomics & Pharmacoproteomics

Research Article Open Access

A Semi-Supervised Pattern-Learning Approach to Extract Pharmacogenomics-Specific Drug-Gene Pairs from Biomedical Literature

Rong Xu[1]* and Quanqiu Wang[2]

[1]Medical Informatics Division, Case Western Reserve University, Cleveland, Oh 44106, USA
[2]Thintec Llc, Palo Alto, Ca 94305, USA

Abstract

Personalized medicine is to deliver the right drug to the right patient in the right dose. Pharmacogenomics (PGx), the studies in identifying genetic variants that may affect drug response, is important for personalized medicine. Computational approaches in studying the relationships between genes and drug response are emerging as an active area of research for personalized medicine. Currently, systematic study of drug-gene relationships is limited because a large-scale machine understandable drug-gene relationship knowledge base is difficult to build and to keep update. Scientific literature contains rich information of drug-gene relationships, therefore is the ultimate knowledge source for PGx studies and for personalized medicine. However, this information is largely buried in free text with limited machine understandability. There is a need to develop automatic approaches to extract structured drug-gene relationships from biomedical literature. In this study, we present a semi-supervised approach to extracting drug-gene relationships from MEDLINE. The technique uses one seed pattern and iteratively learns various ways the relationship may be expressed in 20 million MEDLINE abstracts. Our approach has achieved high precisions (0.961-1.00) in extracting drug-gene relationships from MEDLINE and found many drug-gene pairs that are not available in PharmGKB, a large-scale manually curated PGx knowledge base.

Introduction

We develop a semi-supervised pattern learning method to extract drug-gene relationships from free text. Central to our approach is the observation that: the semantic relationship between a drug and a gene can be expressed in many different ways due to the flexibility and expressive nature of human natural language. However, these patterns are not randomly distributed and there are predominant patterns people use to describe specific types of drug-gene relationships. For example, pattern “DRUG is metabolized by GENE” is typically used to describe metabolism relationship between a drug and a gene. Example sentences include “Quetiapine is metabolized by CYP3A4 and sertindole by CYP2D6” (PMID 10422890), and “Cerivastatin is metabolized by CYP2C8 and CYP3A4, and fluvastatin is metabolized by CYP2C9” (PMID 17178259). On the other hand, pattern “GENE inhibitor DRUG” is typically used to express the inhibition relationships between a drug and a gene.Example sentences include “In addition, the effect of the CYP2C9 inhibitor fluvastatin was evaluated using S-warfarin as a probe” (PMID 16758259) and “The CYP2C8 inhibitor gemfibrozil does not increase the plasma concentrations of zopiclone” (PMID 16832679). In this paper, we use two seed patterns for two types of drug-gene relationship extraction: seed “DRUG is metabolized by GENE” for drug-gene metabolism relationship (i.e. quetiapine-CYP3A4, cerivastatin-CYP2C8) extraction and the seed “GENE inhibitor DRUG” for drug-gene target relationship (i.e. fluvastatin-CYP2C9, gemfibrozil-CYP2C8) extraction. First, we use the seed patterns to find their associated drug-gene pairs. Then we iteratively learn new patterns that are associated with the extracted drug-gene pairs and extract corresponding drug-gene relationships from the newly discovered patterns. The iterative process stops when no additional good patterns are found.

Different person responds differently to the same drug. Genetic factors account for 20 to 95 percent of the drug response variability [1]. Pharmacogenomics (PGx) is the study of how human genetic variations affect an individual’s response to drugs, with focuses on drug metabolism, absorption, distribution and excretion. Understanding of the genetic variants associated with various drug responses is an essential step of personalized medicine [2-4]. PGx research is a knowledge intensive field and its goal is to discover new knowledge and put it to clinical uses for disease treatment. In this field, the research focus is rapidly shifting from studying individual entity (e.g. diseases, drugs, genes, phenotypes) to entire networks of many different biological entities. Computational analysis of the knowledge represented in biomedical networks can uncover important new relationships, generate novel testable hypotheses and provides new insight into biological systems. Systems biology methods for examining drug response with a more network-based view of the genes involved in complex drug responses have been recently investigated [5,6]. The success of computational PGx studies largely depends on the availability of accurate, comprehensive and machine understandable knowledge. Adequate knowledge acquisition and integration are therefore becoming fundamentally important for these studies. The volume of published biomedical research, and therefore the underlying biomedical knowledge base, grows exponentially. Currently, more than 22 million biomedical records are available on MEDLINE, an excellent source of drug-gene relationship knowledge. Clearly with the current rate of growth in published biomedical research, it becomes increasingly likely that important knowledge connecting drugs, genes and diseases is being missed.

