COMBINING HETEROGENOUS DATA FOR PREDICTION OF DISEASE RELATED AND PHARMACOGENES

CHRISTOPHER S. FUNK*, LAWRENCE E. HUNTER, and K. BRETONNEL COHEN

Computational Bioscience Program, University of Colorado School of Medicine,
Aurora, CO 80045, USA
**E-mail: Christopher.Funk@ucdenver.edu, Larry.Hunter@ucdenver.edu, Kevin.Cohen@gmail.com*

Identifying genetic variants that affect drug response or play a role in disease is an important task for clinicians and researchers. Before individual variants can be explored efficiently for effect on drug response or disease relationships, specific candidate genes must be identified. While many methods rank candidate genes through the use of sequence features and network topology, only a few exploit the information contained in the biomedical literature. In this work, we train and test a classifier on known pharmacogenes from PharmGKB and present a classifier that predicts pharmacogenes on a genome-wide scale using only Gene Ontology annotations and simple features mined from the biomedical literature. Performance of F=0.86, AUC=0.860 is achieved. The top 10 predicted genes are analyzed. Additionally, a set of enriched pharmacogenic Gene Ontology concepts is produced.

1. Introduction

One of the most important problems in the genomic era is identifying variants in genes that affect response to pharmaceutical drugs. Variability in drug response poses problems for both clinicians and patients.[1] Variants in disease pathogenesis can also play a major factor in drug efficacy.[2,3] However, before variants within genes can be examined efficiently for their effect on drug response, genes interacting with drugs or causal disease genes must be identified. Both of these tasks are open research questions.

Databases such as DrugBank[4] and The Therapeutic Target DB[5] contain information about gene-drug interactions, but only The Pharmacogenomics Knowledgebase (PharmGKB)[6] contains information about how variation in human genetics leads to variation in drug response and drug pathways. Gene-disease variants and relationships are contained in Online Mendelian Inheritance in Man (OMIM),[7] the genetic association database,[8] and the GWAS catalog.[9] Curated databases are important resources, but they all suffer from the same problem: they are incomplete.[10] One approach to this problem is the development of computational methods to aid in database curation. We explore here a method that takes advantage of the large amount of information in the biomedical literature that is waiting to be exploited.

Having a classifier that is able to predict as-yet-uncurated pharmacogenes would allow researchers to focus on identifying the variability within the genes that could affect drug response or disease, and thus, shorten the time until information about these variants is useful in a clinical setting. (We use the term "pharmacogene" to refer to any gene such that a variant has been seen to affect drug response or is implicated in a disease.) Computational methods have been developed to predict the potential relevance of a gene to a query drug.[11] Other computational methods have been developed to identify genetic causes underlying disorders through gene prioritization, but many of these are designed to work on small sets of disease-specific genes.[12–17] The method which is closest to the one that we present here is described in

Costa *et al.*;[18] they create separate classifiers to predict morbidity-associated and druggable genes on a genome-wide scale. A majority of these methods use sequence-based features, network topology, and other features from curated databases; only a few use information from literature.[12,16,17]

In the work presented here, the goal is to predict pharmacogenes at genome-wide scale using a combination of features from curated databases and features mined from the biomedical literature. We evaluate a number of hypotheses:

(1) There is a set of GO concepts that are enriched when comparing the functions of important pharmacogenes and the rest of the human genome and by examining this set of enriched GO concepts, a classifier can be created to provide hypotheses regarding further genes in which variants could be of importance.
(2) Text-mined features will increase performance when combined with features from curated databases.

2. Methods

2.1. *Pharmacogenes*

By *pharmacogene*, we mean any gene such that a variant of that gene has been seen to affect drug response or such that variants have been implicated in disease. PharmGKB contains over 26,000 genes, with only a few having annotations that signify their importance in disease or drug response. For the experiments reported here, only those genes in which a variant exists in the PharmGKB relationship database, specifically gene-disease or gene-drug relationships, are considered to be gold-standard pharamcogenes. By this definition, 1,124 genes meet the criteria for classification as pharamcogenes and are positively labeled training instances; these make up <5% of all genes in PharmGKB. PharmGKB is constantly being updated, so a snapshot of PharmGKB on May 2, 2013 was taken and is used as the gold standard.

2.2. *Background genes*

The rest of the 25,110 genes in PharmGKB, which do not contain disease or drug relationships, are considered to be background genes and will be used as negatively labeled training instances. We acknowledge the fact that PharmGKB is incomplete and that a missing annotation is not indicative of a gene not being involved in disease or drug relationships, but the fact that they have not been discovered or curated yet. (This is an obvious motivation for the work reported here.) Two data sets were created from the background genes. One consists of all 25,110 genes. This is referred to as the unbalanced set. The second consists of 1,124 background genes that have similar numbers of publications as the known pharamcogenes. This is referred to as the balanced set. That is, the two sets differ in whether or not they result in a balanced set of positive and negative exemplars.

