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/© 2014 The Authors. Published by Elsevier Inc.

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 of databases is Online Mendelian Inheritance in Man (OMIM; http://omim.org/), which we provide hyperlinks to for both genes and diseases. More recent efforts include the Human Gene Mutation Database (HGMD) [30], the Comparative Toxicogenomics Database (CTD) (http://ctdbase.org/) [31,32], and GHR (http://ghr.nlm.nih.gov/) [5]. In addition to these dedicated disease–gene association databases, UniProtKB also annotates diseases associated with each gene [6].

Databases also exist that deal with specific diseases or types of diseases, most notably cancer. The COSMIC database is the most comprehensive source of information on somatic mutations and their frequencies in human cancers [8]. Mutation data is manually curated from the primary literature and annotated according to a histology and tissue ontology.

Over the last decade, GWAS have produced data on thousands of single nucleotide polymorphisms (SNPs) associated with the risk of hundreds of diseases. GWAS data are, however, non-trivial to work with for the non-expert, because they identify marker SNPs that are often not the actual causal SNPs [33,34]. For this reason GWAS results must be analyzed in the context of linkage disequilibrium (LD), which is defined as the non-random association of variants at two or more loci [34,35]. GWAS Central (http://www.gwascentral.org/) is a centralized database that collects the results from genetic association studies [36]. Unfortunately it provides data only for small- to medium-scale investigations and explicitly forbids using the data to create similar public resources. By contrast, the National Human Genome Research Institute (NHGRI) GWAS Catalog (http://www.genome.gov/gwastudies/) is public domain [37]. The latter is thus the basis for the derived databases DistiLD [7] and GWASdb [38] databases, which show disease-associated SNPs and genes in their chromosomal context.

3. Material and methods

3.1. Dictionary construction

For human gene and protein names, we used the alias file from STRING v9.1 [29], which integrates names from Ensembl [39], UniProtKB [6], and HGNC [16]. We orthographically expanded the gene symbols with the prefix ‘h’, which means *human* and is commonly used in the literature to disambiguate a human gene from its identically named orthologs in model organisms.

To construct a dictionary of diseases for use in NER, we extracted all names and synonyms from the Disease Ontology [23]. Comparing these to the dictionary of human gene names revealed that the HGNC gene symbol of a disease gene was in many cases listed in Disease Ontology as a synonym for the disease in which the gene is implicated. For example, BRCA1 and BRCA2 were listed as exact synonyms for *hereditary breast ovarian cancer*. As this would be a major source of ambiguity in the combined dictionary, we explicitly filtered out disease names that are identical to HGNC gene symbols.

To improve recall, we next automatically generated variants of the disease names. Although the terms *disease*, *disorder*, and *syndrome* have separate definitions, we found that they are used inconsistently in the literature when part of disease names; for example, *Alzheimer’s disease* is occasionally referred to as *Alzheimer’s disorder* or *Alzheimer’s syndrome*. To address this we automatically generate the two other variants if either of them is in the dictionary. Similarly, the adjectives *hereditary* and *familial* are used interchangeably, and we thus automatically replace one with the other. We also removed words in parentheses and brackets occurring at the end of disease names, unless this would cause ambiguity.

3.2. Recognition of gene and disease names in text

To match a document against the dictionary, we have developed a highly efficient tagging algorithm, which is implemented in C++. The algorithm is described in full detail elsewhere [40], but is summarized here for completeness. Tests of the tagging speed and memory efficiency of the implementation compared to another popular tagger are also provided in our earlier publication [40].

We first tokenize the text on white space characters and special characters, such as hyphen and slash, and identify the leftmost longest matches by looking up all substrings consisting of up to 15 consecutive tokens. To make these lookups fast while handling character case variation as well as spacing and hyphenation of multiwords, we used a custom hash table to store the dictionary. The hash table is case insensitive, disregards white space characters and hyphens within name, and trims off other punctuation characters, such as quotes and parentheses, at the beginning and end of names. To match also acronyms that are not in the dictionary, we use a regular expression to search definitions of acronyms within the text and look up their long forms in the dictionary. Crucially, we globally block tagging of names that would otherwise give rise to many false positives by manually inspecting the tagging results of all names that occur more than 2000 times in Medline. Many of the blocked names are acronyms; for example, the acronym for *disseminated intravascular coagulation* is DIC, which can also mean *deviance information criteria*, *differential interference contrast*, and *dissolved inorganic carbon*. By keeping track of all names that we have inspected — whether they were blocked or not — we are able to efficiently update the list of blocked names as both Medline and the dictionary grows. For each name recognized in the text we normalize it to the corresponding unique identifier and, in case of diseases, backtrack the term to the root of the ontology through is_a relationships to assign also the identifiers of all parent terms.

3.3. Extraction and scoring of disease–gene associations

We score associations between proteins and diseases using the scoring scheme previously described [41], which is also the basis for the co-occurrence-based text-mining scores in STRING v9.1 [29] and COMPARTMENTS [42]. For completeness we reiterate the scoring scheme here.

