Journal of Biomedical Informatics 45 (2012) 842–850

Contents lists available at SciVerse ScienceDirect

Journal of Biomedical Informatics

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

A SNPshot of PubMed to associate genetic variants with drugs, diseases, and adverse reactions

Jörg Hakenberg[a,b,*], Dmitry Voronov[a,d], Võ Hà Nguyên[a], Shanshan Liang[a], Saadat Anwar[a], Barry Lumpkin[a], Robert Leaman[c], Luis Tari[b], Chitta Baral[a]

[a] *Arizona State University, Computer Science Department, 699 S Mill Ave., Tempe, AZ 85281, USA*
[b] *Hoffmann-La Roche, Inc., Pharma Research and Early Development, 340 Kingsland St., Nutley, NJ 07110, USA*
[c] *Arizona State University, Department of Biomedical Informatics, 425 N Fifth St., Phoenix, AZ 85004, USA*
[d] *Ural Federal University, IT Department, Turgeneva St. 4, Ekaterinburg 620075, Russian Federation*

ARTICLE INFO

Article history:
Received 28 April 2011
Accepted 11 April 2012
Available online 30 April 2012

Keywords:
Text mining
Information extraction
Pharmacogenomics
Databases

ABSTRACT

Motivation: Genetic factors determine differences in pharmacokinetics, drug efficacy, and drug responses between individuals and sub-populations. Wrong dosages of drugs can lead to severe adverse drug reactions in individuals whose drug metabolism drastically differs from the "assumed average". Databases such as PharmGKB are excellent sources of pharmacogenetic information on enzymes, genetic variants, and drug response affected by changes in enzymatic activity. Here, we seek to aid researchers, database curators, and clinicians in their search for relevant information by automatically extracting these data from literature.

Approach: We automatically populate a repository of information on genetic variants, relations to drugs, occurrence in sub-populations, and associations with disease. We mine textual data from PubMed abstracts to discover such genotype–phenotype associations, focusing on SNPs that can be associated with variations in drug response. The overall repository covers relations found between genes, variants, alleles, drugs, diseases, adverse drug reactions, populations, and allele frequencies. We cross-reference these data to EntrezGene, PharmGKB, PubChem, and others.

Results: The performance regarding entity recognition and relation extraction yields a precision of 90–92% for the major entity types (gene, drug, disease), and 76–84% for relations involving these types. Comparison of our repository to PharmGKB reveals a coverage of 93% of gene–drug associations in PharmGKB and 97% of the gene–variant mappings based on 180,000 PubMed abstracts.

Availability: http://bioai4core.fulton.asu.edu/snpshot.

1. Introduction

Genetic factors determine differences in pharmacokinetics, drug efficacy, and drug responses between individuals and sub-populations. A large proportion of such differences is attributable to single nucleotide polymorphisms (SNPs)[1] in drug metabolizing enzymes, drug transporters, drug receptors, and ion channels. An example is the drug-metabolizing enzyme CYP2D6 with more than 70 known allelic variations, at least 15 of which encode for a non-functional enzyme (*null allele*) [1]. Hepatic metabolism of many prescription drugs, such as anti-depressants and beta-AR receptor blockers, is (partially) dependent on CYP2D6 activity. Patients characterized as *poor metabolizers* (null allele or decreased activity) may retain high plasma concentrations of the administered drug when they are treated with typical dosages, and develop side effects. On the other hand, *rapid metabolizers*, with a highly efficient variant of the enzyme and thus a high rate of drug metabolism, may not reach the therapeutically required plasma level—or conversely, reach a toxic level of the drug's metabolites[2] [2].

Determining each individual patient's metabolic capacity[3] thus is an important step towards determining the optimal dosage when moving beyond the still common "one-dose-suits-all" approach. The following examples shall explain some of the effects of wrong dosages. Wuttke et al. [1] presented a survey of 1200 patients treated with metoprolol (for coronary artery disease, chronic heart failure, hypertension) and exhibiting adverse drug reactions. Poor metabolizers had a fivefold higher risk for development of adverse drug

* Corresponding author at: Hoffmann-La Roche, Inc., Pharma Research and Early Development Informatics, 340 Kingsland St., Nutley, NJ 07110, USA.

E-mail address: jorg.hakenberg@roche.com (J. Hakenberg).

[1] Additional differences arise from multigenic and environmental factors.

[2] As is the case, for instance, with ultra-rapid metabolizers of codeine, where its substrate morphine cannot be cleared as fast, resulting in narcosis and apnae.

[3] Here: phenotype determination with respect to drug response. *Phenotypes* for the purpose of this study are clinical symptoms; we exclude molecular phenotypes such as gene expression for now.

1532-0464/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jbi.2012.04.006

responses (ADRs); in co-administration with another drug, propafenone, even leading to severe effects. Kirchheiner et al. [3] performed an extensive literature survey on 32 anti-depressants to find sub-population-specific dosage recommendations, covering 54 studies. Focusing on the cytochrome p450 enzymes CYP2D6 and CYP2C19, they distinguished between poor, intermediate, rapid, and ultra-rapid metabolizers. Depending on the drug, Kirchheiner et al. found recommended dosages ranging from only 20% of the 'usual' dosage for poor metabolizers to 140% for ultra-rapid metabolizers. Lazarou et al. [4] found that 6.7% of hospitalized patients would develop severe ADRs, with 0.3% fatal reactions, resulting in around 100,000 deaths per year in the US.[4]

One aim of pharmacogenetics is to correlate genetic fingerprints to probable drug responses, thus helping to select the most efficient drug and give dosage recommendations. Such information on SNPs, their effect on enzymatic activity, their frequencies in populations, and related data, are spread across a large number of clinical studies and surveys. PharmGKB collects information "about the impact of human genetic variations on drug response," currently listing around 2400 human genetic variants from more than 3,500 articles [5]. All data were curated via literature review and can be searched by gene, drug, disease, etc. PubMed indexes close to 100,000 articles for the query "`(drug OR treatment) AND metabolism AND (allele OR variant OR mutation)`" alone. This later fact hints at a large number published data about genetic variations and pharmacokinetic and -dynamic phenotypes still to be indexed. Altman and Klein [6] suggested related opportunities to unveil pharmacokinetic, pharmacodynamic, and pharmacogenomic associations by mining the literature and integrating them with existing databases and results from other bioinformatics methods.

In this paper, we address the issue of automatically extracting information on genetic variations and their implications for drug responses from PubMed abstracts. We concentrate on clinical/pathological symptoms as phenotypes in this paper, and exclude molecular phenotypes such as gene expression and RNAi. We seek to enrich these former correlations with information on diseases, mutations, RefSNPs (rs-numbers), alleles, and their population-specific frequencies. Here, our focus is on finding data relating genes to drugs to diseases, and providing information on genetic variants, allele names, and/or RefSNPs for each gene in a relation; we add frequencies and populations pertaining to observed genotypes/haplotypes whenever possible. From an application-oriented, pharmacogenetics perspective, we want to provide a repository covering all this information, helping database curators to search and incorporate relevant information, guiding clinicians in their decisions to administer which drugs in which dosages, and aiding drug developers in their studies. Thus, we are trying to help establish the "shift from population-based data toward personal health care" [7]. From a practical, text mining perspective, we want to study how one can effectively put together a text mining system for a more complex task, possibly using existing tools; and what to expect regarding effort, efficiency, and performance.