2.3. *Functional annotations from curated databases*

Links within PharmGKB were used to obtain Entrez Gene (EG) identifiers for both pharmacogenes and background genes. To extract all Gene Ontology (GO)[19] annotated functions associated with these genes, the NIH's gene2go file was used. Only curated evidence codes (EXP, IDA, IPI, IMP, IGI, IEP, TAS, and ISS) were used, in order to ensure high-quality annotations. This dataset will be referred to as the curated dataset. It contains many EGID to GO ID mappings obtained solely from curated GO annotations.

2.4. *Functional annotations from biomedical literature*

Entez Gene IDs and the NIH's gene2pubmed file were used to relate genes to documents of which they are the primary subject. By using the gene2pubmed file, we assume that all information retrieved from the article is associated with the gene that is the primary subject. Note that this is not always true and could introduce noise.

The 26,234 genes are mapped to 379,978 unique PubMed/MEDLINE articles. From these ~380,000 articles, two different textual datasets were created, one consisting only of abstracts and the other containing full text. The abstract dataset consists of all abstracts from all articles. For ~26,000 articles, we were only able to download XML or plain text, because PMC articles are available in any format, with some, such as PDF, not being suitable for natural language processing. The ~26,000 full-text articles constitute our full-text dataset. All full-text documents come from the PubMed Open Access Subset.

To extract gene functions (GO concepts) from these corpora, ConceptMapper, a dictionary-based concept recognizer,[20] was used with parameters tuned for each branch of the Gene Ontology (Molecular Function, Biological Process, and Cellular Component), as seen in Funk *et al.* (under review). Descriptive statistics of the documents and the functional annotations retrieved from them and from the curated database are shown in Table 1.

Table 1. **Summary of gene-document and gene-annotation associations** The number of genes within each dataset along with the mean number of biomedical literature documents associated with each **set of genes** and mean number of GO annotations per gene. (+) denotes that this set of genes is the positive labeled set while (–) denotes the negative training sets. The row labelled "Total Numbers" gives the count, not means, of documents and GO annotations.

		Mean # Docs		**Mean # GO Annotations**		
	# Genes	**Abstracts**	**Full-text**	**GOA curated**	**NLP abstracts**	**NLP full-text**
All genes	26,234	35.5	3.1	8.8	80.1	122.0
Known pharmacogenes (+)	1,124	215.2	15.5	16.3	227.5	220.7
All background genes (–)	25,110	26.7	2.5	8.2	72.8	128.7
Small background genes (–)	1,124	211.1	17.1	20.4	310.0	298.9
Total Numbers	26,234	379,978	25,987	112,356	1,891,566	1,951,982

2.5. *Enrichment of Gene Ontology concepts*

FatiGO[21] was used to test whether there are functional concepts that are enriched when pharamcogenes are compared to background genes. FatiGO is a tool that uses Fisher's exact test to extract over- or under-represented GO concepts from two lists of genes and provides a list of enriched GO concepts and their respective p-values as output. The p-values are corrected for multiple testing as described in Ge *et al.*[22] The gene lists and all three sets of annotations—curated, and text-mined–were provided to FatiGO as custom annotations. Fisher's exact test was conducted between GO concepts annotated to pharmacogenes and those annotated to background genes for all three sets of Gene Ontology concepts (curated, mined from abstracts, and mined from full text).

2.6. *Binary Classification*

All classifiers were implemented in the Weka toolkit, version 3.6.9. Three different baselines were used: OneR, a one node decision tree; Naive Bayes; and randomly assigning class labels. Against these, we compared three systems: Random Forests and two different Support Vector Machine implementations. Random Forests provide fast decision-tree training. Support Vector Machines (SVM) are currently the most popular classifier. The built-in classifiers for OneR (weka.classifiers.rules.OneR), Naive Byes (weka.classifiers.bayes.NaiveBayes), Random Forest (weka.classifiers.trees.RandomForest), and Support Vector Machine (weka.classifiers.functions.SMO) were used with default parameters. LibSVM (weka.classifiers.functions.LibSVM) was used with all but one default parameter. By default LibSVM maximizes accuracy; with the unbalanced dataset, this is not optimal, so weights of 90.0 and 10.0 were assigned to the pharmacogene and background classes, respectively. When using LibSVM with the balanced dataset, equal weights were given to both classes. All numbers reported are from five-fold cross-validation.

Table 2. **Machine learning features per dataset** A breakdown of the number and type of features used.

Dataset	# Genes	# Features	Type
GOA curated	12,704	39,329	Curated GO annotations from the GOA database.
NLP abstract	23,849	39,329	GO annotations recognized from MEDLINE abstracts.
NLP full-text	15,168	39,329	GO annotations recognized from full-text journal articles.
Abstract GO + Bigrams	23,849	858,472	GO annotations and bigrams from MEDLINE abstracts.
Full-text GO + Bigrams	15,168	906,935	GO annotations and bigrams from full-text journal articles.
Combined GO + Bigrams	23,867	1,189,175	Curated and NLP GO annotations and all bigrams.
Abstract GO + Collocations	23,849	346,878	GO annotations and collocations from MEDLINE abstracts.
Full-text GO + Collocations	15,168	54,951	GO annotations and collocations from full-text journal articles.
Combined GO + Collocations	23,867	349,243	Curated and NLP GO annotations and all collocations.