An important feature of the scoring scheme is that it simultaneously takes into account co-occurrences at the level of abstracts as well as individual sentences. To this end, we first calculate a weighted count ($C(G, D)$) for each pair of a gene (G) and a disease (D) over the n abstracts in the text corpus:

$$C(G,D) = \sum_{k=1}^{n} w_s \delta_{sk}(G,D) + w_a \delta_{ak}(G,D)$$

where $w_a = 3$ and $w_s = 0.2$ are the weights for co-occurrence within the same abstract and the same sentence, respectively, and the delta functions $\delta_{ak}(G, D)$ and $\delta_{sk}(G, D)$ signify whether or not G and D co-occur in abstract k or a sentence within it. A co-occurrence score ($S(G, D)$) is calculated from the weighted counts as:

$$S(G,D) = C(G,D)^{\alpha} \left(\frac{C(G,D)C(\cdot,\cdot)}{C(G,\cdot)C(\cdot,D)} \right)^{1-\alpha}$$

where $C(G, \cdot)$ is the sum over all diseases paired with gene G, $C(\cdot, D)$ is the sum over all genes paired with disease D, the normalizing factor $C(\cdot, \cdot)$ is the sum over all pairs of genes and diseases, and the weighting factor $\alpha = 0.6$. All parameters (w_a, w_s, and α) have in earlier work been optimized to give the best possible performance on finding functionally associated genes [29]. An important property of this function is that it not only rewards for the gene and disease being mentioned together, but also penalizes for them being frequently mentioned together with other diseases or genes, respectively.

We next convert the co-occurrence scores ($S(G, D)$) to z-scores ($Z(G, D)$), which are easier to interpret and are robust to changes in the size of the text corpus. We assume that the empirically observed score distribution is a mixture of the true signal and a lower-scoring random background, which we model as a Gaussian distribution. The full details of this score conversion have been published elsewhere [41]. Finally, we calculate the confidence score (stars) as $Z(G, D)/2$, limited to a maximum of four stars to account for automatic text mining never being as reliable as manually curated annotations.

3.4. Integration of curated knowledge

The GHR database does not provide download files for use in large-scale analyses. We thus used an automated crawler to download the web page for each disease and store the disease name, which is part of the uniform resource locator (URL), along with any gene symbols listed on the web page. We were able to map the names of 390 diseases to Disease Ontology using the dictionary we developed for text mining. The pages are regularly recrawled to update with new associations; the numbers used in the manuscript are based on what was downloaded on May 31, 2013.

In case of UniProtKB, associations to diseases can be found in the KW lines through the use of 149 keywords from the UniProtKB controlled vocabulary of keywords. We were able to manually map 132 of the 149 disease keywords to corresponding concepts in the Disease Ontology. Most of the keywords that we could not map, such as *Disease mutation*, were not disease names.

We mapped HGNC gene symbols from GHR and identifiers from UniProtKB to their identifiers in STRING v9.1 using the alias file [29]. We subsequently used the explicitly annotated disease–gene associations from GHR and UniProtKB to infer broader Disease Ontology concepts via the is_a relationships in the ontology. As all disease–gene annotations imported and inferred from the two databases are based on manual curation, we assigned them a confidence score of five stars.

3.5. Benchmark of text-mining results

To assess the quality of the text-mining results, we constructed a reference set based on the manually curated annotations imported from GHR and UniProtKB. Due to the hierarchical nature of the Disease Ontology, it is necessary to select on a subset of terms to be used as the basis for the assessment. To this end, we chose to use the subset of terms that were explicitly annotated in the two databases (as opposed to inferred through is_a relationships). In case one term was a child term of another, we selected the broader parent term. This resulted in a positive reference set of 2780 associations between 2001 genes and 173 diseases. We defined the negative set as all other 343393 possible pairings of the same genes and diseases.

We next sorted the text-mined associations descending by score and compared them to the reference set. We present the results as receiver operating characteristic (ROC) curves by plotting the true positive rate (TPR) as function of false positive rate (FPR), considering either all disease–gene associations or only the best-scoring association per gene (Fig. 1). We compare these results to two random backgrounds. One is simple random shuffling of the disease–gene pairs, which ignores that some diseases are associated with many more genes than others. To correct for this, the second random background is calculated by sorting the disease–gene pairs descending by prior probability of the disease. Because the prior of each disease is estimated based on the reference set itself, this likely overestimates the performance that can be attained by random guessing.

