Bravo *et al. BMC Bioinformatics* (2015) 16:55
DOI 10.1186/s12859-015-0472-9

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research

Àlex Bravo, Janet Piñero, Núria Queralt-Rosinach, Michael Rautschka and Laura I Furlong*

Abstract

Background: Current biomedical research needs to leverage and exploit the large amount of information reported in scientific publications. Automated text mining approaches, in particular those aimed at finding relationships between entities, are key for identification of actionable knowledge from free text repositories. We present the BeFree system aimed at identifying relationships between biomedical entities with a special focus on genes and their associated diseases.

Results: By exploiting morpho-syntactic information of the text, BeFree is able to identify gene-disease, drug-disease and drug-target associations with state-of-the-art performance. The application of BeFree to real-case scenarios shows its effectiveness in extracting information relevant for translational research. We show the value of the gene-disease associations extracted by BeFree through a number of analyses and integration with other data sources. BeFree succeeds in identifying genes associated to a major cause of morbidity worldwide, depression, which are not present in other public resources. Moreover, large-scale extraction and analysis of gene-disease associations, and integration with current biomedical knowledge, provided interesting insights on the kind of information that can be found in the literature, and raised challenges regarding data prioritization and curation. We found that only a small proportion of the gene-disease associations discovered by using BeFree is collected in expert-curated databases. Thus, there is a pressing need to find alternative strategies to manual curation, in order to review, prioritize and curate text-mining data and incorporate it into domain-specific databases. We present our strategy for data prioritization and discuss its implications for supporting biomedical research and applications.

Conclusions: BeFree is a novel text mining system that performs competitively for the identification of gene-disease, drug-disease and drug-target associations. Our analyses show that mining only a small fraction of MEDLINE results in a large dataset of gene-disease associations, and only a small proportion of this dataset is actually recorded in curated resources (2%), raising several issues on data prioritization and curation. We propose that joint analysis of text mined data with data curated by experts appears as a suitable approach to both assess data quality and highlight novel and interesting information.

Keywords: Text mining, Information extraction, Big data, Translational bioinformatics, Biocuration, Disease, Machine learning, Corpus development

* Correspondence: lfurlong@imim.es
Research Programme on Biomedical Informatics (GRIB), IMIM, DCEXS, Universitat Pompeu Fabra, Barcelona, Spain

Background

Due to the increasing size of literature repositories, there is a strong need for tools that firstly, identify and gather the relevant information from the literature, and secondly, place it in the context of current biomedical knowledge. Nowadays, the automatic analysis of the literature by text mining approaches eases the access to information otherwise locked in millions of documents and supports translational research projects [1].

Despite the advances on biomedical text mining, several challenges remain to be solved in the field, such as the identification of complex relationships between entities of biomedical interest and the exploitation of the extracted information in real-case settings for supporting specific research questions in translational research. This is particularly relevant for researchers interested in human diseases, since they are struggling with the large number of publications in their domain. There is a pressing need for methods that: a) can extract information on human diseases and their genes in a precise manner; b) can be applied to large document repositories; c) allow to integrate the extracted data with other information to aid subsequent analysis and knowledge discovery. In particular, text mining tools that help in the identification of the actionable knowledge from the vast amount of data available in document repositories are key for bridging the gap between bench and bedside [2].

In the past, most efforts in text mining of relationships have been devoted to the identification of interactions between proteins, both due to the availability of corpora and the push driven by specific text mining challenges [1]. In contrast, less attention has been paid to the identification of relationships involving entities of biomedical interest such as diseases, drugs, genes and their sequence variants. In the last years, however, this trend has changed and there is much more interest in gathering this kind of information [3,4]. There are examples of systems developed for identification of drug-gene interactions [5,6], drug-drug interactions [7,8], drug-indications [9,10], drug-adverse effect [11,12], gene-disease [13,14], and also systems covering different types of relationships [15].

An important aspect in the field of relation extraction (RE) are the different *perspectives* that can be used to define the relationships between entities. The relationship between two entities might be unqualified or not specified at the semantic level (e.g. “The LOXL1 gene *is associated with* exfoliation glaucoma”), or, on the other hand, semantically specified (e.g. “The LOXL1 gene *is overexpressed in* exfoliation glaucoma”). Moreover, the relationships can also be considered from the perspective of their level of certainty; that is, if the scientific statement is phrased as a fact or proven experimental observation or, alternatively, as a speculation or hypothesis (e.g. “The LOXL1 gene *might be associated with* exfoliation glaucoma”). Research in the area of discourse analysis has been applied to approach this latter perspective of RE [16-18]. Finally, complex representations of relations are the *events* as defined by the BioNLP shared tasks, that involve several participant entities, semantically-defined relationship types and their regulators [19].

A wide range of approaches for RE have been applied in the biomedical field, namely co-occurrence based statistics [20-22], rule-based systems [23,24], machine learning [13,25-27] and NLP-based approaches [28,29]. In particular, supervised learning approaches have shown good performance exploiting both syntactic and semantic information [30]. Most of the studies have focused on kernel based methods to identify associations between entities [31-34]. These methods are able to classify text based on how a relationship between two entities is represented. Different kinds of features, such as word frequencies in the sentences or the relationship between words provided by phrase structure or dependency trees, can be used to represent a relationship between two entities. A common approach involves considering distance criteria like the shortest path between the candidate entities in a parse tree to unravel associations [3,35].

In this paper, we propose the combination of the Shallow Linguistic Kernel (K_{SL}) [33] with a new kernel that exploits deep syntactic information, the Dependency Kernel (K_{DEP}), for the identification of relationships between genes, diseases and drugs. The K_{SL}, which uses only shallow syntactic information, was successfully applied to extract adverse drug reactions from clinical reports [11] and drug-drug interactions [8]. On the other hand, the K_{DEP} exploits the syntactic information of the sentence using the walk-weighted subsequence kernels as proposed by [36]. A major requisite of supervised learning approaches for RE is the availability of annotated corpora for both development and evaluation. Although there are several annotated corpora for identification of PPIs (LLL, AIMed, Bioinfer, HPRD50 and IEPA), manually annotated corpora for other associations are scarce [37]. Our group has developed one of these resources, the EU-ADR corpus, that contains annotations on drugs, diseases, genes and proteins and associations between them [38]. We used this corpus to develop a RE system for the identification of relationships between genes and diseases. In addition, we also evaluated the RE systems for the identification of relationships between diseases, drugs and their targets. As we are particularly interested in identifying associations between genes and diseases, we also developed a new corpus in this domain by a semi-automatic annotation procedure based on the Genetic Association Database (GAD), an archive of human genetic association studies of complex diseases and disorders, as a starting point. The RE module in combination with our previously reported Biomedical Named Entity Recognition or BioNER [39] constitutes the BeFree system (http://ibi.imim.es/befree/).

In addition to the evaluation of the performance of the RE system based on Precision (P), Recall (R) and F-score (F), that is common practice in the text mining domain, we wanted to assess the ability of the BeFree system to identify useful information in the context of concrete biomedical problems. More specifically, we applied the BeFree system for the extraction of associations between genes and diseases to two real-case scenarios: 1) the search for genes associated to one of the most prevalent diseases, depression, and 2) the population of DisGeNET, a database of gene-disease associations [40]. In the first case study we demonstrate the ability of BeFree to identify useful information related to this particular disease. In the second case study we evaluated the application of BeFree to large-scale data extraction and integration with another knowledge source. This resulted in a large dataset on gene-disease associations (approx. 500,000 associations) that raised issues related with the quality of the extracted information. Therefore, we were faced with another challenge, which is the prioritization of the results obtained by large-scale mining of the biomedical literature. Since manual curation is not possible for this kind of large datasets, we performed a series of analysis on the data in order to gain insight on its quality and provide a discussion on the outcome.

In summary, we present a novel text mining system, BeFree, specifically focused on the identification of associations between drugs, diseases and genes. Another important contribution of this work is the GAD corpus for the evaluation of RE systems for gene-disease associations. We focus on the identification of gene-disease relationships, and analyse the outcome of the two case studies that highlight the value of the extracted information, and finally discuss the impact of this kind of approach for translational research. We address some of the current challenges in the field, such as improving RE for entities of biomedical interest, disambiguation between semantically different entities, integration with existing knowledge bases and exploitation of extracted information in real-case scenarios.

The complete set of gene-disease associations extracted by BeFree, with the supporting statements and information on the provenance, is available in DisGeNET (http://www.disgenet.org/). The corpora used in this study, including the new corpus on gene-disease associations, are available at http://ibi.imim.es/befree/#corpora. The BeFree code is available upon request.

Results and discussion

We have developed a new RE system to identify associations between genes, drugs and diseases based on the exploitation of semantic and morpho-syntactic information from the text. We first present the results of its evaluation aimed at assessing the performance of both kernels (K_{SL} and K_{DEP}, see Figures 1 and 2) using morpho-syntactic features on three relationships, drug-target, gene-disease and drug-disease, using the EU-ADR and GAD corpora (see Additional file 1: Table S1 for corpora statistics). The complete set of results is available online at http://ibi.imim.es/befree/#supplbefree. Only a representative set of the results is depicted in the manuscript. We also conducted a series of experiments on the identification of protein interactions in order to evaluate the performance of the K_{DEP} kernel using different features, and compare it with previous results (Additional file 2).

We then focus on the identification of gene-disease associations. We present the results of the real-life performance of the system and discuss its application for the identification of associations between genes and diseases in two different scenarios: a) the research on the genes involved in depression, one of the major health problems in the world, and b) the population of DisGeNET, a public database of gene-disease associations.

