Journal of Biomedical Informatics 2012:(in press).

A mutation-centric approach to identifying pharmacogenomic relations in text

Bastien Rance[1], Emily Doughty[2], Dina Demner-Fushman[1], Maricel G. Kann[2], Olivier Bodenreider[1]

[1] National Library of Medicine, Bethesda, MD 20894, USA

[2] University of Maryland, Baltimore County, Baltimore, MD 21250, USA

Abstract

Objectives. To explore the notion of mutation-centric pharmacogenomic relation extraction and to evaluate our approach against reference pharmacogenomic relations.

Methods. From a corpus of MEDLINE abstracts relevant to genetic variation, we identify co-occurrences between drug mentions extracted using MetaMap and RxNorm, and genetic variants extracted by EMU. The recall of our approach is evaluated against reference relations curated manually in PharmGKB. We also reviewed a random sample of 180 relations in order to evaluate its precision.

Results. One crucial aspect of our strategy is the use of biological knowledge for identifying specific genetic variants in text, not simply gene mentions. On the 104 reference abstracts from PharmGKB, the recall of our mutation-centric approach is 33-46%. Applied to 282,000 abstracts from MEDLINE, our approach identifies pharmacogenomic relations in 4534 abstracts, with a precision of 65%.

Conclusions. Compared to a relation-centric approach, our mutation-centric approach shows similar recall, but slightly lower precision. We show that both approaches have limited overlap in their results, but are complementary and can be used in combination. Rather than a solution for the automatic curation of pharmacogenomic knowledge, we see these high-throughput approaches as tools to assist biocurators in the identification of pharmacogenomic relations of interest from the published literature. This investigation also identified three challenging aspects of the extraction of pharmacogenomic relations, namely processing full-text articles, sequence validation of DNA variants and resolution of genetic variants to reference databases, such as dbSNP.

Keywords

Pharmacogenomics, mutation extraction, biocuration.

1 Introduction

One aspect of personalized medicine is better adaptation of therapeutic drugs to the specific situation of a given patient, part of which is determined by his or her unique genetic make-up. Pharmacogenomics attempts to assess the influence of genetic variation on drug response [10, 28]. One poster child of pharmacogenomics is the drug *warfarin*, an anticoagulant widely prescribed for the prophylaxis and treatment of thromboembolic phenomena in patients with deep vein thrombosis and atrial fibrillation.

Warfarin has a narrow therapeutic index, that is, small changes in the dose result in important variations of the therapeutic effect. For an anticoagulant, this means either insufficient anticoagulation and risk of thrombosis if the dose is too low, or excessive anticoagulation and increased hemorrhagic risk if the dose is too high. Since a large fraction of the therapeutic effect of *warfarin* is dependent on genetic variation, it has been shown that testing patients for variations in specific genes can help determine the initial dose and enhance clinical outcomes [18]. More specifically, CYP2C9 and VKORC1 genotype information can be integrated into algorithms used for the determination of the maintenance dose of *warfarin*, outperforming traditional algorithms (not using genotype information), especially for patients requiring low or high doses [18]. Examples of allelic variants include the point mutation G3673A, associated with response to a lower dose of *warfarin* [25]. The exact place of CYP2C9 and VKORC1 genotyping in anticoagulation with *warfarin* has been subject to debate [11]. However, recent studies have demonstrated lower risk of hospitalization for hemorrhage or thromboembolism in patients for which genetic information had been determined [9]. While genetic information is not yet used routinely with *warfarin* prescription, CYP2C9 and VKORC1 genotyping is widely available, and the Food and Drug Administration's standard product label for *warfarin* now discusses the practical influence of allelic variation on the dose needed by specific patient groups.

The biomedical literature is the primary vehicle for reporting the association between gene variants and drugs. Pharmacogenomic information is generally extracted from text and curated manually in order to create reference knowledge bases, such as PharmGKB [16, 19]. Information extraction can also be automated using natural language processing (NLP) tools [12]. However, text mining approaches to extracting pharmacogenomic knowledge generally show limited precision [13]. Our goal here is to leverage biological knowledge to increase the performance of information extraction methods. In previous work [24], we exploited the biomedical literature using a method based mainly on co-occurrences, with limited success. We now apply the lessons learned from this preliminary work to improve our methods. The singlemost important element is the identification of allelic variants. Towards this end, we introduce EMU [8], an extractor of mutations, to complement our original approach.

The objectives of this study are both to explore the notion of mutation-centric pharmacogenomic relation extraction and to evaluate our approach against reference pharmacogenomic relations. Additionally, we compare our approach to a relation-centric approach and we outline the potential of our approach to support the curation of pharmacogenomic relations.

2 Background

2.1 PharmGKB

PharmGKB [19] aims to collect, gather and communicate the knowledge about the impact of human genetic variations on drug response. PharmGKB curation efforts are concentrated on a small set of very important pharmacogenes (referred to as VIP genes), for which a comprehensive domain expert annotation is provided. For these genes, the specific genetic variants are identified, along with related drugs and phenotypes. The articles from which the information was extracted are listed as evidence. It must be noted, however, that pharmacogenomic knowledge in PharmGKB is curated from a limited set

of high-quality journals and for a small number of drugs and genes of particular interest, and is therefore not comprehensive.

2.2 Approaches to extracting pharmacogenomic information

Extracting pharmacogenomic information from the biomedical literature, i.e., information about drugs, phenotypes, gene variants and their interrelations, can be seen as a specific task in the broader discipline of text mining (See [1, 29] for a review of text mining). A recent review article provides a rich description of the state of the art [13] and we will therefore keep our own review of related work to a minimum.