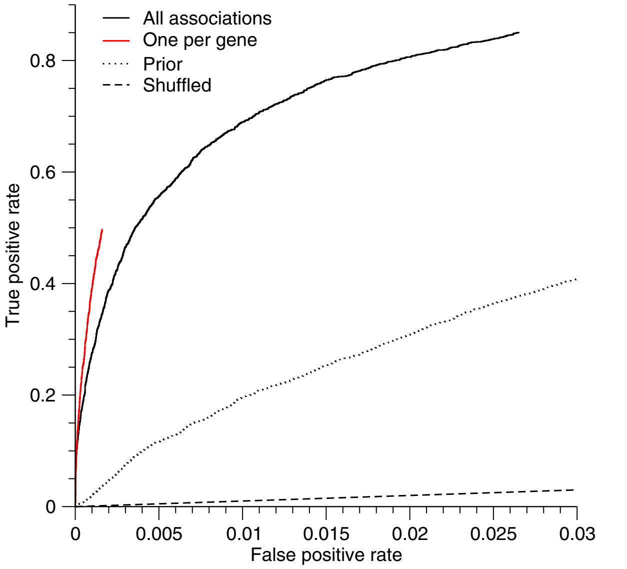


Fig. 1. Benchmark of disease–gene associations obtained through text mining. The receiver operating characteristic (ROC) curves show the true positive rate (TPR) as function of false positive rate (FPR) when considering all associations (black) and when considering only the highest scoring association for each gene (red). The dashed and dotted curves show the random expectations according to simple shuffling and prior-based ranking, respectively. The curves do not intercept TPR = 1 and FPR = 1, because some disease–gene pairs in the benchmark set are not found mentioned together in Medline, for which reason they have no text-mining score.

3.6. Integration of mutation and GWAS data

To integrate cancer mutation data from COSMIC [8], we manually created mappings between terms listed in the fields “Site primary” and “Histology” and Disease Ontology concepts classified under “organ system cancer” and “cell type cancer”, respectively. We mapped the genes to STRING v9.1 identifiers via the Ensembl transcript identifiers provided by COSMIC. For each pair of a gene (G) and a disease (D) we counted the number of disease samples carrying at least one somatic missense or nonsense mutation within the gene ($N(G, D)$). We discarded pairs with a count less than 10 and derived confidence scores (stars) as $\log_{10}(N(G, D)) - 0.5$, limiting it to at most four stars.

To include also GWAS data, we integrated information from the DistiLD database [7], which maps genes and disease-associated SNPs onto so-called LD blocks defined based on data from the HapMap Project [43]. We assigned each SNP with a p-value less than 10^{-5} to the nearest gene within the same LD block. The “Disease/Trait” descriptors from the NHGRI GWAS Catalog were mapped to the corresponding Disease Ontology concepts through the ICD-10 annotations from DistiLD, the Disease Ontology Lite annotations from GWASdb [38], and manual inspection of conflicts. The resulting disease–gene associations were assigned a confidence score (stars) using the formula $3 - \log_{10}(\max(P, P_{\min}))$, where P is the p-value, $P_{\min}$ is the genome-wide GWAS significance threshold ($5 \cdot 10^{-8}$).

4. Results and discussion

4.1. Dictionary-based tagger software

We have developed a highly efficient NER method for diseases and human genes, which are normalized to identifiers from

Disease Ontology [23] and STRING v9.1 [29], respectively. On a server with two Intel E5520 processors and 24 GB of random access memory (RAM), starting the tagger and loading the dictionary took only 4.2 s. Once started, the tagger used 260 MB of RAM and was able to process 360 Medline abstracts per second on a single processor core (measured on a corpus of 100,000 Medline abstracts). The tagger software bundled with a dictionary of disease and human gene names is available for download under the BSD license.

4.2. Cooccurrence-based disease–gene associations

Because the NER task is for us only a step on the way towards the goal of extracting disease–gene associations, we chose to focus our benchmarking effort on assessing the quality of the end result. We therefore compared the text-mined associations to the manually curated associations imported from GHR and UniProtKB in two ways: (1) considering all disease–gene associations, and (2) considering only the highest scoring disease for each gene. The results of these comparisons (Fig. 1) show that our text-mining system is able to extract a large fraction of the known disease–gene associations with high specificity (low FPR). If a user was to simply trust the highest scoring disease association for each gene, 50% of all manually curated disease–gene associations in the benchmark set would be found at a FPR of only 0.16%.

The high quality of text-mining results is reflected by the fact that they are already being used extensively. The text-mined associations from DISEASES are included in the widely used GeneCards database [44]. They have also been used as a basis for inference of disease associations for miRNAs from their predicted target genes [41] and for enrichment analysis of autism-related genes [45].

4.3. Contents of the database

Although we have in this paper placed most emphasis on the text-mining aspects, the DISEASES database integrates disease–gene associations from several sources. This is advantageous, because every source of associations has its shortcomings. Table 1 provides an overview of the total evidence landscape of the database, showing that the text-mining pipeline is indeed the largest single contributor of associations. However, it is important to note that this number depends strongly on the confidence cutoff; indeed the number of associations obtained from the manually curated databases rivals the number of text-mined associations with at least 3 confidence stars. Mutation data from COSMIC and GWAS data from DistiLD also both contribute a sizeable number of associations; however, the former data source only relates genes to cancers.