2.7. *Features derived from natural language processing*

Additional features were extracted from the abstract and full-text document collections using natural language processing. (This is in addition to the automatically extracted Gene Ontology annotations, which are also produced by natural language processing.) These features were word bigrams and collocations. Collocations, or sets of words that co-occur more often than expected, have not been commonly used in text classification, but provide a better reflection of the semantics of a text than bigrams. Both bigrams and collocations were extracted using the Natural Language Tool Kit (NLTK).[23] Any bigram or collocation where one of the tokens only contained punctuation was removed. Additionally, only those features that appear in three or more documents were retained. Six different NLP-derived feature sets were created by combining the three datasets (abstract, full-text, curated + abstract + full-text) along with the two different types of surface linguistic features (bigrams and collocations); these feature sets were tested and trained on both the balanced and unbalanced datasets.

2.8. *Machine learning input*

A breakdown of the kind and number of features used in each dataset can be seen in Table 2.

2.9. *Evaluation metrics*

The performance of our classifier was assessed by estimating precision (P), recall (R), and F-measure (F). The area under the receiving operator characteristic curve (AROC) is reported, as it allows for comparison against other classifiers, but with a word of caution interpreting the unbalanced dataset: inflated AROCs have been seen when working with skewed class distributions.[24] All scores were determined by taking the average of 5-fold cross-validation for all datasets.

3. Results and Discussion

3.1. *Enriched Gene Ontology concepts*

To assess the viability of a machine learner separating background and pharmacogenes, we first determine whether functional differences between the pharamcogenes and background genes exist. At least one curated or text-mined functional annotation was retrieved for 23,647 out of 26,236 total genes (90% of all genes in PharmGKB). The details of obtaining the annotations are given in Sections 2.3 and 2.4. The gene sets and their annotations were passed to FatiGO, a web tool that extracts over- and under-represented GO concepts from two lists of genes, and a list of enriched GO concepts and probabilities was returned as output. Examining the output from FatiGO, we found that, depending on the dataset, between 800-4000 GO concepts were enriched, consistent with our hypothesis that there are enriched pharmacogenetic functions. The top 10 enriched GO concepts for Molecular Function and Biological Process can be seen in Tables 3 and 4, respectively. These lists were obtained by comparing the annotations from all pharmacogenes to all background genes. To ensure that bias was not introduced solely because there is a large difference in the number of genes and the number of annotations between the two sets, another comparison was done between all pharamacogenes and the set of 1,124 background genes with equal representation in the biomedical literature. The enriched GO concepts returned are similar the concepts returned when comparing against all background genes, and therefore we can conclude that no bias is introduced. Because 800-4000 statistically enriched GO concepts were returned for each dataset, we can conclude that there are functional differences between the set of pharmacogenes and background genes.

Many of the enriched GO concepts can be categorized as playing a role in pharmacodynamics (PD) or pharmacokinetics (PK). Pharmacodynamics is the study of the activity of a drug in the body, e.g. its binding and effect on the body. Examples of PD concepts are “integral to plasma membrane” (GO:0005887), “drug binding” (GO:0008144), and “positive regulation of protein phosphatase type 2B activity” (GO:0032514)—they are either associated with receptors that drugs bind to, or refer to the possible effect that a drug has on the body. Pharmacokinetics is the study of drug absorption, distribution, metabolism, and excretion. Examples of PK concepts are “xenobiotic metabolic process” (GO:0006805), “small molecule metabolic process” (GO:0044281), and “active transmembrane transporter activity”

(GO:0022804)—they refer to metabolism of a molecule or are involved in the metabolism or transportation of a molecule.

Table 3. **Top 10 enriched GO concepts from the Molecular Function hierarchy.** The enriched GO concepts from the Molecular Function branch of Gene Ontology obtained when comparing pharmacogenes versus all background genes using FatiGO.

	GOA curated	
Concept ID	**Concept name**	**Adj. P-value**
GO:0005515	protein binding	$< 1.0 \times 10^{-8}$
GO:0019899	enzyme binding	$< 1.0 \times 10^{-8}$
GO:0042803	protein homodimerization activity	$< 1.0 \times 10^{-8}$
GO:0046982	protein heterodimerization activity	$< 1.0 \times 10^{-8}$
GO:0004497	monooxygenase activity	$< 1.0 \times 10^{-8}$
GO:0005245	voltage-gated calcium channel activity	$< 1.0 \times 10^{-8}$
GO:0020037	heme binding	$< 1.0 \times 10^{-8}$
GO:0004713	protein tyrosine kinase activity	$< 1.0 \times 10^{-8}$
GO:0004674	protein serine/threonine kinase activity	$< 1.0 \times 10^{-8}$
GO:0003677	DNA binding	$< 1.0 \times 10^{-8}$