As summarized in [13], approaches to extracting pharmacogenomic relations share many features. Common to all approaches is the identification of named entities of interest (drugs, diseases, genes and their variants) from the biomedical literature, relying on dictionaries, rules or machine learning techniques. Analogously, the identification of the relations among these entities generally relies on the presence of these entities within a given span of text, i.e., co-occurrence. Additional cues are used to avoid false positive relations, including statistical cues (e.g., frequency of co-occurrence) and linguistic cues (e.g., syntactic dependencies among entities). Models of such relations can also be identified through machine learning approaches.

Some of the systems developed recently were presented at the workshop on "Mining the pharmacogenomics literature" organized at the Pacific Symposium on Biocomputing 2011. Interestingly, the following systems all leverage dependency graphs, i.e., graphs representing the syntactic structure of sentences, to extract pharmacogenomic relations, and can be thought of as "relation-centric approaches". Initially developed for event extraction in the BioNLP Shared Task, JReX was adapted to gene-drug relation extraction [5]. Analogously, the OntoGene Relation Miner developed for extracting protein-protein interactions was extended to support the extraction of pharmacogenomic relations [26]. Not surprisingly, the largest pharmacogenomic text mining effort was done by the researchers associated with PharmGKB curation at Stanford University [7]. Since we compare our results to theirs, their approach is described later in the Discussion section.

Although not specific to the extraction of pharmacogenomic relations, some recent work in biomedical information extraction is relevant to our investigation. OpenDMAP, the Open source Direct Memory Access Parser, developed at the University of Colorado is an ontology-driven system that achieves high precision [17]. The use of biological knowledge has been shown to increase the performance of information extraction systems [20]. In particular, the use of nucleotide and amino acid sequences, leveraged by EMU for extracting mutations from text, has also been identified as a key element for the normalization of ambiguous gene names among species in the GNAT system [15].

The specific contribution of this work is not the development of a new method, but rather the combination of existing tools and techniques (MetaMap, RxNorm filtering, EMU) into a novel strategy for identifying pharmacogenomic relations. While based on co-occurrence between genes and drugs, our strategy is characterized by the use of biological knowledge for identifying specific genetic variants in text. Our approach can therefore be termed a "mutation-centric approach".

3 Datasets and methods

Our approach to extracting pharmacogenomic information from text takes advantage of biological knowledge to increase the precision of the simple co-occurrence approach. Drug mentions and genetic variants are identified in MEDLINE abstracts in order to establish a list of <article, drug, genetic variant> relations. We present an example illustrating the identification of point mutations and drug mentions. The applicability of our high-throughput method is verified on a large set of MEDLINE abstracts relevant to genetic variation. Finally, our results are evaluated against reference relations curated manually in PharmGKB.

3.1 Identifying drugs

We identify drugs in text by using a generic biomedical entity recognition system, MetaMap, and restricting its output with the drug-specific resource RxNorm.

MetaMap [2] is a medical concept recognizer developed by the National Library of Medicine. MetaMap annotates biomedical text with concepts from the Unified Medical Language System (UMLS) [3]. In practice, MetaMap associates a UMLS concept unique identifier with any medical concept recognized in the text. Moreover, it is important to map drugs to a reference terminology in order to facilitate comparisons across resources. Therefore, we filter and normalize the concepts extracted by MetaMap using a drug-specific resource, RxNorm.

RxNorm [23] is a standardized nomenclature for clinical drug entities developed by the National Library of Medicine. RxNorm is one of a suite of designated standards for use in U.S. Federal Government systems for the electronic exchange of clinical health information. The RxNorm model distinguishes between various types of drug entities (e.g., ingredient, precise ingredient, brand name) and asserts relations among these types, making it possible to navigate among them. RxNorm is integrated in the UMLS. In practice, we used the RxNorm API [4] to perform the filtering and normalization steps.

Filtering. From all the biomedical concepts extracted by MetaMap, only those concepts present in RxNorm as drug ingredients are selected, ensuring that only drug concepts are retained. Molecules such as amino acids, simple sugars (e.g., *glucose*) and inorganic elements (e.g., *calcium*) are listed as ingredients in RxNorm. However, when mentioned in the biomedical literature, these molecules rarely refer to clinical drugs. Therefore, we eliminate them from the list of drugs identified by RxNorm.

Normalization. Variation in drug names is generally captured by the UMLS, where synonymous names are associated with the same concept. In contrast, salt and non-salt ingredients (*atorvastatin* and *atorvastatin calcium*) denote different entities in the UMLS. For the purpose of relating drugs to genetic variants, it is preferable to ignore such differences. We leverage RxNorm relations to aggregate drug entities at the appropriate ingredient level.

3.2 Identifying genetic variants

We identify point mutations in text and attempt to resolve them to reference Single Nucleotide Polymorphisms (SNPs) in dbSNP [21, 27] whenever possible. While methods such as MutationFinder [6] have been developed to extract mutational information from biomedical text with high precision, these

methods lack the mutation-gene associations required here in order to relate drugs to specific genetic variants. In prior work, we developed ***EMU*** [8] (Extractor of Mutations) to identify point mutations and their associated genes in biomedical text. (A detailed example of identification of a point mutation by EMU is presented in the next section.) One original feature of EMU compared to other such tools is to leverage biological information for the validation of the mutations extracted against reference sequences. More specifically, for candidate mutations identified in abstracts with regular expressions specifically crafted for capturing the many ways in which mutations are expressed in text, the protein products of the corresponding genes are checked for the possible existence of a mutation at the location indicated (i.e., we verify that the wild type amino acid recorded in the given mutation corresponds to the actual amino acid in the specified protein sequence position). Finally, EMU attempts to resolve each variant to an identifier (rsid) from the reference database of SNPs, dbSNP. Of note, while EMU is designed to identify mentions of mutations referring to either nucleotide or amino acid sequences, only the protein mutations can be validated against reference sequences.