Identification of drug-target, gene-disease and drug-disease relationships

We assessed the performance of the K_{SL} and K_{DEP} kernels on the relationships available in the EU-ADR corpus [38]. We used different combination of features to represent the associations, but only a selection of the better results is shown in the BeFree webpage (http://ibi.imim.es/befree/#supplmaterial) and some of them transcribed here (Additional file 3: Table S2). In the case of the drug-disease associations, the best performance both in terms of F-score and Recall is obtained with the K_{DEP} kernel (Experiment 3: P 70.2%, R 93.2%, F 79.3%), using stems on the v-walk feature, while in terms of Precision the best result is obtained using POS tags on both the e-walk and v-walk features (Experiment 19: P 74.5%, R 71.5%, F 72.3%). The best results obtained by combining both kernels did not improve the performance of the dependency kernel alone (Experiment 75: P 72.0%%, R 84.0%, F 77.0%). Similar results were obtained for the gene-disease association classification, where the K_{DEP} kernel alone achieved the best performance. The best performance in terms of F-score and Recall was obtained using stem or lemma over the v-walk features (Experiments 5 and 3: P 75.1%, R 97.7%, F 84.6%), while the best performance in terms of Precision was obtained when using lemma in v-walk and role in e-walk (Experiment 30: P 83.8%, R 71.0%, F 75.6%). Finally, for target-drug relationship, the highest Precision was obtained by the K_{DEP} kernel with role and POS features over the v-walk and e-walk, respectively (Experiment 21: P 75.2%, R 68.1%, F 70.2%), while the highest Recall was obtained when using a combination of K_{SL} and K_{DEP} (Experiment 80: P 73%, R 98%, F 82.8%). The best classification in terms of F-score is achieved when using different

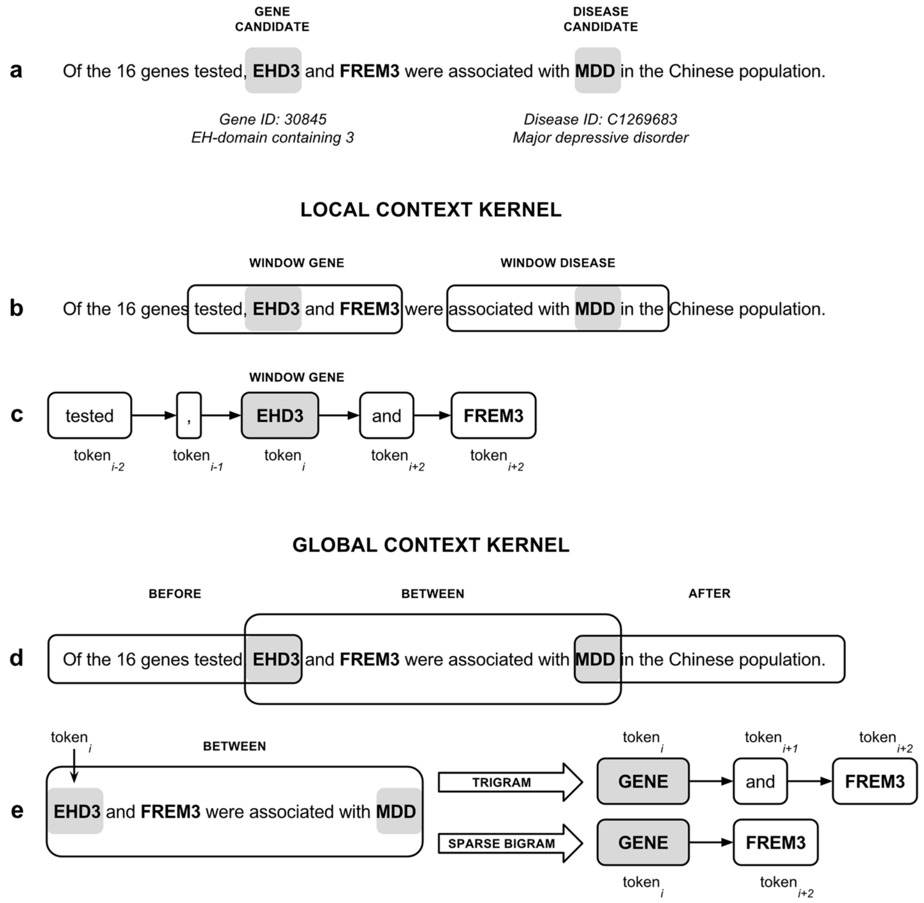

Figure 1 Global and local context kernels to represent a gene-disease association. a) The sentence extracted form a MEDLINE abstract (PMID:22337703) expresses the association between the disease MMD (Major Depressive Disorder) and the genes EHD3 and FREM3. We will focus in the association between EHD3 and MMD to illustrate the features considered in each kernel. **b** and **c)** The local context kernel (K_{LC}) uses orthographic and shallow linguistic features (POS, lemma, stem) of the tokens located at the left and right (window size of 2) of the candidate entities (EHD3 and MDD). **d)** The global context kernel (K_{GC}) is based on the assumption that an association between two entities (in this case EHD3 and MDD) is more likely to be expressed within on of three patterns (fore-between, between, between-after). In this example the association between EHD3 and MDD is expressed in the between pattern. **e)** In the global context kernel (K_{GC}) we consider both trigrams and sparse bigrams in each pattern.

combination of features with both kernels (see for instance Experiment 102: P 74.2%, R 97.4%, F 83.3%). Nevertheless, it is worth mentioning that the K_{SL} kernel, which only uses shallow linguistic information, achieves competitive results in the classification of sentences containing drug-disease, gene-disease and drug-gene associations (F-score: 76.7%, 80.9%, 81.1% respectively).

In order to evaluate the results of BeFree in the context of other approaches, we evaluated the performance of SemRep [28,41] for the identification of the three relationship types using the EU-ADR corpus. SemRep is quite different than BeFree because it has been designed to identify a large variety of semantic predications taking into account the hierarchical relationships between concepts. Nevertheless, we decided to use SemRep for comparison because: 1) it is publicly available, and 2) among all the relationship types covered, some of them can be mapped to the three relationship types covered in this study (gene-disease, drug-target and drug-disease, see Methods). SemRep identified these relationship types with high precision but lower Recall than BeFree, achieveing 96% P, 36% R and 52% F1 for gene-disease associations, 95% P, 39% R and 55% F1 for drug-target and finally, 100% Precision, 40% R and 57% F1 for drug-disease (Additional file 3: Table S2). Thus, compared to SemRep, BeFree achieves more balanced results in terms of P and R for the identification of the three entity types.

We can also analyze the results obtained by BeFree in the context of recent work in the field (Table 1). Note that the studies cited in the Table 1 define in different ways the relationships, use different benchmarks for evaluation and sometimes different metrics. Therefore, the results of this comparison has to be taken with caution. For gene-disease associations, F-scores of 78% [13]

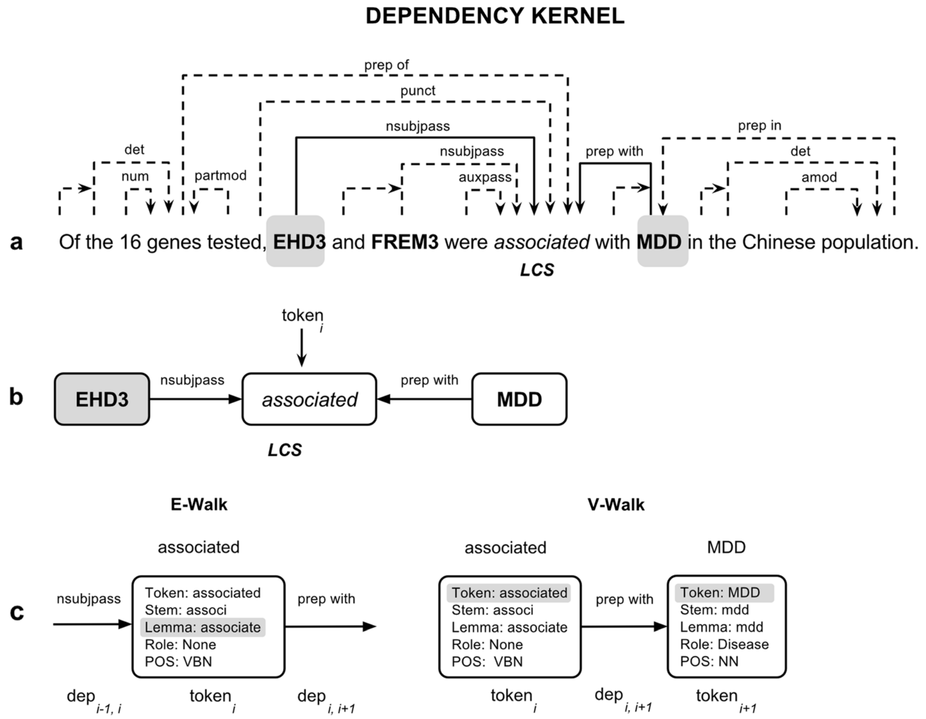

Figure 2 Dependency graph representation of a gene-disease association. a) Dependency graph representation of the sentence. Solid lines represent the shortest path between the two candidates. The token "associated" is the Least Common Subsumer (LCS) node of both candidates. **b)** Subgraph representing the shortest path between EHD3 and MDD, where syntactic dependencies are represented as edges and tokens as nodes. **c)** The e-walk and v-walk features for the node "association" and the syntactic (token, stem, lemma, POS) and semantic features (role) considered in the K_{DEP} kernel.

and 76% [42] have been reported. For drug-disease, F-scores of 87% [11], 79% [42], 69% [10] and 50.5% [12] were reported, while for drug-target the values are around 80% [42]. In summary, BeFree achieves results that are comparable to previous work on the field.

The EU-ADR corpus is a valuable resource because it contains annotations for different types of associations, but its main drawback is its small size. There are only a limited number of corpora for entities of biomedical interest (see http://corpora.informatik.hu-berlin.de/ for a recent update). In order to test the feasibility of using a semi-automatic annotated corpus for biomedical RE, we developed a corpus from the GAD database to have a large benchmark of gene-disease associations in which to train and evaluate gene-disease classifiers. Then, we tested the classifier for gene-disease relationships on the

Table 1 Comparison of BeFree performance to previous work

Method	Drug-disease	Gene-disease	Drug-target
Chun et al. 2006 [25]	-	83%	-
Bundschus et al. 2008 [13]	-	78%	-
Gurulingappa et al. 2012 [11]	87%	-	-
Kang et al. 2014 [12]	54%	-	-
Névéol and Lu. 2010 [10]	-	-	69%
Xu and Wang 2012 [6]	-	-	40%
Xu and Wang 2013 [43]	23%	-	-
Hakenberg et al. 2012 [15]	76%	84%	83%
Buyko et al. 2012 [42]	79%	76%	82%
BeFree	79%	85%	83%

Performance of different approaches including BeFree in terms of F for the three association types is presented. Note that the results corresponding to the work of others are quoted verbatim from the literature, and are therefore not directly comparable.