	NLP abstracts	
Concept ID	**Concept name**	**Adj. P-value**
GO:0022804	active transmembrane transporter activity	$< 1.0 \times 10^{-8}$
GO:0005322	low-density lipoprotein	$< 1.0 \times 10^{-8}$
GO:0005321	high-density lipoprotein	$< 1.0 \times 10^{-8}$
GO:0005320	apoplioprotein	$< 1.0 \times 10^{-8}$
GO:0005179	hormone activity	$< 1.0 \times 10^{-8}$
GO:0005041	low-density lipoprotein receptor activity	$< 1.0 \times 10^{-8}$
GO:0005215	transporter activity	$< 1.0 \times 10^{-8}$
GO:0016088	insulin	$< 1.0 \times 10^{-8}$
GO:0004697	protein kinase C activity	$< 1.0 \times 10^{-8}$
GO:0045289	luciferin monooxygenase activity	$< 1.0 \times 10^{-8}$

	NLP full-text	
Concept ID	**Concept name**	**Adj. P-value**
GO:0042031	angiotensin-converting enzyme inhibitor activity	$< 1.0 \times 10^{-8}$
GO:0005262	calcium channel activity	$< 1.0 \times 10^{-8}$
GO:0016088	insulin	$< 1.0 \times 10^{-8}$
GO:0022804	active transmembrane transporter activity	$< 1.0 \times 10^{-8}$
GO:0005179	hormone activity	$< 1.0 \times 10^{-8}$
GO:0004872	receptor activity	$< 1.0 \times 10^{-8}$
GO:0005215	transporter activity	$< 1.0 \times 10^{-8}$
GO:0016791	phosphatase activity	$< 1.0 \times 10^{-8}$
GO:0008083	growth factor activity	$< 1.0 \times 10^{-8}$
GO:0004601	peroxidase activity	$< 1.0 \times 10^{-8}$

There are interesting differences when examining the top enriched concepts between the different datasets (curated, abstracts, and full text). Impressionistically, curated annotations seem to be more specific, while NLP annotations appear to be more general (especially evident when examining Biological Processes, Table 4). This may be the case because there are limitations to the depth in GO that concept recognizers can identify; a large gap exists between how near-terminal concepts are stated in the ontology and their expression in free text.

3.2. *Classification of pharmacogenes*

Having established that the functions of pharmacogenes are different from background genes, the next step is to test the ability of machine learning to differentiate between them. Our goal is to predict at genome-wide scale pharmacogenes that are not currently known in PharmGKB to have drug or disease relationships. We approach the problem as binary classification, where the classifier separates pharmacogenes from the rest of the genes.

3.3. *Classification using Gene Ontology concepts*

To see how well known pharmacogenes can be classified through their functional annotation similarity, five classifiers were created using the manually curated and text-mined functional annotations on both the unbalanced and balanced datasets. Baselines for comparison against are a one-node decision tree (OneR), Naive Bayes, and randomly assigning class labels. Performance of all classifiers and baselines can be seen in Table 5. A breakdown of features used for each dataset can be seen in Table 2 and a summary of functional annotations is seen in Table 1.

Table 4. **Top 10 enriched GO concepts from the Biological Process hierarchy.** The enriched GO concepts from the Biological Process branch of the Gene Ontology obtained when comparing pharmacogenes versus all background genes using FatiGO.

	GOA curated	
Concept ID	**Concept name**	**Adj. P-value**
GO:0044281	small molecule metabolic process	$< 1.0 \times 10^{-8}$
GO:0007596	blood coagulation	$< 1.0 \times 10^{-8}$
GO:0030168	platelet activation	$< 1.0 \times 10^{-8}$
GO:0006805	xenobiotic metabolic process	$< 1.0 \times 10^{-8}$
GO:0048011	neurotrophin TRK receptor signaling pathway	$< 1.0 \times 10^{-8}$
GO:0007268	synaptic transmission	$< 1.0 \times 10^{-8}$
GO:0008543	fibroblast growth factor receptor signaling pathway	$< 1.0 \times 10^{-8}$
GO:0007173	epidermal growth factor receptor signaling pathway	$< 1.0 \times 10^{-8}$
GO:0045087	innate immune response	$< 1.0 \times 10^{-8}$
GO:0055085	transmembrane transport	$< 1.0 \times 10^{-8}$

	NLP abstracts	
Concept ID	**Concept name**	**Adj. P-value**
GO:0007568	aging	$< 1.0 \times 10^{-8}$
GO:0009405	pathogenesis	$< 1.0 \times 10^{-8}$
GO:0046960	sensitization	$< 1.0 \times 10^{-8}$
GO:0008152	metabolic process	$< 1.0 \times 10^{-8}$
GO:0006629	lipid metabolic process	$< 1.0 \times 10^{-8}$
GO:0007610	behavior	$< 1.0 \times 10^{-8}$
GO:0006810	transport	$< 1.0 \times 10^{-8}$
GO:0014823	response to activity	$< 1.0 \times 10^{-8}$
GO:0006280	mutagenesis	$< 1.0 \times 10^{-8}$
GO:0042638	exogen	$< 1.0 \times 10^{-8}$