In our mutation-centric approach, specific genetic variants, not only gene mentions, are identified in text. Moreover, whenever possible, the genetic variants are validated against reference sequences and resolved to the reference database dbSNP.

In summary, after applying the drug and gene variant identification methods to a set of MEDLINE abstracts, we obtain a smaller set of abstracts in which we have identified pharmacogenomic relations of the form <article, drug, genetic variant>.

3.3 Extended example

In order to illustrate how pharmacogenomic relations are identified in text, we use the following text fragment from the abstract of a PubMed article (PMID 12492608) [14]: *A noncoding single nucleotide polymorphism (SNP) in exon 26* ***3435C > T*** *of the highly polymorphic* ***MDR1*** *gene has been demonstrated to alter* ***digoxin*** *absorption […]*.

Genetic variant. In this sentence, EMU identified a DNA mutation. "3435C > T" denotes a single nucleotide polymorphism (SNP) in which the nucleotide C is substituted by T in position 3435 on the reference sequence NM_000927.3 (chromosomal position 87138645 on chromosome 7), corresponding to the human gene ABCB1 (for which *MDR1* is a synonym). This SNP can be resolved to rs1045642 in the reference database of SNPs, dbSNP, which validates the mutation identified by EMU. (Since EMU does not support sequence validation for DNA mutations, resolution to dbSNP of the genetic variant identified by EMU was performed manually in this case).

Drug. MetaMap identified the drug *digoxin* in the sentence, which it mapped to the UMLS concept C0012265. Since this drug is listed as an ingredient in RxNorm (RxCUI 3407), it was selected as a valid drug.

Overall, our approach identified the relation <3435C>T/rs1045642, digoxin(RxCUI:3407)> from article PMID:12492608 as a potential pharmacogenomic relation. Of note, this relation is also among the gene variants curated in PharmGKB (*ABCB1*:3435T>C).

3.4 Application to a large set of MEDLINE abstracts

In order to restrict the large MEDLINE corpus (over 20M citations) to a manageable dataset, we first use the PubMed search engine to select those abstracts in which pharmacogenomic information is most likely to be identified. In prior work, we determined that two Medical Subject Headings (MeSH) descriptors, "Mutation" OR "Polymorphism, Genetic", were used most frequently as indexing terms in articles referenced in the PharmGKB VIP (very important pharmacogenes) dataset. Here, we further constrain the search by restricting it to articles published since 2000 (earlier abstracts are less likely to contain genomic information), in English, and for which an abstract is available. This PubMed search was performed in January 2011 and yielded 281,947 abstracts identified by their PubMed identifier (PMID). We also considered processing full-text articles, but despite the growth of PubMed Central over the past few years, a relatively small proportion of articles is publicly accessible and amenable to processing by our tools. (We come back to full-text processing in the discussion). Each abstract in this set was submitted to the drug and gene variant identification processes presented above.

3.5 Evaluation

3.5.1 Recall

For evaluation purposes, we use a reference dataset of 104 articles corresponding to the PharmGKB VIP (very important pharmacogenes) dataset. We compare the <article, drug, genetic variant> relations we extracted to similar reference relations curated manually in PharmGKB. Here we evaluate recall, i.e., the ability of our methods to identify these relations from a set of reference documents. In PharmGKB, we concentrate on those <drug, genetic variant> relations that have undergone in-depth curation ("VIP annotations"). For example, for the drug *warfarin*, VIP annotations are provided only for the gene VKORC1, for which three variants are listed (G3673A, C6484T, G9041A). For each variant, MEDLINE abstracts are cited in reference (e.g., PMID: 16270629 for all 3 variants). In practice, PharmGKB provides <drug, allelic variant, article> relations. For the identification of allelic variants, we use the identifier from dbSNP (rsid) listed in PharmGKB. Drugs listed in PharmGKB are normalized with RxNorm, as was done for the drugs extracted from the literature.

3.5.2 Precision

Precision cannot be evaluated with this reference dataset, because PharmGKB is not exhaustive (i.e., other valid <drug, genetic variant> relations may be present in the reference set of PharmGKB abstracts, but not curated). In order to estimate precision, i.e., the ability of our approach to identify only pharmacogenomic relations, we selected a random sample of 180 relations extracted from the large set of MEDLINE abstract by our high-throughput approach. Each relation extracted was reviewed independently by two authors with expertise in medicine and bioinformatics. All differences were reconciled by consensus. We considered false positives those relations where the mutation does not correspond to the gene (e.g., non-human variants) and where the drug is not mentioned in a clinical context (e.g., *folate* used as a reagent).

4 Results

The contribution of each step of the mutation-centric approach applied to a large set of MEDLINE abstracts is briefly presented, followed by the evaluation of our approach in terms of recall and precision.