Table 1
Overview of disease–gene association evidence. Each row shows the number of genes, diseases and associations between them that are supported by a given type, confidence level (in case of text mining), or source (in case of Knowledge and Experiments). The numbers in parentheses specify the counts prior to backtracking of Disease Ontology terms through is_a relationships.

Evidence	Genes	Diseases	Associations
Text mining	15631	4598	478407
4 star confidence	478	662	1044
3 star confidence	3207	2267	15226
2 star confidence	12706	4354	142892
Knowledge	2001	735 (453)	15231 (2953)
Genetics Home Reference	965	671 (390)	7551 (1169)
UniProtKB	1651	271 (120)	11576 (2187)
Experiments	10711	423 (264)	89073 (20206)
COSMIC	8786	142 (76)	55791(13050)
DistiLD	4315	351 (210)	36650 (7185)
Total	17606	4610	543405

All disease–gene associations from all evidence sources are available for bulk download in tab-delimited format under the Creative Commons Attribution (CC-BY) license.

4.4. The DISEASES web interface

Whereas tab-delimited files are convenient for bioinformaticians wanting to perform large-scale analyses or create derived resources, a web interface better caters to researchers interested in individual genes or diseases. We have thus developed a web interface for the DISEASES resource that allows users to either query for a gene to find associated diseases or query for a disease to find associated genes (Fig. 2). In either case, the user will be presented with three tables called Knowledge, Experiments, and Text mining. These show the manually curated associations from GHR and UniProtKB, the mutation and association data from COSMIC and DistiLD, and the text-mined associations, respectively. Besides summarizing the imported information, the Knowledge and Experiments tables provide direct hyperlinks to the source entries in the external databases.

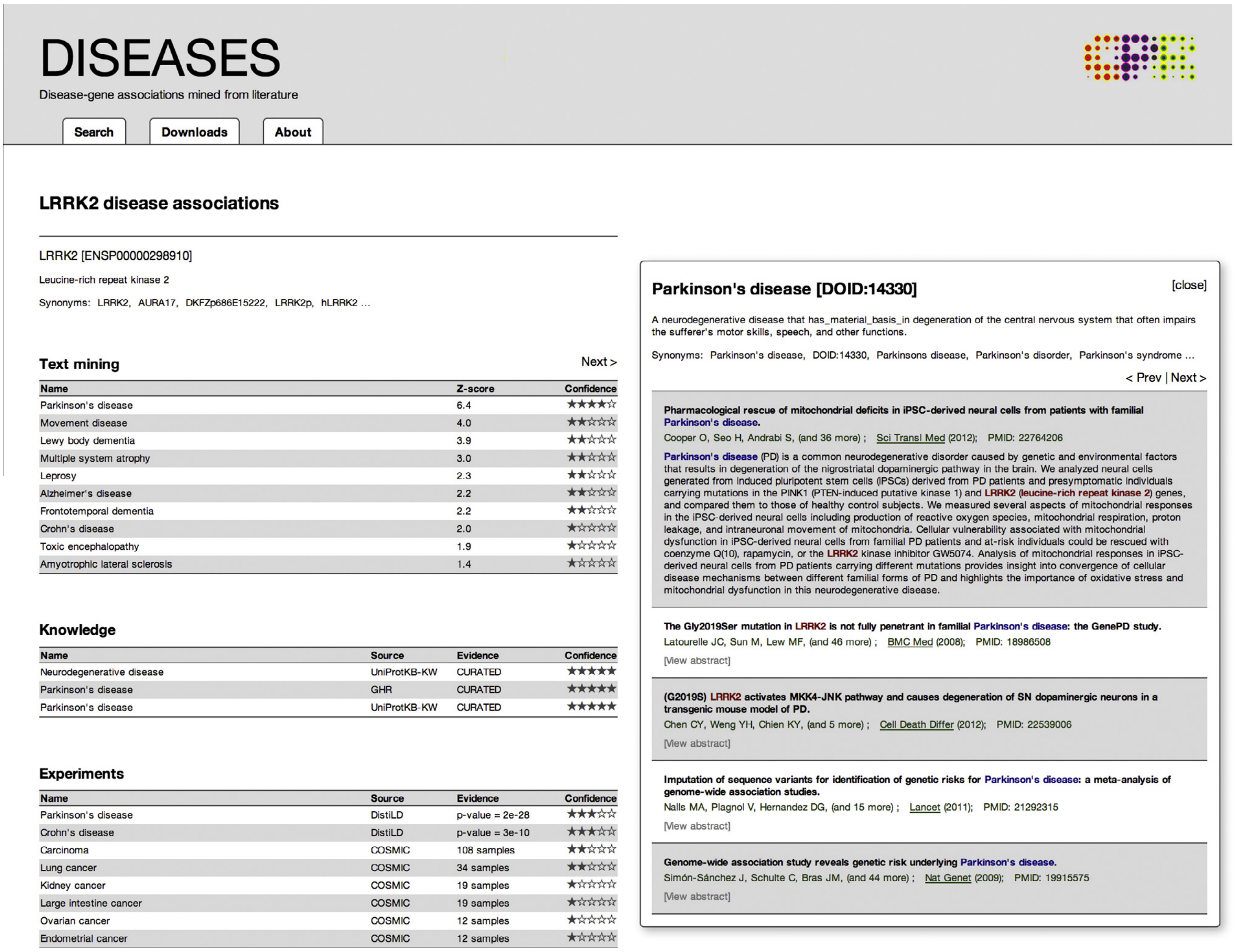
The table summarizing the text-mined evidence deserves special attention. As the text-mining method correctly takes into account information from the narrower child terms of each disease, the text-mined disease associations for a gene have inherent redundancy. When showing the list of diseases associated with a gene of interest, the web interface thus dynamically filters out redundant Disease Ontology terms for which better alternatives are present. The web interface also gives the user the possibility to inspect the text-mining evidence behind any disease–gene association by viewing the underlying abstracts with the gene and disease names highlighted.

