Journal of Biomedical Informatics 43 (2010) 1009–1019

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Using text to build semantic networks for pharmacogenomics

Adrien Coulet [a,b], Nigam H. Shah [a], Yael Garten [c], Mark Musen [a], Russ B. Altman [a,b,c,d,*]

[a] Department of Medicine, 300 Pasteur Drive, Room S101, Mail Code 5110, Stanford University, Stanford, CA 94305, USA
[b] Department of Genetics, Mail Stop-5120, Stanford University, Stanford, CA 94305, USA
[c] Stanford Biomedical Informatics, 251 Campus Drive, MSOB, Room X215, Mail Code 5479, Stanford University, Stanford, CA 94305, USA
[d] Department of Bioengineering, 318 Campus Drive, Room S172, Mail Code 5444, Stanford University, Stanford, CA 94305, USA

ARTICLE INFO

Article history:
Received 13 May 2010
Available online 17 August 2010

Keywords:
Relationship extraction
Pharmacogenomics
Natural Language Processing
Ontology
Knowledge acquisition
Data integration
Biological network
Text mining
Information extraction

ABSTRACT

Most pharmacogenomics knowledge is contained in the text of published studies, and is thus not available for automated computation. Natural Language Processing (NLP) techniques for extracting relationships in specific domains often rely on hand-built rules and domain-specific ontologies to achieve good performance. In a new and evolving field such as pharmacogenomics (PGx), rules and ontologies may not be available. Recent progress in syntactic NLP parsing in the context of a large corpus of pharmacogenomics text provides new opportunities for automated relationship extraction. We describe an ontology of PGx relationships built starting from a lexicon of key pharmacogenomic entities and a syntactic parse of more than 87 million sentences from 17 million MEDLINE abstracts. We used the syntactic structure of PGx statements to systematically extract commonly occurring relationships and to map them to a common schema. Our extracted relationships have a 70–87.7% precision and involve not only key PGx entities such as genes, drugs, and phenotypes (*e.g.*, VKORC1, warfarin, clotting disorder), but also critical entities that are frequently modified by these key entities (*e.g.*, VKORC1 polymorphism, warfarin response, clotting disorder treatment). The result of our analysis is a network of 40,000 relationships between more than 200 entity types with clear semantics. This network is used to guide the curation of PGx knowledge and provide a computable resource for knowledge discovery.

1. Introduction

Most biological knowledge exists in published scientific text. In order to support the creation of databases and to enable the discovery of new relationships, there is great interest in extracting relationships automatically. Several successful efforts use manually created rules to define patterns of relationships between entities. These approaches are efficient when used in domains that are of limited scope, such as protein–protein interactions or protein transport. However, the complexity and diversity of the semantics used to describe relationships in broad or evolving domains, such as pharmacogenomics (PGx), are harder to capture. Thus, no general set of rules exists for extracting the relationships relevant to such fields, and creating/maintaining them manually would be tedious and time consuming.

Syntactic sentence parsers can identify the *subject*, *object* and *type* of relationships using grammatical rules. General statistical parsing techniques have recently emerged, and there are several general-purpose parsers that yield reasonable results when applied to scientific text. These parsers depend on the need for good domain-specific lexicons of key entities, since named-entity recognition for particular fields in science can be difficult. We consider named-entity recognition as the process of identifying members of the lexicon within the text, amidst other words. With such lexicons, there is an opportunity to use syntactic sentence parsers to identify rich rule sets automatically. These rule sets take advantage of sentence structure and grammar to extract more precise information. In addition, these rule sets can be organized in an ontology that allows normalization of relationships and inference over them.

Pharmacogenomics (PGx) is the study of how individual genomic variations influence drug–response phenotypes. PGx knowledge exists for the most part in the scientific literature in sentences that mention relationships. We can represent a large fraction of this knowledge as binary relationships R (*a*, *b*), where *a*, and *b* are *subjects* and *objects* related by a relationship of type R. Sometimes, *a* and *b* are instances of a gene (*e.g.*, *VKORC1 gene*), drug (*e.g.*, *warfarin*), or phenotype (*e.g.*, *clotting disorder*). As we shall demonstrate later, very often *a* and *b* are entities that are modified by genes (*e.g.*, VKORC1 polymorphism), drugs (*e.g.*, warfarin dose) or phenotypes (*e.g.*, clotting disorder treatment). R is a type of relation described by words such as "inhibits", "transports", or "treats" and their synonyms. Thus, although the three *key*

* Corresponding author at: Department of Bioengineering, 318 Campus Drive, Room S172, Mail Code 5444, Stanford University, Stanford, CA 94305, USA. Fax: +1 650 723 8544.
E-mail address: russ.altman@stanford.edu (R.B. Altman).

1532-0464/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2010.08.005

entities in PGx (genes, drugs, and phenotypes) can be target nouns for relation extraction, they are more often indicators of latent PGx knowledge, as they modify other concepts to create a second set of entities required to precisely describe PGx relationships. We call these *modified entities* in contrast with the key entities that modify and expand them. These modified entities can be any biomedical entity, such as a gene variation, drug effect, or disease treatment. For example, the gene entity *VKORC1* (a key entity) is used as a modifier of the concept *polymorphism* in "VKORC1 polymorphisms affect warfarin response," indicating that *VKORC1 polymorphism* is a critical (composite) PGx entity. This sentence also indicates that a modified entity, *warfarin response*, will be important as well.