4.1 Application to a large set of MEDLINE abstracts

The number of abstracts selected at each step of our approach is shown in Figure 1. The initial PubMed search yielded 281,947 abstracts. Of these, 35,926 (12.7%) were identified by EMU as containing mention of some point mutation with its associated gene. A drug was identified by MetaMap and RxNorm in 63,027 abstracts (22.3%), for a total of 1970 unique drugs. (The number of unique mutations is not known as the mutations have not been normalized.) Overall, we found a total of 12,590 <drug, genetic variant> relations in 4,534 abstracts.

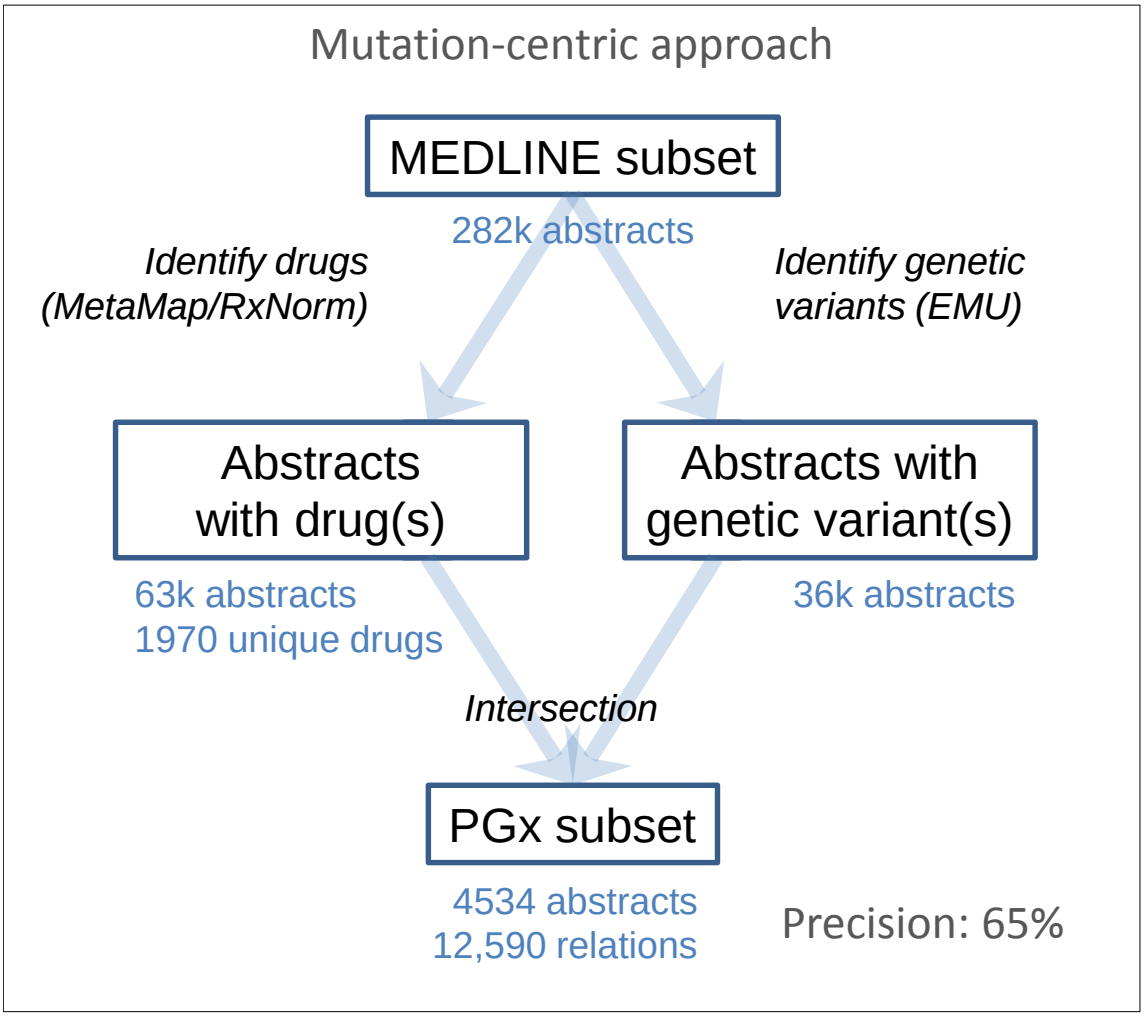

Figure 1. Mutation-centric approach applied to a large set of MEDLINE abstracts

4.2 Evaluation

4.2.1 Recall

The proportion of reference abstracts identified as relevant by our approach measures the recall of our approach. Of the 104 reference abstracts, 34 (33%) were identified by our approach as containing mention of both a drug and a mutation. Recall can also be measured for the <drug, genetic variant> relations from the reference documents, using an rsid from dbSNP as genetic variant identifier. Of such 441 reference relations, 57 (13%) were identified by our approach (with automatic or manual resolution to dbSNP). The 441 reference relations correspond to 85 unique drugs, 29 unique genetic variants and 420 unique <drug, genetic variant> relations. A failure analysis is presented in the discussion.

4.2.2 Precision

Of the 180 <drug, genetic variant> relations randomly selected from the 12,590 relations extracted from the large set of MEDLINE abstracts, 65% have been identified as true positives by manual review.

5 Discussion

We first point out salient elements of the results and discuss their significance, before comparing our approach to the relation-centric approach developed at Stanford. Then we review some of the remaining challenges in extracting pharmacogenomic relations. Finally we present the application of this work to the prioritization of pharmacogenomic relations for biocurators.

5.1 Findings and significance

With a recall of 33% for the abstracts and 13% for the <drug, genetic variant> relations in comparison to the reference abstracts curated in PharmGKB, the performance of our approach seems inadequate. We performed an analysis and identified the following reasons for failure, illustrated in Figure 2.