4.5. Generality of the approach

The approach to text mining described in this paper is readily applicable to recognize other types of named entities in text and extract associations among them. Using the same tagger with a dictionary constructed from the NCBI Taxonomy [46], we were able to accurately identify taxonomic names in the biomedical literature [40]. We are currently extending that work to identify environments from the Environment Ontology [47] in text, for example, from the Encyclopedia of Life [48]. We have even used a slightly modified version of the tagger as part of a method for recognition of adverse drug events in Danish clinical narratives [49]. This illustrates the flexibility of a simple dictionary-based NER approach in terms of applicability to new knowledge domains.

Combining the tagger with the co-occurrence scoring scheme for the purpose of IE is equally flexible. As previously mentioned, the scoring scheme was originally developed to extract functional associations between proteins for use in the STRING database based on co-occurrence of gene names within biomedical literature [29]. In addition to using it for disease–gene associations as described here, we have since applied the same scoring scheme to extract information on protein–small molecule associations in the STITCH database [50], protein subcellular localization in the COMPARTMENTS database [42], and tissue distribution of proteins in the TISSUES database (http://tissues.jensenlab.org/).

Besides using the same methods for NER and IE, DISEASES and the other resources mentioned above have in common that they integrate heterogeneous evidence from many sources. This sets them aside from the many resources that use text mining to extract associations between a wide variety of named entities and concepts. As tool developers, it is easiest and most efficient to be technology-centric and apply a single technology, such as text mining, to a wide range of topics. However, from a user's perspective, a resource that integrates many sources of information



LRRK2 disease associations

LRRK2 [ENSP00000298910]

Leucine-rich repeat kinase 2

Synonyms: LRRK2, AURA17, DKFZp686E15222, LRRK2p, hLRRK2 ...

Text mining Next >

Name	Z-score	Confidence
Parkinson's disease	6.4	★★★★☆
Movement disease	4.0	★★☆☆☆
Lewy body dementia	3.9	★★☆☆☆
Multiple system atrophy	3.0	★★☆☆☆
Leprosy	2.3	★★☆☆☆
Alzheimer's disease	2.2	★★☆☆☆
Frontotemporal dementia	2.2	★★☆☆☆
Crohn's disease	2.0	★☆☆☆☆
Toxic encephalopathy	1.9	★☆☆☆☆
Amyotrophic lateral sclerosis	1.4	★☆☆☆☆

Knowledge

Name	Source	Evidence	Confidence
Neurodegenerative disease	UniProtKB-KW	CURATED	★★★★★
Parkinson's disease	GHR	CURATED	★★★★★
Parkinson's disease	UniProtKB-KW	CURATED	★★★★★

Experiments

Name	Source	Evidence	Confidence
Parkinson's disease	DistiLD	p-value = 2e-28	★★★☆☆
Crohn's disease	DistiLD	p-value = 3e-10	★★★☆☆
Carcinoma	COSMIC	108 samples	★★☆☆☆
Lung cancer	COSMIC	34 samples	★★☆☆☆
Kidney cancer	COSMIC	19 samples	★☆☆☆☆
Large intestine cancer	COSMIC	19 samples	★☆☆☆☆
Ovarian cancer	COSMIC	12 samples	★☆☆☆☆
Endometrial cancer	COSMIC	12 samples	★☆☆☆☆

Parkinson's disease [DOID:14330] [close]

A neurodegenerative disease that has_material_basis_in degeneration of the central nervous system that often impairs the sufferer's motor skills, speech, and other functions.

Synonyms: Parkinson's disease, DOID:14330, Parkinsons disease, Parkinson's disorder, Parkinson's syndrome ...