1.1. Related work

In work related to ours, Wang et al. [8] recently conducted a feasibility study on mining literature for pharmacokinetics (PK) numerical data. They searched for PK parameters AUC, half-life, bio-availability, clearance, dose, and administration route (oral *vs.* intravenous), as well as subject type (sex, age, ethnicity). For selecting abstracts relevant to the drug midazolam, they achieved an *f*-score between 68% and 78%; for midazolam clearance data, they achieved an *f*-score of 90% from 41 relevant abstracts. Overall, for seven different drugs, literature mining generated up to 4-fold higher information content on clearance data as compared to the DiDB database.[5]

The method presented by Theobald and colleagues starts with a set of known relations between genes, drugs, and diseases to generate Bayesian networks [9]. Conditional probabilities were generated using co-occurrence statistics of these relations gathered from all of PubMed. Overall, Theobald et al. found data on 1730 gene/drug/disease relations that were contained in PharmGKB to build this network. In the end, one can analyze the network to find hypotheses on drug mechanisms, biomarkers for treatment, or markers of genetic diseases.

Doughty et al. presented a study in which they map mutations to diseases, namely forms of prostate and breast cancer [10]. Mutations were extracted based on a set of regular expressions, wrapped in the "extractor of mutations" (EMU) tool [10]; to recognize gene names, a gazetteer lookup based on HUGO and NCBI was developed; and MetaMap [11] was employed after narrowing its UMLS vocabulary down to concepts denoting forms of prostate and breast cancer, respectively. The authors evaluated three types of extracted data (mutations, mutations mapped to genes, and gene + mutation + disease associations). EMU extracted mutations with a precision of 94–97%; initial mapping of mutations to genes achieved a precision of 42–53%; after a filtering steps that cross-checked the predicted gene's sequence with the mutation (residue found at given position in the sequence), EMU's precision reached 74–80%. Associating mutations with genes with diseases yielded a precision of 55–77% after filtering. Doughty et al. [10] cross-checked their extracted associations with OMIM, dbSNP, and SwissProt, showing that, for instance, 86 out of 144 breast cancer-related mutations were not yet contained in OMIM.

Lauria and coworkers presented a rule-based approach to extract mutations and map them to impacts on protein properties, such as function, stability, and enzyme kinetics [12]. For recognition of mutations and their mapping to canonical forms, the authors utilized MutationFinder [13]. Grounding mutations to protein sequences started with candidate proteins extracted from the entire document the mutation occurred in. In a filtering step, residues and positions involved were compared to the candidates' amino acid sequences in UniProtKB. To find protein properties, the authors scanned the text for head nouns such as "activity" and "binding". Tokens of the encompassing noun-phrases are stemmed and the resulting phrase compared to terms in the Molecular Function branch of the Gene Ontology. In addition, the authors searched for impact directions, such as "increase", "abolish", and negations. For the final mapping of mutations to impacts, Laurila et al. [12] employ a score measuring the sentence distance between a mutation occurrence and an impact occurrence. Precision for grounding mutations was given as 83% at 83% recall; the mapping of mutants to impacts achieved 86% precision at 34% recall; evaluation was performed on a corpus of 13 full text documents, with 54 unique mutations and 73 unique mutant–impact relations.

Tari et al. studied the synthesis of drug-metabolic networks (mapping each drug to its metabolizing enzymes) by extracting these relations from the literature [14]. They parsed sentences with a dependency parser, establishing syntactic dependencies between enzymes and drugs mentioned together in a sentence. Using pre-formulated queries against these syntactic structures, they search for patterns like "DRUG $*$ metabolized by $*$ ENZYME", where the wild-cards allow for certain dependencies connecting the parts, for example, "`predicate` → `object`". Tari et al. populated a database with about 13,000 PubMed abstracts, in their parsed form, so

[4] Meta-analysis of 39 studies from 1998 and older.

[5] DiDB—see http://www.druginteractioninfo.org/.

similar queries can be issued by users to find these types of relations. Their system achieves a precision of 92% for gene–drug relations, albeit estimated on a low number of predicted relations.

Coulet and colleagues built a semantic network of pharmacogenomic relationships between genes, drugs, phenotypes, and entities that are modified by the former, such as drug response [15]. Their method employs dependency graphs obtained from the Stanford Parser, representing the syntactical structure of a sentence [16]. Sentences that contained either one gene and one drug, or one gene and one phenotype[6], were retrieved from PubMed abstracts and then parsed. The authors extracted all paths that link two entities using an intermediary (possibly nominalized) verb and map each instance onto an ontology of normalized pharmacogenomics relationships. For instance, the phrase "differences in warfarin requirements" was mapped to the concepts of *drug* ("warfarin"), *dose* ("requirements"), and *variation* ("differences"), resulting in the overall notion of "warfarin dose variation". From these normalized relations, the authors were able to build semantic networks surrounding entities of interest to pharmacogenomics (such as individual genes).

1.2. Outline

The remainder of this paper is organized as follows. We start with a description of our methods for named entity recognition (Section 2.1), normalization Section 2.2, and relationship extraction Section 2.3. In Section 3 we present statistics on the content of the automatically filled repository SNPshot. We continue with an evaluation of our methods by comparing SNPshots content with data from PharmGKB and DrugBank to estimate coverage; and by manual inspection to obtain precision/recall values. Section 4 describes the user interface and examples available at http://bioai4core.fulton.asu.edu/snpshot. We conclude this paper with a discussion in Section 5.

2. Methods

2.1. Recognition of named entities

The entity types we are trying to find are genes (incl. proteins, enzymes), drugs, diseases, adverse drug reactions (ADRs), mutations/SNPs, RefSNPs (rs-numbers), names of alleles, populations, and frequencies, also listed in Table 1. For some of them, we consider simple rule-based approaches to be sufficient; for others, we use dictionaries generated from relevant databases; for genes and diseases, we train an off-the-shelf tagger as these names come in too many variations, many of which will not be captured by dictionary-matching. We will present the method for each entity type in the following. Running named entity recognition for various types in parallel will typically result in overlapping entities—a gene name might contain or be equal to the name of a disease ("breast cancer associated gene 1", "Neurofibromatosis 2"), a gene name might be identical to a drug ("insulin"), etc. In these cases, we decided on a heuristic to replace drugs with genes, diseases with genes, and genes with mutations for exact matches; we keep overlapping names as separate entities ("breast cancer" within "breast cancer associated gene 1").

