Published in final edited form as:
Pac Symp Biocomput. 2010 ; : 305–314.

IMPROVING THE PREDICTION OF PHARMACOGENES USING TEXT-DERIVED DRUG-GENE RELATIONSHIPS

YAEL GARTEN[§],
Stanford Biomedical Informatics Training Program, Stanford University, Stanford, CA 94305, USA

NICHOLAS P TATONETTI[§], and
Stanford Biomedical Informatics Training Program, Stanford University, Stanford, CA 94305, USA

RUSS B ALTMAN[†]
Departments of Bioengineering & Genetics Stanford University, Stanford, CA 94305, USA

Abstract

A critical goal of pharmacogenomics research is to identify genes that can explain variation in drug response. We have previously reported a method that creates a genome-scale ranking of genes likely to interact with a drug. The algorithm uses information about drug structure and indications of use to rank the genes. Although the algorithm has good performance, its performance depends on a curated set of drug-gene relationships that is expensive to create and difficult to maintain. In this work, we assess the utility of text mining in extracting a network of drug-gene relationships automatically. This provides a valuable aggregate source of knowledge, subsequently used as input into the algorithm that ranks potential pharmacogenes. Using a drug-gene network created from sentence-level co-occurrence in the full text of scientific articles, we compared the performance to that of a network created by manual curation of those articles. Under a wide range of conditions, we show that a knowledge base derived from text-mining the literature performs as well as, and sometimes better than, a high-quality, manually curated knowledge base. We conclude that we can use relationships mined automatically from the literature as a knowledgebase for pharmacogenomics relationships. Additionally, when relationships are missed by text mining, our system can accurately extrapolate new relationships with 77.4% precision.

1. Introduction

Individuals have variable response to drug treatment[1,2]. The assumption underlying personalized medicine and pharmacogenetics is that an individual's genotype can be used to predict variable drug response[3]. Understanding and describing this variation is an essential first step of personalized medicine[2,4,5]. Pharmacogenomics investigates how genes and their variation impact drug response. Such research has historically been pharmacogenetic, focusing on smalls set of genes or proteins[6]. However, in this new age of high throughput technologies, the research has become increasingly pharmaco*genomic*, involving multiple genes. Pharmacogenomics (PGx) knowledge has expanded rapidly, as we uncover new connections between genes and the effects of their variants on drug response. Simply determining the genes that are important for drug response is a critical requirement. Recently, high throughput technologies such as genome wide association studies have

[†]Corresponding author, russ.altman@stanford.edu.
[§]First author

Author Contributions
YG and NPT designed the study and carried out the analysis. YG wrote the manuscript. NPT contributed to the manuscript. RBA provided critical guidance on the study design, interpretation of results, and preparation of the manuscript.

yielded important new insights, however these technologies are plagued with high false positive rates, and statistical analysis of the data does not take advantage of existing biomedical knowledge[7,8]. Hansen et al. recently described an algorithm that uses existing knowledge in order to rank 12,460 genes in the genome on the basis of their potential relevance to a specific drug of interest[9]. This algorithm can prioritize genes in high throughput data sets, thus removing some false positives. The Hansen algorithm, called PGxPipeline, uses two knowledge bases of known drug-gene relationships, the Pharmacogenomics Knowledge Base (PharmGKB)[10] and DrugBank[11]. While these knowledge bases are extremely useful for pharmacogenomics they are also created manually by a staff of curators, who read the literature and annotate the PGx information. Thus, they are expensive to maintain and difficult to update, particularly as the volume of pharmacogenomic literature increases.

Therefore, there is a need for a scalable, inexpensive way to generate a comprehensive knowledge base of drug-gene relationships that can be used as input to the PGxPipeline algorithm. PharmGKB contained knowledge about 404 drugs and 585 genes at the time of download, however the literature contains an order of magnitude more of both drugs and genes[10]. Automatic methods of monitoring this space are necessary.

Text mining techniques allow us to survey the literature in an automated fashion, and extract information from the unstructured scholarly literature[12], into a structured format in a database. As described by Hunter and Cohen, today's interdisciplinary research scientist has an increasingly overwhelming amount of literature to assimilate[13]. Only through efficient text mining techniques can the data in the literature be extracted and rendered most useful. We previously described Pharmspresso[14], which performs the task of extracting pharmacogenomic relationships from sentences. In this work we combined Pharmspresso and PGxPipeline to assess the suitability of automatically derived knowledge in training a gene-ranking algorithm. Thus, we can compare the performance of the text-mining-based knowledge source to the curation-based knowledge source utilized by Hansen, et al. If successful, we can contemplate using a continuously updated and expanded network of drug-gene relationships as the literature expands. This will clearly improve the results as Hansen et al. showed that the performance of the ranking algorithm depends critically on the size of the set of input drug-gene relationships[9]. Additionally, this work also serves as an external validation for the Pharmspresso automated text-mining algorithm, which, until now, has only been validated on a small set of relationships.

2. Methods

2.1. Generation of corpus for text mining algorithm

We used the QUOSA desktop application[15] to automatically download the full text PDFs of all articles that were manually curated by PharmGKB curators. At the time the PharmGKB relationships were extracted from the database, 2202 articles had been curated. Of these, 1731 articles had available full text and this set was used as our corpus.

2.2. Generation of PharmGKB set of drug-gene relationships for training

We extracted drug-gene relationships from the core tables of PharmGKB[10], for all 1731 articles for which we had full text. Of these articles, 964 contained drug-gene relationships (the remaining 767 contained drug-disease or gene-disease relationships). A total of 1782 unique drug-gene relationships are found in these 1731 articles. For articles that contain more than one gene or more than one drug we relate all possible combinations of genes and drugs.

2.3. Generation of gold standard drug-gene relationships

The PharmGKB staff curate the literature at the article level, annotating the genes and drugs discussed in the article. In order to obtain a gold standard of manually curated relationships at the single relationship level, we used only those articles that mention at most one gene and one drug, to ensure a direct pharmacogenomic relationship between gene and drug. There were 916 such articles, containing a total of 682 unique drug-gene relationships. This was used as our gold standard for evaluation. We use this validation set derived from PharmGKB data to compare the performance of a classifier trained on PharmGKB data and a classifier trained on text-mined data; it is important to note that this validation set is not included when training either classifier.