< Prev | Next >

Pharmacological rescue of mitochondrial deficits in iPSC-derived neural cells from patients with familial Parkinson's disease.
Cooper O, Seo H, Andrabi S, (and 36 more) ; Sci Transl Med (2012); PMID: 22764206

Parkinson's disease (PD) is a common neurodegenerative disorder caused by genetic and environmental factors that results in degeneration of the nigrostriatal dopaminergic pathway in the brain. We analyzed neural cells generated from induced pluripotent stem cells (iPSCs) derived from PD patients and presymptomatic individuals carrying mutations in the PINK1 (PTEN-induced putative kinase 1) and LRRK2 (leucine-rich repeat kinase 2) genes, and compared them to those of healthy control subjects. We measured several aspects of mitochondrial responses in the iPSC-derived neural cells including production of reactive oxygen species, mitochondrial respiration, proton leakage, and intraneuronal movement of mitochondria. Cellular vulnerability associated with mitochondrial dysfunction in iPSC-derived neural cells from familial PD patients and at-risk individuals could be rescued with coenzyme Q(10), rapamycin, or the LRRK2 kinase inhibitor GW5074. Analysis of mitochondrial responses in iPSC-derived neural cells from PD patients carrying different mutations provides insight into convergence of cellular disease mechanisms between different familial forms of PD and highlights the importance of oxidative stress and mitochondrial dysfunction in this neurodegenerative disease.

The Gly2019Ser mutation in LRRK2 is not fully penetrant in familial Parkinson's disease: the GenePD study.
Latourelle JC, Sun M, Lew MF, (and 46 more) ; BMC Med (2008); PMID: 18986508
[View abstract]

(G2019S) LRRK2 activates MKK4-JNK pathway and causes degeneration of SN dopaminergic neurons in a transgenic mouse model of PD.
Chen CY, Weng YH, Chien KY, (and 5 more) ; Cell Death Differ (2012); PMID: 22539006
[View abstract]

Imputation of sequence variants for identification of genetic risks for Parkinson's disease: a meta-analysis of genome-wide association studies.
Nalls MA, Plagnol V, Hernandez DG, (and 15 more) ; Lancet (2011); PMID: 21292315
[View abstract]

Genome-wide association study reveals genetic risk underlying Parkinson's disease.
Simón-Sánchez J, Schulte C, Bras JM, (and 44 more) ; Nat Genet (2009); PMID: 19915575
[View abstract]

Fig. 2. The DISEASES web resource. The figure shows how the disease–gene associations are presented in the web interface, exemplified by the LRRK2 gene. The three tables provide the user with an overview of the evidence from text mining, curated knowledge, and experimental data. Clicking on an association, e.g. to Parkinson's disease, in the text mining table gives access to the underlying abstracts with the co-occurring gene and disease highlighted. The two other tables provide hyperlinks to the relevant entries in the source databases.

pertaining to a single topic of interest is usually what is sought after. We attempt to find a compromise by creating a general framework, which allows us to set up resources that each integrate information on a different topic but are maintainable, because they share software infrastructure.

5. Conclusions

We have developed a dictionary-based NER tool for Disease Ontology concepts and combined it with a co-occurrence scoring scheme to efficiently and accurately extract disease–gene associations from Medline. We have integrated these with manually curated associations from the GHR and UniProtKB databases as well as somatic mutation and GWAS data from COSMIC and DistiLD, respectively. We make the resulting database available as a searchable web resource at http://diseases.jensenlab.org/, where bulk datasets and the NER software are also available for download.

Acknowledgments

This work was in part funded by the Novo Nordisk Foundation Center for Protein Research (NNF14CC0001) and by the European Union's Seventh Framework Programme (FP7/2007-2013) under Grant agreement n259348. The authors thank Andreas Bok Andersen for help with developing the web interface.