	NLP full-text	
Concept ID	**Concept name**	**Adj. P-value**
GO:0009626	plant-type hypersensitive response	$< 1.0 \times 10^{-8}$
GO:0007568	aging	$< 1.0 \times 10^{-8}$
GO:0016311	dephosphorylation	$< 1.0 \times 10^{-8}$
GO:0032514	positive regulation of protein phosphatase type 2B activity	$< 1.0 \times 10^{-8}$
GO:0008152	metabolic process	$< 1.0 \times 10^{-8}$
GO:0009405	pathogenesis	$< 1.0 \times 10^{-8}$
GO:0042592	homeostatic process	$< 1.0 \times 10^{-8}$
GO:0046960	sensitization	$< 1.0 \times 10^{-8}$
GO:0006810	transport	$< 1.0 \times 10^{-8}$
GO:0050817	coagulation	$< 1.0 \times 10^{-8}$

The results are shown in Table 5. A clear effect of balance versus imbalance in the data is evident. F-measure increases between 0.29 and 0.53 when using a balanced training set. Examining performance across unbalanced training sets, we notice that Naive Bayes produces the highest recall (0.68) but the lowest precision (0.17), whereas Random Forest produces highest precision (0.69) but lowest recall (0.11). The same trends do not hold for the balanced training sets. On both training sets, it is the SVM-based classifiers that balance precision and recall and produce the highest F-measures. The highest F-measures of 0.81 and 0.78, are produced by LibSVM and SMO, respectively, on the balanced NLP abstract annotations. Naive Bayes and Random Forrest perform poorly in comparison to the SVM classifiers, but better than a single-node decision tree or random assignment; OneR performs slightly better than random assignment.

For a majority of the classifiers, GO annotations from literature produce the best performance—surprisingly, text-mined annotations seem to be better features than those from curated datasets. This could be explained by the difference in number of annotations, there are 15 times more text-mined annotations than curated ones (Table 1). Another explanation could be that more information is encoded in text-mined annotations than just gene function. From this set of experiments, we can conclude that using only Gene Ontology concepts, we are able classify pharmacogenes on the balanced training set but it remains unclear, because of poor performance, whether it is sufficient to use only GO concepts with an unbalanced training set. We can also conclude that LibSVM should be used for the next set of experiments because it is best performing and was the fastest to train (training time not shown).

3.4. *Classification using GO concepts and literature features*

To test the hypothesis that features derived from surface linguistic features can increase performance over conceptual features alone, we trained classifiers with two additional feature types: bigrams and collocations. Bigrams consist of every sequence of two adjacent words in a document and are commonly used in text classification. Collocations are a subset of bigrams, containing words that co-occur more frequently than expected. They are a better representation of the semantics of a text than bigrams alone. The methods for extracting these features are described above in Section 2.7. Adding bigrams and collocations introduces up to 30x more features than functional annotations alone (Table 2).

The performance of LibSVM with GO annotations and bigrams/collocations on both training sets can be seen in Table 6. Baselines are the same.

On the unbalanced training set, the maximum F-measure seen is 0.57, obtained by using text-mined functional annotations and bigrams extracted from abstracts. By using bigrams in addition to GO annotations, precision is increased by 0.17 while recall is decreased by 0.02, resulting in an increase in F-measure of 0.09 (Table 5 versus Table 6). On the balanced training set, the maximum F-measure seen is 0.81, also obtained by using text-mined functional annotations and bigrams from abstracts. With the addition of bigrams, both precision and recall are increased by 0.06 and 0.03,respectively, resulting in an increase in F-measure of 0.06 (comparing Table 5 to Table 6).

Table 5. **Classification using Gene Ontology concepts** Five-fold cross validation performance of five binary classifiers when providing Gene Ontology concepts as features. Results from both unbalanced and balanced training sets are shown. The highest F-measure is bolded. The baselines provided are OneR (one-node decision tree), Naive Bayes, and randomly assigning classes (median of 5 random assignments).

	GOA curated	**NLP abstracts**	**NLP full-text**
Classifier	P/R/F	P/R/F	P/R/F
	Unbalanced Training		
Random	0.05/0.50/0.09	0.07/0.50/0.12	0.05/0.50/0.09
OneR	0.57/0.01/0.03	0.56/0.17/0.25	0.80/0.10/0.18
Naive Bayes	0.17/0.60/0.26	0.17/0.68/0.27	0.17/0.59/0.26
Random Forest	0.53/0.17/0.25	0.69/0.12/0.21	0.58/0.11/0.18
SMO	0.43/0.31/0.36	0.39/0.41/0.40	0.37/0.34/0.35
LibSVM	0.29/0.55/**0.38**	0.41/0.58/**0.48**	0.37/0.52/**0.42**
	Balanced Training		
Random	0.50/0.50/0.50	0.50/0.50/0.50	0/50/0.50/0.50
OneR	0.71/0.41/0.52	0.68/0.51/0.59	0.73/0.48/0.56
Naive Bayes	0.65/0.72/0.68	0.75/0.70/0.72	0.67/0.70/0.68
Random Forest	0.63/0.71/0.67	0.72/0.77/0.74	0.67/0.73/0.69
SMO	0.64/0.66/0.65	0.79/0.77/0.78	0.70/0.73/0.72
LibSVM	0.71/0.71/**0.71**	0.83/0.80/**0.81**	0.76/0.79/**0.78**