2.4. Extraction of drug-gene relationships from text by Pharmspresso

We used the Pharmspresso system described previously[14] to extract the drug-gene relationships from the corpus. Pharmspresso extracts all sentences that contain co-occurrences of a gene and a drug (see Figure 2). To allow direct comparison of performance of the PGxPipeline using the text-mining-based drug-gene network to the curation-based drug- gene network, we used only genes and drugs found in the PharmGKB database when running the Pharmspresso algorithm: a total of 585 genes and 404 drugs.

2.5. Generation of scores for drug-gene relationships by PGxPipeline algorithm

The PGxPipeline algorithm, as presented by Hansen et al., assigns scores to 12,460 genes representing their propensity to modulate drug response for a query drug. Figure 3 illustrates this method. Briefly, the algorithm derives the scores by using two knowledge bases, (1) a gene-gene interaction network and (2) a drug-gene relationship network. These two networks are combined to make a gene-gene-drug network. For a query drug, the PGxPipeline scores each gene by comparing the query drug to drugs connected to that gene in the gene-gene-drug network, and assigning a score based on this similarity. Drug similarity is measured by both structural similarity and similarity of indications. As described in Hansen et al. structural drug similarity is defined as the Tanimoto coefficient of 166 structural features. The Tanimoto coefficient is also used as a metric to compare the similarity of the indication sets of two drugs. We trained a logistic regression classifier on the input positive examples and random negative examples using these two types of features, structural similarity and indication set similarity. The more similar the drugs in the local network of the gene are to the query drug, the higher the score the gene will receive.

The original PGxPipeline[9] used PharmGKB as a source of “genetic” drug-gene relationships, DrugBank as a source of “physical” drug-gene interactions, and the InWeb interactome[16] as a source of gene-gene interactions. InWeb is a protein-protein interaction network created with data from experiments. In this work, we explored the use of drug-gene relationships mined from the literature, as an alternative knowledge source to the algorithm, in place of using the combination of relationships provided by the PharmGKB curation process and the physical interactions from DrugBank.

2.6. Refining the source of negative relationships

In order to provide negative relationships to the logistic classifier during training, the PGxPipeline matches each positive relationship from its gold standard with relationships between that same drug and three randomly selected genes. The set of genes it samples from is the entire set of genes in InWeb, PharmGKB, and DrugBank (12,460 genes). The PGxPipeline knowledge base contains approximately 400 drugs and 12,460 genes. However, only approximately 1,000 of those genes have relationships (genetic or physical) with drugs. Given a gene chosen at random from the set 12,460 genes, the chance that the gene will have

any drug relationships is quite low. The consequence of this is that it can be relatively easy to differentiate between a positive and negative relationship, since most negative examples have no important relationships with any drugs.

In order to make our classification task more challenging we select negative examples from among known pharmacogenes—genes that in fact have at least some known relationship with a drug. Therefore, we replaced the pool of genes from which negative examples are selected with only the set of genes that exists in PharmGKB (585 genes), a much smaller gene set of known pharmacogenes. This allows us to more stringently evaluate the classifier while still maintaining the power to predict potential drug relationships with unknown pharmacogenes.

2.7. Comparison of the two drug-gene knowledge sources: Curated versus Text-Mined

To facilitate comparison between the text-mining-based classifier and the curation-based classifier we trained a logistic regression classifier in a similar manner, and validated with fivefold cross-validation. Therefore, in each of the folds, all knowledge about the relationships in the validation set (1/5 of the data) is dropped from the training set (the 4/5 of the data used for training). The performance of each classifier on this task is a metric of how accurate the model is in classifying known pharmacogenetic relationships.

2.8. Using text-mining-derived relationships combined with PGxPipeline scores to extrapolate; discovering additional drug-gene relationships

Pharmspresso identified 5,312 pharmacogenomic relationships, PharmGKB contained 1,782 relationships, with an overlap of 1,157 between the two sources (Figure 4). As expected and previously described[14], Pharmspresso is a very sensitive test for pharmacogenomic relationships, while PharmGKB is a highly specific one. There are 625 relationships in PharmGKB that Pharmspresso does not identify when searching for co-occurrence at the sentence level (the lexical names of the gene and drug may not occur in the same sentence). To test whether we can use the PGxPipeline scores to recognize true relationships not directly found in literature by Pharmspresso, we did the following: We trained the classifier with the 5,132 drug-gene relationships found by text-mining and applied the classifier to all of the 625 drug-gene relationships in PharmGKB that were not found by text-mining, to get a pharmacogene score for each relationship. For comparison we also applied the classifier to a randomly generated set of drug-gene relationships to get a pharmacogene score for each relationship. We then investigated our ability to use the pharmacogene score to distinguish between the relationships that were in PharmGKB versus the randomly created relationships. (To find relationships in region titled “Extrapolated Knowledge” in Figure 4.)

2.9. External validation: New relationships registered by the PharmGKB staff

During the time since we first downloaded the pharmacogenomic relationships from PharmGKB an additional 1,462 articles were curated resulting in an additional 1,636 drug-gene relationships. This set of relationships was used as an external validation set. For each drug-gene relationship we used the trained classifier to score the relationship and randomly sampled three more genes to pair with the drug as a source of negative relationships.