There are substantial research efforts in constructing PGx-specific drug-gene relationship knowledge bases using both manual and automatic approaches. Biocuration is the activity of transforming the information buried in human natural language into machine

***Corresponding author:** Rong Xu, Medical Informatics Division, Case Western Reserve University, Cleveland, Oh 44106, USA, E-mail: Rxx@Case.Edu

Received May 19, 2013; **Accepted** June 28, 2013; **Published** July 06, 2013

Citation: Xu R, Wang Q (2013) A Semi-Supervised Pattern-Learning Approach to Extract Pharmacogenomics-Specific Drug-Gene Pairs from Biomedical Literature. J Pharmacogenom Pharmacoproteomics 4: 117. doi:10.4172/2153-0645.1000117

understandable knowledge by curators reading scientific reports and extracting knowledge from published literature [7]. Biocuration has become an essential part of biological discovery and biomedical research. Substantial manual curation efforts have been used to extract PGx knowledge from literature. For example, The Pharmacogenomics Knowledge Base (PharmGKB) currently is the largest manually created resource about how variation in human genetics leads to variation in response to drugs [8]. However, to extract biomedical information from published literature manually and to transform it into machine understandable knowledge is a difficult task, since biomedical terminologies and knowledge are huge, dynamic, diversified and complex. In addition, human curators are liable to error and subjective bias. Therefore, any manually curated terminology and knowledge base is deemed to be incomplete [9]. Automated information extraction of structured knowledge from natural language text is important for biomedical researchers to find up-to-date knowledge from published scientific reports.

Developing automatic approaches to extract PGx drug-gene relationships from free text is an active research area.Hahn et al have recently surveyed the state of the art in mining the pharmacogenomics literature [10]. In general, statistical learning, machine learning, rule-based approaches and natural language processing (NLP) methods have been used [11-19]. Recently, we have developed a knowledge-driven 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 F1 measure when compared with the unconditioned approach [18]. We also developed an iterative learning approach to iteratively extract and rank drug-gene pairs according to their relevance to drug pharmacogenomics [19]. That study was 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.

Approach

Extracting PGx-specific drug-gene relationships from free text is challenging. Firstly, there are different types of drug-gene relationships. Two main types are drug-gene metabolism relationship and drug-gene target (inhibition or induction) relationship. The semantic relationship between drug diclofenac and gene CYP2C9 is related to metabolism as shown in below sentences.

1. The relationship between gemfibrozil and gene CYP2C8 is related to drug inhibition.
2. The relationship between gene CYP3A4 and drugs rifampin, carbamazepine, omeprazole, phenobarbital, and phenytoin is related to drug induction.
3. Note that all three genes mentioned above are PGx-related genes: CYP2C9, CYP2C8 and CYP3A4. However, the relationships between PGx-specific genes and drugs are not necessarily restricted to drug metabolism.

1. "Evidence exists to suggest that diclofenac is metabolized by CYP2C9" (PMID 10853880).
2. "The CYP2C8 inhibitor gemfibrozil does not increase the plasma concentrations of zopiclone" (PMID 16832679).
3. "LIPA metabolism in human hepatocytes was found to be induced by the treatment of human hepatocytes with the prototypical CYP3A4 inducers rifampin, carbamazepine, omeprazole, phenobarbital, and phenytoin but not by the CYP1A2 inducer 3-methylcholanthrene" (PMID 19451401).

As shown in above three sentences, it is challenging for many statistical approaches to differentiate the drug-gene metabolism relationship from drug-gene target (inhibition or induction) relationships. However, as shown in above sentences, researchers often use specific textual patterns, such as "DRUG is metabolized by GENE", "GENE inhibitor DRUG" or "GENE inducers DRUGs", in describing the relationships between drugs and genes. Even though there exist many textual patterns specific for drug-gene semantic relationships, manually identifying these patterns will be challenging. In this study, we use one specific seed pattern to iteratively find and rank other patterns that are similar to the seed pattern. Then using the newly learned textual patterns, we extract their associated drug-gene pairs from MEDLINE sentences. The process is semi-supervised since it requires no additional domain knowledge except the seed pattern, therefore maximally reducing the labor-intensive annotation effort required in many supervised machine learning approaches in extracting biomedical relationships from free text.

Data and Methods

Local MEDLINE search engine

For the text corpus, we used 20 million MEDLINE abstracts (around 100 million of sentences) published from 1965 to 2010. We downloaded the MEDLINE data from the U.S. National Library of Medicine (http://mbr.nlm.nih.gov/Download/index.shtml). We used the publicly available information retrieval library Lucene (http://lucene.apache.org) to create a local MEDLINE search engine with index created on sentences. Sentences were annotated with any drug terms or gene terms using the following drug lexicon and gene lexicon.