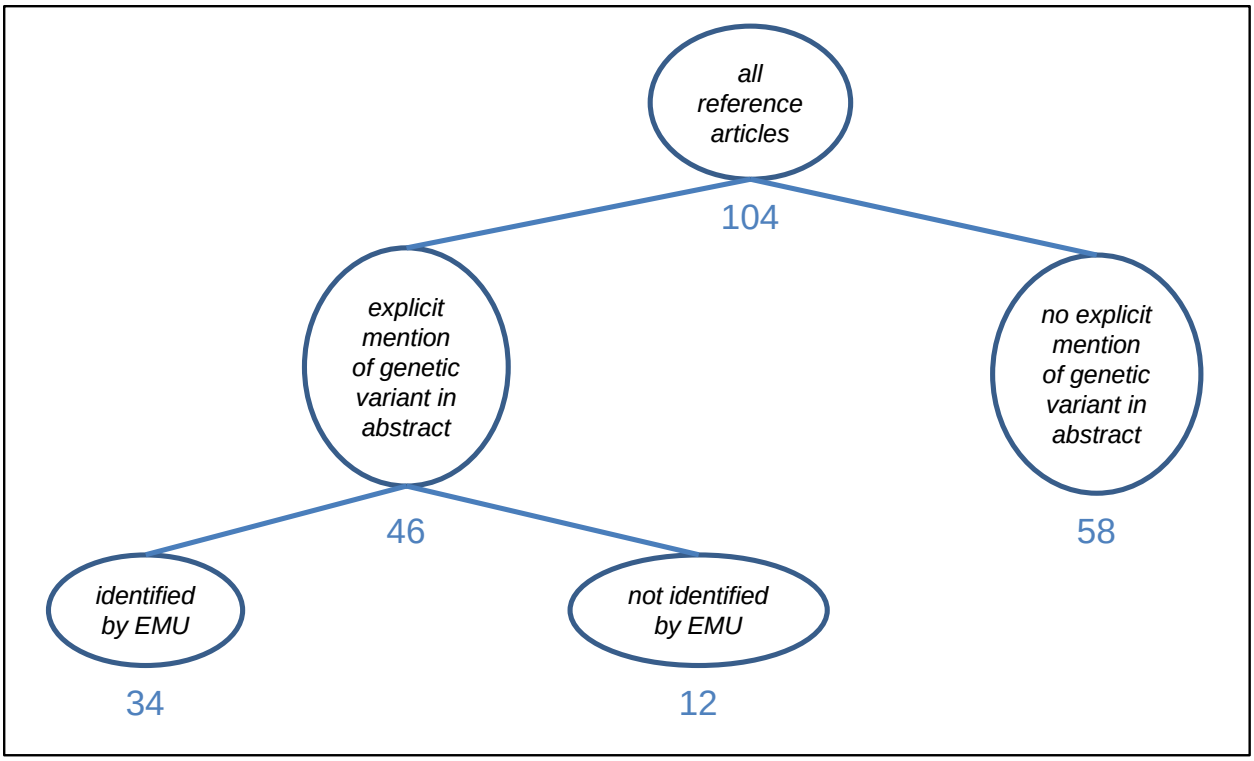

Figure 2. Failure analysis: Number of abstracts from the reference dataset for which an explicit mention of genetic variant is found in the abstract and identified by EMU.

Upon manual review of the 104 abstracts from the reference set, we determined that a genetic variant was explicitly mentioned in the abstract (or title) in only 46 cases (44%). In other words, in over half of the cases, biocurators have relied on information from full-text articles or external sources for extracting pharmacogenomic relations. (Issues in processing full-text articles are discussed later).

From the 46 abstracts whose abstracts contain explicit mentions of genetic variants, EMU failed to identify the mutation in 12 cases. For example, from the text fragment *the Gly49Arg389/Ser49Gly389 diplotype* in article PMID:12844134, EMU missed the two mutations: *Gly49Ser* (rs1801252) and *Arg389Gly* (rs1801253) in the human gene ADRB1. Although EMU handles concatenated mutations, here the concatenation pattern is unusual and ambiguous with more common patterns (confusion with *Gly49Arg* and *Ser49Gly*).

The recall observed on the reference dataset arguably corresponds to the lower bound of the performance of our method. The reference dataset contains an uncharacteristically large proportion of older abstracts (30 of the 104 abstracts published before 2000), in which the precise genotypic information is less likely to be available and where genetic variation is less likely to be expressed in a standard manner. When measured on the subset of the reference abstracts published in 2000 or after, the performance of our method increases significantly. In fact 34 of the 74 recent abstracts are identified as relevant, increasing the recall to 46% (from 33%).

5.2 Comparison to Stanford's relation-centric approach

5.2.1 The relation-centric approach

The Stanford group applied a relation-centric approach to extracting pharmacogenomic relations from MEDLINE [7]. They processed the whole MEDLINE dataset, parsed 87 million sentences in order to identify syntactic dependencies between two entities, one representing gene variation (e.g., *VKORC1 polymorphisms*) and the other related to a drug (e.g., *warfarin dose*) or phenotype (e.g., *thrombophlebitis*). They used a simple lexicon-based method for identifying genes, drugs and phenotypes in text, but a linguistically-motivated method for identifying associations between entities in text from the syntactic structure of the sentence. They created an ontology to organize and normalize the types of entities and relationships encountered. However, no attempt was made to systematically identify gene variants or to resolve them to reference databases. They extracted over 41,000 <gene variation entity, relationship, drug/phenotype entity> relations, with a precision of 88% (evaluated on a sample of 220 relations). The Stanford group shared with us the list of abstracts from which a pharmacogenomic relation had been identified, but not the relations themselves. Applied to our subset of 282,000 MEDLINE abstracts, the Stanford approach identified 2764 abstracts containing pharmacogenomic relations.