3. Results

3.1. Comparison of the two drug-gene knowledge sources: Curated versus Text-Mined

To evaluate the use of a text-mining based network as a pharmacogenomic relationship knowledge base we compared the performance of the text-mining-based classifier with that of the curation-based classifier (Methods 2.7) using 5-fold cross validation on the gold standard set of drug-gene relationships (Methods 2.3). We find that the text-mining-based

classifier out-performs the curation- based classifier, with receiver operator characteristic (ROC) curves with area under the curve (AUC) of 0.701 and 0.672 respectively (see Figure 5). Besides having an overall AUC that is slightly higher, the text-mining- based classifier achieves high sensitivity in the region of high specificity (FPR ≤ 0.2). Achieving a greater AUC in this area alone is often desirable by experimentalists as the algorithm can ensure a very low false positive rate, even though it may not have high recall. In addition we tested the two classifiers under the exact conditions described by Hansen (negative set genes selected from InWeb and broader definition of gold standard from PharmGKB data). This yields ROC curves with AUC values of 0.814 and 0.799 for the curation-based classifier and the text-mining based classifier respectively, and so under those conditions the performance of the two classifiers is comparable. The 0.814 AUC of the curation-based classifier is slightly lower than the 0.82 as reported by Hansen et al., presumably because the input knowledgebase is smaller—it is based on the subset of 1731 articles for which we obtained full text to allow fair comparison with Pharmspresso.

3.2. Using text-mining-derived relationships combined with PGxPipeline scores to extrapolate; discovering additional drug-gene relationships

We observe that there are relationships in PharmGKB that Pharmspresso does not discover when searching for co-occurrence at the sentence level and thus not used to train the text-mining-based classifier. We explored whether we can detect these relationships by using the scores assigned by the classifier. As described in Section 2.8, we selected a balanced set of positive and negative relationships and tested which relationships lie in the region titled “Extrapolated Knowledge” in Figure 4 (relationships positively scored by the classifier, not found by text-mining, and in PharmGKB). We call these “extrapolated” since they were not identified by the text-mining algorithm and so not part of the input knowledge base of drug-gene relationships. We validated against the PharmGKB relationships and found 3.44 fold enrichment (134/39) with a cutoff score of zero. That is, of the set of 173 relationships that score positively, we have 134 true positives that are in PharmGKB and presumably 39 false positives, a false discovery rate (FDR) of 22.5%.

Figure 6 describes the contribution of the local network to the score for a given pharmacogenomic relationship extrapolated from the text-mining-derived relationships by the text-mining-based classifier, for three examples. These examples represent a known pharmacogenetic relationship (they appear in PharmGKB) where the drug and gene did not co-occur at the sentence level in the literature, yet the text-mining-based classifier assigns the relationship a high score.

Figure 6A shows the underlying evidence for the predicted relationship between the drug trimipramine and gene SLC6A4, based on the text-mining-derived network of drug-gene relationships, and the gene-gene interactions in InWeb. SLC6A4 is a sodium-dependent serotonin transporter. The drug trimipramine is a tricyclic antidepressant[17]. The support for the predicted relationship between SLC6A4 and trimipramine stems from the similarity to the two drugs directly related to the gene, in the text-mining-derived network (mirtazapine and clomipramine), and from the similarity to the drug carbamazapine, which is related in the text-mining-derived network to the CALR gene, found to interact with SLC6A4 in the InWeb network. Mirtazapine is an antidepressant used for the treatment of moderate to severe depression, and has a very similar set of indications as trimipramine [18]. Of the drugs that co-occur in the literature with SLC6A4, the one that is most similar in structure to trimipramine is clomipramine. Of the drugs related to the CALR gene in the text-mining-derived network, the one that is most structurally similar to trimipramine, the query drug, is carbamazapine. It is also the most similar in its indications: both carbamazapine and trimipramine are used to treat depression, the indirect connection to carbamazapine via CALR boosts the prediction[10].

Figure 6B shows the support for the relationship between doxorubicin and BAK1, a BCL2-antagonist/killer 1. The InWeb interactome connects BAK1 to two genes, TP53 and BCL2, each of which has a literature co-mention with docetaxel, an anti-mitotic chemotherapy medication. Both doxorubicin and docetaxel are cancer treatments as well, and so the similarity of indications plays a role in uncovering the relationship between BAK1 and doxorubicin. Doxorubicin is a type of anthracycline, which is the most active group of cytotoxic agents for the treatment of breast cancer. Docetaxel with anthracyclines are sometimes used together and share structural similarity[19]. BAK1 "borrows" drug relationships from its neighbors to boost the likelihood of sharing a relationship with doxorubicin.

Figure 6C shows the predicted relationship between diltiazem and CYP2C8. Diltiazem is a calcium channel blocker, a member of the benzothiazepine class that reduce blood pressure through vasodilation[20]. It is used to treat hypertension and rhabdomyolysis, as is verapamil. InWeb connects CYP2C8 to 3 genes: ALDH7A1, ALDH2, and PON1. Each of these interact with drugs that have substantial structural or indication overlap with diltiazem[10]. CYP2C8 itself is found in the literature with 2 drugs; pitavastatin has the highest indications similarity to diltiazem and repaglinide has the highest structural similarity to diltiazem.

3.3. External Validation of the text-mining-based pharmacogene classifier

As an external validation of the text-mining-based classifier 1,636 drug-gene relationships added to the PharmGKB after we established the training set as well as three times that many randomly chosen drug-gene relationships were scored. The ROC curve has an area under the curve of 0.78, as shown in Figure 7. The text-mining-based classifier had comparable performance to the curation-based classifier on the same external validation set, which produced a ROC curve with an AUC of 0.8.

There are relationships found by Pharmspresso that do not appear in PharmGKB. We scored each of these relationships by leaving out the knowledge about the relationship during training of the text-mining based classifier. The relationships that receive positive scores appear in the intersection of the green and blue circles of Figure 4 – that is, the intersection of regions "Text-mining relationships" and "Text-mining based high scoring relationships". Within this intersection, those that do not appear in PharmGKB are titled "Putative Relationships" and are sent to the curators for potential insertion into PharmGKB. For example, the relationship between *CYP3A5* and cyclosporine scored highly, was not in the original set of PharmGKB relationships (thus appears in region titled "Putative Relationships" in Figure 4), and in fact PharmGKB now has three articles supporting this relationship [21–23].