GAD corpus by 10-fold cross-validation. Compared to the gene-disease set from the EU-ADR corpus, the GAD corpus is larger and contains a different ratio of true/false associations. Thus, it is interesting to see how the different combination of kernels and features behave in this benchmark. In addition, it contains a larger fraction of negative sentences, allowing the classification of positive (PA) and negative (NA) sentences pertaining gene-disease associations. Although these annotations are available in the EU-ADR corpus, due to its small size, it was not possible to train a classifier to distinguish between positive and negative sentences. When assessing the classification over the class TRUE, the best results where those obtained with the K_{SL} (1: P 77.8%, R 87.2% F 82.2%). Contrasting with the results obtained on the EU-ADR corpus for gene-disease associations, K_{DEP} alone did not work very well on the GAD corpus, and the combination of both kernels showed an improvement of the performance but was always lower than the ones obtained with K_{SL} alone (see http://ibi.imim.es/befree/#supplbefree). In the scenario of the classification over three classes (PA vs NA vs FALSE), although the best performance in terms of Precision or Recall is obtained with combination of kernels, the best F-score is achieved by the K_{SL} kernel with sparse bigrams, where the Precision and Recall values although not optimal are competitive (2: P 66.0%, R 73.8% F 69.6%). In summary, these results show that a corpus developed by automatic annotation from an expert-curated database on gene-disease associations can produce competitive classifiers, and that the K_{SL} kernel with shallow linguistic information performs quite well in the classification.

Evaluation on real-life case studies

Case study on genetic basis of depression

Depression is a chronic, recurring, life-threatening disease and the second cause of morbidity worldwide, costing billions of dollars per year to the society [44]. It is currently accepted that a variety of genetic, environmentally-driven epigenetic changes and neurobiological factors play a role in the development of depression; however the exact mechanisms that lead to the disease and affect therapy efficacy are still poorly understood. MEDLINE currently indexes more than 100,000 publications on depression, thus it is a good resource to gather information on genetic determinants of this illness. We performed an evaluation on a real-life setting to test the performance of BeFree in identifying genes associated to depression. We evaluated the results in terms of Precision, Recall and F-score of the predictions. Next, we evaluated the quality of extracted information comparing it with what is available in curated resources. We defined a document set of 270 abstracts pertaining to depression that were published during 2012. This document set was processed to identify genes and depression terms with BioNER (see Methods) and the associations between them using gene-disease models trained in EU-ADR and GAD corpora. In both cases, we used the model that in cross-validation achieved the best F-score (for EU-ADR, experiment 3; for GAD, experiment 1 and 2, see http://ibi.imim.es/befree/#supplbefree, Additional file 3: Table S2). From a total number of 830 gene-disease associations predicted by the models, we manually reviewed a subset of 100 selected at random to estimate the performance of each model. In the case of the model trained on the EU-ADR, although the Recall was almost perfect (96.6%), we observe a decrease in the Precision of the classification compared to the cross-validation scenario (59.4%). On the other hand, the model trained on the GAD corpus performed better in terms of Precision (70%) but worst in terms of Recall (59.3%) when compared to the cross-validation scenario. A model trained in the GAD corpus is also able to classify sentences containing gene-disease associations as positive, negative and false with F-score of 53.7% (data not shown). All in all, the model trained in the EU-ADR corpus, despite its small size, performed a better classification of gene-disease sentences in a real case setting (F-score 73.5%).

We then carried out a qualitative analysis of the information extracted by BeFree. We compared the genes identified as related to depression using BeFree with the genes already known to be associated to depression available from DisGeNET. The BeFree model trained on the EU-ADR corpus identified 170 genes from the full set of publications, 41 of them available in DisGeNET, whereas the model trained on the GAD corpus retrieved 106 genes, 37 of them were already reported in DisGeNET (Figure 3). More interestingly, the EU-ADR and the GAD models found 129 and 69 genes respectively, not present in DisGeNET, which might represent novel findings that could be introduced in the database. We analysed more deeply the set of genes that were predicted by both methods and were not present in DisGeNET (59 genes). For this purpose we used PsyGeNET, an expert-curated database on psychiatric diseases and their genes. In particular, we assessed if the set of 59 genes identified by both methods were present in PsyGeNET. Thirty seven% of the genes (22 genes) were present in this database, indicating that these genes are known players in depression. Next, we characterized the set of 59 genes by functional enrichment analysis with GO terms using DAVID [45], in order to gain insight into their biological function. We found significant annotations for terms like synaptic transmission, transmission of nerve impulse, biogenic amine catabolic process, regulation of neurological system process, regulation of cell cycle, regulation of inflammatory response, which are also found for the list of genes from DisGeNET, and are representative of the biology of depression. More

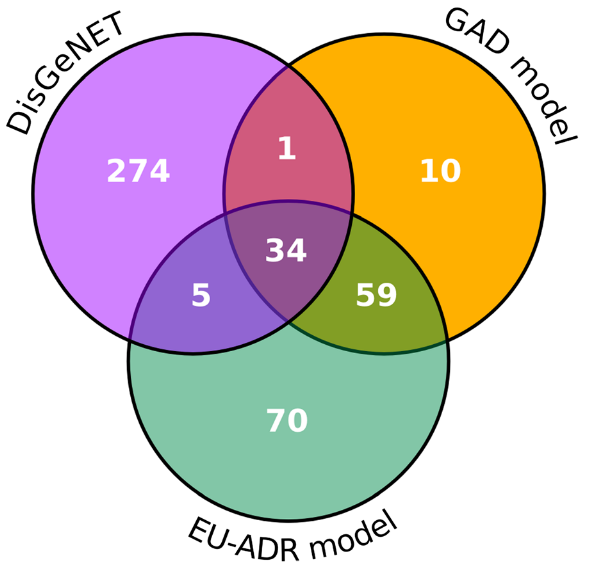

Figure 3 Depression genes identified by BeFree and their overlap with genes available in other repositories. Venn diagram showing the overlap for the depression genes identified by BeFree trained in GAD or EU-ADR corpora, and the depression genes present in DisGeNET.

interestingly, some of the genes identified by text mining are putatively involved in RNA regulation, RNA splicing and epigenetic regulation, such as *MEG3* (GeneId: 55384), *BDNF antisense RNA* (GeneId: 497258), *DGCR8* (GeneId: 54487), *EXOSC6* (GeneId: 118460), and *GEMIN4* (GeneId: 50628). This is noteworthy since there is an increasing interest in the relationship between the aforementioned processes and the physiopathology of depression. In summary, the application of the BeFree system achieves competitive performance in a real-case scenario and allows the identification of genes related to depression, not previously associated to the disease in specialized databases. More importantly, some of these genes represent novel aspects of the physiopathology of depression.

Large-scale analysis of gene-disease associations from the literature

We applied the BeFree system on a set of 737,712 abstracts pertaining to human diseases and their genes (see Methods for details on document selection) to identify relationships between genes and diseases. Note that our approach for NER takes into account the existing ambiguity in the nomenclature between entities of different semantic types, such as genes and diseases (see Methods for more details). This resulted in 530,347 gene-disease associations between 14,777 genes and 12,650 diseases, which were reported in 355,976 publications. DisGeNET, a database that integrates associations between genes and diseases from several sources, includes 372,465 gene-disease associations at the time of this analysis. Thus, the data identified by BeFree represent a very large dataset on gene-disease associations. Some concerns on the quality of the extracted information could be raised such as 1) errors in the text mining approach, both at the level of NER and RE, and 2) quality of the experimental evidence supporting the association. A simple way to identify both types of error would be to manually curate all the associations, but this is not a feasible task.

Thus, before delivering the data to the public through the DisGeNET knowledge portal, we conducted a series of analysis to learn more about the data and its provenance.

Data analysis and filtering

We first analysed the frequency distribution of the number of publications or PubMed IDs (PMIDs) that support each disease association (Figure 4). As can be observed from the figure, 68.5% of the associations (363,382 associations) are supported by only one publication and 72,693 by two publications. On the other extreme of the distribution, there are approximately 900 associations supported by more than 200 publications (0.16%). On average, each association is supported by 2.8 publications. We then inspected in more detail the associations supported by only one publication. These might be associations that could not be reproduced again by any other research group. Another reason for the low publication number could be that the related research area is not a hot topic and therefore it is more difficult to publish in this specific domain. Moreover, these associations could be new findings from recent publications, which might be in the future reproduced or followed up in other publications.

We analysed in more detail the set of associations supported by only one publication, looking at their publication

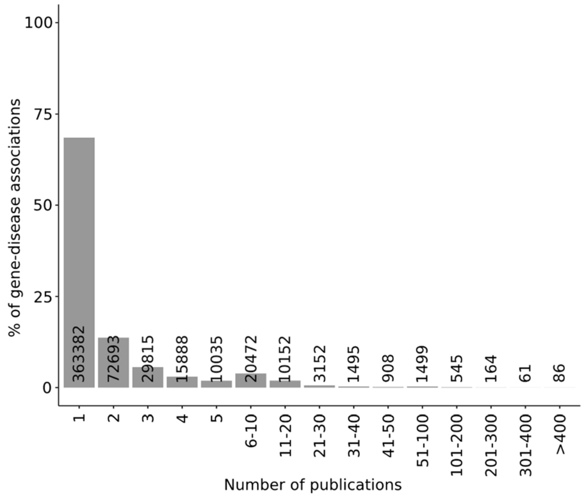

Figure 4 Number of gene-disease associations as a function of the number of PMIDs that support each association.

dates (Figure 5) and the Impact Factors (IF) of the journals (Figure 6), evaluating also their coverage in the DisGeNET database. Figure 5 shows that most of the associations supported by only one publication have been published in the last 15 years. Notably, 40,000 associations (11% of all the associations supported by one PMID) have been published during 2013. These associations represent newly open research areas on the genetic basis of diseases that might point out potential candidates for biomarkers or therapeutic targets. Interestingly, almost none (97%) of these associations are present in the DisGeNET database. Moreover, 35% of the associations have been published in journals without IF and 16% in journals with IF between 0 and 2.5 (Figure 6). Remarkably, a very small fraction of the associations with one supporting publication have been published in journals with the higher IF, while the majority of the associations are reported in journals with IF between 2.5 and 5.

We then inspected the distribution of the number of gene-disease associations reported per MEDLINE abstract (Figure 7). As can be observed, most of the abstracts (47%) report 1–2 gene-disease associations and on average each abstract reports 1.5 gene-disease associations. However, there is a subset of 15 abstracts that report more than 100 gene-disease associations, with one extreme case of 372 gene-disease associations. Manual inspection of the 15 abstracts that report more than 100 associations indicated that in most of the cases, these abstracts report associations for a number of genes to a disease, using long sentences with coordination structures. In order to avoid possible sources of errors during text mining processing of these long, complex sentences, we decided to remove those abstracts that report more than 20 associations.