Drug lexicon and gene lexicon

Both the drug lexicon and gene lexicon were obtained from PharmGKB. We downloaded a total of 10,898 drug-gene pairs from PharmGKB (data accessed in 10/2010). Our drug lexicon was consisted of 918 drugs appeared in PharmGKB drug-gene pairs. The gene lexicon was consisted of 2,388 genes from the drug-gene pairs in PharmGKB. The reason of using drugs and genes from PharmGKB is that we will use the drug-gene pairs from PharmGKB as performance comparison in the subsequent relationship extraction. Using the same drugs and genes will allow for direct comparison.

Semi-supervised drug-gene relationship extraction: The semi-supervised drug-gene relationship extraction algorithm is depicted in Figure 1 and can be formulated as follows:

Given: (1) a seed pattern such as "*DRUG is metabolized by GENE*" or "*Drug inhibitor GENE*", where DRUG and GENE are terms from the input drug lexicon and gene lexicon; (2) a text corpus of MEDLINE sentences; (3) a drug lexicon and a gene lexicon.

Do: starting from the seed pattern, which represents a typical way of expressing specific drug-gene semantic relationship, iteratively discover new patterns ("Pattern Extraction") and extract new pairs with newly discovered patterns ("Pair Extraction"). The iterative process

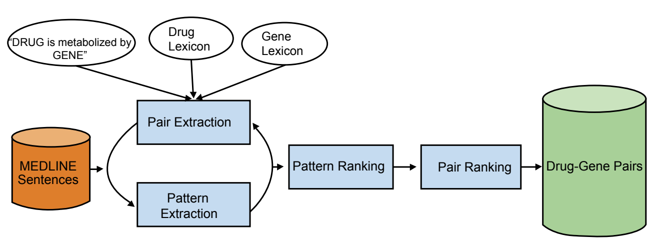

Figure 1: The semi-supervised pattern-learning approach for extracting PGx-specific drug-gene relationship from MEDLINE.

stops when no significant number of new patterns is discovered (two iterations in this study). We then rank extracted patterns, and rank extracted pairs.

Pair Extraction Seed pattern or textual patterns extracted from the previous iteration were used as search queries to the local MEDLINE search engine. Sentences that contain these patterns were retrieved. We extracted drug-gene pairs from the retrieved sentences if the drug-gene pairs and the textual pattern followed the following format: "DRUG pattern GENE" or "GENE pattern DRUG", wherein DRUG and GENE are from the lexicons. For example, the seed patterns we used for drug metabolism relationship extraction were "DRUG is metabolized by GENE" and "GENE substrate DRUG". The seed patterns for drug gene inhibition relationship extraction were "DRUG inhibited GENE" and "GENE inhibitor DRUG".

Pattern Extraction Drug-gene pairs extracted from previous iteration were used as search queries to the local MEDLINE search engine. Corresponding sentences were retrieved. Textual patterns between a drug and a gene were extracted if drug-gene pairs and the pattern conformed to the following format: "DRUG pattern GENE" or "GENE pattern DRUG", where the pattern was the search query. The iterative pair extraction and pattern extraction process ran until no significant number of new patterns was discovered (two iterations in this study).

Pattern Ranking After the iterative pattern extraction and pair extraction step, we ranked extracted patterns in order to find PGx-specific textual patterns. Each pattern was ranked based on how similar its output (its associated drug-gene pairs) was to the output of the eed pattern. Using the output of the seed pattern (p0) as gold standard, we developed three pattern-ranking algorithms: (1) Precision-based ranking, wherein patterns were ranked based on pattern specificity; (2) Recall-based ranking, wherein patterns were ranked based on pattern generality; and (3) F1-based pattern ranking algorithm, wherein both pattern specificity and generality were taken into account. We define ins(p) to be the set of pairs matched by pattern p, and the intersection ins(p) $\cap$ ins(p0) as the set of pairs matched by both pattern p and see pattern p0. Then the Precision-based, Recall-based, and F1-based ranking pattern ranking scores are defined as below:

Pair Ranking After pattern ranking, we then ranked extracted pairs based on their associated pattern scores and their frequency counts in MEDLINE. A reliable D1->D2 pair is one that is associated with reliable patterns many times. The ranking score of a pair is defined as following: where the score of its associated patterns and the number of times that the pair is associated with the pattern.