References

[1] B.T.F. Alako, A. Veldhoven, S. van Baal, R. Jelier, S. Verhoeven, T. Rullmann, et al., BMC Bioinform. 6 (2005) 51, http://dx.doi.org/10.1186/1471-2105-6-51.
[2] W.W.M. Fleuren, S. Verhoeven, R. Frijters, B. Heupers, J. Polman, R. van Schaik, et al., Nucleic Acids Res. 39 (2011) W450–W454, http://dx.doi.org/10.1093/nar/gkr310.
[3] Y. Tsuruoka, J. Tsujii, S. Ananiadou, Bioinformatics 24 (2008) 2559–2560, http://dx.doi.org/10.1093/bioinformatics/btn469.
[4] Y. Tsuruoka, M. Miwa, K. Hamamoto, J. Tsujii, S. Ananiadou, Bioinformatics 27 (2011) i111–i119, http://dx.doi.org/10.1093/bioinformatics/btr214.
[5] J.A. Mitchell, J. Med. Libr. Assoc. 94 (2006) 336–342.
[6] The UniProt Consortium, Nucleic Acids Res. 42 (2014) D191–D198, http://dx.doi.org/10.1093/nar/gkt1140.
[7] A. Pallejà, H. Horn, S. Eliasson, L.J. Jensen, Nucleic Acids Res. 40 (2012) D1036–D1040, http://dx.doi.org/10.1093/nar/gkr899.
[8] S.A. Forbes, G. Bhamra, S. Bamford, E. Dawson, C. Kok, J. Clements, et al., Curr. Protoc. Hum. Genet. (2008), http://dx.doi.org/10.1002/0471142905.hg1011s57 (chapter 10, Unit 10.11).
[9] L.J. Jensen, J. Saric, P. Bork, Nat. Rev. Genet. 7 (2006) 119–129, http://dx.doi.org/10.1038/nrg1768.
[10] L. Chen, H. Liu, C. Friedman, Bioinformatics 21 (2005) 248–256, http://dx.doi.org/10.1093/bioinformatics/bth496.
[11] K. Fukuda, A. Tamura, T. Tsunoda, T. Takagi, Pac. Symp. Biocomput. (1998) 707–18. <http://www.ncbi.nlm.nih.gov/pubmed/9697224> (accessed 16.01.14).
[12] B. Settles, Bioinformatics 21 (2005) 3191–3192, http://dx.doi.org/10.1093/bioinformatics/bti475.

[13] G. Zhou, D. Shen, J. Zhang, J. Su, S. Tan, BMC Bioinform. 6 (Suppl. 1) (2005) S7, http://dx.doi.org/10.1186/1471-2105-6-S1-S7.

[14] D. Hanisch, K. Fundel, H.-T. Mevissen, R. Zimmer, J. Fluck, BMC Bioinform. 6 (Suppl. 1) (2005) S14, http://dx.doi.org/10.1186/1471-2105-6-S1-S14.

[15] S. Gaudan, H. Kirsch, D. Rebholz-Schuhmann, Bioinformatics 21 (2005) 3658–3664, http://dx.doi.org/10.1093/bioinformatics/bti586.

[16] K.A. Gray, L.C. Daugherty, S.M. Gordon, R.L. Seal, M.W. Wright, E.A. Bruford, Nucleic Acids Res. 41 (2013) D545–D552, http://dx.doi.org/10.1093/nar/gks1066.

[17] P.B. Jensen, L.J. Jensen, S. Brunak, Nat. Rev. Genet. 13 (2012) 395–405, http://dx.doi.org/10.1038/nrg3208.

[18] F.S. Roque, P.B. Jensen, H. Schmock, M. Dalgaard, M. Andreatta, T. Hansen, et al., PLoS Comput. Biol. 7 (2011) e1002141, http://dx.doi.org/10.1371/journal.pcbi.1002141.

[19] A.R. Aronson, F.-M. Lang, J. Am. Med. Inform. Assoc. 17 (2010) 229–236, http://dx.doi.org/10.1136/jamia.2009.002733.

[20] C. Friedman, H. Liu, L. Shagina, S. Johnson, G. Hripcsak, Proc. AMIA Symp. (2001) 189–93. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2243298&tool=pmcentrez&rendertype=abstract>.

[21] G.K. Savova, J.J. Masanz, P.V. Ogren, J. Zheng, S. Sohn, K.C. Kipper-Schuler, et al., J. Am. Med. Inform. Assoc. 17 (2010) 507–513, http://dx.doi.org/10.1136/jamia.2009.001560.

[22] H. Kilicoglu, G. Rosemblat, M. Fiszman, T.C. Rindflesch, BMC Bioinform. 12 (2011) 486, http://dx.doi.org/10.1186/1471-2105-12-486.

[23] L.M. Schriml, C. Arze, S. Nadendla, Y.-W.W. Chang, M. Mazaitis, V. Felix, et al., Nucleic Acids Res. 40 (2012) D940–D946, http://dx.doi.org/10.1093/nar/gkr972.

[24] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, et al., Nat. Biotechnol. 25 (2007) 1251–1255, http://dx.doi.org/10.1038/nbt1346.

[25] J.D. Osborne, J. Flatow, M. Holko, S.M. Lin, W.A. Kibbe, L.J. Zhu, et al., BMC Genomics 10 (Suppl. 1) (2009) S6, http://dx.doi.org/10.1186/1471-2164-10-S1-S6.

[26] J. Ding, D. Berleant, D. Nettleton, E. Wurtele, Pac. Symp. Biocomput. (2002) 326–37. http://www.ncbi.nlm.nih.gov/pubmed/11928487 (accessed 16.01.14).

[27] J.D. Wren, H.R. Garner, Bioinformatics 20 (2004) 191–198, http://dx.doi.org/10.1093/bioinformatics/btg390.

[28] T.K. Jenssen, A. Laegreid, J. Komorowski, E. Hovig, Nat. Genet. 28 (2001) 21–28, http://dx.doi.org/10.1038/88213.