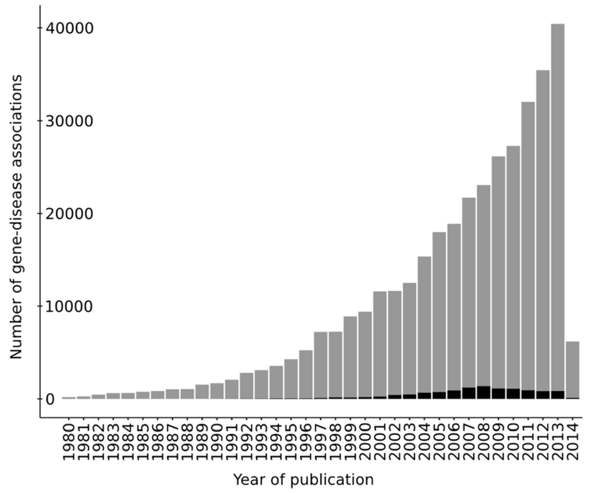

Figure 5 Number of gene-disease associations reported by only one PMID in each calendar year. In red we show the number of associations present in DisGeNET.

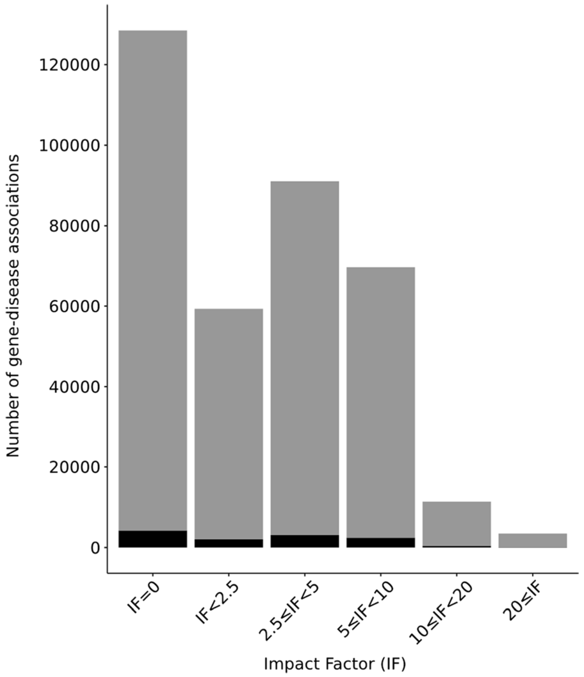

Figure 6 Number of gene-disease associations reported by only one PMID in journals classified by their Impact Factor. In red we show the number of associations present in DisGeNET.

Based on this preliminary data analysis, we developed a decision tree workflow on the BeFree data that takes into account the number of publications supporting the gene-disease association, the overlap with DisGeNET and the IF of the journals (Figure 8). After applying this workflow, we obtained 330,888 gene-disease associations (62% of the original data set) between 13,402 genes and 10,557 diseases.

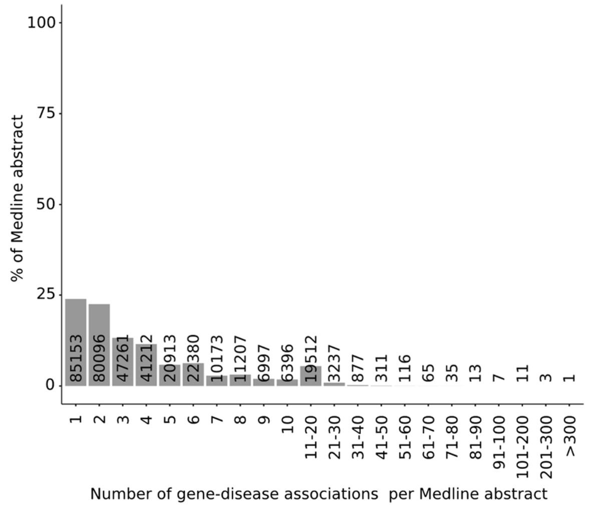

Figure 7 Distribution of the number of gene-disease associations reported per MEDLINE abstract.

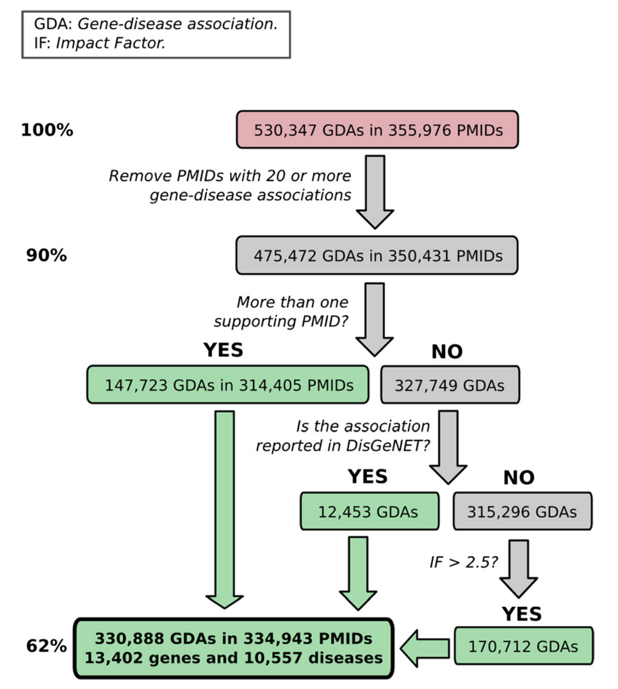
GDA: Gene-disease association.
IF: Impact Factor.
100%
530,347 GDAs in 355,976 PMIDs
Remove PMIDs with 20 or more gene-disease associations
90%
475,472 GDAs in 350,431 PMIDs
More than one supporting PMID?
YES
NO
147,723 GDAs in 314,405 PMIDs
327,749 GDAs
Is the association reported in DisGeNET?
YES
NO
12,453 GDAs
315,296 GDAs
IF > 2.5?
YES
62%
330,888 GDAs in 334,943 PMIDs
13,402 genes and 10,557 diseases
170,712 GDAs

Figure 8 Decision Tree Workflow for selection of BeFree dataset on gene-disease associations.

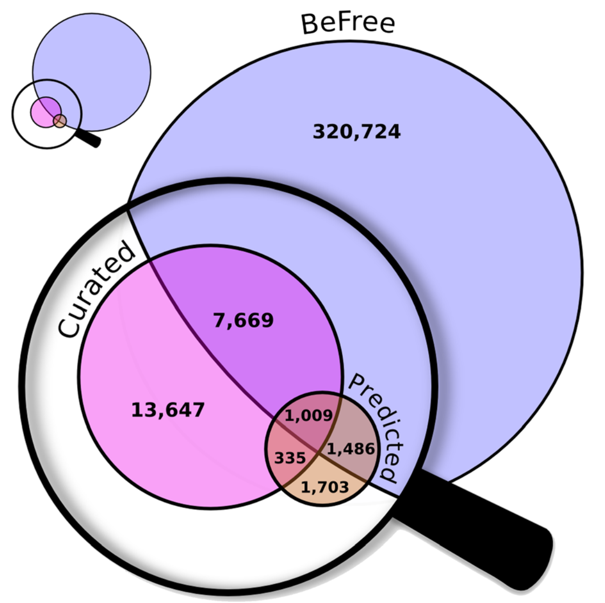

Figure 9 Overlap of the gene-disease associations identified by BeFree with the associations available in DisGeNET curated and predicted sources. DisGeNET information coming from expert curated sources such as UniProt are classified as curated, whereas information coming from model animals such as mouse are classified as predicted. For more information see http://www.disgenet.org/.

Integration with DisGeNET and data prioritization

A pragmatic way to assess the quality of the extracted information is to contrast it to the information present in expert curated resources. Thus, we integrated the data extracted by Befree with expert reviewed DisGeNET sources (curated and predicted, see http://www.disgenet.org/web/DisGeNET/v2.1/dbinfo#sources for more details on DisGeNET datasets) in order to perform this comparison. Only 7,669 gene-disease associations (2% of BeFree associations) are in common between expert curated associations from DisGeNET and BeFree, while the overlap between the two sources is quite small (0.3% of BeFree associations, Figure 9). Remarkably, from all the gene-disease associations (curated, predicted, and BeFree) 3.9% are only reported by curated sources, and 92.5% only provided by BeFree. The gene-disease associations present in DisGeNET but not recovered by BeFree might be examples of associations mentioned in the full-text and supplementary material of articles and not present in the abstract, or derived from publications not retrieved by our PubMed query used for document selection. Alternatively, they might be false negatives from our text mining approach. The high percentage of associations recovered by text mining and not present in the curated resources highlight the difficulty in collating all this putative useful information in curated databases. It is important to note that in our approach we do not mine the full MEDLINE repository but only a small, but significant in terms of content, fraction (approx. 3% of current MEDLINE database).

We computed a score for the gene-disease association based on the number of data sources that report the association, the level of curation of each source and number of supporting publications (see Methods) in order to analyse in an integrative manner the data extracted by text mining. Figure 10 shows the DisGeNET score for the BeFree associations versus the number of supporting publications for each association. Most of the associations (99%) have less than 200 publications and have a wide range of scores, reflecting the fact that they are reported in one or several sources with different levels of curation. Moreover, the analysis of this plot let us identify some interesting outliers. First, the associations *APP*-Alzheimer disease and *CFTR*-Cystic Fibrosis receive a very high score because they are also reported in all the DisGeNET sources, and represent examples of very well studied gene-disease associations. Notably, there are 22 associations with very low score (0.06, meaning that they are only reported by BeFree) but are reported by more than 1000 publications. It is intriguing why these associations, that seem to be very well studied as reported in thousands of papers, are not present in any other DisGeNET source. A closer look to some of them (Table 2) indicate that they represent very well studied gene disease associations between breast and ovarian

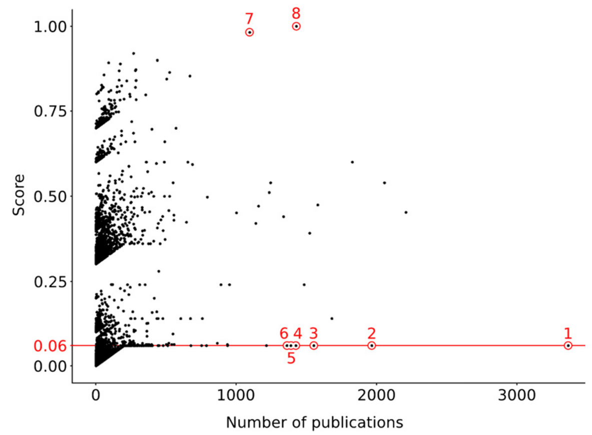

Figure 10 DisGeNET score vs number of supporting publications for the gene-disease associations identified by BeFree. The selected examples discussed in the text are: 1) TP53-Malignant Neoplasm; 2) BRCA1-Breast Carcinoma; 3) ESR1-Breast Carcinoma; 4) ERBB2-Breast Carcinoma; 5) BRCA1-Ovarian Carcinoma; 7) APP-Alzheimer disease; 8) CFTR-Cystic Fibrosis.

cancer with *TP53*, *BRCA1*, *BRCA2*, *ESR1*, *ERBB2*, and also associations of specific genes to generic cancer terms (neoplasms). For all these cases we find the corresponding gene-disease association in DisGeNET but with a different, yet closely related, UMLS concept. The diversity, both at expressivity and granularity levels of disease terminologies used in biomedical sources, is highlighted by the large number of CUIs normalized (aprox. 10,000) in associations unlocked by text mining. This opens the question about disease terminologies standardization, specially to ensure interoperability in translational research. We observed differences in the disease terminology used by database curators and the literature. In general, there is a preference for using disease concepts that contain MeSH terms by database curators (at least for the databases included in DisGeNET). It is interesting to note that most of the disease concepts present in the curated sources in DisGeNET contain MeSH terms, while this is not the case for the data extracted by BeFree (see Figure 11).