Evaluation

In this study, we have extracted two types of drug-gene relationships: drug-gene metabolism relationship and drug-gene target (inhibition) relationship. For each relationship, we selected the two top-ranked patterns and extracted their associated drug-gene pairs from MEDLINE.For each pattern, we retrieved their corresponding MEDLINE sentences and manually examined the correctness of drug-gene relationships using these sentences as evidences. Three evaluators with graduate degrees in biomedical science performed the manual evaluation.The drug-gene pairs that all three evaluators agreed upon were determined as assigned as true positive.

Results

Many top patterns associated with drug-gene pairs in PharmGKB are not necessarily specific for PGx-specific relationship.

To get an empirical sense of the variability of natural language used to express pharmacogenomic drug gene relationships in MEDLINE, we studied lexical textual patterns associated with drug-gene pairs in PharmGKB. We used the drug-gene pairs in PharmGKB as queries to the local MEDLINE search engine. We then extracted the text string between the drug-gene pairs. We counted the number of drug-gene pairs that were associated with each of the patterns. Among all 10,898 drug-gene pairs from PharmGKB, only 2,596 have even co-occurred in MEDLINE sentences. Figure 2 shows the top-50 ranked patterns in the format of "DRUG pattern GENE", where DRUG-GENE pair is from PharmGKB and DRUG is in front of the GENE.Examples include "DRUG (GENE", "DRUG and GENE" and DRUG, GENE". Figure 3 shows the top-50 ranked patterns in the format of "GENE pattern DRUG", where DRUG-GENE pair is from PharmGKB and GENE is in front of the DRUG. Examples include "GENE (DRUG", "GENE by DRUG" and "GENE and DRUG".

Several observations can be made from Figures 2 and 3. First, most of the textual patterns associated with drug-gene pairs from PharmGKB are highly specific and were associated one drug-gene pair. For example, using drug-gene pairs from PharmGKB as search queries, we extracted a total of 34,141 textual patterns that were in the format "DRUG pattern GENE". Among them, 33,488 (98%) patterns were only associated with one drug-gene pair in MEDLINE. Similarly, a total of 41,468 patterns in the format "GENE pattern DRUG" were extracted, among which 40,656 (98%) patterns were associated with only one drug-gene pair.

Second, many of the top-ranked patterns (patterns that were associated with many drug-gene pairs from PharmGKB) are in fact not PGx-specific patterns, such as "DRUG and GENE" or "DRUG (GENE)". Third, the drug-gene pairs from PharmGKB are of many different semantic types, including both drug-gene metabolism relationship and drug-gene target relationship. For example, the top 9 pattern ("GENE inhibited DRUG") and the top10 pattern ("GENE, an inhibitor for DRUG") in Figure 2 are patterns specific for drug-gene target relationship. However, these patterns were associated with many drug-gene pairs from PharmGKB, which is primarily a knowledge base for PGx-specific drug-gene pairs. Finally, as seen from both figures, there exist some representative patterns among top-ranked patterns, implying that researchers indeed used specific patterns to describing drug-gene semantic relationships, which is the critical assumption for our pattern-based relationship extraction approach.

Semi-supervised pattern learning approach is able to find specific patterns for drug-gene relationships.

In this study, we learned two different types of patterns: patterns specific for drug-gene metabolism relationship and patterns specific for drug-gene target relationship. The two seed patterns for drug-gene