4. Discussion

In this work, we explored the use of a text-mining-derived network of drug-gene relationships, as a knowledge base to replace human-curated literature relationships in the PGxPipeline. The PGxPipeline uses the knowledge base to predict pharmacogenes for an input query drug. While the human curated data are high quality, they are much less abundant. In this application, it is apparent that the improved coverage afforded by automatic detection outweighs the introduction of noise and errors because of imperfect text-mining extraction. Of course, the benefits are substantial: curation is a very expensive process (in terms of time and money), whereas text-mining is inexpensive and scalable[14]. In addition, PharmGKB staff curators often only read the abstract of articles because of the large volume of papers they must annotate. Abstracts do not necessarily contain the pharmacogenomic drug-gene relationship reported by the article, whereas the Pharmspresso system analyzes the full text of an article.

The task of Pharmspresso is really to identify relationships between genes and drugs, within the small scope of a single sentence. This is not what PharmGKB curators have been tasked with; they curate articles with respect to the genes and drugs that are mentioned without specifically asserting which genes and drugs relate. Therefore, it is not surprising that when using the high-quality gold standard the text-mining-based classifier actually performs slightly better than the curation-based classifer (0.701 AUC vs. 0.672 respectively, Figure 5). These results demonstrate that the drug-gene network derived by Pharmspresso can be used in place of manually curated data in the PGxPipeline algorithm, which may allows us to enlarge the drug-gene network to millions of articles in the scientific literature. Our results also provide an independent, large-scale, external validation of the usefulness and accuracy of Pharmspresso. Pharmspresso had previously been validated on a small evaluation set. Finally, we have demonstrated that the scores assigned by the PGxPipeline can be used to detect new relationships.

The Hansen et al. algorithm relied on a manually-curated network of pharmacogenomic drug-gene relationships derived from experimental or clinical data, as reported in the literature. Our results show, however, that a text-mined sentence co-occurrence drug-gene network can perform as well and even better under some circumstances. We acknowledge that the co-occurrence drug-gene network contains noise. Nonetheless, it is more likely that a co-occurrence is a meaningful relationship than a random one. Text mining allows us to generate a large network of relationships, thereby increasing the signal-to-noise ratio. This implies that the likelihood of a pharmacogenomic drug-gene relationship increases in proportion to the number of similar drugs that co-occur in the literature with the gene or genes in its pathway. We can therefore expect that as our knowledge base increases by mining more pharmacogenomic articles, so will our power to predict pharmacogenomic relationships. Additionally, because the method is general, this logic may apply to other types of pharmacogenomic relationships such as drug-SNP relationships, a hypothesis which we are currently investigating.

4.1. Limitations

While we have shown that we can predict pharmacogenomic drug-gene relationships based on a corpus of pharmacogenomic articles, it not yet clear that this methodology will work for other interesting biological problems, such as deriving drug-SNP relationships or gene regulatory networks. The generality of our methodology may be limited since our analysis is based on a corpus that is highly enriched for pharmacogenomics articles. We plan to investigate how dependent the performance of the algorithm is on this specialized corpus. One other limitation of using simple co-occurrence is the inability to derive the type of drug-gene relationship text-mined from the literature[24]. For example, it would be advantageous to know if a drug and gene have a positive or negative relationship or whether the gene is pharmacokinetic or pharmacodynamic for the drug. This type of characterization of edges in the network requires more sophisticated text mining.

Our next step is to expand the corpus of articles available to Pharmspresso to include a larger pharmacogenomics literature. Methods that classify publications likely to contain pharmacogenomic information, such as MScanner, can be used to filter Medline in order to identify pharmacogenomic articles[25]. This expansion of the drug-gene relationship network will greatly improve the performance of the PGxPipeline. The PGxPipeline relies on other types of relationship networks in addition to the drug-gene network, namely a gene-gene network and a drug-disease network. Mining these relationships from the literature may also increase the predictive power of the algorithm as well as keep the knowledge base scalable and up-to-date. Finally, we plan to incorporate high-scoring predictions into the curation pipeline at PharmGKB, to prioritize these predictions for curator review and subsequent insertion into the knowledge base.

Pharmacogenomics is not only concerned with the important genes but also with their particular variations that impact drug response. For example, variations in the *VKORC1* and *CYP2C6* genes are critical for determining warfarin dose, and can be used to predict the optimal dose of warfarin[26–29]. The Pharmspresso algorithm can detect genetic variations, and can be used to create a network linking specific variations to specific drugs[14]. Such a network might be useful in refining the PGxPipeline to weight pharmacogene predictions based on this additional knowledge source. The text-mining-based PGxPipeline classifier produced a substantial number of high scoring drug-gene relationships that were not found to be in PharmGKB. A high-throughput biological assay could be employed to test these relationships for their validity.

Acknowledgments

YG and NPT are supported by the Graduate Training in Biomedical Informatics grant (T15 LM007033) from the National Library of Medicine. YG and RBA are supported by NIH/NIGMS Pharmacogenetics Research Network and Database and the PharmGKB resource (NIH U01GM61374). The authors would like to thank R. Whaley for PharmGKB data. We thank our anonymous reviewers for their constructive comments.

References

1. Evans, Mcleod. Pharmacogenomics — Drug Disposition, Drug Targets, and Side Effects TV. Engl J Med. 2003; 348:538.
2. Swen, et al. Translating Pharmacogenomics: Challenges on the Road to the Clinic. PLoS Med. 2007; 4:e209. [PubMed: 17696640]
3. Weiss, et al. Creating and evaluating genetic tests predictive of drug response. Nature reviews Drug discovery. 2008; 7:568.
4. Davis, et al. The microeconomics of personalized medicine: today's challenge and tomorrow's promise. Nature reviews Drug discovery. 2009; 8:279.
5. Kirchheiner, et al. Pharmacogenetics-based therapeutic recommendations - ready for clinical practice? Nature reviews Drug discovery. 2005; 4:639.
6. Goldstein, et al. Pharmacogenetics goes genomic. Nat Rev Genet. 2003; 4:937. [PubMed: 14631354]
7. Dollery CT. Beyond genomics. Clin Pharmacol Ther. 2007; 82:366–70. [PubMed: 17851575]
8. McCarthy MI, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet. 2008; 9:356–69. [PubMed: 18398418]
9. Hansen NT, Brunak S, Altman RB. Generating genome-scale candidate gene lists for pharmacogenomics. Clin Pharmacol Ther. 2009; 86:183–9. [PubMed: 19369935]
10. Klein TE, et al. Integrating genotype and phenotype information: an overview of the PharmGKB project. Pharmacogenetics Research Network and Knowledge Base. The Pharmacogenomics Journal. 2001; 1:167–70. [PubMed: 11908751]
11. Wishart DS, et al. DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Research. 2008; 36:D901–6. [PubMed: 18048412]
12. Müller HM, Kenny EE, Sternberg PW. Textpresso: an ontology-based information retrieval and extraction system for biological literature. PLoS Biol. 2004; 2:e309. [PubMed: 15383839]
13. Hunter L, Cohen KB. Biomedical language processing: what's beyond PubMed? Mol Cell. 2006; 21:589–94. [PubMed: 16507357]
14. Garten Y, Altman RB. Pharmspresso: a text mining tool for extraction of pharmacogenomic concepts and relationships from full text. BMC Bioinformatics. 2009; 10 (Suppl 2):S6. [PubMed: 19208194]
15. Quosa. The total solution for efficiently managing and monitoring scientific literature. 2009. http://www.quosa.org
16. Lage K, et al. A human phenome-interactome network of protein complexes implicated in genetic disorders. Nat Biotechnol. 2007; 25:309–16. [PubMed: 17344885]