5.2.2 Contrasting the two approaches

The identification of the relation between genetic variants and drugs differs in two respects between the two approaches, namely in scope and extraction method. The mention of a genetic variant is identified directly in the mutation-centric approach, while the relation-centric approach first detects a gene name and then the indication of variation through a modifier (e.g., VKORC1 SNP). Drug identification relies on dictionaries in both cases, but the relation-centric approach identifies not only drugs, but also phenotypes related to genetic variants. Finally, the relation is approximated by simple co-occurrence within the abstract in the mutation-centric approach, while it is confined to a sentence and derived from its syntactic structure in the relation-centric approach.

5.2.3 Comparison

The comparison assesses whether a given abstract selected as the source of pharmacogenomic relations by one approach is also selected by the other. The relations themselves are not compared, because only the set of abstracts in which relations were identified by Stanford was made available to us.

Both approaches show limited recall. Of the 104 abstracts of the reference dataset, 34 (33%) were identified by our mutation-centric approach, while 36 (35%) were identified by the relation-centric approach. In both cases, recall is inadequate to support automatic biocuration. The advantage of both approaches, however, is that they are fully automated and can be used to scan the biomedical literature systematically.

Both approaches show state-of-the-art precision. In the absence of a gold standard for pharmacologic relations, both groups relied on manual review of a limited set of relations to evaluate the precision of their approach. Not surprisingly, more false positives are identified by our co-occurrence-based approach than by Stanford's approach where linguistic cues are required to support the relation.

However, with a precision of 65% for the mutation-centric approach and 88% for the relation-centric approach, the performance of both approaches reflects the state of the art in relation extraction.

Complementarity between the two approaches. Of the 104 abstracts of the reference dataset, eight abstracts (7.7%) are found by both approaches. The overlap between the two approaches is also limited on the set of 282,000 MEDLINE abstracts, where only 224 abstracts are identified by both methods, representing 4.9% of the abstracts identified by the mutation-centric approach and 8.1% of the abstracts identified by the relation-centric approach.

Given the differences between the two approaches, we did not expect a large overlap between the two result sets. However, the proportion of abstracts identified by both approaches is extremely limited. Since both approaches have reasonable precision, the two approaches are complementary and can be used jointly to help identify pharmacogenomic relations from the biomedical literature. For example, our mutation-centric approach identified 34 of the 104 reference abstracts, while Stanford's relation-centric approach identified 36, with an overlap of eight. Therefore the two approaches identified 62 distinct abstracts and the recall of the combined approaches on the reference abstracts is 60%, i.e., significantly higher than the recall of 35% obtained by each approach taken in isolation.

5.3 Remaining challenges

Processing full-text articles, sequence validation of DNA variants and resolution of genetic variants to reference databases are three aspects of the extraction of pharmacogenomic relations that remain particularly challenging.

5.3.1 Processing full-text articles

As mentioned earlier, we chose to process abstracts rather than full-text articles in this study, mainly because a majority of full-text articles are still not available in public access repositories, such as PubMed Central [22]. In fact, of the 104 articles in the reference dataset, only 33 (32%) are available there. Moreover, when available, full-text articles often need to be converted from PDF format, which is suboptimal as it may result in loss of the document structure.

Using our mutation-centric approach, we processed the 33 articles available in full text downloaded from PubMed Central. While 11 of these 33 articles (33%) had already been identified as a source of pharmacogenomic relations based on the abstract alone, five additional articles were identified when processing the full text, increasing recall to 48%. This limited experiment suggests that two thirds of the relevant articles can be identified based solely on the abstracts.

5.3.2 Sequence validation of DNA variants

EMU failed to provide sequence validation for 26 of the 34 mutations it identified from the reference dataset, because these mutations were described in reference to nucleotide sequences, rather than amino acid sequences. While we were able to confirm the validity of these 26 mutations manually using dbSNP as a reference, automatic sequence validation remains challenging for DNA mutations, because comparing nucleotide sequences (combinations of four nucleotides) is likely to yield incorrect associations more frequently than when comparing amino acid sequences (combinations of 20 amino acids). Overall, sequence validation could be obtained for less than one third of the mutations detected

by EMU. While we showed in earlier work that sequence validation significantly contributed to the precision of EMU [8], such requirement also significantly decreases recall. In practice, sequence validation was not required as part of the identification of genetic variants in this study.

5.3.3 Resolution of genetic variants to reference databases

References databases of point mutations, such as dbSNP, are still incomplete, i.e., not all genetic variants described in the biomedical literature have been recorded in dbSNP. Therefore, only a fraction of the genetic variants identified by EMU can be resolved into an entry in dbSNP and associated with an rsid. Examples of mutations automatically resolved to dbSNP by EMU include *Ala893Ser* identified in gene MDR1 from the article PMID:11503014 and resolved to the variant rs2032582. In contrast, EMU failed to resolve the DNA mutation *C3435T* in the same gene from PMID:10716719 (resolved manually to rs1045642). However, we showed that failure by EMU to resolve a given mutation to dbSNP (for a nucleotide or amino acid sequence) was not indicative of invalid mutation identification.