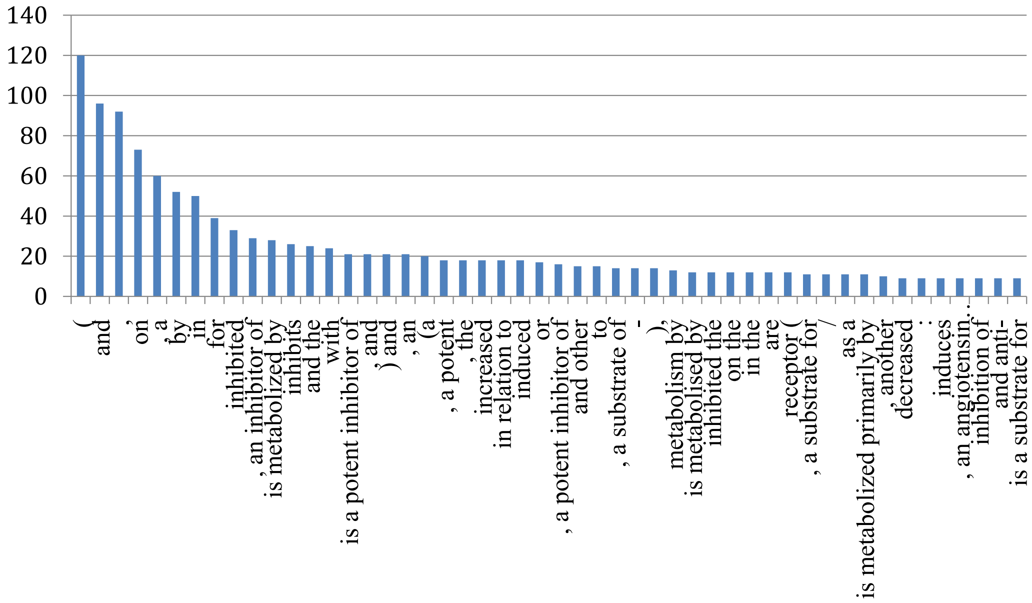

Figure 2: Top 50 textual patterns (“DRUG pattern GENE”) associated with drug-gene pairs from PharmGKB. X-axis shows one of the top 50 ranked textual patterns. Y-axis represents the number of distinct drug-gene pairs (from PharmGKB) associated with each pattern.

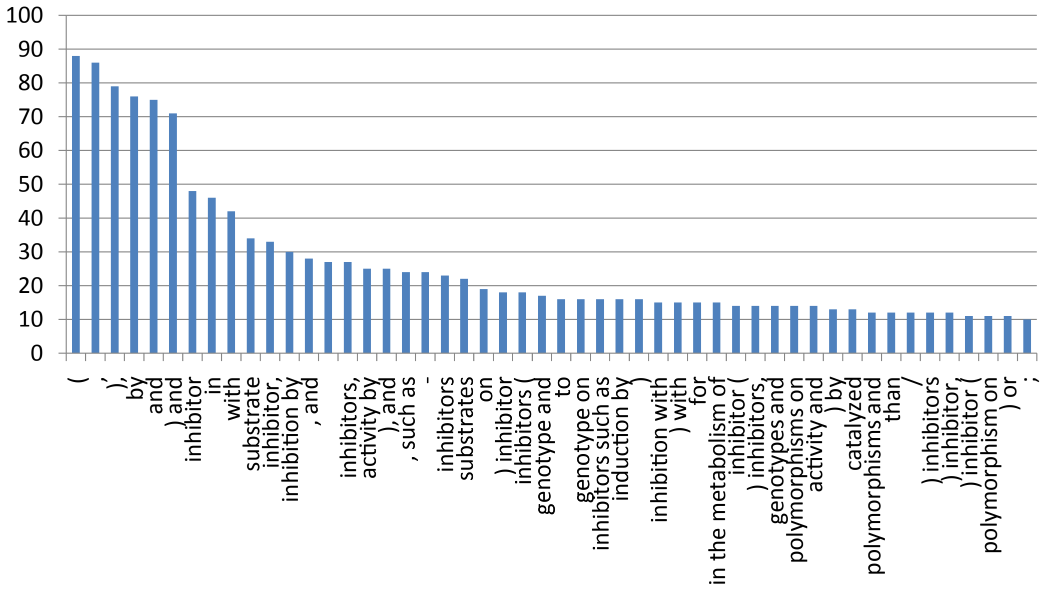

Figure 3: Top 50 textual patterns (“GENE pattern DRUG”) associated with drug-gene pairs from PharmGKB. X-axis shows one of the top 50 ranked textual patterns. Y-axis represents the number of distinct drug-gene pairs (from PharmGKB) associated with each pattern.

metabolism-specific relationship extraction are “DRUG is metabolized by GENE” and “GENE substrate DRUG”. The two seed patterns for drug-gene target-specific relationship are “DRUG inhibited GENE” and “GENE inhibitor DRUG”. After two iterations, we ranked extracted patterns using three pattern-ranking approaches: the Precision-based approach, the Recall-based approach and the F1-based approach. By comparing the top-ranked patterns by each of these approaches, we found out that the F1-based approach worked better than the other two in ranking patterns with both good precision and recalls highly among all patterns. The Precision-based patterns tended to rank very specific patterns (patterns associated with only one drug-gene pair) highly, while the Recall-based approached worked the opposite way in ranking many overly general patterns highly.

The top 10 ranked patterns for drug-gene metabolism relationships using seed pattern “DRUG is metabolized by GENE” and “GENE substrate DRUG” are shown in Table 1. The pattern ranking orders were determined by the F1-based ranking approach. Metabolism-specific patterns are highlighted. As shown in the table, starting with one seed pattern, our iterative pattern extraction and pattern ranking approaches were able to rank many metabolism-specific patterns highly among all extracted patterns. Note that many of these patterns don't necessarily following specific syntactic patterns.