2.1.1. Genes

For the recognition of gene names, we trained BANNER [17] on the BioCreative II GM training set [18]. BANNER runs a 2^{nd}-order conditional random field to assign a sequence of labels to a sequence of tokens (a sentence), where labels are in BIO-format: "not part of a gene name" (O), "inside of a gene name" (I), and "at the beginning of a gene name" (B). It uses feature classes such as token surface, character bi- and tri-grams, prefixes, suffixes, part-of-speech tag, as well as features taken from a window surrounding the current token. On the BioCreative II GM training set, this results in a collection of about 1.5 million individual features.

Table 1
Numbers of entities found in 179,935 abstracts.

	Total[a]	Unique[b]	Avg.[c]	DrugBank/PharmGKB[d]
Genes/proteins	1,346,395	15,794	55	1018/3029
Drugs	277,258	1693	164	3797/646
Diseases	1,098,883	9539	101	–/705
Adverse effects	525,705	1414	372	
Mutations	92,173	20,339	4	
RefSNP	15,883	6208	2	–/1462
Alleles	52,732	1099	39	
Populations	139,986	208	663	
Frequencies	36,260			

[a] Total number of occurrences, including duplicates, including entities with no identifier or canonical name.

[b] Unique entities include only those entities that were assigned to an identifier or canonical name, excluding duplicates; entities of the type 'Frequencies' (example: "8%") were not mapped to canonical names.

[c] Average number of occurrences of a particular *identified* entity; includes multiple occurrences of that entity per abstract, therefore the average number of abstracts that discuss an entity is lower.

[d] Number of unique entities found in DrugBank and PharmGKB.

2.1.2. Genotypes: variants, mutations, SNPs, alleles, haplotypes

Although there have been recommendations for the description of sequence variants [19],[7] favoring the form "c.76A>T", they occur in much more variations in free text. We use MutationFinder [13] as a starting point, adding more variations frequently occurring in PubMed abstracts ("-3402C>T", "c.183 A>C"), which we found searching PubMed for related keywords (allele, mutation, gene, etc.) and then manually enumerated all variations not yet covered. We wrote additional components to also find non-single nucleotide variants ("c.76_78delACT", "MV324KF") and insertions/deletions ("1707 del T", "c.76_77insG"), RefSNP (rs) numbers, as well as names of alleles/haplotypes (like in "CYP2D6*4" "T allele", "null allele", "GC/GC genotype"). For simplicity, in the remainder we refer to variants, mutations, alleles, and haplotypes together as *variants* when applicable.

2.1.3. Diseases and adverse drug reactions

To find names referring to diseases, we pursue two strategies. The first uses BANNER, trained on a corpus with disease annotations, consisting of about 3000 sentences[8] [20]. As only about 10% of the sentences in this corpus do not contain any disease, we added 200 random sentences from the BioCreative II GM training data. The second strategy builds on a dictionary extracted from UMLS, consisting of about 162 k terms for 54 k concepts from the six categories "Disease or Syndrome", "Neoplastic Process", "Congenital Abnormality", "Mental or Behavioral Dysfunction", "Experimental Model of Disease", and "Acquired Abnormality". We manually went through the list of names to remove unspecific or spurious disease names, such as "symptoms", "disease", "disorder", "dependence", and "deficiency", as well as "endocrine", "cortex", "AA", and "dissociation". This task was alleviated by sorting terms according to occurrence frequency in a PubMed sample.

Our dictionary for adverse drug reactions originates from the SIDER Side Effect Resource [21]. SIDER provides a mapping of

[6] From a fixed set of 41 genes, 3007 drugs, and 4204 phenotypes (diseases and adverse reactions).

[7] HGVS—http://www.hgvs.org/mutnomen/recs.html.

[8] AZDC—http://diego.asu.edu/downloads/AZDC.

ADR terminology to UMLS CUIs; we were thus able to expand each ADR's synonym set based on the respective UMLS entry. The final dictionary has 1676 UMLS concepts with 6497 terms.

2.1.4. Drugs

The dictionary for drugs is based on DrugBank [22], containing around 29,000 different drug names (incl. generic names, brand names). As drug names do not appear in many variations – this holds in particular for brand names – we can apply a case-insensitive but otherwise exact matching. DrugBank cross-links with PharmGKB and other resources, so we were able to collect additional synonyms and PharmGKB IDs, as well as PubChem Compound and Substance identifiers to provide hyperlinks to additional information.

2.1.5. Populations and frequencies

We use close to 950 terms referring to ethnicities, regions, countries and their inhabitants, mostly based on WikiPedia entries for countries and ethnic groups, and match them (case insensitive) against the text. Examples include "Caucasian", "Ashkenazi Jews", "Italian", "North African". Based on observations, we filter out some occurrences in phrases such as "Chinese [hamster]" and "United States [Food and Drug Administration]". For frequencies, we extract all numerals (decimal values from 0..1 and percentages, including ranges such as "7–12%") from sentences that contain any of the words "allele", "variant", "mutation", or "population". We remove typical false positives, for example, referring to p-values, odds ratios, and confidence intervals, also using trigger words such as 'CI', 'OR', and '$p \leqslant \ldots$'.

2.2. Entity grounding and normalization

Names referring to genes, diseases, and drugs come in many variations: "CYP2D6", "Cytochrome p450 2D6", "P450 IID6" all refer to the same enzyme (EntrezGene ID 1565). Knowledge integration (collecting different facts on the same gene), evaluation (compare our results to curated facts in databases), and search (search within our collected information) all require to map each occurrence to a unique identifier, which is then used to refer to the entity. To assign EntrezGene IDs to each individual gene, we ran GNAT on the recognized genes [23], with the difference that we limit the candidate genes to human, murine, and rat genes. GNAT first tries to normalize each gene name recognized by BANNER, for instance, removing modifiers such as "wild-type" or species and tissue from the beginning of a name, and others such as "gene" and "isozyme" from the end. It then tries to match the resulting name to a dictionary of gene names from EntrezGene (here, restricted to the species human, mouse, and rat) to find candidate IDs for each gene name. In case of multiple candidates, disambiguation first considers the species, and then compares the text surrounding the gene mention with the gene's annotations known from EntrezGene and UniProt: Gene Ontology terms, interaction partners, location and tissue specificity, functional descriptions, etc., as obtained from EntrezGene and UniProt. This comparison yields a score per gene and text, converted into a likelihood and used to rank candidate IDs accordingly.

Note that for our purposes, "genes" includes proteins/enzymes. When we use dictionary-based approaches to find drugs and diseases/ADRs, the dictionaries explicitly link each match to an identifier referring to the respective database (DrugBank, using DB and PA accession codes; UMLS, using CUIs). In the case of diseases, we match each mention found by BANNER to the dictionary to assign an ID (case-insensitive, independent of token order). In case an ID was found, we represent genes, drugs, and diseases/ADRs using an official symbol or preferred term; in the remaining cases, we pick a canonical term based on occurrence frequency, token surface (preferring "Ocular Hypertension" over "ocular hypertension"), and length (preferring shorter terms).