Characterization of BeFree data

We analysed the frequency distribution of the number of associated diseases per gene (Figure 12) and the number of associated genes per disease (Figure 13). The plots show that there are very “promiscuous” genes regarding their association to diseases (e.g. *VEGFA*, *IL6*, *TNF* and *TP53*), whereas other genes seem to be more specific as they are reported as associated with one or two diseases only. The same can be observed when analysing diseases and their associated genes (Figure 13). In this case, it is expected that neoplastic diseases occupy the extremes of the distribution, both for their genetic heterogeneity or for being very well studied. We inspected the therapeutic areas (Figure 11) and the protein classes covered by the gene-disease associations identified by BeFree (Figure 14). The coverage of diseases by therapeutic areas according to the MeSH classification from BeFree paralleled the one obtained for the diseases in DisGeNET. It is important to note that a large fraction (more than 40%) of diseases identified by BeFree cannot be assigned to a MeSH disease class, while this is not the case for DisGeNET diseases. Thus, the five most covered MeSH disease classes in BeFree, following the not classified diseases, are “Congenital, Hereditary, and Neonatal Diseases Abnormalities”, “Nervous System Diseases”, “Pathological Conditions, Signs and Symptoms”, “Nutritional and Metabolic Diseases and Neoplasms”. The disease genes identified by BeFree are classified as protein-coding (89%), ncRNA (2%), pseudogenes (2%), rRNA, tRNA, snoRNA and snRNA (0.5%), other (6%). Again, regarding the classification of the

Table 2 Examples of gene-disease associations from BeFree with low score and supported by a large number of publications

Gene	Disease	Source	Score	Number of PMIDs (Number of PMIDS provided by BeFree)
TP53	Malignant Neoplasms (C0006826)	BeFree	0.06	3365
TP53	Neoplasms (C0027651)	DisGeNET	0.401	475 (401)
BRCA1	Breast Carcinoma (C0678222)	BeFree	0.06	1966
BRCA1	Malignant neoplasm of breast (C0006142)	DisGeNET	0.702	2123 (2056)
ESR1	Breast Carcinoma (C0678222)	BeFree	0.06	1553
ESR1	Malignant neoplasm of breast (C0006142)	DisGeNET	0.3	1690 (1681)
ERBB2	Breast Carcinoma (C0678222)	BeFree	0.06	1425
ERBB2	Malignant neoplasm of breast (C0006142)	DisGeNET	0.4	1493 (1484)
BRCA1	Ovarian carcinoma (C0029925)	BeFree	0.06	1389
BRCA1	Malignant Neoplasm of Ovary (C1140680)	DisGeNET	0.6	1365 (1337)
BRCA1	Ovarian Neoplasm (C0919267)	DisGeNET	0.6	144 (71)

Similar associations provided by DisGeNET are indicated.

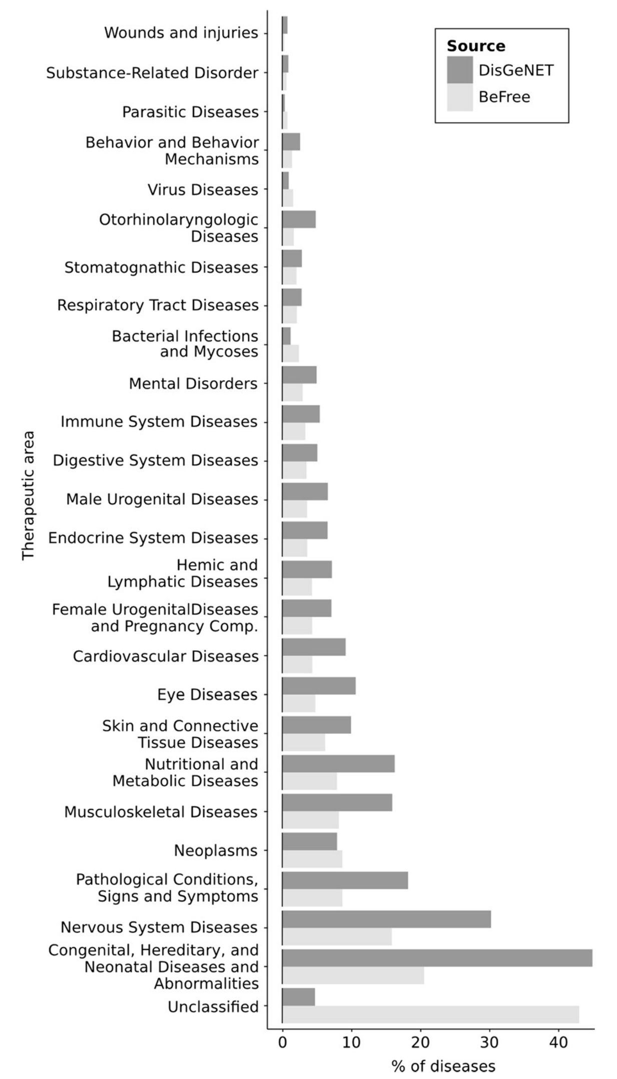

Figure 11 Distribution of diseases according to the MeSH disease classification in the BeFree and DisGeNET datasets. Note that more than 40% of diseases in BeFree do not contain a MeSH disease class.

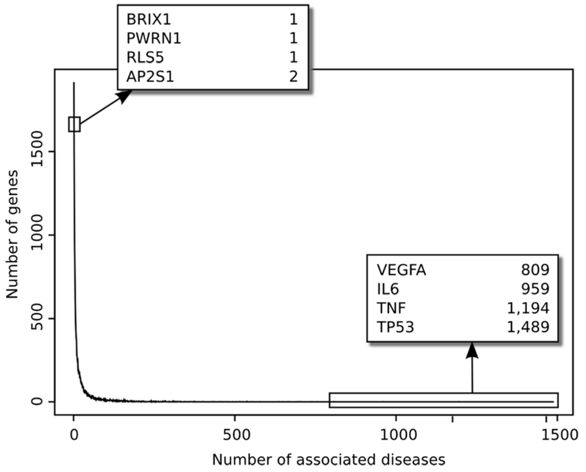

Figure 12 Frequency distribution of the number of associated diseases per gene.

proteins encoded by disease genes according to the Panther protein classification, we observe again that disease proteins identified by BeFree have a similar class distribution than those present in DisGeNET (Figure 14).

Conclusions

Our results show that a kernel based approach using both morpho-syntactic and dependency information performs competitively for the identification of drug-disease, drug-target and gene-disease relationships from free text. Although the exact combination of features that yield better results depends both on the association type and the corpus used for training the system, the use of shallow linguistic information is enough to produce accurate RE classifiers to recognize these associations. A supervised learning RE system for gene-disease associations trained on different corpora (EU-ADR, GAD) with very different characteristics is able to identify gene-disease associations in real-case scenarios with good performance. As previously suggested by others [5], a corpus developed by semi-automatic annotation is a good resource for developing a RE system in biomedicine.

We evaluated the value of the information extracted by BeFree for specific case studies in translational research. Particularly, the results obtained in the case study on depression indicated that BeFree is able to identify genes associated to depression that are not present in public databases and support novel hypothesis

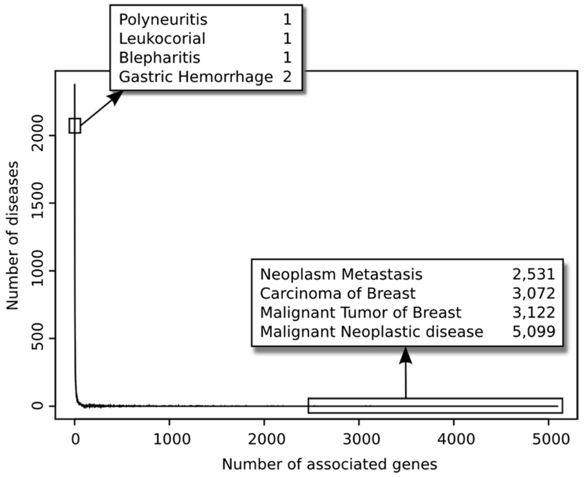

Figure 13 Frequency distribution of the number of associated genes per disease.

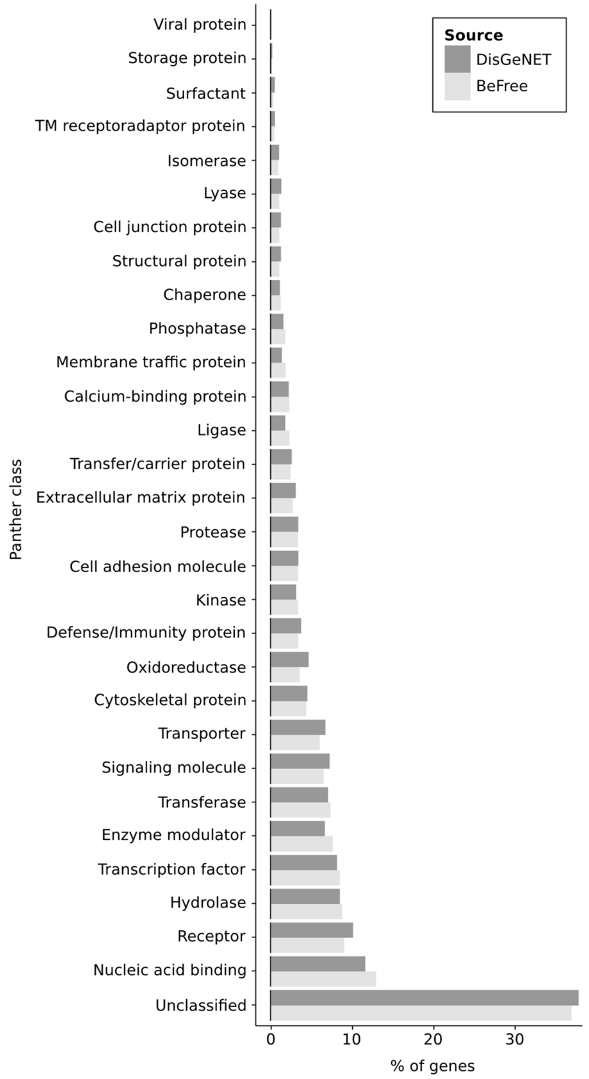

Figure 14 Distribution of disease proteins according to the Panther Protein classification. Data from Panther (http://www.pantherdb.org/) was used to annotate disease proteins from BeFree and DisGeNET. Note that more than 37% of proteins in BeFree cannot be classified according to Panther.

