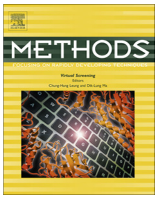
Methods 74 (2015) 83–89

Contents lists available at ScienceDirect

Methods

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



DISEASES: Text mining and data integration of disease–gene associations

Sune Pletscher-Frankild [a], Albert Pallejà [a,b], Kalliopi Tsafou [a], Janos X. Binder [c,d], Lars Juhl Jensen [a,*]

[a] Department of Disease Systems Biology, Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark
[b] Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark
[c] Structural and Computational Biology Unit, European Molecular Biology Laboratory (EMBL), Heidelberg, Germany
[d] Bioinformatics Core Facility, Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, Luxembourg

ARTICLE INFO

Article history:
Received 21 January 2014
Received in revised form 15 November 2014
Accepted 25 November 2014
Available online 5 December 2014

Keywords:
Text mining
Named entity recognition
Information extraction
Data integration
Web resource

ABSTRACT

Text mining is a flexible technology that can be applied to numerous different tasks in biology and medicine. We present a system for extracting disease–gene associations from biomedical abstracts. The system consists of a highly efficient dictionary-based tagger for named entity recognition of human genes and diseases, which we combine with a scoring scheme that takes into account co-occurrences both within and between sentences. We show that this approach is able to extract half of all manually curated associations with a false positive rate of only 0.16%. Nonetheless, text mining should not stand alone, but be combined with other types of evidence. For this reason, we have developed the DISEASES resource, which integrates the results from text mining with manually curated disease–gene associations, cancer mutation data, and genome-wide association studies from existing databases. The DISEASES resource is accessible through a web interface at http://diseases.jensenlab.org/, where the text-mining software and all associations are also freely available for download.

1. Introduction

Linking human genes to the diseases in which they are involved lies at the very heart of molecular medicine. Such links can be made through a variety of different types of studies, including classical pedigree-based genetics studies of Mendelian and complex diseases, genome-wide association studies (GWAS), somatic mutation frequencies, transcriptomics and proteomics studies, and detailed molecular biology studies of individual proteins. Because the relevant data come from so many types of experiments performed by researchers working in different disciplines, such as geneticists and molecular biologists, all the relevant data are not collected in a single place, making it difficult to get a comprehensive overview of which genes are involved in which diseases. However, due to the vast amount of research being performed on the topic, much has been written in the biomedical literature about the associations between genes and diseases. Extracting disease–gene associations from text is thus an obvious use case for text mining, and disease–gene associations have indeed previously been extracted by generalized co-occurrence-based text-mining systems [1–4].

Besides addressing the technical tasks of text mining, which we outline in the next section, it is important to consider how to make the text-mining solution as useful as possible to biologists. To this end, we believe it is crucial to view text mining, not as an isolated problem, but as a means to integrate the literature with other relevant data. A major challenge here is to handle the heterogeneity, varied quality, and scattered nature of the data in a manner that brings together the available evidence for disease–gene associations. Moreover, it is important to ensure that the resource does not become a silo, but that it instead is integrated with related resources, in particularly established resources that have a broad user base that reaches beyond bioinformatics and text-mining experts.

Here we describe the DISEASES resource, which aims to be the most comprehensive freely available database of disease–gene associations. To this end, we have developed open-source text-mining software that recognizes diseases and human genes in text and extracts disease–gene associations. We integrate the

---

Abbreviations: COSMIC, Catalog of Somatic Mutations in Cancer; GHR, Genetics Home Reference; GWAS, Genome Wide Association Study; HGNC, HUGO Gene Nomenclature Committee; HUGO, Human Genome Organization; ICD, International Classification of Diseases; LD, linkage disequilibrium; MeSH, Medical Subject Headings; NER, named entity recognition; SNOMED CT, Systematized Nomenclature of Medicine – Clinical Terms; UMLS, Unified Medical Language System.

* Corresponding author at: Novo Nordisk Foundation Center for Protein Research, Blegdamsvej 3b, 2200 Copenhagen N, Denmark.

E-mail address: lars.juhl.jensen@cpr.ku.dk (L.J. Jensen).

http://dx.doi.org/10.1016/j.ymeth.2014.11.020
1046-2023/

associations extracted through automatic text mining with evidence from databases with permissive licenses, namely manually curated associations from Genetics Home Reference (GHR) [5] and UniProt Knowledgebase (UniProtKB) [6], GWAS results from DistiLD [7], and mutation data from Catalog of Somatic Mutations in Cancer (COSMIC) [8]. To make the data easy to use for large-scale analyses, we map all sources of evidence to common identifiers, assign them comparable quality scores, and make them available for bulk download. We also make the information available as a web resource (http://diseases.jensenlab.org/) aimed at end users interested in individual diseases or genes.

2. Background and related work

2.1. Named entity recognition (NER)

Recognizing named entities and concepts, such as genes and diseases, in text is the basis for most biomedical applications of text mining [9]. NER is sometimes divided into two subtasks, namely recognition and normalization (also known as identification or grounding), the former being to recognize the words of interest and the latter being to map them to the correct identifiers in databases or ontologies. However, as recognition without normalization has very limited practical use, the normalization step is now often implicitly considered part of the NER task.