Mentions of genetic variants, alleles, and populations are grounded to canonical names in the following ways. For genetic variants, we identify the position/range and the nucleotides/residues involved to map them to the format recommended by HGVS [19], available for DNA, RNA, and amino acid sequences for mutations, insertions, and deletions affecting single or multiple positions (ranges). We did not yet implement the recommendations for repeats, inversions, conversions, and translocations, due to the infrequent occurrences of either in our data. As canonical names for alleles we use the star notation ('*1'), and the genotype ("TT allele") or fixed terms such as "null allele" and "variant allele" when no specific information is given. Mentions of populations are mapped to a controlled vocabulary, to make orthographic and lexical variations.

2.3. Extraction of relations

Our system extracts twelve types of binary relations between the aforementioned entities, as listed in Table 2. Driven by examples and preliminary evaluation, we decided on a list of seven methods to extract relations. Whether a method gets applied depends on (i) the relation type we are looking for (one of the aforementioned list), (ii) the basic structure of the sentence, and (iii) if another method was able to extract a relation beforehand. We start with sentence-based co-occurrence, which yields good precision for gene–drug, gene–disease, and drug–disease associations already, in particular when refined by using relation-specific keywords (see section on evaluation). For other types of relations, co-occurrence does not yield sufficient precision, and therefore we implement additional extraction methods (detailed in the following):

1. high confidence co-occurrence that includes keywords,
2. co-occurrence without keywords,
3. 1:n co-occurrence,
4. enumerations with matching counts,
5. LCA sub-tree,
6. m:n co-occurrence,
7. low confidence co-occurrence.

Based on results from manual evaluation of predictions, see Section 3.4.1, we sorted these methods by descending expected

Table 2
Numbers of relations found in 179,935 abstracts. Total numbers include duplicate occurrences, in the same and across abstracts. Unique relations include only those in which both entities could be assigned to an identifier or to a canonical name. Relations found in PharmGKB, including/excluding duplicates.

	Total	Unique	PharmGKB
Gene–drug	191,054	31,593	6820/3014
Gene–disease	709,987	102,881	8147/4478
Drug–disease	117,834	26,268	4343/939
Drug–adverse effect	73,696	16,569	
Gene–variant	101,477	21,704	645/516
Gene–allele	65,569	6802	146/99
Gene–RefSNP	12,881	5748	1820/1125
Allele–population	7181	1891	
Variant–population	12,897	6765	
Allele frequency	6,893	279	
Variant frequency	6,646	1654	
Population frequency	8404	144	
Drugs–populations	12,849	4388	
Drugs–alleles	6,778	1858	
Drug–RefSNP	1,161	721	
Drug–mutation	10,809	5491	
Sum	1,315,811	233,964	

confidence, which corresponds to the precision measured. We apply methods 1–7, in this order, to each sentence that has entities of the desired types; once one method extracted a relationship between two entities, we would stop there and not run any of the downstream methods to check the same pair of entities again. To each relationship we therefore assign a confidence score that corresponds to the measured precision of the highest-ranking method that could extract that relation. In the worst case, if methods 1–6 fail, we still extract a relation based on mere sentence-level co-occurrence, but assign a low confidence score.

(1) "High confidence co-occurrence" refers to co-occurrence that includes keywords referring to the type of relation. We apply this strategy to gene–drug, gene–disease, drug–ADR, drug–disease, and mutation–disease associations, and use keywords from PolySearch [24] as well as our own.[9] Every entity of one type is paired with every entity of another type when they co-occur in the same sentence. "High confidence" results from our own estimates for these three types, see our evaluation in Section 3, combined with previous results [25,26]; also see Section 3.4.1 on manual evaluation.

(2) For allele–population and variant–population relationships, we use sentence-level co-occurrence without keywords:

"The LRRK2 G2019S mutation is frequent in apparently sporadic PD in North Africans."

We found only very few instances were such pairings where invalid for these two relation types; and most of those were negations. For some gene–drug, gene–disease, and drug–disease co-occurrences, no keyword was found in the sentence, so they fall into this second category, giving lower confidence than ones found by the first step.

(3) For some types of associations, a relation between all possible pairs can be predicted when one of the entity types has only a single instance in the given sentence, and the other occurs one or more times, called 1:n co-occurrence here. Genetic variants are an example: if a single gene is mentioned in a sentence, together with one or more mutations, most likely all mutations mentioned in the same sentence will refer to this gene. Note that "single instance" includes repetitions of the same entity. Whenever we find multiple instances for both entity types, we would apply one of the following methods (4)–(6). In many cases, in particular for allele and variant frequencies in populations, associations are discussed in either of the following way:

(4) Many associations occur as lists of entities of one type are followed by lists of entities of the other type, with matching counts. An example is the sentence

"The frequencies of CYP1B1*1, *2, *3, and *4 alleles were 0.087, 0.293, 0.444, and 0.175, respectively."

Here, we assign the first frequency mentioned to the first allele, the second to the second, and so on.

(5) If the sentence consists of a list of associations, for instance,

"The allele frequencies were 18.3% (-24T) and 21.2% (1249A)",

we assign associations based on distance in the parse tree, essentially preferring pairs of entities that occur in smaller phrases. The Stanford parser [16] provides us with the grammatical structures of such sentences, represented as a tree. The tree reflects dependencies between phrases, such as from a verb to its subject, or from nouns to their modifiers[10] Our method utilizes the sub-trees of all lowest common ancestors (LCAs) for all potential pairings (in the example above, each potential allele–frequency pair) to subsequently pick the closest pair and exclude those two entities from further pairings. We settled on a maximum distance (edges in the sub-tree connecting both leaf nodes) of ten that showed the best balance between precision and recall.

(6) If none of the criteria for (3)–(5) apply, we build associations between all pairs and call them "m:n co-occurrences". In those, we have only an intermediate confidence, since most likely some predicted associations will be false.

(7) Finally, if all previous filtering steps fail (that is, they do not apply to the pair nor predict an association), we still store each pair in the repository, but assign a low confidence to them.

In addition to relations extracted from sentences, we add abstract-level co-occurrence for completeness. Such associations often provide useful hints on potential relations, although not explicitly mentioned. All information derived in this way are marked as such and appear in dedicated sections of SNPshot to not get mixed with the more certain single sentence-based associations. We also require each abstract-level relation to occur in at least five different abstracts in order to appear in SNPshot.