5.4 Application to support the curation of pharmacogenomic relations

Our approach to identifying pharmacogenomic relations was never envisioned as a solution to automatic curation of pharmacogenomic knowledge. The identification of pharmacogenomic relations is only one element of the development of a resource, such as PharmGKB, as biocurators generally collect additional information, including allele frequency and odds-ratios, in order to precisely characterize the influence of a given genetic variant on drug effect. However, we argue that high-throughput approaches, such as ours, can help support and prioritize biocuration efforts by providing enhanced information retrieval and quantification of the frequency of the pharmacogenomic relations.

Enhanced information retrieval. Biocuration efforts are limited by the resources of teams, such as PharmGKB. Therefore, biocurators typically restrict their effort to a small number of articles from selected journals and provide in-depth curation only for a limited set of genetic variants. In contrast, high-throughput approaches including our mutation-centric approach and Stanford's relation-centric approach can be used for scanning the literature systematically and regularly. Even if their recall is limited, the precision of these approaches is sufficient to make useful recommendations to biocurators. Recall can be increased by combining several complementary high-throughput approaches and by processing full-text articles when available. Gains in precision can be obtained through additional filtering (e.g., on journal, publication date and MeSH indexing). Moreover, our mutation-centric approach would help biocurators identify all genetic variants mentioned in the literature for a drug of interest.

Quantification of the frequency of the pharmacogenomic relations. One issue for biocurators is to prioritize the drugs and genetic variants on which to concentrate their efforts. High-throughput approaches can help perform automatic "surveillance" of the drugs and genetic variants discussed in the literature, as well as quantify the frequency of the pharmacogenomic relations for certain drugs or drug classes. For example, the top 15 drugs we identified in pharmacogenomic relations extracted from the 290,000 MEDLINE abstracts, participate in 3197 pharmacogenomic relations. Among these 15 drugs, three tyrosine kinase inhibitors, gefitinib, imatinib and erlotinib, account for 18% of the 3197 relations

(and 4.6% of all relations). This example illustrates how drug classes of interest can easily be screened for their association with genetic variants and possibly given a higher priority in the curation process.

6 Conclusion

Several approaches to identifying pharmacogenomic relations from the biomedical literature have been investigated recently. In contrast to methods relying on sophisticated NLP techniques (e.g., Stanford's relation-centric approaches), we propose a mutation-centric approach in which specific genetic variants, not only gene mentions, are identified in text and validated against reference sequences whenever possible. When evaluated against a reference set of abstracts from PharmGKB, our approach exhibited a recall of 33-46%, which is similar to the performance of relation-centric approaches. The precision of our approach is 65%. Moreover, we showed that mutation-centric and relation-centric approaches are complementary. This investigation identified three challenging aspects of the extraction of pharmacogenomic relations, namely processing full-text articles, sequence validation of DNA variants and resolution of genetic variants to reference databases, such as dbSNP. Given the limited performance of automatic approaches to identifying pharmacogenomic relations, the principal interest of these methods is their ability to process vast amounts of biomedical text automatically. Rather than a solution for the automatic curation of pharmacogenomic knowledge, we see these high-throughput approaches as tools to assist biocurators in the identification of pharmacogenomic relations of interest from the published literature.

7 Acknowledgments

The authors want to thank Adrien Coulet, Nigam H. Shah, Yael Garten, Mark Musen and Russ B. Altman (Stanford University) for sharing with us the list of MEDLINE abstracts in which they have identified pharmacogenomic relations using the relation-centric method reported in [7]. We would also like to thank Jim Mork for his help in the preparation of the MEDLINE dataset. This work was supported by the National Institutes of Health (NIH) through grant 1K22CA143148 (MGK, ED), and by the Intramural Research Program of the NIH, National Library of Medicine (OB, BR, DDF).

8 Bibliography

[1] S. Ananiadou, J. McNaught, Text mining for biology and biomedicine, Artech House, Boston, 2006.

[2] A.R. Aronson, F.M. Lang, An overview of MetaMap: historical perspective and recent advances, J Am Med Inform Assoc 17 (2010) 229-236.

[3] O. Bodenreider, The Unified Medical Language System (UMLS): integrating biomedical terminology, Nucleic Acids Res 32 (2004) D267-270.

[4] O. Bodenreider, L. Peters, RxNav: Browser and application programming interfaces for RxNorm, AMIA Annual Symposium, Washington, 2010, pp. 1330.

[5] E. Buyko, K. Hornbostel, U. Hahn, Extended study for extracting events between genes/proteins and drugs/active pharmaceutical ingredients (API) with JReX, PSB Workshop on Mining the Pharmacogenomics Literature, 2011, pp. Electronic proceedings.

[6] J.G. Caporaso, W.A. Baumgartner, Jr., D.A. Randolph, K.B. Cohen, L. Hunter, MutationFinder: a high-performance system for extracting point mutation mentions from text, Bioinformatics 23 (2007) 1862-1865.

[7] A. Coulet, N.H. Shah, Y. Garten, M. Musen, R.B. Altman, Using text to build semantic networks for pharmacogenomics, J Biomed Inform 43 (2010) 1009-1019.

[8] E. Doughty, A. Kertesz-Farkas, O. Bodenreider, G. Thompson, A. Adadey, T. Peterson, M.G. Kann, Toward an automatic method for extracting cancer- and other disease-related point mutations from the biomedical literature, Bioinformatics 27 (2010) 408-415.