17. Hebert C, Habimana A, Elie R, Reader TA. Effects of chronic antidepressant treatments on 5-HT and NA transporters in rat brain: an autoradiographic study. Neurochem Int. 2001; 38:63–74. [PubMed: 10913689]

18. Marcus SC, Hassan M, Olfson M. Antidepressant switching among adherent patients treated for depression. Psychiatric services (Washington, DC). 2009; 60:617–23.

19. Jones S, et al. Docetaxel With Cyclophosphamide Is Associated With an Overall Survival Benefit Compared With Doxorubicin and Cyclophosphamide: 7-Year Follow-Up of US Oncology Research Trial 9735. J Clin Oncol. 2009; 21:1177–83. [PubMed: 19204201]

20. Neuvonen PJ, Niemi M, Backman JT. Drug interactions with lipid-lowering drugs: mechanisms and clinical relevance. Clin Pharmacol Ther. 2006; 80:565–81. [PubMed: 17178259]

21. Kreutz R, et al. CYP3A5 genotype is associated with longer patient survival after kidney transplantation and long-term treatment with cyclosporine. The Pharmacogenomics Journal. 2008; 8:416–22. [PubMed: 18180803]

22. Kreutz R, et al. The effect of variable CYP3A5 expression on cyclosporine dosing, blood pressure and long-term graft survival in renal transplant patients. Pharmacogenetics. 2004; 14:665–71. [PubMed: 15454731]

23. Fanta S, et al. Pharmacogenetics of cyclosporine in children suggests an age-dependent influence of ABCB1 polymorphisms. Pharmacogenet Genomics. 2008; 18:77–90. [PubMed: 18192894]

24. Ahlers, CB.; Fiszman, M.; Demner-Fushman, D.; Lang, F-M.; Rindflesch, TC. Extracting semantic predications from Medline citations for pharmacogenomics; Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing; 2007. p. 209-20.

25. Poulter GL, Rubin DL, Altman RB, Seoighe CM. Scanner: a classifier for retrieving Medline citations. BMC Bioinformatics. 2008

26. Rieder, et al. Effect of VKORC1 Haplotypes on Transcriptional Regulation and Warfarin Dose. N Engl J Med. 2005; 352:2285. [PubMed: 15930419]

27. Takeuchi, et al. A genome-wide association study confirms VKORC1, CYP2C9, and CYP4F2 as principal genetic determinants of warfarin dose. PLoS Genet. 2009; 5:el000433.

28. Rajanayagam, Rajanayagam. Pharmacogenetics: Optimizing warfarin therapy. Nature Reviews Cardiology. 2009; 6:324.

29. Cooper GM, et al. A genome-wide scan for common genetic variants with a large influence on warfarin maintenance dose. Blood. 2008; 112:1022–7. [PubMed: 18535201]

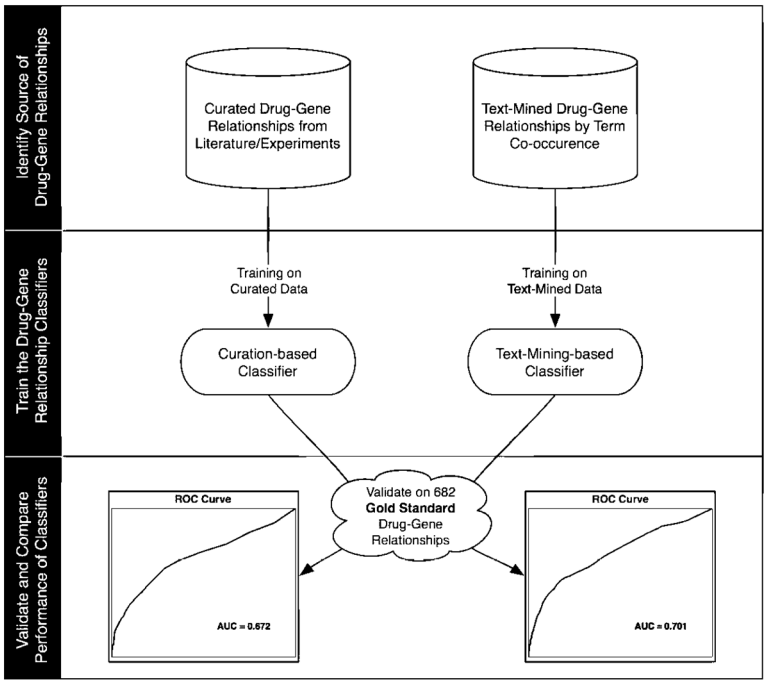

Figure 1.

Methods overview: A knowledge base of drug-gene relationships is extracted from a curated source (PharmGKB and DrugBank) as well as from an automatic text- mining source (Pharmspresso). One classifier is trained using each of the two types of knowledge sources and then validated against a gold standard set of drug-gene relationships to allow comparison of the two sources.