3.4.1. *Comparison with other methods*

As mentioned in the introduction, there are very few methods against which our method can be compared. Most gene-disease or gene prioritization methods are designed to work on small sets of disease-specific genes,[12–14] while our method predicts pharmacogenes on a genome-wide scale. One method, Garten *et al.*,[25] utilizes text mining to extract drug-gene relationships from the biomedical literature, also using PharmGKB as a gold standard, with an AUC of 0.701. The closest methods to ours do not predict pharmacogenes as defined here, but only predict disease genes. CIPHER[26] predicts human disease genes with precision of ∼0.10 using protein-protein interaction networks and gene-phenotype associations. PROSPECTR[27] uses

Table 6. **Classification with GO concepts and natural language processing** Five-fold cross-validation performance of LibSVM when combining Gene Ontology concepts and literature-based features. Both the balanced and unbalanced training results are shown. The highest F-measure and AROC are bolded. The baselines provided are OneR (one-node decision tree), Naive Bayes, and randomly assigning classes (median of 5 random assignments).

	Abstract GO + Bigrams		**Full-Text GO + Bigrams**		**Combined GO + Bigrams**	
Classifier	**P/R/F**	**AUC**	**P/R/F**	**AUC**	**P/R/F**	**AUC**
			Unbalanced Training			
Random	0.07/0.50/0.12	0.501	0.05/0.50/0.09	0.501	0.05/0.50/0.09	0.499
LibSVM	0.58/0.56/**0.57**	**0.771**	0.50/0.46/0.48	0.711	0.50/0.54/0.52	0.756
			Balanced Training			
Random	0.50/0.50/0.50	0.500	0.50/0.50/0.50	0.500	0.50/0.50/0.50	0.500
OneR	0.75/0.59/0.66	0.696	0.71/0.53/0.61	0.663	0.79/0.50/0.61	0.685
LibSVM	0.89/0.83/**0.86**	**0.860**	0.79/0.82/0.80	0.807	0.86/0.83/0.85	0.848

	Abstract GO + Collocations		**Full-Text GO + Collocations**		**Combined GO + Collocations**	
Classifier	**P/R/F**	**AUC**	**P/R/F**	**AUC**	**P/R/F**	**AUC**
			Unbalanced Training			
Random	0.07/0.50/0.12	0.501	0.05/0.50/0.09	0.501	0.05/0.50/0.09	0.499
LibSVM	0.54/0.56/**0.55**	**0.767**	0.41/0.52/0.46	0.730	0.47/0.56/0.51	0.763
			Balanced Training			
Random	0.50/0.50/0.50	0.500	0.50/0.50/0.50	0.500	0.50/0.50/0.50	0.500
OneR	0.78/0.46/0.58	0.664	0.67/0.64/0.66	0.675	0.75/0.59/0.66	0.698
LibSVM	0.87/0.82/**0.85**	**0.850**	0.77/0.80/0.78	0.786	0.85/0.81/0.83	0.833

23 sequence-based features and predicts disease genes from OMIM with precision = 0.62 and recall = 0.70 with an AUC of 0.70.The most directly comparable method, presented in Costa *et al.*,[18] utilizes topological features of gene interaction networks to predict both morbidity genes (P=0.66, R=0.65, AUC=0.72) and druggable genes (P=0.75, R=0.78, AUC=0.82). While the majority of other methods utilize sequence-based features, protein interactions, and other genomic networks, our method requires only Gene Ontology annotations and simple bigrams/collocations extracted from biomedical literature. Precision and recall for our classifier trained on the unbalanced dataset with GO annotations and bigrams from abstracts are slightly lower than both PROSPECTR and the method presented in Costa *et al.*, our AUC (0.771) is higher than all but the predicted druggable genes from Costa *et al.* Performance on the balanced training set using GO concepts and bigrams extracted from abstracts (F=0.86, AUC=0.860) are higher than any of the methods presented here.

3.4.2. *Limitations*

There are two major limitations of our work. The first is that we grouped together all pharmacogenes, while it may have been more useful to differentiate between disease-associated and drug-response-associated variant. The other limitation is that we don't provide a ranking, but rather just a binary classification.

3.5. *Prediction of pharmacogenes*

Now that classifiers have been created and evaluated, we can analyze the predicted pharmacogenes. 141 genes were predicted to be pharmacogenes by all six unbalanced datasets seen in Table 6. Predictions from unbalanced models were analyzed because the models produced through balanced training were unknowingly weighted for recall. For example, the balanced model trained on abstract GO and bigrams produces a recall of 0.99 and precision of 0.10

Table 7. **Top 10 predicted pharmacogenes** Top 10 pharmacogenes predicted by all combined classifiers and ranked by functional similarity to the known pharmacogenes. All information from PharmGKB and OMIM is presented along with the class that was predicted by Costa *et al.*[18] (Morbid: mutations that cause human diseases, Druggable: protein-coding genes whose modulation by small molecules elicits phenotypic effects).