in the pathophysiology of depression. The large-scale analysis of gene-disease associations provided interesting insights on the kind of information that can be found in the literature about gene-disease associations, and raised some issues regarding data prioritization and curation. The conclusions of the analysis of the provenance of the gene-disease associations identified by BeFree can be summarized in the following points:

- The scientific literature is a rich resource for extracting gene-disease associations, even considering only abstracts from a specific subset of MEDLINE.
- Only a small proportion of the gene-disease associations discovered by text mining are collected in expert curated databases. There is a pressing need to find alternative strategies to manual curation to review, prioritize and curate these associations and incorporate them into domain-specific databases.
- A first and important step is to extract this information and put it in a standardized format to allow its integration with other data sources and their subsequent analysis for different purposes.
- Joint analysis of data derived by text mining with data curated by experts appears as a suitable approach to assess data quality and identify novel and interesting information.
- A large proportion of the associations are supported by only one publication, raising concerns on data reproducibility but also pointing out novel putative targets for research and innovation.
- There are important differences in the use of disease terminologies between database curators and the authors of publications, and also in the level of granularity of disease concepts to describe a disease phenotype. This is a current challenge for large-scale disease data integration that aims to gather a comprehensive coverage of disease and ensure systematic interoperability across biomedical domains.

Biocuration of large data sets, a.k.a. big data, is becoming a bottleneck for biomedical research. Recently, the crowdsourcing approach has attracted interest in the bioinformatic domain and holds promise for biocuration tasks [46]. As more and more scientific groups are extracting knowledge blocked in free text by text mining and exposing it to the public domain, another upcoming question is the meta-curation of such deluge of data. In this regard, the nanopublication concept, based on Semantic Web triple-assertions, is a promising approach to aid the prioritization of associations based on the supporting evidence [47]. This kind of approaches could be applied to large datasets such as the gene-disease associations extracted by BeFree.

In summary, the study presented here highlights the importance of performing several steps of data analysis on large data sets, before using the data for further bioinformatic analysis and even to feed it to the curation pipeline of a database. We suggest that this kind of iterative process of data extraction, analysis and refinement of data extraction methodology should be applied to other approaches aimed at extracting large-scale information from the literature.

One of the main limitations of this work is that, due to the type of annotations currently available in the corpora for gene-disease, drug-disease and drug-target associations, the relationships identified are not semantically typed. For instance, for a drug-target relationship, it is not possible to know if the drug is an agonist, or if it inactivates the target. Availability of the corpora for these relationships annotated at the semantic level will allow the development of a system able to type these relationships. Another limitation is the focus on relationships that are stated at the sentence level, not handling anaphora to detect associations that go beyond the boundaries of a sentence. Future work in both these directions will provide a system with a higher Recall that provides better definition of relationships from the semantic point of view. Another aspect that we will like to explore is the identification of contextual information of the relationship. For example, in the case of gene-disease associations, we would like to know the experimental method used to detect this association (GWAS, exome sequencing, transcriptomic analysis, etc.), information on cell-type or tissue (e.g. adrenal cells), population tested (e.g. caucasic males), etc.

Methods

We present the development of the BeFree system, composed of a biomedical named entity recognition system (BioNER, presented in [39]) and a kernel-based KE.

Kernel based RE

In order to implement a RE for different relationships (drug-target, drug-disease, gene-disease), we propose the combination of the Shallow Linguistic Kernel (K_{SL}) based on the system originally proposed by [33] and our Dependency Kernel (K_{DEP}). Both kernels are described in the next two sections.

Shallow Linguistic Kernel (K_{SL})

The Shallow Linguistic Kernel (K_{SL}), developed by [33] has been successfully applied to PPI, drug-side effects [11] and drug-drug interaction extraction [8]. Here, we propose its application for the identification of the relationships gene-disease, drug-disease and drug-target. K_{SL} is composed of a linear combination of the kernels K_{GC} and K_{LC} that provide different representations of the association between two candidate entities. The global context kernel (K_{GC}) is based on the assumption that an association between two entities is more likely to be expressed within one of three patterns (fore-between, between, between-after, see Figure 1d). Three term frequency vectors are obtained based on the bag-of-words approach using trigrams of tokens. Sparse bigrams were included to improve the classification performance, as suggested in the original implementation. The local context kernel (K_{LC}) uses orthographic and shallow linguistic features (POS, lemma, stem) of the tokens located at the left and right of the candidate entities (window size of 2). Figure 1 shows the features considered by each kernel using an example sentence.

Dependency Kernel (K_{DEP})

We developed the Dependency Kernel (K_{DEP}) to train a model to recognize relationships between the entities of interest using walk features [35]. The syntactic dependencies of the words within a sentence can be represented as dependency graphs. Figure 2a shows the dependency graph of an example sentence extracted from MEDLINE as obtained by the Stanford parser (http://nlp.stanford.edu/software/lex-parser.shtml). The shortest path between the two candidate entities can be extracted from the dependency graph (highlighted with a solid line in the example, Figure 2b), which includes the Least Common Subsumer (LCS) node (common governor node between the two candidates, subgraph detailed in Figure 2b). In the K_{DEP}, two types of walk features are used, the v-walk feature that is composed of $node_{(i)}$-$edge_{(i,\ i+1)}$-$node_{(i+1)}$, and the e-walk feature that is composed of $edge_{(i-1,\ i)}$-$node_{(i)}$-$edge_{(i,\ i+1)}$ (both illustrated in Figure 2c). For the edges we consider the dependency relation type, while for the nodes we consider different features of the token, such as the token itself, its stem, lemma, role (if this token is candidate or not) and part-of-speech (POS) tag.

Corpora

We used two manually annotated corpora: AIMed for PPIs and EU-ADR for gene-disease, drug-target and disease-drug associations. In addition, we developed a semi-automatically annotated corpus for gene-disease associations based on the GAD database. All datasets were pre-processed with a combination of tools to extract the features required by the RE system. More specifically, sentence boundaries were identified by NLTK (http://www.nltk.org/), tokens and part-of-speech (POS) tags were obtained using UIMA modules (http://www.julielab.de/), lemmas were obtained with Biolemmatizer (http://biolemmatizer.sourceforge.net/) and stems were identified with the Porter's algorithm. Syntactic dependencies were obtained with the Stanford parser (http://nlp.stanford.edu/software/lex-parser.shtml). Both EU-ADR and GAD corpora are publicly available (http://ibi.imim.es/befree/#corpora).

AIMed

The AIMed corpus is widely used for PPI extraction (ftp://ftp.cs.utexas.edu/pub/mooney/bio-data/). The AImed corpus consists of 225 MEDLINE abstracts, of which 200 abstracts describe interactions between human proteins and 25 do not refer to any interaction. There are 5625 annotated sentences, 1008 containing a true PPI (TRUE) and 4617 not containing a true PPI (FALSE).

EU-ADR

The EU-ADR corpus contains annotations of different entities (drugs, diseases, and genes/proteins) and the relationships between them [38]. In particular, it contains annotations of relationships between drug and diseases (drug-disease set), drug and their protein targets (drug-target set) and genes/proteins and their association to diseases (gene-disease set). In addition, each relationship is classified according to its level of certainty as: positive association (PA), negative association (NA), speculative association (SA) and false association (FA). The EU-ADR corpus is composed of 100 MEDLINE abstracts for each relationship set, and its annotation was performed by three experts. In this study we considered the relationships that result from the consensus annotation of two experts. Additional file 1: Table S1 shows the number of relationships for each set.

GAD

The Genetic Association Database (GAD) is an archive of human genetic association studies of complex diseases, including summary data extracted from publications on candidate gene and GWAS studies (http://geneticassociationdb.nih.gov/). We use GAD for the development of a corpus on associations between genes and diseases (downloaded on January 21st, 2013). We considered the annotations of relationships between a gene and a disease in a single sentence as a reference set to build this corpus. GAD contains over 130,000 records with different types of information. We selected the records satisfying the following requirements: (i) the association between gene and disease is annotated as positive or negative, (ii) the association is expressed in one sentence and (iii) the Entrez Gene identifier for the gene is provided. Although GAD provides the sentence in which a gene-disease association is stated, there is no information on the exact location of the gene and disease entities in the text. In order to develop a corpus suitable for training a gene-disease RE system, the exact location of the interacting entities in the text is required. To achieve that, we applied our own NER system (BioNER, see below) to identify the gene and disease entities in the text and normalize them to NCBI Gene and UMLS identifiers, respectively. Then, the sentences in which a given gene was found together with a specific disease, and this gene-disease association was annotated by GAD curators as positive or negative, were labelled as TRUE. In order to create a dataset containing false associations (FALSE) between a gene and a disease, that is, a gene and a disease that co-occur in a sentence but are semantically not associated, we selected the sentences with co-occurrences between a disease and a gene found by the BioNER system that were not annotated by GAD curators as gene-disease associations. Additional file 3: Table S2 shows the number of TRUE and FALSE associations that represent the GAD corpus.

Evaluation of Kernel based RE

The performance of each model for association classification was evaluated by sentence-level 10-fold cross validation in each corpus. The classifiers' performances were assessed using P, R and F-score over the class TRUE. TRUE sentences contain real relationship between the entities analysed, in contrast with FALSE sentences where the two entities co-occur, but there is no semantic relationship between them. In the case of the GAD corpus, we also trained a classifier that distinguishes between positive, negative and false associations, and therefore the performance was assessed over the class positive (PA) and negative (NA) separately. Due to the nature of the annotations available in both corpora used, the focus of this work is on associations that are unqualified (not defined at the semantic level).