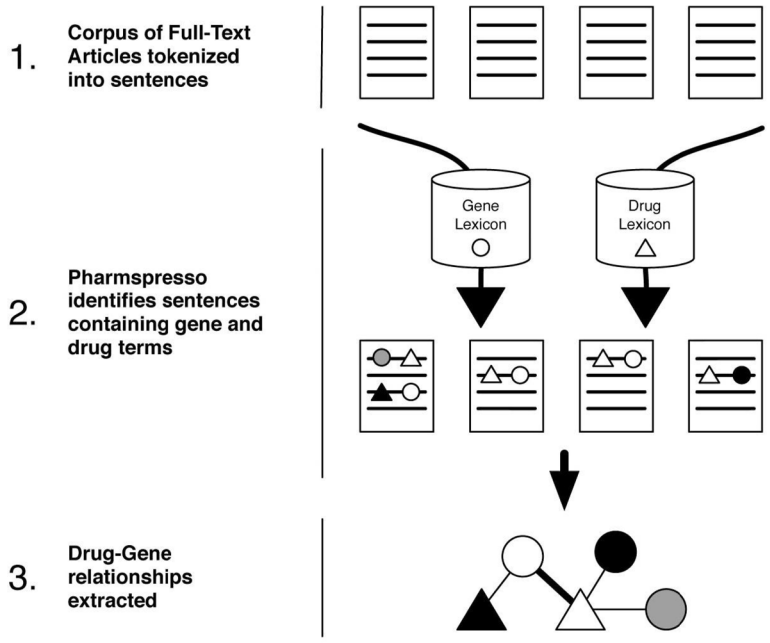

Figure 2.

Description of Pharmspresso system for relationship extraction at the sentence level. A corpus of full-text articles is first tokenized into sentences. Pharmspresso then marks up the sentences by identifying terms associated with genes and drugs. A drug-gene network is then created by drawing edges between genes and drugs that co-occur at the sentence level. The width of the edge corresponds to the number of articles that support the relationship.

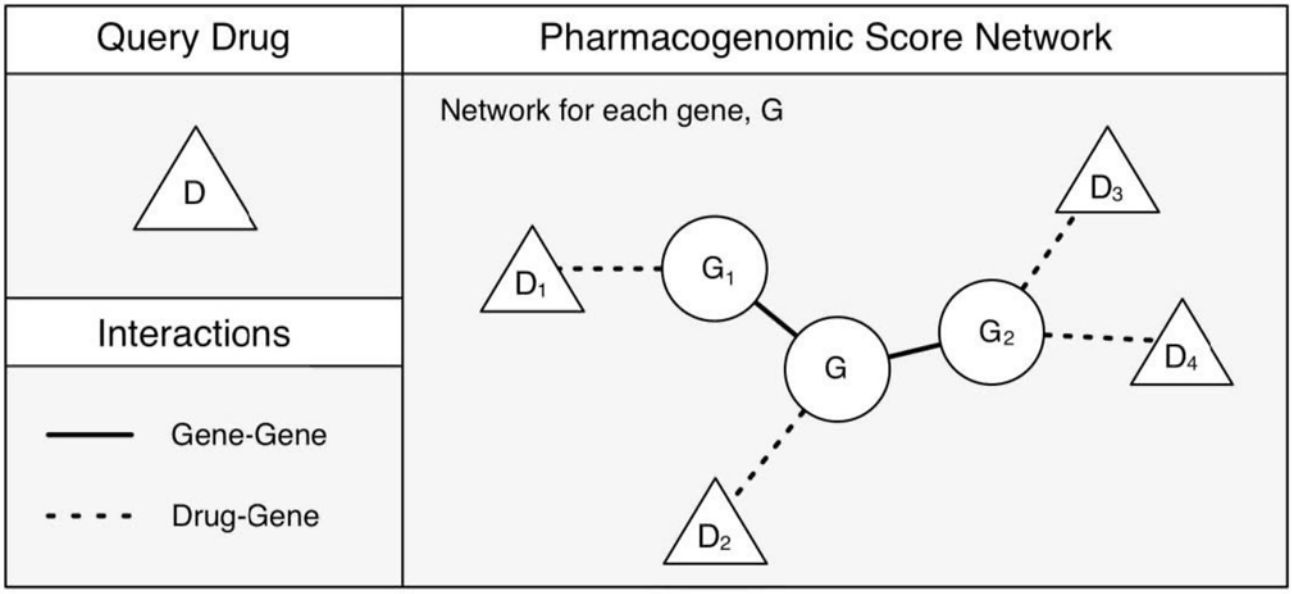

Figure 3.
The Pharmacogenomics Pipeline. Given a drug, D, each gene in the genome is scored based on the similarity of the neighboring drugs to the query drug. A neighboring drug may interact directly with the gene (D2) or indirectly (D1, D3, D4) through neighboring genes (G1, G2).