3. Evaluation and results

3.1. Datasets

We ran two experiments to evaluate intrinsic performances of our approach. (1) To estimate precision and recall of individual extraction components (entity recognition and normalization, relation extraction), we processed more than 3500 PubMed abstracts found via PharmGKB relations and manually checked 2500 predictions. (2) To compare our extracted data to DrugBank and PharmGKB[11] to estimate coverage, we expanded the first set in two steps: (i) we collected all PubMed IDs annotated for genes important for pharmacogenomics from PharmGKB (356 articles for 40 "Annotated PGx genes")[12] and fetched PubMed's "Related Articles", resulting in around 26,000 abstracts and (ii) we searched PubMed with a query that returns abstract likely to discuss variants of human genes together with drug responses and increased risk for disease:

(phenotype OR haplotype OR genotype OR mutation OR SNP.
OR variant OR allele OR polymorphism) AND (disease OR risk.
OR disorder OR malfunction) AND (drug OR bioavailability.
OR metabolize OR inhibit OR orally OR pharmacological).
AND human[MH] AND hasAbstract.

This query yielded more than 35,000 additional abstracts, for each of which we then collected the top 20 related articles using PubMed's ELink utility.[13] All in all, SNPshot currently contains 179,935 PubMed abstracts.

3.2. Data in the SNPshot repository

Tables 1 and 2 summarize all entities and relations found in the two datasets; Table 3 shows the number of relations per extraction

[9] Gazetteers—http://bioai4core.fulton.asu.edu/snpshot/download/keywords.html.

[10] For more information and discussion of a related task, please refer to Coulet et al. [15].

[11] DrugBank files: all_enzyme_protein.fasta, all_target_protein.fasta; PharmGKB: relations.zip, variantAnnotations.zip.

[12] PharmGKB VIP genes—http://www.pharmgkb.org/search/annotatedGene/.

[13] http://eutils.ncbi.nlm.nih.gov/corehtml/query/static/elink_help.html.

Table 3
Numbers of relations found per method, including duplicates and non-identified entities. See column ‘Total’ in Table 2 for types of relations. M*x* refers to the respective number for each method as discussed in Section 2.3.

Method	Total number of relations
M1: Co-occ with keywords	890,828
M2: Simple co-occ	223,113
M3: Co-occ for 1:n	107,408
M7: Low confidence co-occ	44,249
M6: Co-occ for m:n	28,655
M5: LCA sub-tree	19,343
M4: Enumerations	2215
Total	1,315,811

method. We provide more detailed statistics at http://bioai4core.fulton.asu.edu/snpshot/Statistics. For comparison, Table 1 also shows the numbers of unique entities contained in DrugBank and PharmGKB, and Table 2 the relations in PharmGKB. DrugBank contains 12,110 gene–drug relations: 1,039 enzyme–drug and 11,121 drug–target pairs. PharmGKB and DrugBank are manually curated, so more precise information can be expected, but for a lower total number of entities and relations; DrugBank also contains automatically and non-approved relations, see Section 3.4.2.

Analyzing all 179,935 abstracts took around 180 h on an eight core machine with 32 GB of RAM, where we split the data into two batches of six chunks each. The largest proportion of time went into deep parsing with the Stanford parser [16], and recognition and normalization of gene and disease names. We split abstracts into sentences using the Julielab Sentence Boundary Detector,[14] resulting in 1,890,624 sentences.

3.3. Evaluation of entity recognition and normalization

For some components, performance was evaluated previously on external data sets. BANNER’s performance yields an *f*-score of 86% for gene mention recognition, as evaluated on the BioCreative 2 GM test data. On the AZDC corpus for gene-disease associations (see previous section), the disease recognition component yields 77% *f*-score. The gene name normalization component, GNAT, currently achieves 81% *f*-score. Narrowing down the task to identify only genes of a single species, performance varies between 75% and 89% in *f*-score, as evaluated on BioCreative 1 and 2 GN data; GNAT achieves 85% for human genes, which were the focus of this paper.

65% of the gene names could be mapped to an EntrezGene ID (872 k out of 1346 k instances). 4% were either gene lists or protein complexes that would require at least two IDs (“CYP3A4/5”) and were not split properly by our method. The remaining 30% consist of too unspecific gene mentions to assign an ID (for example, a name of a gene family, “major histocompatibility complex”), not human, murine, or rat genes, or false positive gene mentions (“mutation”, “q21”).

3.4. Evaluation of relation extraction

As mentioned above, we performed two evaluations: the first resulted from manual inspection of predicted results; the second evaluation was a comparison of our predictions to the data in DrugBank and PharmGKB, yielding an estimate about coverage. We present details on each evaluation and results next.

3.4.1. Manual evaluation

For the manual evaluation, we ran our system on 3614 abstracts that were discussed with the 40 VIP genes of PharmGKB or associated with these as determined by PubMed’s “Related articles” functionality (see Section 3.1 for details). Five annotators (LT, JH, SL, NV, DV) annotated sentences that had at least one predicted relation, by marking true positive, false positive, and adding false negative entities as well as relations (of all types considered in this paper). Note that this manual annotation covered only sentences for which at least one relation (of any type) was predicted already: our aim was mostly to get an understanding of expected precision, which we then use to assign confidence scores to each relation in our database. All predictions were sorted by sentence, to get a balanced assortment of entities and relations. Tables 4–6 show the performances for extracted entities, extracted relations, and per relation extraction method, respectively. From the manual evaluation, we see that most entities can be found with a high precision, especially the types allele, population, frequency, and mutation. For genes, drugs, and diseases, precision lies between 89% and 92%; here, we note that the system often marks diseases also as adverse effects, so precision for ADRs is low. F-scores for all entities range from 85% to 92% for entity types with reasonable numbers of occurrences. For relations, *f*-score is between 62 and 84%. Here,

Table 4
Performance per entity type, as estimated by manual evaluation. The penultimate row shows macro-averaged precision, recall, and *f*-score; the last row shows the respective micro-averaged results. TP, true positive; FP, false positive; FN, false negative; precision, recall, *f*-score in %.

Entity type	TP	FP	FN	Total	*P*	*R*	*F*
Genes	509	54	40	603	90.4	92.7	91.5
Drugs	117	10	33	160	92.1	78.0	84.5
Diseases	283	32	13	328	89.8	95.6	92.6
Adverse effect			– *Evaluated with diseases* –				
Mutations	53	0	19	72	100.0	73.6	84.8
RefSNP	12	0	0	12	100.0	100.0	100.0
Alleles	102	5	13	120	95.3	88.7	91.9
Populations	30	0	7	37	100.0	81.1	89.6
Frequencies	28	0	0	28	100.0	100.0	100.0
Total	1134	101	125	1360	96.0	88.7	91.9
					91.8	90.1	91.0

Table 5
Performance per type of relation, estimated by manual evaluation.