[29] A. Franceschini, D. Szklarczyk, S. Frankild, M. Kuhn, M. Simonovic, A. Roth, et al., Nucleic Acids Res. 41 (2013) D808–D815, http://dx.doi.org/10.1093/nar/gks1094.

[30] P.D. Stenson, M. Mort, Hum. Genet. (2013), http://dx.doi.org/10.1007/s00439-013-1358-4.

[31] A.P. Davis, C.G. Murphy, R. Johnson, J.M. Lay, K. Lennon-Hopkins, C. Saraceni-Richards, et al., Nucleic Acids Res. 41 (2013) D1104–D1114, http://dx.doi.org/10.1093/nar/gks994.

[32] A.P. Davis, T.C. Wiegers, P.M. Roberts, B.L. King, J.M. Lay, K. Lennon-Hopkins, et al., Database (Oxford) 2013 (2013) bat 080, http://dx.doi.org/10.1093/database/bat080.

[33] M.I. McCarthy, G.R. Abecasis, L.R. Cardon, D.B. Goldstein, J. Little, J.P.A. Ioannidis, et al., Nat. Rev. Genet. 9 (2008) 356–369, http://dx.doi.org/10.1038/nrg2344.

[34] M. Slatkin, Nat. Rev. Genet. 9 (2008) 477–485, http://dx.doi.org/10.1038/nrg2361.

[35] D. Altshuler, M.J. Daly, E.S. Lander, Science 322 (2008) 881–888, http://dx.doi.org/10.1126/science.1156409.

[36] G.A. Thorisson, O. Lancaster, R.C. Free, R.K. Hastings, P. Sarmah, D. Dash, et al., Nucleic Acids Res. 37 (2009) D797–D802, http://dx.doi.org/10.1093/nar/gkn748.

[37] L.A. Hindorff, P. Sethupathy, H.A. Junkins, E.M. Ramos, J.P. Mehta, F.S. Collins, et al., Proc. Natl. Acad. Sci. U.S.A. 106 (2009) 9362–9367, http://dx.doi.org/10.1073/pnas.0903103106.

[38] M.J. Li, P. Wang, X. Liu, E.L. Lim, Z. Wang, M. Yeager, et al., Nucleic Acids Res. 40 (2012) D1047–D1054, http://dx.doi.org/10.1093/nar/gkr1182.

[39] P. Flicek, I. Ahmed, M.R. Amode, D. Barrell, K. Beal, S. Brent, et al., Nucleic Acids Res. 41 (2013) D48–D55, http://dx.doi.org/10.1093/nar/gks1236.

[40] E. Pafilis, S.P. Frankild, L. Fanini, S. Faulwetter, C. Pavloudi, A. Vasileiadou, et al., PLoS ONE 8 (2013) e65390, http://dx.doi.org/10.1371/journal.pone.0065390.

[41] S. Mørk, S. Pletscher-Frankild, A. Palleja, Bioinformatics (2013), http://dx.doi.org/10.1093/bioinformatics/btt677.

[42] J.X. Binder, S. Pletscher-Frankild, K. Tsafou, C. Stolte, S.I. O'Donoghue, R. Schneider, et al., Database (Oxford) 2014 (2014) bau012, http://dx.doi.org/10.1093/database/bau012.

[43] The International HapMap Consortium, Nature 437 (2005) 1299–1320, http://dx.doi.org/10.1038/nature04226.

[44] M. Safran, I. Dalah, J. Alexander, N. Rosen, T. Iny Stein, M. Shmoish, et al., Database (Oxford) 2010 (2010) baq020, http://dx.doi.org/10.1093/database/baq020.

[45] B.E. Eisinger, M.C. Saul, T.M. Driessen, S.C. Gammie, BMC Neurosci. 14 (2013) 147, http://dx.doi.org/10.1186/1471-2202-14-147.

[46] E.W. Sayers, T. Barrett, D.A. Benson, S.H. Bryant, K. Canese, V. Chetvernin, et al., Nucleic Acids Res. 37 (2009) D5–D15, http://dx.doi.org/10.1093/nar/gkn741.

[47] P.L. Buttigieg, N. Morrison, B. Smith, C.J. Mungall, S.E. Lewis, J. Biomed. Semantics 4 (2013) 43, http://dx.doi.org/10.1186/2041-1480-4-43.

[48] E.O. Wilson, Trends Ecol. Evol. 18 (2003) 77–80, http://dx.doi.org/10.1016/S0169-5347(02)00040-X.

[49] R. Eriksson, P.B. Jensen, S. Frankild, L.J. Jensen, S. Brunak, J. Am. Med. Inform. Assoc. 20 (2013) 947–953, http://dx.doi.org/10.1136/amiajnl-2013-001708.

[50] M. Kuhn, D. Szklarczyk, S. Pletscher-Frankild, T.H. Blicher, C. von Mering, L.J. Jensen, et al., Nucleic Acids Res. 42 (2014) D401–D407, http://dx.doi.org/10.1093/nar/gkt1207.