In this paper we present a method for using a syntactical parser to identify recurrent binary relationships that express PGx knowledge. Many of these relationships use genes, drugs and phenotypes as modifiers of other entities. We organized these relationships and the associated entities in an ontology that maps diverse sentence structures and vocabularies to a common semantics. We processed 87 million sentences using this ontology to capture and normalize more than 40,000 specific PGx relationships. These relationships are summarized in the form of a semantic network (*i.e.*, a network where entities (nodes) and relationships (edges) are associated with the semantics defined in our ontology). We anticipate that they will be useful to assist database curation and as a foundation for knowledge discovery and data mining.

2. Related work

Our work is partially motivated by our efforts building the Pharmacogenomic Knowledge Base, PharmGKB (http://www.pharmgkb.org/) [1]. PharmGKB aims to catalog all knowledge of how human genetic variation impacts drug–response phenotypes, and is a manually curated database that summarizes published gene–drug–phenotype relationships. The rapidly increasing size of the pharmacogenomic literature threatens to overwhelm the PharmGKB curators. Automatic approaches using NLP techniques are therefore promising. Methods based on co-occurrence assume that entities occurring together in a sentence are related, but the semantics of the relationships are not typically captured. Nevertheless, these approaches efficiently identify potential relationships that can subsequently be evaluated manually. For example, the Pharmspresso system uses co-occurrence to group frequently co-mentioned genes, genomic variations, drugs, and diseases [2]. These groups are then used to assist manual curation. Li et al. used the co-occurrence of drug and disease names in MEDLINE abstracts to derive drug–disease relations and to build a disease-specific drug–protein network [3]. Blaschke et al. and Rosario et al. expanded this co-occurrence approach to extract more complete relations by searching for "tri-co-occurrence" [4,5]. Tri-co-occurrence refers to the co-occurrence of two named entities and one type of relationship in a unique piece of text. Statistical analysis of co-occurrence can help derive semantic similarities between entities [6].

In contrast to co-occurrence, syntactic parsing can explicitly identify relationships between two entities in text [7]. Hand-coded parsing rules can extract protein–protein interactions and protein transport relationships [8,9]. Fundel et al. defined three general patterns of relations (specifying the semantic type of subjects and objects, and using a lexicon of association words) to identify protein–protein interactions [10]. For example their pattern "effector – relation – effectee" enables the capture of relationships of the form "protein A activates protein B". The OpenDMAP system also uses patterns to identify protein interaction and transport [11]. Ahlers et al. used vocabularies and semantic types of the UMLS (Unified Medical Language System) to specify patterns to extract gene–disease and drug–disease relationships [12]. Several groups have used extracted relationships to create networks, including molecular interaction networks [13], gene–disease networks [14], regulatory gene expression networks [15], and gene–drug–disease networks [16]. In order to be efficient, these syntactic approaches often rely on large sets of patterns and stable ontologies to guarantee performance on diverse sentence structures. Unfortunately, a systematic catalog of patterns for pharmacogenomics is not available [17,18].

The Semantic Web community has developed methods for learning ontologies from text using unsupervised approaches [19,20]. Most of these efforts focus on learning hierarchies of concepts. Ciaramita et al. studied unsupervised learning of relationships between concepts [21]. Their method produces a network of concepts where edges are associated with precise semantics (*e.g.*, Virus encodes Protein). Other efforts have focused on enriching existing ontologies for NLP using Web content [22]. Cilibrasi and Vitányi proposed a method to automatically learn the semantics of processed words, hypothesizing that semantically related words co-occur more frequently in Web pages than do unrelated words [23]. Gupta and Oates used Web content to identify concept mappings for previously unrecognized words discovered while processing text [24].

We describe here our method of relationship extraction that uses (1) syntactic rules to extract relationships and (2) a learned ontology to normalize those relationships.

3. Methods

Fig. 1 gives an overview of the four steps of our method, described in the following sub-sections. The first input is a corpus of article abstracts split into individual sentences. We benefit from previous work that made such a corpus available and also provides a convenient way to retrieve the sentences [25]. We use lexicons of PGx key entities (drugs, genes, and phenotypes) from PharmGKB[1] to retrieve sentences mentioning pairs of key entities. We parse retrieved sentences with the Stanford Parser and represent the sentence using a convenient data structure called a "Dependency Graph" [26]. Each retrieved sentence is analyzed to extract the raw relationships between key entities themselves or other entities that they modify. After applying this procedure to many pairs of key entities, we gather the raw relationships and entities and manually map them to a much smaller set of "normalized" relationships and entities based on synonymy, arranged hierarchically in an OWL ontology.[2] We assume that this ontology is representative of PGx relationships mentioned in our corpus. This ontology can then be applied to all raw relationship instances in the corpus to create a very large set of normalized relationships representing the semantic content of the corpus.

3.1. Sentence parsing of MEDLINE into Dependency Graphs

The goal of the first step is to provide, in a format easy to process, the syntactical structure of sentences that potentially mention a PGx relationship. We focus on sentences that mention at least two PGx key entities. We used an index of individual sentence of MEDLINE abstracts published before 2009 (17,396,436 abstracts and 87,806,828 sentences) processed by Xu et al. [25]. This index has been built on the full text of sentences with the Lucene library and can consequently be queried with any term [27]. It returns sentences that have been indexed with the query terms and also returns "parse trees" that correspond to retrieved sentences. A

[1] http://www.pharmgkb.org/resources/downloads_and_web_services.jsp.
[2] OWL (Web Ontology Language): http://www.w3.org/TR/owl-features/.