[9] R.S. Epstein, T.P. Moyer, R.E. Aubert, O.K. DJ, F. Xia, R.R. Verbrugge, B.F. Gage, J.R. Teagarden, Warfarin genotyping reduces hospitalization rates results from the MM-WES (Medco-Mayo Warfarin Effectiveness study), J Am Coll Cardiol 55 (2010) 2804-2812.

[10] W.E. Evans, M.V. Relling, Pharmacogenomics: translating functional genomics into rational therapeutics, Science 286 (1999) 487-491.

[11] D.A. Flockhart, D. O'Kane, M.S. Williams, M.S. Watson, B. Gage, R. Gandolfi, R. King, E. Lyon, R. Nussbaum, K. Schulman, D. Veenstra, Pharmacogenetic testing of CYP2C9 and VKORC1 alleles for warfarin, Genet Med 10 (2008) 139-150.

[12] Y. Garten, R.B. Altman, Teaching computers to read the pharmacogenomics literature ... so you don't have to, Pharmacogenomics 11 (2010) 515-518.

[13] Y. Garten, A. Coulet, R.B. Altman, Recent progress in automatically extracting information from the pharmacogenomic literature, Pharmacogenomics 11 (2010) 1467-1489.

[14] T. Gerloff, M. Schaefer, A. Johne, K. Oselin, C. Meisel, I. Cascorbi, I. Roots, MDR1 genotypes do not influence the absorption of a single oral dose of 1 mg digoxin in healthy white males, Br J Clin Pharmacol 54 (2002) 610-616.

[15] J. Hakenberg, C. Plake, R. Leaman, M. Schroeder, G. Gonzalez, Inter-species normalization of gene mentions with GNAT, Bioinformatics 24 (2008) i126-132.

[16] T. Hernandez-Boussard, M. Whirl-Carrillo, J.M. Hebert, L. Gong, R. Owen, M. Gong, W. Gor, F. Liu, C. Truong, R. Whaley, M. Woon, T. Zhou, R.B. Altman, T.E. Klein, The pharmacogenetics and pharmacogenomics knowledge base: accentuating the knowledge, Nucleic Acids Res 36 (2008) D913-918.

[17] L. Hunter, Z. Lu, J. Firby, W.A. Baumgartner, Jr., H.L. Johnson, P.V. Ogren, K.B. Cohen, OpenDMAP: an open source, ontology-driven concept analysis engine, with applications to capturing knowledge regarding protein transport, protein interactions and cell-type-specific gene expression, BMC Bioinformatics 9 (2008) 78.

[18] T.E. Klein, R.B. Altman, N. Eriksson, B.F. Gage, S.E. Kimmel, M.T. Lee, N.A. Limdi, D. Page, D.M. Roden, M.J. Wagner, M.D. Caldwell, J.A. Johnson, Estimation of the warfarin dose with clinical and pharmacogenetic data, N Engl J Med 360 (2009) 753-764.

[19] T.E. Klein, J.T. Chang, M.K. Cho, K.L. Easton, R. Fergerson, M. Hewett, Z. Lin, Y. Liu, S. Liu, D.E. Oliver, D.L. Rubin, F. Shafa, J.M. Stuart, R.B. Altman, Integrating genotype and phenotype information: an overview of the PharmGKB project. Pharmacogenetics Research Network and Knowledge Base, Pharmacogenomics J 1 (2001) 167-170.

[20] K. Livingston, H. Johnson, K. Verspoor, L. Hunter, Leveraging Gene Ontology annotations to improve a memory-based language understanding system, IEEE Fourth International Conference on Semantic Computing (ICSC) (2010) 40-45.

[21] dbSNP, http://www.ncbi.nlm.nih.gov/projects/SNP/, (last accessed: 5/18/2012).

[22] PubMedCentral, http://www.ncbi.nlm.nih.gov/pmc/, (last accessed: 5/18/2012).

[23] S.J. Nelson, K. Zeng, J. Kilbourne, T. Powell, R. Moore, Normalized names for clinical drugs: RxNorm at 6 years, J Am Med Inform Assoc (2011).
[24] B. Rance, D. Demner-Fushman, T. Rindflesch, O. Bodenreider, Exploring automatic approaches to extracting pharmacogenomic information from the biomedical literature, PSB Workshop on Mining the Pharmacogenomics Literature, 2010, pp. Electronic proceedings.
[25] M.J. Rieder, A.P. Reiner, B.F. Gage, D.A. Nickerson, C.S. Eby, H.L. McLeod, D.K. Blough, K.E. Thummel, D.L. Veenstra, A.E. Rettie, Effect of VKORC1 haplotypes on transcriptional regulation and warfarin dose, N Engl J Med 352 (2005) 2285-2293.
[26] F. Rinaldi, G. Schneider, S. Clematide, Mining complex Drug/Gene/Disease relations in PubMed, PSB Workshop on Mining the Pharmacogenomics Literature, 2011, pp. Electronic proceedings.
[27] S.T. Sherry, M.H. Ward, M. Kholodov, J. Baker, L. Phan, E.M. Smigielski, K. Sirotkin, dbSNP: the NCBI database of genetic variation, Nucleic Acids Res 29 (2001) 308-311.
[28] L. Wang, H.L. McLeod, R.M. Weinshilboum, Genomics and drug response, N Engl J Med 364 (2011) 1144-1153.
[29] P. Zweigenbaum, D. Demner-Fushman, H. Yu, K.B. Cohen, Frontiers of biomedical text mining: current progress, Brief Bioinform 8 (2007) 358-375.