The main challenges in NER are the poor standardization of names and the fact that a name of, for example, a gene or disease may have other meanings [10]. To recognize names in text, many systems thus make use of rules that look at features of names themselves, such as capitalization and word endings, as well as contextual information from nearby words. In early methods the rules were hand crafted [11], whereas newer methods make use of machine learning [12,13], relying on the availability of manually annotated text corpora.

Dictionary-based methods instead rely—as the name suggests—on matching a dictionary of names against text. For this purpose the quality of the dictionary is obviously very important; the best performing methods for NER according to blind assessments rely on carefully curated dictionaries to eliminate synonyms that give rise to many false positives [14,15]. Moreover, dictionary-based methods have the crucial advantage of being able to normalize names. Whether or not one makes use of machine learning, a high-quality, comprehensive dictionary of gene and disease names is thus a prerequisite for mining disease–gene associations from the biomedical literature.

2.2. Controlled vocabularies of diseases

It is fairly straightforward to find a good starting point for a dictionary of human gene names due to efforts such as the Human Genome Organization (HUGO) Gene Nomenclature Committee (HGNC) [16] and UniProtKB [6]. It is less obvious to find a good dictionary of disease names, as there are several competing classifications and ontologies, which are designed for different purposes, mutually inconsistent, and thus poorly integrated with each other.

In a clinical setting, various versions of the International Classification of Diseases (ICD; http://www.who.int/classifications/icd/) are almost ubiquitously used for coding diagnoses in electronic health records (EHRs) and derived health registries [17]. European countries, Canada, and Australia use revision 10 (ICD-10), whereas the United States still use revision 9 (ICD-9). ICD-10 is not just an update to ICD-9; it is a restructured diagnosis classification, and no official mapping exists between the two revisions. Because ICD is designed for clinical coding and billing purposes, its structure and disease names are poorly suited for biomedical literature mining. It is, however, useful for text mining of clinical narrative in EHRs, especially because it has been translated to many languages [18].

A newer alternative is the Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT; http://www.ihtsdo.org/snomed-ct/). It cross maps to several revisions of ICD and has a considerably broader scope than just diseases. SNOMED-CT is one of many terminologies combined in the even broader Unified Medical Language System (UMLS) Metathesaurus; another is Medical Subject Headings (MeSH; http://www.ncbi.nlm.nih.gov/mesh/). Dictionaries based on subsets of UMLS have been used for recognition of disease names with varying success in text-mining tools, such as MetaMap [19], Medical Language Extraction and Encoding (MedLEE) [20], and the Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES) [21]. However, because UMLS contains many distinct concepts that are very close in meaning even human annotation of UMLS concepts in text is problematic [22]. Licenses for SNOMED-CT and other terminologies in UMLS further restrict their use in resources intended for redistribution.

In contrast to these, the Disease Ontology [23] is part of the Open Biomedical Ontologies (OBO) Foundry initiative [24]. It cross maps to UMLS and has extensive annotation of synonyms. Consequently, Disease Ontology works well for recognition of disease names mentioned in Gene Reference Into Function (GeneRIF; http://www.ncbi.nlm.nih.gov/gene/about-generif/) entries [25].

2.3. Information extraction (IE)

Having addressed the NER task using appropriate dictionaries of gene and disease names, the next task is to extract information on associations between genes and diseases. There are two fundamentally different approaches to IE: natural language processing (NLP), using a grammar to parse the syntax of each sentence, and statistical co-occurrence methods [9]. We focus on the latter approach, which is highly flexible and generally gives better recall, but worse precision, than NLP [1,26,27]. Other disadvantages of co-occurrence methods are that they are unable to extract the direction of an association and have difficulty distinguishing between direct and indirect associations [9]. However, neither of these disadvantages is important with respect to extracting disease–gene associations.

Almost all co-occurrence methods implement a frequency-based scoring scheme to account for the fact that a pair of entities or concepts may co-occur a few times without being in any way related [3,27,28]. These scoring schemes have traditionally counted either the number of sentences or the number of abstracts in which the pair co-occurred, and both sizes of text units have merit [26]. We have therefore recently introduced a scoring scheme that simultaneously takes into account both sentence-level and abstract-level co-occurrences [29].

Disease–gene associations extracted from Medline abstracts can already be searched through generalized co-occurrence tools such as CoPub [1,2] and FACTA+[3,4]. However, as these resources are technology-centric — focusing on text mining — they do not take into account any other types of evidence. This limitation is aggravated by the fact that neither resource allows bulk download of all associations, making it difficult for others to integrate additional evidence.

2.4. Disease–gene association databases

Several existing databases focus on or contain disease–gene associations, mainly obtained through manual curation of the biomedical literature. Unfortunately, most of these use an in-house controlled vocabulary of diseases and are subject to restrictive licenses, which makes it difficult to integrate them both from a technical and from a legal standpoint. The oldest and most famous