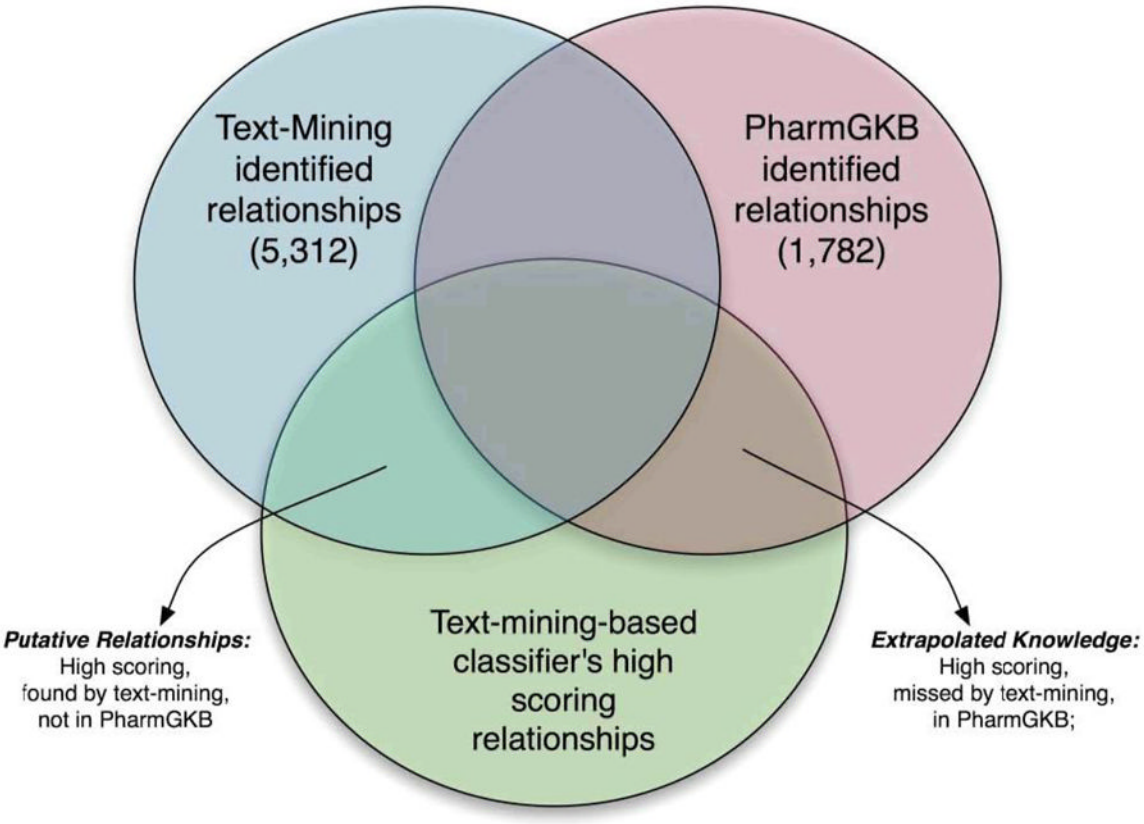

Figure 4.

The intersection of drug-gene interactions identified by Pharmspresso text-mining or by PharmGKB curators, and those interactions receiving high scores when applying the text-mining- based classifier. Pharmspresso identified 5,312 pharmacogenomic interactions, PharmGKB contained 1782 interactions, with an overlap of 1.157 between the two sources.

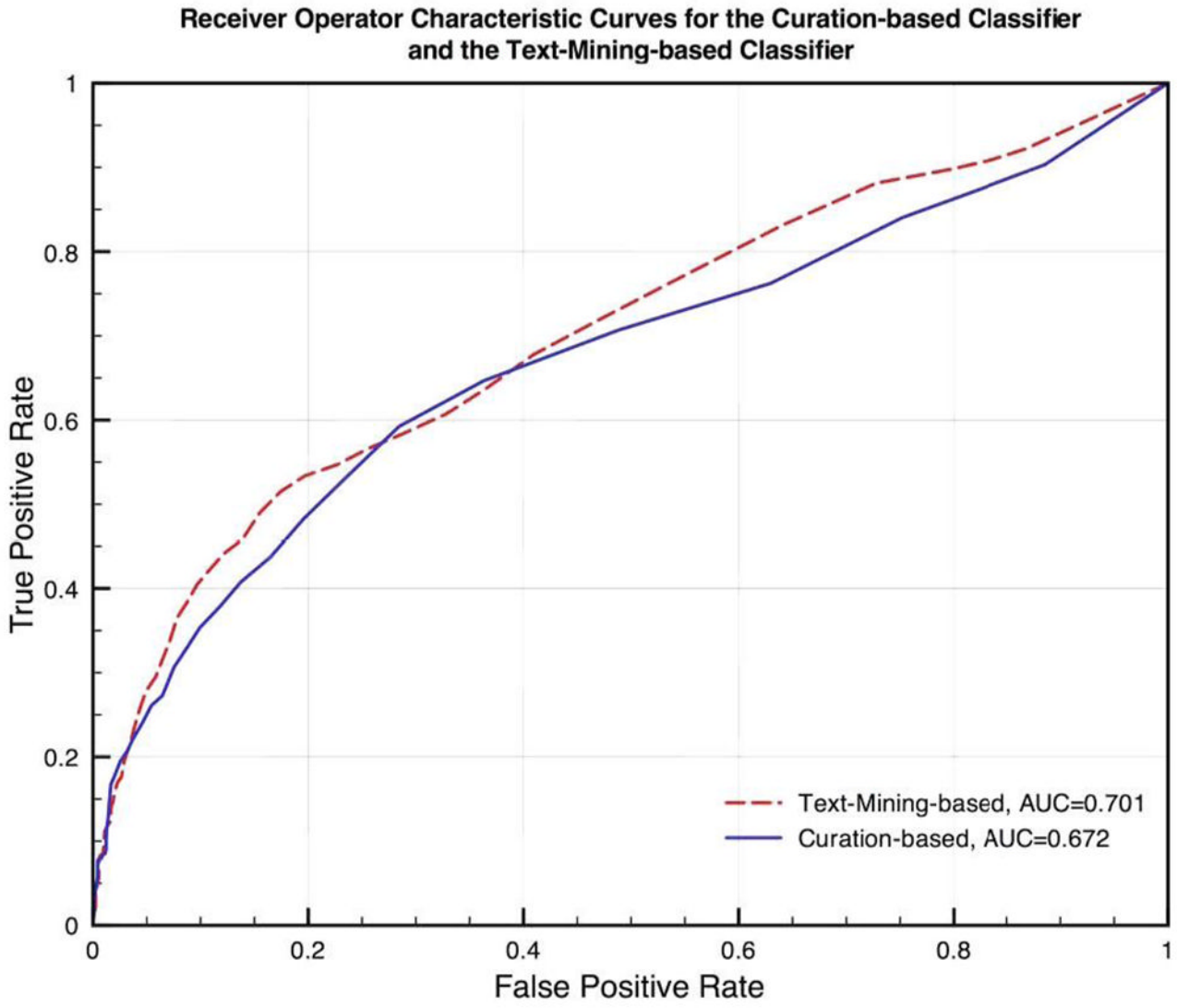

Figure 5.
The ROC curves for the curation-based classifier and text- mining-based classifier validated on the gold standard. The text- mining-based classifier out-performs the curation-based classifier.