The top 10 ranked patterns for drug-gene target-specific relationships using seed pattern "DRUG inhibited GENE" and "GENE inhibitor DRUG" are shown in Table 2 with target-specific patterns highlighted. As shown in the table, starting with one seed pattern, our iterative pattern extraction and pattern ranking approaches were able to rank many drug-gene inhibition-specific patterns highly among all extracted patterns. As seen in the table, many patterns such as "DRUG inhibited GENE" and "DRUG inhibits GENE" can be combined to significantly reduce the number of patterns.

IN summary, the semi-supervised pattern-learning approach starting with one single seed were able to automatically find patterns for specific drug-gene semantic relationships. In addition, these patterns were associated with more drug-gene pairs than those included in PharmGKB, therefore we can used these newly learned patterns to extract additional drug-gene pairs from MEDLINE. For example, a total of 34 drug-gene pairs from PharmGKB were associated with pattern "GENE substrate DRUG" (Figure 2). However, when using the pattern "GENE substrate DRUG" as search query, we extracted a total of 50 distinct drug-gene pairs from MEDLINE sentences, among which only 34 pairs were included in PharmGKB.

Pattern-based drug gene relationship extraction

We selected four sets of specific patterns from the top-ranked patterns and used these selected patterns to extract drug-gene pairs from MEDLINE. We selected six metabolism-specific patterns in the format of "DRUG pattern GENE" from top-ranked patterns for seed "DRUG is metabolized by GENE". These six patterns are "DRRUG is metabolized by GENE", "DRUG metabolism by GENE", "DRUG, a substrate for GENE", "DRUG is a substrate for GENE", "DRUG is metabolised by GENE", and "DRUG is metabolized primarily by". These six patterns were associated with a total of 76 drug-gene pairs.We also selected 10 metabolism-specific pairs in the format of "GENE pattern DRUG" from top-ranked patterns for seed "GENE substrate DRUG". These patterns were associated with 124 drug-gene pairs. Similarly for target-specific patterns, we selected 12 patterns from top-ranked patterns for seed "DRUG inhibited GENE" and these patterns were associated with a total of 832 drug-gene pairs. We selected 10 patterns for seed "GENE inhibitor DRUG", which were associated with a total of 193 drug-gene pairs. For each set of patterns, we manually evaluated the precision of the extracted pairs and calculated how many of these extracted pairs were not included in PharmKGB.

As shown in Table 3, all four set of selected patterns were highly precise (precision ranging from 0.96 to 1.00) in extracting drug-gene pairs from MEDLINE. In addition, many of these pairs were not included in PharmGKB. Examples of extracted drug gene metabolism pairs that were not in PharmGKB included sudofetilide-CYP3A4, eplerenone-CYP3A4, estradiol-CYP3A4, ganciclovir-ABCG2, lidocaine-CYP3A, and terfenadine-CYP2C9. Note that recalls were not calculated. First, there is no good standard that represents all drug-gene pairs that appeared in MEDLINE. Second, the main goal of this study was to demonstrate that the semi-supervised pattern learning approach is able to accurately extract many additional drug-gene pairs that have not included in the currently most comprehensive PGx-specific drug-gene relationship knowledge base. Unlike many supervised machine learning approach, the semi-supervised learning approach required only a single seed pattern, therefore maximally minimized the human curation effort involved.

Rank	"DRUG *is metabolized by* GENE"	"GENE *substrate* DRUG"
1	**"DRUG *is metabolized by* GENE"**	**"GENE *substrate* DRUG"**
2	**"DRUG *is metabolised by* GENE"**	**"GENE *mediated* DRUG"**
3	**"DRUG *is metabolized* GENE"**	**"GENE *catalyzed* DRUG"**
4	**"DRUG *metabolism by* GENE"**	**"GENE *substrates* DRUG"**
5	**"DRUG, *a substrate for* GENE"**	"GENE *dependent* DRUG"
6	"DRUG, *which is* GENE"	"GENE *activity (*DRUG"
7	"DRUG *in relation to* GENE"	"GENE *genotype and* DRUG"
8	"DRUG *on the activity of* GENE"	**"GENE *probe drug* DRUG"**
9	**"DRUG *oxidation (*GENE"**	**"GENE *substrates such as* DRUG"**
10	**"DRUG *is substrate for* GENE"**	"GENE *activity as* DRUG"

Table 1: Top 10 ranked patterns (as determined by the F1-based ranking)using seed patterns "DRUG *is metabolized by* GENE" and "GENE *substrate* DRUG".