Relation type	TP	FP	FN	Total	*P*	*R*	*F*
Gene–drug	136	39	18	199	77.7	88.3	82.7
Gene–disease	356	113	22	499	75.9	94.2	84.1
Drug–disease	42	8	15	65	84.0	73.7	78.5
Gene–variant	47	33	17	98	58.8	73.4	65.3
Gene–allele	111	45	12	170	71.2	90.2	79.6
Gene–RefSNP	11	6	0	17	64.7	100.0	78.6
Allele frequency	13	9	2	24	59.1	86.7	70.3
Variant frequency	2	2	0	4	50.0	100.0	66.7
Popul. frequency.	14	15	2	31	48.3	87.5	62.2
Variant–popul.	6	3	0	9	66.7	100.0	80.0
Allele–popul.	13	9	2	25	59.1	86.7	70.3
Total	751	282	90	1141	65.0	89.2	74.4
					72.7	89.3	80.2

Table 6
Performance per method, estimated by manual evaluation.

Method	TP	FP	Relations	P
M1: Co-occ with keywords	500	147	651	77.3
M2: Simple co-occ	53	25	78	67.9
M4: Enumerations	2	0	2	100.0
M5: LCA sub-tree	32	7	39	82.1
M3: Co-occ for 1:n	125	61	186	67.2
M6: Co-occ for m:n	16	11	27	59.3
M7: Low confidence co-occ	23	31	54	42.6

[14] JSBD—https://www.julielab.de/Resources/Software/NLP+Tools.html.

Table 7
Comparison of relations in PharmGKB to the results automatically extracted from 179,935 PubMed abstracts. TP: true positives (found in SNPshot and PharmGKB), FN: false negatives (not found in SNPshot).

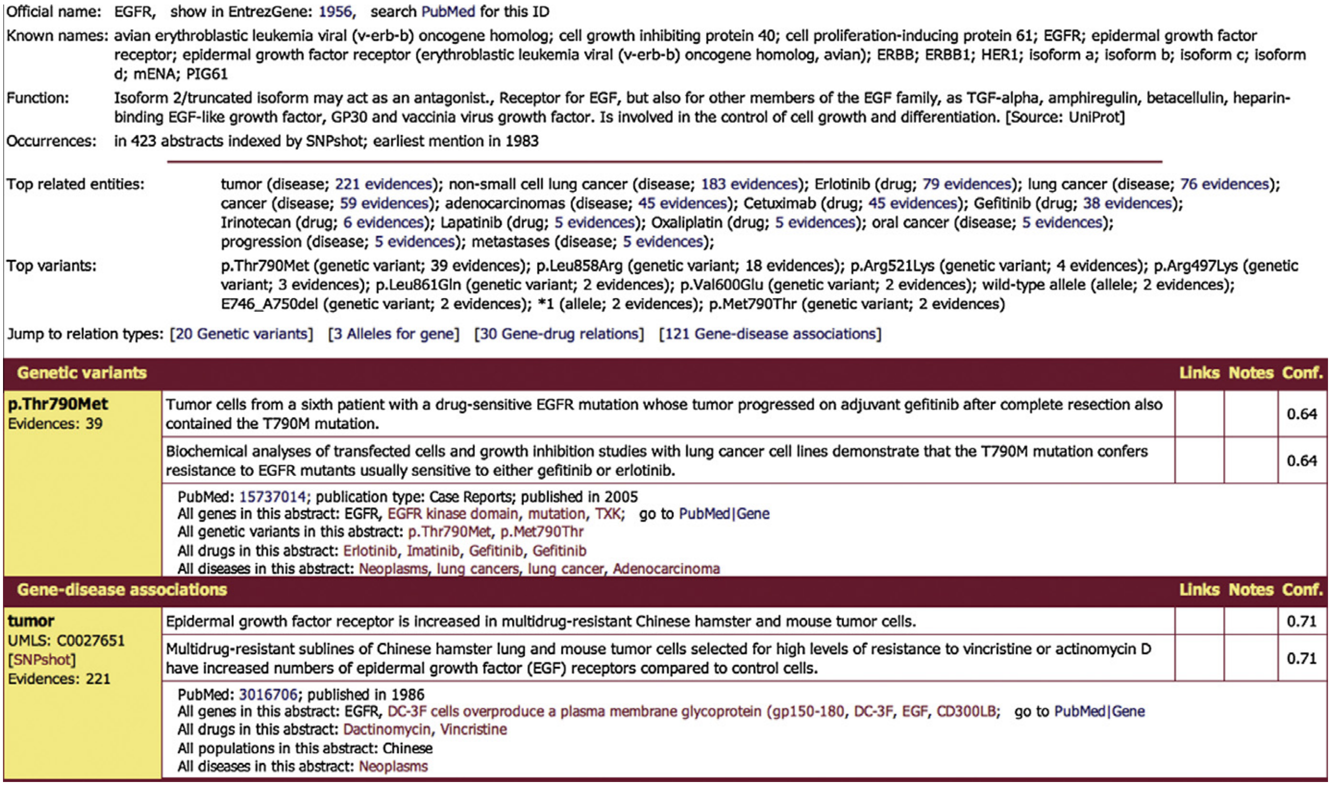
Relation type	Coverage (%)	TP	FN
Gene–drug	93.7	2946	199
Gene–disease	94.7	4683	260
Drug–disease	78.0	981	276
Gene–variant	96.5	505	18
Gene–allele	100.0	100	0
Gene–RefSNP	65.4	744	394

surprisingly, relations between "complex" entities (where NER is concerned) score higher than relations involving the simple types. This can in part be explained by the keyword-filtering we use for gene–drug, gene–disease, and drug–disease relations. These numbers also point out that co-occurrence can still be considered a precise (and not only high recall) baseline technique for extracting certain types of relations, which concurs with findings by others. Chun et al. [25] found that associations of genes with diseases can be predicted with 94% precision using sentence co-occurrence. This holds under the assumption that both entities are extracted correctly by named entity recognition. Chang and Altman [26] estimated a precision of 70% for gene-drug co-occurrence based on 100 predicted pairs. Note that for other types of relations, such as protein–protein interactions, co-occurrence yields less than 50% precision. In SNPshot, entity extraction is aided by entity normalization, namely mapping each entity to a database identifier (EntrezGene ID, DrugBank ID, etc.), which further increases confidence in individual entities as well as relations if a mapping could be found.

3.4.2. Automated evaluation—comparison to DrugBank and PharmGKB

Comparing our predictions based on around 180,000 PubMed abstracts to DrugBank [22] and PharmGKB [5], we can estimate a coverage provided by our method. This enables us to tell how many relations currently stored in DrugBank and PharmGKB can be extracted from PubMed, with confidences per relation type and method as discussed in the previous section. Note that we are not able to properly estimate precision using this method: for a relation predicted by our method that is not contained in either DrugBank or PharmGKB, we cannot automatically decide whether the relation is a false positive, or just not yet contained in either database. In addition, about 50% of the drug targets in DrugBank are not approved and many of them likely false positives; they were predicted by the PolySearch engine based on sentence-level co-occurrence [cmp. 22,Table 1]. We therefore also presented the manual evaluation in the previous section.