Evaluation of SemRep for identification of drug-target, gene-disease and drug-disease relationships

We performed an evaluation of the SemRep system [41] for identification of drug-target, gene-disease and drug-disease relationships using the EU-ADR corpus. Since the scope and types of associations covered by SemRep are quite different than the ones covered by the EU-ADR corpus, we selected a subset of the association types retrieved by SemRep and mapped them to each of the association types in the EU-ADR corpus. For gene-disease associations, we selected the following SemRep association types: AFFECTS, ASSOCIATED_WITH, AUGMENTS, CAUSES, PREDISPOSES, COEXISTS_WITH, NEG_ASSOCIATED_WITH. For drug-target, the SemRep association types selected were: AFFECTS, ASSOCIATED_WITH, AUGMENTS, DISRUPTS, DISRUPTS(SPEC), INHIBITS, INHIBITS(SPEC), INTERACTS_WITH, NEG_INTERACTS_WITH, PRODUCES, STIMULATES. For drug-disease, we selected these SemRep association types: AFFECTS, ASSOCIATED_WITH(INFER), CAUSES, NEG_AFFECTS, NEG_TREATS, PREDISPOSES, PREVENTS, TREATS, TREATS(INFER), TREATS(SPEC), USES.

We used the batch mode of the SemRep program available at http://skr3.nlm.nih.gov/. For concept recognition we used the 2012AA version of the knowledge sources, that is the most recent version available within SemRep.

Identification of entities

We identified gene and disease mentions in free text using the BioNER system [39]. During the initial phase of the project, we evaluated several NER tools publicly available. Our requirements were that the NER tool had to be able to detect and normalize to database identifiers two types of entities: genes/proteins and diseases. After an initial evaluation, the decision was to develop our own system because none of the tools evaluated work properly for us (a brief description of the tools evaluated is presented in the Suppl. File S1). Another reason for developing our own tool was to be able to regularly update the dictionaries used by the NER to keep the data up-to-date, and to perform curation of the dictionaries to reduce errors. We also invested efforts on the problem of the ambiguities between entities (see below), which we think is not currently addressed by any other tool. We believe that this point is important for subsequent steps in the text-mining workflow, such as the identification of relationships between entities.

BioNER uses gene and disease dictionaries with fuzzy and pattern matching methods to find and uniquely identify these entity mentions in the literature. During initial analysis of the RE results we observed that a source of error was the wrong identification of entities due to ambiguities in the terminologies for diseases and genes. This is particularly problematic in the case of acronyms, where the same token can be used to refer to a disease or a gene. Thus, we introduced a series of modifications on BioNER in order to address ambiguities in the identification of a single entity type (e.g. a gene) and between different entity types (genes and diseases).

Frequently, an acronym appears after the long term is defined in the text. In this case, we compare the list of concept identifiers of both mentions (acronym, long form) to determine if the acronym refers to the long form, using an in-house developed tool. For example, in the sentence "Selective gene targeting using the carcinoembryonic antigen (CEA) promoter is useful in gene therapy for gastrointestinal cancer" (from PMID 11053994), BioNER detects the long form expression "carcinoembryonic antigen" as a gene with NCBI Gene Id 1048, and the acronym "CEA" as four different gene entities (with four NCBI Gene Ids 1087, 5670, 1084 and 1048). The concept identifier in common between the two entities (NCBI Gene Id 1048) is kept as the right annotation. If there is more than one concept identifier in common between the two entities, we look at the similarity of the terms of each concept to select the right identifier. The file gene2pubmed source from Entrez Gene was also used to select the correct identifier in these ambiguous cases. Evaluation of BioNER for gene normalization using the BioCreative II Gene Normalization (BC2GN) [48] resulted in very low Precision (P: 48.1% R: 80.1% F: 60.1%), which could be improved considerably when applying the above mentioned strategies to handle the ambiguities between genes (P: 74.0%, R: 76.2%, F: 75.0%).

The other type of ambiguity arises when one candidate entity in a sentence can refer to different semantic types (disease and gene). For example the symbol "APC" can refer to the gene "adenomatous polyposis coli" or to the disease "atrial premature complex". To properly recognize the identity of the mention, we take into account the contextual information of the candidate entity. For instance, to disambiguate a candidate entity to a gene, we look for keywords such as "gene", "protein", "factor", "target", "biomarker", etc., whereas to disambiguate a candidate entity to a disease, we look for keywords like "disease", "disorder", "condition", "syndrome", etc. We also looked at the MeSH Disease annotations of the corresponding abstract to decide if a candidate entity refers to a gene or a disease. We compared the terms of the candidate entity to the terms of the MeSH disease concepts annotated to the abstract using a soft-matching approach, and if a match was found, we annotated the candidate entity as a disease.

In addition, we performed an evaluation of the performance of BioNER in the identification and normalization of disease entities using the Arizona Disease Corpus achieving competitive (P: 72.1% R: 64.4% F: 68.0%) results compared to previous approaches [49,50].

Case study on genetic basis of depression

We defined a PubMed query to retrieve a set of document to depression and published in 2012 as follows:

("Depression" [Mesh] OR "Depressive Disorder" [Mesh]) AND "genetics" [Subheading] AND (hasabstract[text] AND ("2012" [PDAT]) AND English[lang] AND "humans" [MeSH Terms]) NOT ("Case Reports" [PT] OR "Clinical Trial" [PT] OR "Clinical conference" [PT] OR "Clinical Trial, Phase I" [PT] OR "Clinical Trial, Phase II" [PT] OR "Clinical Trial, Phase III" [PT] OR "Clinical Trial, Phase IV" [PT] OR "Controlled Clinical Trial" [PT] OR "Randomized Controlled Trial" [PT] OR "Meta-Analysis" [PT]).

This query resulted in 270 citations (date of search March 19, 2013). The abstracts were processed with BeFree trained on GAD and EU-ADR corpora to find gene-disease associations.

Case study on large-scale analysis of gene-disease associations from the literature

We defined a PubMed query to retrieve documents pertaining to human diseases and their associated genes published from 1980:

("Psychiatry and Psychology Category" [Mesh] AND "genetics" [Subheading]) OR ("Diseases Category" [Mesh] AND "genetics" [Subheading]) AND (hasabstract[text] AND ("1980" [PDAT] : "2014" [PDAT]) AND "humans" [MeSH Terms] AND English[lang]).

This query retrieved 737,712 citations (date of search February 25, 2014), which were processed by Befree trained on the EU-ADR corpus to identify relationships between genes and diseases.

DisGeNET score

The DisGeNET score is described at the DisGeNET web page (http://www.disgenet.org/web/DisGeNET/v2.1/dbinfo#score).

Here we reproduce its formulation in order to help in the interpretation of the results. Briefly, we assign a score to each gene-disease association in DisGeNET [51] according to the source in which this association is reported (CURATED, PREDICTED, LITERATURE), the level of curation of each source, and the number of publications that report each association in the case of LITERATURE sources. DisGeNET is a database on gene-disease associations covering all therapeutic areas, that integrates information from resources curated by human experts (UniProt and CTD), from orthologous genes from mouse and rat (MGD, RGD and CTD), and from literature repositories by text mining (LHGDN and GAD). Thus, the DisGeNET source databases are classified accordingly in CURATED, PREDICTED and LITERATURE reflecting the different sources where each association is reported. The gene-disease associations extracted by BeFree are then classified as LITERATURE once integrated in DisGeNET. For the associations reported in LITERATURE sources, we can rank the associations based on the number of publications that support each association. The DisGeNET score is defined as follows:

$$S = S_{CURATED} + \ S_{PREDICTED} + S_{\mathrm{LITERATURE}}$$

$$\begin{aligned} S = {} & (W_{UniProt} + W_{CTDhuman}) + \ (W_{Rat} + W_{Mouse}) \\ & + \ (W_{GAD} + W_{LHGDN} + \ W_{BeFree}) \end{aligned}$$

Where

$$W_{UniProt} \ = \begin{cases} 0.3 \ \mathit{if\ the\ association\ is\ reported\ by\ UniProt} \\ \qquad 0 \ \mathit{otherwise} \end{cases}$$

$$W_{CTDhuman} = \begin{cases} 0.3 \ \mathit{if\ the\ association\ is\ reported\ by} \\ \quad \mathit{CTDhuman} \\ \qquad 0 \ \mathit{otherwise} \end{cases}$$

$$W_{Rat} = \begin{cases} 0.1 \ \mathit{if\ the\ association\ is\ reported\ by} \\ \quad \mathit{CTDRat\ or\ RGD} \\ \qquad 0 \ \mathit{otherwise} \end{cases}$$

$$W_{Mouse} = \begin{cases} 0.1 \ \mathit{if\ the\ association\ is\ reported\ by} \\ \quad \mathit{CTDMouse\ or\ MGD} \\ \qquad 0 \ \mathit{otherwise} \end{cases}$$

$$W_{LITERATURE} \ = \begin{cases} \mathit{maximum\ if} \ \dfrac{n_{gd} \times 100}{N_{LITERATURE}} \geq \mathit{maximum} \\ \dfrac{n_{gd} \times 100}{N_{LITERATURE}} \ \mathit{if} \ \dfrac{n_{gd} \times 100}{N_{LITERATURE}} < \mathit{maximum} \end{cases}$$

$$\mathrm{maximum} \ = \begin{cases} 0.08 \ \mathit{if\ source} = \mathit{GAD} \\ 0.06 \ \mathit{if\ source} = \mathit{LHGDN\ or\ BeFree} \end{cases}$$

$W_{LITERATURE}$ *is the weight of source GAD, LHGDN and BeFree*

$N_{LITERATURE}$ *is the number of publications in source*

n_{dgj} *is the number of publications reporting a gene–disease association in source j*

For more details on the DisGeNET score visit the DisGeNET web page (http: //www.disgenet.org/).

Availability

The complete set of gene-disease associations extracted by BeFree, with the supporting statements and information on the provenance, are available in DisGeNET (http://www.disgenet.org). The corpora used in this study are available at http://ibi.imim.es/befree/#corpora. DisGeNET data is distributed under the Open Database License (http://opendatacommons.org/licenses/odbl/).

Additional files

Additional file 1: Table S1. Statistics of the EU-ADR and GAD corpora. The Association type classifies the association according to the level of certainty: TRUE (positive (PA), negative (NA) and speculative (SA)) and FALSE (FA).

Additional file 2: In addition, this article contains supplementary information available online (http://ibi.imim.es/befree/#supplmaterial).