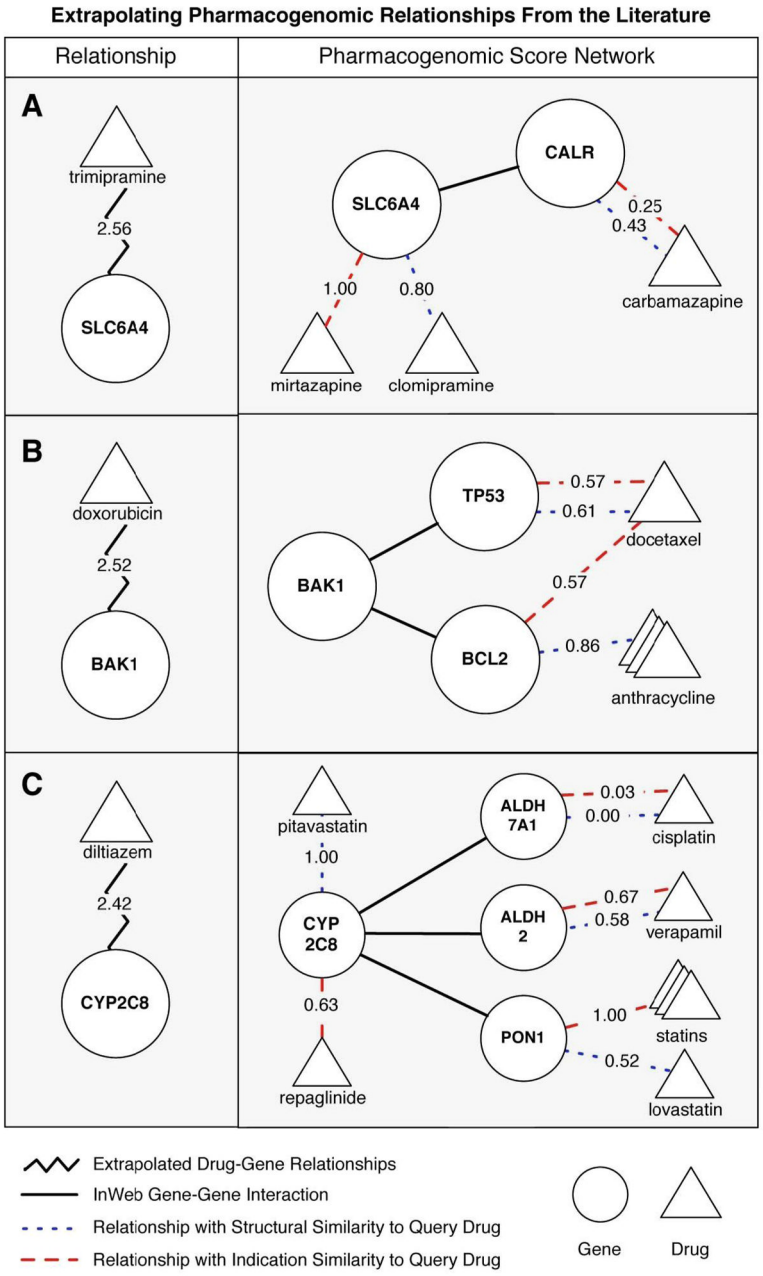

Figure 6.

Examples of extrapolation of drug-gene interactions using the text-mining- based PGxPipeline classifier. All examples are in fact found in PharmGKB; meaning there is literature support for these relationships recorded manually by curators. Although Pharmspresso misses them, they are recovered by the PGxPipeline scoring mechanism. Zigzag line: suggested, positive-score interaction (left panel). This interaction is not found directly in the literature by Pharmspresso, but receives a positive PGxPipeline score (score appears on the line). Solid lines: gene-gene relationships from InWeb. Dashed/dotted lines: drug-gene relationship found in literature by Pharmspresso. Dashed red - indication similarity, Dotted blue - structural similarity. The score shown on the edge represents the similarity score of the edge's drug, to the query drug.

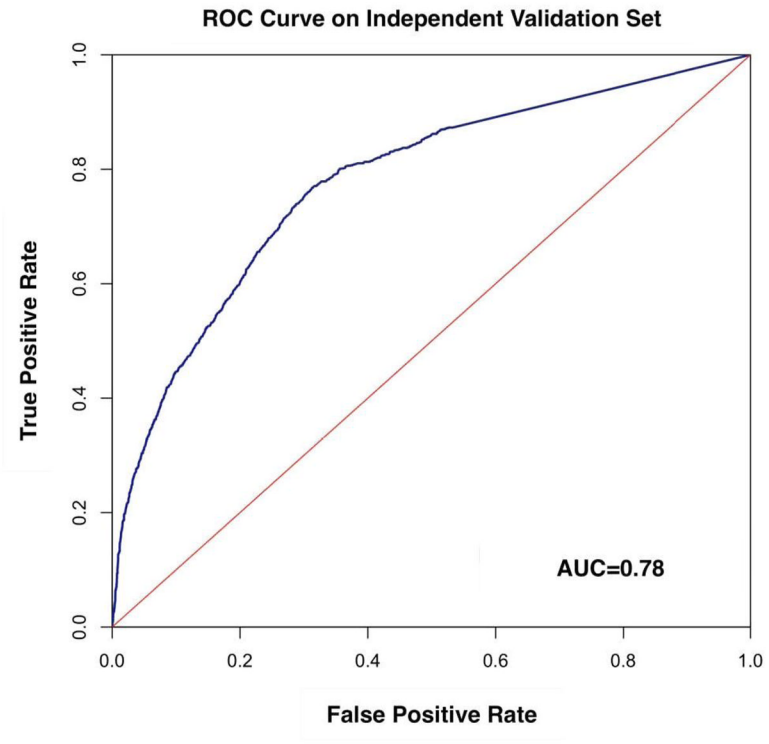

Figure 7.
The ROC curve for the literature-based classifier on the external validation set of 1,636 drug-gene interactions not included in the training set. This performance was achieved under the same conditions as presented in the Hansen paper.