For the automated comparison, we considered all relations in the two databases for which a unique identifier or canonical name was available for each partner in the sought relation. Table 7 lists coverages for the types of relations found in the PharmGKB knowledge base. Based on 180,000 PubMed abstracts, SNPshot contains almost 94% of the gene–drug relations in PharmGKB. We found the highest coverage with gene–variant annotations in PharmGKB (96.5%), and the lowest for genes mapped to rs-numbers (65.4%), which could be explained by the low number of total relations of this kind in SNPshot (compare Table 2). Comparing SNPshot with DrugBank, we found that SNPshot can recover 91% of the enzyme–metabolite and drug–target relations in DrugBank; but note that many of the drug–gene relations in DrugBank are also automatically extracted and not approved.



Official name: EGFR, show in EntrezGene: 1956, search PubMed for this ID
Known names: avian erythroblastic leukemia viral (v-erb-b) oncogene homolog; cell growth inhibiting protein 40; cell proliferation-inducing protein 61; EGFR; epidermal growth factor receptor; epidermal growth factor receptor (erythroblastic leukemia viral (v-erb-b) oncogene homolog, avian); ERBB; ERBB1; HER1; isoform a; isoform b; isoform c; isoform d; mENA; PIG61
Function: Isoform 2/truncated isoform may act as an antagonist., Receptor for EGF, but also for other members of the EGF family, as TGF-alpha, amphiregulin, betacellulin, heparin-binding EGF-like growth factor, GP30 and vaccinia virus growth factor. Is involved in the control of cell growth and differentiation. [Source: UniProt]
Occurrences: in 423 abstracts indexed by SNPshot; earliest mention in 1983

Top related entities: tumor (disease; 221 evidences); non-small cell lung cancer (disease; 183 evidences); Erlotinib (drug; 79 evidences); lung cancer (disease; 76 evidences); cancer (disease; 59 evidences); adenocarcinomas (disease; 45 evidences); Cetuximab (drug; 45 evidences); Gefitinib (drug; 38 evidences); Irinotecan (drug; 6 evidences); Lapatinib (drug; 5 evidences); Oxaliplatin (drug; 5 evidences); oral cancer (disease; 5 evidences); progression (disease; 5 evidences); metastases (disease; 5 evidences);
Top variants: p.Thr790Met (genetic variant; 39 evidences); p.Leu858Arg (genetic variant; 18 evidences); p.Arg521Lys (genetic variant; 4 evidences); p.Arg497Lys (genetic variant; 3 evidences); p.Leu861Gln (genetic variant; 2 evidences); p.Val600Glu (genetic variant; 2 evidences); wild-type allele (allele; 2 evidences); E746_A750del (genetic variant; 2 evidences); *1 (allele; 2 evidences); p.Met790Thr (genetic variant; 2 evidences)

Jump to relation types: [20 Genetic variants] [3 Alleles for gene] [30 Gene-drug relations] [121 Gene-disease associations]

Genetic variants		Links	Notes	Conf.
p.Thr790Met Evidences: 39	Tumor cells from a sixth patient with a drug-sensitive EGFR mutation whose tumor progressed on adjuvant gefitinib after complete resection also contained the T790M mutation.			0.64
	Biochemical analyses of transfected cells and growth inhibition studies with lung cancer cell lines demonstrate that the T790M mutation confers resistance to EGFR mutants usually sensitive to either gefitinib or erlotinib.			0.64
	PubMed: 15737014; publication type: Case Reports; published in 2005 All genes in this abstract: EGFR, EGFR kinase domain, mutation, TXK; go to PubMed|Gene All genetic variants in this abstract: p.Thr790Met, p.Met790Thr All drugs in this abstract: Erlotinib, Imatinib, Gefitinib, Gefitinib All diseases in this abstract: Neoplasms, lung cancers, lung cancer, Adenocarcinoma			

Gene-disease associations		Links	Notes	Conf.
tumor UMLS: C0027651 [SNPshot] Evidences: 221	Epidermal growth factor receptor is increased in multidrug-resistant Chinese hamster and mouse tumor cells.			0.71
	Multidrug-resistant sublines of Chinese hamster lung and mouse tumor cells selected for high levels of resistance to vincristine or actinomycin D have increased numbers of epidermal growth factor (EGF) receptors compared to control cells.			0.71
	PubMed: 3016706; published in 1986 All genes in this abstract: EGFR, DC-3F cells overproduce a plasma membrane glycoprotein (gp150-180, DC-3F, EGF, CD300LB; go to PubMed|Gene All drugs in this abstract: Dactinomycin, Vincristine All populations in this abstract: Chinese All diseases in this abstract: Neoplasms			

Fig. 1. Excerpt from the data sheet for the human epidermal growth factor receptor, EGFR, showing a summary of information on the gene, lists of predicted related entities, and examples for a genetic variant and disease association. View the entire entry at http://bioai4core.fulton.asu.edu/snpshot/FactSheet?id=1956&type=GENE.

4. Visualization

SNPshot is available at http://bioai4core.fulton.asu.edu/snpshot. Users can search the interface by gene, drug, disease, mutation, adverse reaction, RefSNP number, allele, population, Entrez Gene ID, DrugBank accession code, PharmGKB ID, and PubChem Substance and Compound IDs. SNPshot shows a "Fact sheet" reflecting all extracted relations concerning this entity. All facts are linked to DrugBank and PharmGKB when possible, and backed up by evidence sentences, shown together with sentences that contain related information from the same source. SNPshot thus provides a quick overview over genetic variations, implicated in drug response and/or diseases. The fact sheets, when searched by drugs, can help guiding clinical decisions or research questions by first scanning the literature available on the topic in a condensed manner. They provide a quick overview of how genetic variants affect the drug response of individual patients. We also envision that SNPshot could first be searched based on populations (that is, for race/ethnicity of the patient), even before genetic finger-printing takes place, to first estimate the likelihood that the patient might carry the suspicious allele. Fig. 1 shows an exemplary repository entry for the EGFR gene, together with two drugs.

For summaries and to download and copy-and-paste information, two additional views are also available from each "Fact sheet"; one displays a tabular summary of all related entities, the other returns a tab-separated file. Downloads from SNPshot data can also be triggered by parameterized HTTP requests. For instance, the request http://QuickOverview?id=6557&type=GENE&rettype=tsv will fetch all information on the gene SLC12A1 (EntrezGene ID 6557) in a tab-separated form. More information on formats and parameters are available in the "About" section of the website.

5. Discussion and conclusions

In this paper, we presented automated methods to populate a repository, called SNPshot, related to genetic variants and their associations with drugs and diseases. SNPshot links facts covering genes, variants/mutations, alleles, frequencies, and populations, with drug and disease/adverse reactions data. It can be used to guide decision processes in personalized medicine, linking drug treatments to known impacts of "non-wild-type" enzymatic activity. SNPshot indexes around 180,000 abstracts from PubMed and thus allows to quickly search and browse these types of information, in addition to manually curated databases such as PharmGKB and DrugBank. It can also function as an aid to populate these renowned collections.