Rank	"DRUG *inhibited* GENE"	"GENE *inhibitor* DRUG"
1	**"DRUG *inhibited* GENE"**	**"GENE *inhibitor* DRUG"**
2	**"DRUG *inhibits* GENE"**	**"GENE *inhibitor,* DRUG"**
3	**"DRUG *decreased* GENE"**	**"GENE *inhibitors* DRUG"**
4	**"DRUG, *an inhibitor of* GENE"**	**"GENE*) inhibitor* DRUG"**
5	**"DRUG *reduced* GENE"**	**"GENE *inhibitors,* DRUG"**
6	"DRUG *on* GENE"	**"GENE *inhibition by* DRUG"**
7	**"DRUG *suppressed* GENE"**	**"GENE *inhibitor (*DRUG"**
8	**"DRUG *induced* GENE"**	**"GENE *inhibitors such as* DRUG"**
9	**"DRUG *inhibition of* GENE"**	**"GENE *inhibition with* DRUG"**
10	"DRUG, *a potent* GENE"	**"GENE) *inhibitors* DRUG"**

Table 2: Top 10 ranked patterns (as determined by the F1-based ranking) using seed patterns "DRUG *inhibited* GENE" and "GENE *inhibitor* DRUG".

Relationship	Pattern	Pairs (n)	Precision	Pairs not in PharmGKB (n)
Metabolism	"DRUG pattern GENE"	76	0.973	14
	"GENE pattern DRUG"	124	1.000	28
Target	"DRUG pattern GENE"	832	0.961	708
	"GENE pattern DRUG"	193	1.000	116

Table 3: Precision of four sets of selected patterns in extracting drug-gene pairs from MEDLINE and the numbers of additional pairs extracted compared to ones in PharmGKB.

Some of the false positives are caused by gene symbol ambiguity. For example, incorrect pair glucose-MRS was extracted from sentence "Also the study of cerebral glucose metabolism by MRS is very promising, allowing a resolution and sensitivity comparable to PET" (PMID 9029941), where MRS is not a gene symbol. In this sentence, it represents "magnetic resonance spectroscopy". For the same reason, false pair glucose-DO was extracted from sentence "Pyrroline carboxylate reduced [5-3H] glucose metabolism by DO…" (PMID 10330104), where DO represent "denuded oocytes", not gene. Therefore, to further improve the precision of our methods, gene disambiguation is necessary.

Discussion and Conclusions

We have developed an iterative pattern learning approach for extracting precise drug gene relationships from free text. Our method achieves high precision in extracting specific types of drug gene semantic relationships. In addition, our method is able to extract many drug gene relationships currently not included in PharmGKB. One of the advantages of our method is that it is highly efficient and does not involve sentence parsing, therefore avoiding many errors introduced by parsing complicated biomedical text. However, there is still significant space in which to seek improvement in increasing the coverage of methods and the quality of our patterns.

First of all, the pattern-based approach only worked on extracted drug-gene pairs that co-occurred in the sentences. Pairs that appeared in abstracts but not in sentences will be missed. Even though important drug-gene pairs often appeared in the same sentences, there will be pairs only appeared in abstracts. Second, our study only used the abstracts, not the full-text. Even though many of full-text articles related to pharmacogenomics are often not publically available, they may contain richer set of information. It will be interesting to systematically compare drug-gene pairs that appeared in abstracts to those that appeared in the full-text articles for the same set of abstracts. Third, our restriction on patterns (“DRUG pattern GENE” or “GENE pattern DRUG”) limited the space of patterns that we could potentially examine. In some cases, it is necessary to consider patterns before and after drug or gene entities, in the format “pattern DRUG pattern GENE” or “DRUG pattern GENE pattern” or “pattern DRUG pattern GENE pattern”. For example, in sentence “R(+)XK469 inhibits hydroxylation of S-warfarin by CYP2C9” (PMID 19464879). Pattern “hydroxylation of DRUG by GENE’ is a highly precise pattern for drug gene metabolism relationship. However, as the amount of data increase, the relationships will appear with typical pattern in the format of “DRUG pattern GENE” as shown in sentence “OBJECTIVE: The aim of this study was to determine whether folic acid supplementation increases the dosage requirement of the CYP2C9 substrate warfarin” (PMID 20206792). To learn more complex patterns such as “pattern DRUG pattern GENE” or “DRUG pattern GENE pattern” or “pattern DRUG pattern GENE pattern”, parsing will be needed since our methods cannot decide the boundaries of the prefix and postfix patterns.