EG ID	**Symbol**	**PharmGKB Annotations**	**OMIM Phenotype**	**Costa *et al.*[18] predicted**
2903	*GRIN2A*	None	Epilepsy with neurodevelopment defects	Druggable
7361	*UGT1A*	None	None	Not tested
2897	*GRIK1*	None	None	Druggable
1128	*CHRM1*	None	None	Druggable
1131	*CHRM3*	Member of Proton Pump Inhibitor Pathway	Eagle-Barrett syndrome	Druggable
3115	*HLA-DPB1*	None	Beryllium disease	Morbid/Druggable
6571	*SLC18A2*	Member of Nicotine, Selective Serotonin Reuptake Inhibitor, and Sympathetic Nerve Pathway	None	Morbid/Druggable
477	*ATP1A2*	None	Alternating hemiplegia of childhood, Migraine (familial basilar and familial hemiplegic)	Morbid/Druggable
3643	*INSR*	Member of Anti-diabetic Drug Potassium Channel Inhibitors and Anti-diabetic Drug Repaglinide Pathways	Diabetes mellitus, Hyperinsulinemic hypoglycemia, Leprechaunism, Rabson-Mendenhall syndrome	Morbid/Druggable
2905	*GRIN2C*	None	None	Druggable

when the classifier is applied to all genes in PharmGKB; this is not informative and further work and error analysis will be conducted to examine why this is.

The top 10 predicted genes, ranked by functional similarity (as calculated by ToppGene) to the known pharmacogenes, along with all known information from PharmGKB and Online Mendelian Inheritance in Man (OMIM),[7] and if/what the gene was predicted to be by Costa *et al.* can be seen in Table 7. We first notice that there are no gene-disease or gene-drug relationships in PharmGKB for these predicted genes, but a few of them participate in curated pathways. We expand our search to see if other databases have drug or disease information about them. OMIM provides insight into genetic variation and phenotypes; half of the predicted genes have a variant that plays a role in a mutant phenotype. We also looked up our predicted genes in the results from a previous study on predicting morbid and druggable genes, and 90% (9 out of 10) of our predicted pharmacogenes were also predicted to be morbid (variations cause hereditary human diseases) or druggable.[18]

To assess the hypothesized pharmacogenes further, PubMed and STITCH[28] were used to find any known drug or disease associations not in PharmGKB or OMIM. The top-ranked gene, *GRIN2A*, seems to play a part in schizophrenia and autism spectrum disorders[29] along with binding to memantine, a class of Alzheimer's medication blocking glutamate receptors. Interestingly, *UGT1A* is unable to be found in STITCH or OMIM, but an article from May 2013 introduces a specific polymorphism that suggests that it is an important determinant of acetaminophen glucuronidation and could affect an individual's risk for acetaminophen-induced liver injury.[30] It is also known to be linked to irinotecan toxicity. We also find genetic variations in *GRIK1* have been linked to schizophrenia[31] and down syndrome.[32] Even only examining the top three predicted pharmacogenes, there is evidence in other databases and literature that suggests these should be further examined by the PharmGKB curators for possible annotation.

4. Conclusions

One of the surprising findings of this study was that features extracted from abstracts performed better than features extracted from full text. Since full text was available for a smaller number of genes, the comparison may not be appropriate. Pursuing this remains for further research.

The collocation features performed almost as well as the bigrams, despite the fact that we took a poor approach to extracting them, since we did collocation recognition on the document level, rather than on the level of the document collection as a whole. With a better approach to collocation extraction, performance of the collocation features might have been much higher.

The fact that features derived from text-mined functional annotations outperformed manually curated annotations was a surprise. In this work, we did not evaluate the correctness of text-mined functional annotations. Therefore, the performance of the text-mined functional annotation features is the only indication of how well the actual Gene Ontology concept recognition worked. Based on the fact that they performed higher than the manually curated Gene Ontology concepts, it appears that the performance of the ConceptMapper approach was at minimum good enough for this task.

In this paper we identified a set of functions enriched in known pharmacogenes. This list could be used to rank genes predicted by our classifier, but also has usefulness beyond the work presented here. The list could prove useful in literature-based discovery by providing linkages to identify gene-drug or gene-disease relationships from disparate literature sources.

We also present a classifier that is able to predict pharmacogenes at a genome wide scale (F=0.86, AUC=0.860). The top 10 hypothesized pharmacogenes predicted by our classifier are presented; 50% contain allelic variations in OMIM and 90% were previously predicted but remain unannotated in PhamGKB. Additionally, using other sources at least the top three genes predicted are known to bind a drug or to be associated with a disease. Other methods attempting similar problems, utilize sequence based features and genomic networks; only a few incorporate literature features. Our method, on the other hand, uses mainly features mined from the biomedical literature along with functional annotations from databases. Because our method offers comparable performance to others utilizing sequence and network based features, this work illustrates the importance of incorporating curated databases with information available in the biomedical literature for biomedical discovery.

Acknowledgments

This work was supported by NIH grants 5R01 LM009254-07, 5R01 LM008111-08, and 2T15LM009451 to LEH.