Based on the results from manual evaluation, we assigned confidence values to each method per relation type, which are reflected in the repository to indicate how certain each predicted association is. This allows to sort the repository by certainty, to first use relations that are more likely to be correct for further analysis, annotation, or integration. Due to the limited size of our manual evaluation, it has to be noted that confidence scores for the more prevalent relation types, such as gene–drug and gene–allele, are more reliable than others, including variants in populations. From the larger-scale automated comparison against other databases, we found that SNPshot covers between 65% (RefSNPs mapped to genes) and 95% (genes associated with diseases) of the relations in PharmGKB, based on 180,000 PubMed abstracts. For certain types of relations, the coverage reaches 100%, but the number of relations to compare with is too low to draw definite conclusions (such as for the 100 gene to allele mappings). Regarding gene–drug associations, we were able to recover 91% of the data in DrugBank, and 94% of PharmGKB. Starting with 31,000 unique gene–drug associations (enzyme-drug or drug-target), and taking into account false positive rates for genes and drugs (around 10% each) and gene–drug relations (25%), there are more than 10,000 potentially relevant gene–drug associations left in SNPshot that are yet to be (hand-) curated in other databases. SNPshot can thus suggest those for curation, and we are looking into mechanisms for ranking entities and associations for novelty, coverage in databases, and relevance to disease, and relevance to pharmacokinetics and -dynamics.

We seek to further integrate our data with facts known about the "druggable genome" [27], to better identify potential drug targets and annotate known targets with variants. Comparing the data in SNPshot to DrugBank reveals that our selection procedure to find relevant PubMed abstracts focuses more on drug–metabolizing enzymes than a larger number of known drug targets, which we want to address in the future. We categorize each gene according to its Gene Ontology annotations, searching for the closes parent in the GO hierarchy that is either "metabolic process', "transporter activity", "ion channel activity", "receptor activity", or "other". This helps in identifying the potential role of the gene product, by also looking for corresponding evidence in the sentence a gene–drug relation was found in.

For future work we envision mapping of gene/disease entities and relations to OMIM, and further integration with PharmGKB and SIDER. Following Chang and Altman [26], it would be valuable to automatically assign evidences to categories such as *Clinical outcome*, *Pharmacokinetics*, and *Genotype* (already used in PharmGKB) to "enhance retrieval and stimulate hypothesis generation" [26]. GeneOntology employs a system of evidence codes, ranging from *Author statement* and *Automatically assigned evidence* to *Experimental evidence* codes, and we want to add similar annotations to SNPshot relations based on the respective source of information to reflect reliability. Precision for named entity recognition, grounding, and normalization is high (above 90%) for all entities expect for adverse drug reactions, which were often confused by the system with diseases, as these overlap to a large extent. Adding supervised classification on top of recognition of disease and ADR mentions to disambiguate between these two classes is one of the next steps under development, since training data is now becoming available through in-house evaluation and external user input. As next steps concerning types of information, we want to concentrate on adding data on cell lines and additional kinds of phenotypes, including gene expression and RNA interference. In the same way that we categorize proteins, see above, it will be useful to include drug classes, such as antipsychotics and steroids, in the displayed information as well as allow them in searches.

Acknowledgments

We kindly acknowledge funding by the National Science Foundation (VN, SL, BL), Science Foundation Arizona (LT, RL), Fulbright International Student Program Russia (DV), and Arizona State University (JH, CB). We would like to thank the anonymous reviewers whose suggestions helped to improve this manuscript.

References

[1] Wuttke H, Rau T, Heide R, Bergmann K, Böhm M, Weil J, et al. Clin Pharmacol Ther 2002;72:429–37.
[2] Zhou S-F. Clin Pharmacokinet 2009;48:689–723.
[3] Kirchheiner J, Brøsen K, Dahl ML, Gram LF, Kasper S, Roots I, et al. Acta Psychiatr Scand 2001;104:173–92.
[4] Lazarou J, Pomeranz BH, Corey PN. JAMA 1998;279:1200–5.
[5] Klein TE, Chang JT, Cho MK, Easton KL, Fergerson R, Hewett M, et al. Pharmacogenomics J 2001;1:167–70.
[6] Altman RB, Klein TE. Ann Rev Pharmacol Toxicol 2002;42:113–33.
[7] Deisboeck T. Mol Syst Biol 2009;5:249.

[8] Wang Z, Kim S, Quinney SK, Guo Y, Hall SD, Rocha LM, et al. J Biomed Inform 2009;42:726–35.

[9] Theobald M, Shah N, Shrager J. In: Proc AMIA translat bioinf; 2009.

[10] Doughty E, Kertesz-Farkas A, Bodenreider O, Thompson G, Adadey A, Peterson T, et al. Bioinformatics 2011;27:408–15.

[11] Aronson AR, Lang F-M. J Am Med Inform Assoc 2010;17:229–36.

[12] Laurila JB, Naderi N, Witte R, Riazanov A, Kouznetsov A, Baker CJ. BMC Genomics 2010;11:S24.

[13] Caporaso JG, Baumgartner JWilliamA, Randolph DA, Cohen KB, Hunter L. Bioinformatics 2007;23:1862–5.

[14] Tari L, Hakenberg J, Gonzalez G, Baral C. In: Proc Pacific symposium on biocomputing, Kona, Hawaii, USA;2010.

[15] Coulet A, Shah NH, Garten Y, Musen M, Altman RB. J Biomed Inform 2010;43:1009–19.

[16] Klein D, Manning C. Accurate unlexicalized parsing. In: Proceedings of the 41st meeting of the association for computational linguistics; 2003. p. 423–30.

[17] Leaman B, Gonzalez G. In: Proc Pac symp biocomput, Hawaii; 2008.

[18] Smith L, Tanabe LK, Johnson R, Kuo C-J, Chung I-F, Hsu C-N, et al. Genome Biol 2008;8:S2.

[19] den Dunnen JT, Antonarakis SE. Human Mutation 2000;15:7–12.

[20] Leaman R, Miller C, Gonzalez G. In: Proc int symp languages in biology and medicine, South Korea; 2009. p. 82–9.

[21] Kuhn M, Campillos M, Letunic I, Jensen LJ, Bork P. Mol Syst Biol 2010;6:343.

[22] Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, et al. Nucl Acids Res 2008;36:D901-6.

[23] Hakenberg J, Plake C, Leaman R, Schroeder M, Gonzalez G. Bioinformatics 2008;24:126–32.

[24] Cheng D, Knox C, Young N, Stothard P, Damaraju S, Wishart DS. Nucl Acid Res 2008;36:W399–405.

[25] Chun H-W, Tsuruoka Y, Kim J-D, Shiba R, Nagata N, Hishiki T, et al. In: Proc Pac symp biocomput; 2006.

[26] Chang JT, Altman RB. Pharmacogenetics 2004;14:577–86.

[27] Russ AP, Lampel S. Drug Discov Today 2005;10:1607–10.