Some of the patterns in the tail have very typical syntactic feature. For example in sentence “Warfarin, the principal oral anticoagulant used in the treatment and prevention of thromboembolic disease, is primarily metabolized by CYP2C9” (PMID 15341704), there is a common syntactic pattern “DRUG is metabolized by GENE” while the lexical pattern, “the principal oral anticoagulant used in the treatment and prevention of thromboembolic disease, is primarily metabolized by” occurs only once in MEDLINE. Our method can be combined with NLP-intensive method to further improve recall. Pattern generalization would significantly fatten the head; however, we believe that no amount of generalization will eliminate the tail.

Our method is simple and will miss some drug gene pairs. However, as the corpus of literature increases, redundancy will increase the likelihood of a drug gene pair being matched by a simple lexical pattern. The rapid growth of biomedical knowledge and literature, which makes automatic extraction of drug gene relationships necessary, can also act to increase its coverage over time. Although our ultimate goal would be to be able to reliably reproduce the relationship annotations provided by human, in practice we do not expect automatic tools to be able to fully replace the work of human curators. When automatic methods are intelligently combined with human curation process, they can reduce curators’ workload and bias, and increase the completeness and recency of manually curated pharmacogenomics knowledge base.

References

1. Evans WE, McLeod HL (2003) Pharmacogenomics--drug disposition, drug targets, and side effects. N Engl J Med 348: 538-549.
2. Weiss ST, McLeod HL, Flockhart DA, Dolan ME, Benowitz NL, et al. (2008) Creating and evaluating genetic tests predictive of drug response. Nat Rev Drug Discov 7: 568-574.
3. Swen JJ, Huizinga TW, Gelderblom H, de Vries EG, Assendelft WJ, et al. (2007) Translating pharmacogenomics: challenges on the road to the clinic. PLoS Med 4: e209.
4. Davis JC, Furstenthal L, Desai AA, Norris T, Sutaria S, et al. (2009) The microeconomics of personalized medicine: today’s challenge and tomorrow’s promise. Nat Rev Drug Discov 8: 279-286.
5. Ma’ayan A, Jenkins SL, Goldfarb J, Iyengar R (2007) Network analysis of FDA approved drugs and their targets. Mt Sinai J Med 74: 27-32.
6. Yildirim MA, Goh KI, Cusick ME, Barabási AL, Vidal M (2007) Drug-target network. Nat Biotechnol 25: 1119-1126.
7. Howe D, Costanzo M, Fey P, Gojobori T, Hannick L, et al. (2008) Big data: The future of biocuration. Nature 455: 47-50.
8. Klein TE, Chang JT, Cho MK, Easton KL, Fergerson R, et al. (2001) Integrating genotype and phenotype information: an overview of the PharmGKB project. Pharmacogenetics Research Network and Knowledge Base. Pharmacogenomics J 1: 167-170.
9. Hahn U, Wermter J, Blasczyk R. & Horn P A (2007) Text mining: powering the database revolution. Nature 448: 130-130.
10. Hahn U, Cohen KB, Garten Y, Shah NH (2012) Mining the pharmacogenomics literature--a survey of the state of the art. Brief Bioinform 13: 460-494.
11. Chang JT, Altman RB (2004) Extracting and characterizing gene-drug relationships from the literature. Pharmacogenetics 14: 577-586.
12. Garten Y, Altman RB (2009) Pharmspresso: a text mining tool for extraction of pharmacogenomic concepts and relationships from full text. BMC Bioinformatics 10 Suppl 2: S6.
13. Theobald M, Shah N, Shrager J (2009) Extraction of Conditional Probabilities of the Relationships Between Drugs, Diseases, and Genes from PubMed Guided by Relationships in PharmGKB. Summit on Translat Bioinforma 2009: 124-128.
14. Hansen NT, Brunak S, Altman RB (2009) Generating genome-scale candidate gene lists for pharmacogenomics. ClinPharmacolTher 86: 183-189.
15. Ahlers CB, Fiszman M, Demner-Fushman D, Lang FM, Rindflesch TC (2007) Extracting semantic predications from Medline citations for pharmacogenomics. Pac SympBiocomput.
16. Coulet A, Shah NH, Garten Y, Musen M, Altman RB (2010) Using text to build semantic networks for pharmacogenomics. J Biomed Inform 43: 1009-1019.
17. Xu R, Wang Q (2013) An iterative searching and ranking algorithm for prioritising pharmacogenomics genes. Int J ComputBiol Drug Des 6: 18-31.
18. Xu R, Wang Q (2012) A knowledge-driven conditional approach to extract pharmacogenomics specific drug-gene relationships from free text. J Biomed Inform 45: 827-834.
19. Xu R, Wang Q (2013) A semi-supervised approach to extract pharmacogenomics-specific drug-gene pairs from biomedical literature for personalized medicine. J Biomed Inform.