Additional file 3: Table S2. Evaluation of BeFree and SemRep for identification of drug-target, gene-disease and drug-disease relationships using the EU-ADR corpus. A selection of the results obtained by BeFree by 10-fold cross-validation on the EU-ADR corpus and the performance of SemRep on the same corpus are shown. The first column indicates the number of the experiment as it appears in http://ibi.imim.es/befree/#supplbefree, Table 1. The second column shows if K_{SL} is used with (TG +SBG) or without (TG) sparse bigrams, or if it is not used (-). The next two columns focus on K_{DEP} walk features indicating the use of one of the following features: token (T), stem (S), lemma (L), POS-tag (P), role (R) or none (-). Finally, the last columns show the result obtained in each experiment indicating Precision (P), Recall (R) and f-measure (F) in percentage (%). *In the case of SemRep, note that the results were not obtained by cross-validation.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

AB developed the system and performed the evaluation, and participated in data analysis and writing the manuscript. MR contributed to the development of the system. NQ and JP performed data analysis and participated in writing the manuscript. LIF conceived the study, coordinated the work, contributed to data analysis and wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We thank Dr. Olga Valverde (UPF) and Dr. Marta Torrens (IMIM) for their expertise in the use case of depression. Funding: The research leading to these results has received support from Instituto de Salud Carlos III-Fondo Europeo de Desarrollo Regional (PI13/00082 and CP10/00524), the Innovative Medicines Initiative Joint Undertaking under grants agreements n° [115002] (eTOX) and n° [115191] (Open PHACTS)], resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and EFPIA companies' in kind contribution. À.B. and L.I.F received support from Instituto de Salud Carlos III Fondo Europeo de Desarollo Regional (CP10/00524). The Research Programme on Biomedical Informatics (GRIB) is a node of the Spanish National Institute of Bioinformatics (INB).

Received: 24 July 2014 Accepted: 19 January 2015
Published online: 21 February 2015

References

1. Rebholz-Schuhmann D, Oellrich A, Hoehndorf R. Text-mining solutions for biomedical research: enabling integrative biology. Nat Rev Genet. 2012;13:829–39.
2. Cases M, Furlong LI, Albanell J, Altman RB, Bellazzi R, Boyer S, et al. Improving data and knowledge management to better integrate health care and research. J Intern Med. 2013;274:321–8.
3. Hahn U, Cohen KB, Garten Y, Shah NH. Mining the pharmacogenomics literature–a survey of the state of the art. Brief Bioinform. 2012;13:460–94.
4. Arighi CN, Wu CH, Cohen KB, Hirschman L, Krallinger M, Valencia A, et al. BioCreative-IV virtual issue. Database. 2014;2014:bau039–9.
5. Pakhomov S, McInnes BT, Lamba J, Liu Y, Melton GB, Ghodke Y, et al. Using PharmGKB to train text mining approaches for identifying potential gene targets for pharmacogenomic studies. J Biomed Inform. 2012;45:862–9.
6. Xu R, Wang Q. A knowledge-driven conditional approach to extract pharmacogenomics specific drug-gene relationships from free text. J Biomed Inform. 2012;45:827–34.
7. Percha B, Garten Y, Altman RB. Discovery and explanation of drug-drug interactions via text mining. Pac Symp Biocomput 2012:410–21.
8. Segura-Bedmar I, Martínez P, de Pablo-Sánchez C. Using a shallow linguistic kernel for drug-drug interaction extraction. J Biomed Inform. 2011;44:789–804.
9. Fiszman M, Demner-Fushman D, Kilicoglu H, Rindflesch TC. Automatic summarization of MEDLINE citations for evidence-based medical treatment: a topic-oriented evaluation. J Biomed Inform. 2009;42:801–13.
10. Névéol A, Lu Z. Automatic integration of drug indications from multiple health resources. In: Proc ACM Int Conf Heal informatics - IHI '10. New York, New York, USA: ACM Press; 2010. p. 666.
11. Gurulingappa H, Mateen-Rajput A, Toldo L. Extraction of potential adverse drug events from medical case reports. J Biomed Semantics. 2012;3:15.
12. Kang N, Singh B, Bui C, Afzal Z, van Mulligen EM, Kors JA. Knowledge-based extraction of adverse drug events from biomedical text. BMC Bioinformatics. 2014;15:64.
13. Bundschus M, Dejori M, Stetter M, Tresp V, Kriegel H-P. Extraction of semantic biomedical relations from text using conditional random fields. BMC Bioinformatics. 2008;9:207.
14. Ozgür A, Vu T, Erkan G, Radev DR. Identifying gene-disease associations using centrality on a literature mined gene-interaction network. Bioinformatics. 2008;24:i277–85.
15. Hakenberg J, Voronov D, Nguyên VH, Liang S, Anwar S, Lumpkin B, et al. A SNPshot of PubMed to associate genetic variants with drugs, diseases, and adverse reactions. J Biomed Inform. 2012;45:842–50.
16. Vincze V, Szarvas G, Farkas R, Móra G, Csirik J. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. BMC Bioinformatics. 2008;9 Suppl 11:S9.
17. Kilicoglu H, Bergler S. Recognizing speculative language in biomedical research articles: a linguistically motivated perspective. BMC Bioinformatics. 2008;9 Suppl 11:S10.
18. Nawaz R, Thompson P, Ananiadou S. Negated bio-events: analysis and identification. BMC Bioinformatics. 2013;14:14.
19. Ananiadou S, Thompson P, Nawaz R, McNaught J, Kell DB. Event-based text mining for biology and functional genomics. Brief Funct Genomics 2014:elu015–.
20. Jenssen T-K, Laegreid A, Komorowski J, Hovig E. A literature network of human genes for high-throughput analysis of gene expression. Nat Genet. 2001;28:21–8.
21. Hoffmann R, Valencia A. Implementing the iHOP concept for navigation of biomedical literature. Bioinformatics. 2005;21 Suppl 2:ii252–8.
22. Wang X, Hripcsak G, Markatou M, Friedman C. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. J Am Med Inform Assoc , 16:328–37.
23. Fundel K, Küffner R, Zimmer R. RelEx–relation extraction using dependency parse trees. Bioinformatics. 2007;23:365–71.
24. Caporaso JG, Baumgartner WA, Randolph DA, Cohen KB, Hunter L. Rapid pattern development for concept recognition systems: application to point mutations. J Bioinform Comput Biol. 2007;5:1233–59.
25. Chun H, Tsuruoka Y, Kim J, Shiba R, Nagata N, Hishiki T, et al. Extraction of gene-disease relations from Medline using domain dictionaries and machine learning. Pac Symp Biocomput 2006;4.
26. Liu H, Hunter L, Kešelj V, Verspoor K. Approximate Subgraph Matching-Based Literature Mining for Biomedical Events and Relations. PLoS One. 2013;8:e60954.
27. McClosky D, Riedel S, Surdeanu M, McCallum A, Manning CD. Combining joint models for biomedical event extraction. BMC Bioinformatics. 2012;13 Suppl 11:S9.
28. Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. J Biomed Inform. 2003;36:462–77.
29. Friedman C, Liu H, Shagina L, Johnson S, Hripcsak G. Evaluating the UMLS as a source of lexical knowledge for medical language processing. Proc AMIA Symp 2001;189–193.
30. Kim J-D, Ohta T, Pyysalo S, Kano Y, Tsujii J. Overview of BioNLP'09 shared task on event extraction. In: BioNLP '09 Proc Work Curr Trends Biomed Nat Lang Process Shar Task. Association for Computational Linguistics. 2009. p. 1–9.
31. Chowdhury MFM, Lavelli A. Combining tree structures, flat features and patterns for biomedical relation extraction. In EACL '12 Proc 13th Conf Eur Chapter Assoc Comput Linguist. Association for Computational Linguistics; 2012:420–429.
32. Culotta A, Sorensen J. Dependency tree kernels for relation extraction. In: Proc 42nd Annu Meet Assoc Comput Linguist - ACL '04. Morristown, NJ, USA: Association for Computational Linguistics; 2004. p. 423–es.
33. Giuliano C, Lavelli A, Romano L. Exploiting Shallow Linguistic Information for Relation Extraction from Biomedical Literature. In: 11th Conf Eur Chapter Assoc Comput Linguist (EACL '06). 2006. p. 401–8.
34. Miwa M, Saetre R, Miyao Y, Tsujii J. Protein-protein interaction extraction by leveraging multiple kernels and parsers. Int J Med Inform. 2009;78:e39–46.
35. Kim S, Yoon J, Yang J, Park S. Walk-weighted subsequence kernels for protein-protein interaction extraction. BMC Bioinformatics. 2010;11:107.
36. Kim S, Yoon J, Yang J. Kernel approaches for genic interaction extraction. Bioinformatics. 2008;24:118–26.
37. Hahn U, Cohen K. Mining the pharmacogenomics literature—a survey of the state of the art. Brief Bioinform. 2012;13(4):460–94.
38. Van Mulligen EM, Fourrier-Reglat A, Gurwitz D, Molokhia M, Nieto A, Trifiro G, et al. The EU-ADR corpus: annotated drugs, diseases, targets, and their relationships. J Biomed Inform. 2012;45:879–84.
39. Bravo A, Cases M, Queralt-Rosinach N, Sanz F, Furlong LI. A knowledge-driven approach to extract disease-related biomarkers from the literature. Biomed Res Int. 2014;2014:253128.
40. Bauer-Mehren A, Rautschka M, Sanz F, Furlong LI. DisGeNET: a Cytoscape plugin to visualize, integrate, search and analyze gene-disease networks. Bioinformatics. 2010;26:2924–6.
41. Semantic Knowledge Representation. [http://semrep.nlm.nih.gov/]
42. Buyko E, Beisswanger E, Hahn U. The extraction of pharmacogenetic and pharmacogenomic relations–a case study using PharmGKB. Pac Symp Biocomput 2012;376–87.
43. Xu R, Wang Q. Large-scale extraction of accurate drug-disease treatment pairs from biomedical literature for drug repurposing. BMC Bioinformatics. 2013;14:181.
44. Albert PR, Benkelfat C, Descarries L. The neurobiology of depression–revisiting the serotonin hypothesis. I. Cellular and molecular mechanisms. Philos Trans R Soc Lond B Biol Sci. 2012;367:2378–81.
45. Dennis G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, et al. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol. 2003;4:P3.
46. Good BM, Su AI. Crowdsourcing for bioinformatics. Bioinformatics. 2013;29:1925–33.
47. Mons B, van Haagen H, Chichester C, Hoen P-B 't, den Dunnen JT, van Ommen G, et al. The value of data. Nat Genet. 2011;43:281–3.
48. Smith L, Tanabe L, Ando R, Kuo C, Chung I, Hsu C, et al. Overview of BioCreative II gene mention recognition. Genome Biol. 2008;9:S2.
49. Kang N, Singh B, Afzal Z, van Mulligen EM, Kors JA. Using rule-based natural language processing to improve disease normalization in biomedical text. J Am Med Inform Assoc. 2012;20:876–81.
50. Leaman R, Miller C. Enabling Recognition of Diseases in Biomedical Text with Machine Learning : Corpus and Benchmark. In: Proc 3rd Int Symp Lang Biol Med. 2009. p. 82–9.
51. DisGeNET - a database of gene-disease associations. [http://www.disgenet.org/]