References

1. W. E. Evans and M. V. Relling, *Science* **286**, 487 (1999).
2. J. Poirier, M.-C. Delisle, R. Quirion, I. Aubert, M. Farlow, D. Lahiri, S. Hui, P. Bertrand, J. Nalbantoglu and B. M. Gilfix, *Proceedings of the National Academy of Sciences* **92**, 12260 (1995).
3. J. A. Kuivenhoven, J. W. Jukema, A. H. Zwinderman, P. de Knijff, R. McPherson, A. V. Bruschke, K. I. Lie and J. J. Kastelein, *New England Journal of Medicine* **338**, 86 (1998).

4. D. S. Wishart, C. Knox, A. C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang and J. Woolsey, *Nucleic acids research* **34**, D668 (2006).
5. X. Chen, Z. L. Ji and Y. Z. Chen, *Nucleic acids research* **30**, 412 (2002).
6. M. Hewett, D. E. Oliver, D. L. Rubin, K. L. Easton, J. M. Stuart, R. B. Altman and T. E. Klein, *Nucleic acids research* **30**, 163 (2002).
7. A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini and V. A. McKusick, *Nucleic acids research* **33**, D514 (2005).
8. K. G. Becker, K. C. Barnes, T. J. Bright and S. A. Wang, *Nature genetics* **36**, 431 (2004).
9. L. A. Hindorff, P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins and T. A. Manolio, *Proceedings of the National Academy of Sciences* **106**, 9362 (2009).
10. W. A. B. Jr., K. B. Cohen, L. Fox, G. K. Acquaah-Mensah and L. Hunter, *Bioinformatics* **23**, i41 (2007).
11. N. T. Hansen, S. Brunak and R. Altman, *Clinical Pharmacology & Therapeutics* **86**, 183 (2009).
12. S. Aerts, D. Lambrechts, S. Maity, P. Van Loo, B. Coessens, F. De Smet, L.-C. Tranchevent, B. De Moor, P. Marynen, B. Hassan *et al.*, *Nature biotechnology* **24**, 537 (2006).
13. O. Vanunu, O. Magger, E. Ruppin, T. Shlomi and R. Sharan, *PLoS computational biology* **6**, p. e1000641 (2010).
14. J. E. Hutz, A. T. Kraja, H. L. McLeod and M. A. Province, *Genetic epidemiology* **32**, 779 (2008).
15. J. Chen, B. J. Aronow and A. G. Jegga, *BMC bioinformatics* **10**, p. 73 (2009).
16. L.-C. Tranchevent, R. Barriot, S. Yu, S. Van Vooren, P. Van Loo, B. Coessens, B. De Moor, S. Aerts and Y. Moreau, *Nucleic acids research* **36**, W377 (2008).
17. G. Gonzalez, J. C. Uribe, L. Tari, C. Brophy and C. Baral, in *Incorporating Interactions, Connectivity, Confidence, and Context Measures. in Pacific Symposium in Biocomputing. 2007. Maui*, 2007.
18. P. R. Costa, M. L. Acencio and N. Lemke, *BMC genomics* **11**, p. S9 (2010).
19. T. G. O. Consortium, *Genome Research* **11**, 1425 (2001).
20. M. Tanenblatt, A. Coden and I. Sominsky, in *International Conference on Language Resources and Evaluation*, 2010.
21. F. Al-Shahrour, R. Díaz-Uriarte and J. Dopazo, *Bioinformatics* **20**, 578 (2004).
22. H. Ge, A. J. Walhout and M. Vidal, *TRENDS in Genetics* **19**, 551 (2003).
23. S. Bird, in *Proceedings of the COLING/ACL on Interactive presentation sessions*, 2006.
24. U. Kaymak, A. Ben-David and R. Potharst, *Engineering Applications of Artificial Intelligence* **25**, 1082 (2012).
25. Y. Garten, N. P. Tatonetti and R. B. Altman, in *Pac Symp Biocomput*, 2010.
26. X. Wu, R. Jiang, M. Q. Zhang and S. Li, *Molecular Systems Biology* **4** (2008).
27. E. A. Adie, R. R. Adams, K. L. Evans, D. J. Porteous and B. S. Pickard, *BMC bioinformatics* **6**, p. 55 (2005).
28. M. Kuhn, C. von Mering, M. Campillos, L. J. Jensen and P. Bork, *Nucleic acids research* **36**, D684 (2008).
29. J. Tarabeux, O. Kebir, J. Gauthier, F. Hamdan, L. Xiong, A. Piton, D. Spiegelman, E. Henrion, B. Millet, F. Fathalli *et al.*, *Translational psychiatry* **1**, p. e55 (2011).
30. M. Freytsis, X. Wang, I. Peter, C. Guillemette, S. Hazarika, S. X. Duan, D. J. Greenblatt, W. M. Lee *et al.*, *Journal of Pharmacology and Experimental Therapeutics* **345**, 297 (2013).
31. Y. Hirata, C. C. Zai, R. P. Souza, J. A. Lieberman, H. Y. Meltzer and J. L. Kennedy, *Human Psychopharmacology: Clinical and Experimental* **27**, 345 (2012).
32. D. Ghosh, S. Gochhait, D. Banerjee, A. Chatterjee, S. Sinha and K. Nandagopal, *Genetic Testing and Molecular Biomarkers* **16**, 1226 (2012).