BeFree: a text mining system to extract relations between genes, diseases and drugs for translational research

Àlex Bravo
GRIB, IMIM, DCEXS, UPF, Spain
abravo@imim.es

Janet Piñero
GRIB, IMIM, DCEXS, UPF, Spain
jpinero@imim.es

Núria Queralt
GRIB, IMIM, DCEXS, UPF, Spain
nqueralt@imim.es

Michael Rautschka
GRIB, IMIM, DCEXS, UPF, Spain
rautschy@gmail.com

Laura I. Furlong
GRIB, IMIM, DCEXS, UPF, Spain
lfurlong@imim.es

Abstract

Current biomedical research needs to leverage the large amount of information reported in publications. Text mining approaches aimed at finding relationships between entities are key for identification of actionable knowledge from free text repositories. We report on the development of the BeFree system designed to identify relationships between biomedical entities with a special focus on genes and their associated diseases. BeFree, by exploiting morpho-syntactic information of the text, performs competitively not only for the identification of gene-disease relationships, but also for drug-disease and drug-target associations. The application of BeFree to a real-case scenario shows its potentiality in extracting relevant information for translational research.

1 Introduction

Due to the increasing size of scientific 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. Text mining approaches ease the access to information otherwise locked in millions of documents (Rebholz-Schuhmann, Oellrich, & Hoehndorf, 2012), but several challenges still remain, 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. In the last years there has been an increasing demand on the identification of relationships involving entities of biomedical interest such as diseases, drugs, genes and their sequence variants (Hahn, Cohen, Garten, & Shah, 2012). Regarding relation extraction (RE), supervised learning approaches have shown good performance exploiting both syntactic and semantic information (J.-D. Kim, Ohta, Pyysalo, Kano, & Tsujii, 2009). Most of the studies have focused in kernel based methods to identify associations between entities (Chowdhury & Lavelli, 2012; Culotta & Sorensen, 2004; Giuliano, Lavelli, & Romano, 2006; Miwa, Saetre, Miyao, & Tsujii, 2009). 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 interactions between proteins (LLL, AIMed, Bioinfer, HPRD50 and IEPA), manually annotated corpora for other associations are scarce (Hahn et al., 2012). We present BeFree, a supervised text mining system to identify relationships involving genes, diseases, and drugs from free text. BeFree is composed of a biomedical Named Entity Recognition module named BioNER (Bravo, Cases, Queralt-Rosinach, Sanz, & Furlong, 2014), and a kernel-based RE module. We performed an evaluation of BeFree kernel-based RE models obtained by the combination of the Shallow Linguistic Kernel (K_{SL}) (Giuliano et al., 2006) with the Dependency Kernel (K_{DEP}), a new kernel that exploits deep syntactic information, for the identification of relationships between genes, diseases and drugs. In addition, we assessed the potential of the BeFree system to identify relevant information in a real-case scenario: the search for genes associated to depression, one of the most prevalent diseases nowadays. As we focused our case studies in identifying associations between genes and diseases, we also developed a new corpus by a semi-automatic annotation procedure based on the Genetic Association Database (GAD), an archive of genetic association studies of complex diseases. In summary, we addressed some of the current challenges in the field, such as improving RE for entities of biomedical interest, integration with existing knowledge bases and exploitation of the extracted information in real-case scenarios.

2 Methods

BeFree is a supervised text mining system composed of a Named Entity Recognition (NER) module named BioNER (Bravo, Cases, et al., 2014), and a kernel-based RE module.

2.1 NER module: BioNER is based on dictionaries and fuzzy matching methods, more details in (Bravo, Cases, et al., 2014).

2.2 RE module: In order to implement a RE for different relationships (drug-target, drug-disease, gene-disease), we combine the Shallow Linguistic Kernel (K_{SL}) described by (Giuliano et al., 2006) and our Dependency Kernel (K_{DEP}) that exploits the syntactic information of the sentence using the walk-weighted subsequence kernels as proposed by (S. Kim, Yoon, & Yang, 2008). For more details see (Bravo, Pinero, Queralt, Rautschka, & Furlong, 2014).

2.3 Corpora: We used the EU-ADR corpus (van Mulligen et al., 2012) 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 (http://geneticassociationdb.nih.gov/), using the annotations of relationships between a gene and a disease in a single sentence. We applied BioNER to these sentences to identify the gene and disease entities and normalize them to NCBI Gene and UMLS identifiers, respectively. 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 labeled as TRUE (2800 sentences). In order to create a dataset containing false associations between a gene and a disease (labeled as FALSE, gene and disease co-occur in a sentence but are not semantically associated, 2529 sentences), 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.

2.4 Evaluation: 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 (real relationship between the entities, in contrast with FALSE sentences where two entities co-occur, but there is no semantic relationship between them).

3 Results and Discussion

We developed a RE system for 3 types of relationships between entities relevant in translational research: genes, drugs and diseases. We assessed the performance of the K_{SL} and K_{DEP} kernels using different morphosyntactic features on the relationships available in the EU-ADR corpus (see Table 1 in http://ibi.imim.es/befree/#supplmaterial for the full set of features evaluated). 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 (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 part-of-speech tags on both the e-walk and v-walk features (P 74.5%, R 71.5%, F 72.3%). Similar results were obtained for the gene-disease associations, where the K_{DEP} kernel alone achieved the best performance, using stem or lemma over the v-walk features (P 75.1 %, R 97.7%, F 84.6%), or when using lemma in v-walk and role in e-walk (P 83.8 %, R 71.0%, F 75.6%). Finally, for target-drug relationship, the best classification in terms of F score is achieved when using different combination of features with both kernels (P 74.2%, R 97.4%, F 83.3%). Nevertheless, it is worth to mention that the K_{SL} kernel, which only uses shallow linguistic information, achieves competitive results in the classification of drug-disease, gene-disease and drug-gene associations (F score: 76.7 %, 80.9%, 81.1%, respectively). Although it is not possible to do a comparison to other approaches due to different benchmarks used for evaluation, our results are comparable to state-of-the-art approaches. For instance, for gene-disease associations F scores of 78 % (Bundschus, Dejori, Stetter, Tresp, & Kriegel, 2008) and 76 % (Buyko, Beisswanger, &

Hahn, 2012) have been reported. For drug-disease, F scores of 87 % (Gurulingappa, Mateen-Rajput, & Toldo, 2012), 79 % (Buyko et al., 2012) and 50.5 % (Kang et al., 2014) were reported, while for drug-target the values are around 80 % (Buyko et al., 2012).

In order to test the feasibility of using a semi-automatic annotated corpus for biomedical relation extraction, we developed a corpus from the GAD database. Compared to the gene-disease set from the EU-ADR corpus, this is larger and contain a different ratio of true/false associations, thus it is interesting to see how the different combination of kernels and features behaves in this benchmark. The best results where those obtained with the K_{SL} (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.

Case study on genetic basis of depression

Depression is a chronic, life-threatening disease and the second cause of morbidity worldwide, costing billions of dollars per year to the society (Albert, Benkelfat, & Descarries, 2012). 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 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. Thus, we evaluated the performance of BeFree to identify genes associated to depression. Using a set of 270 abstracts pertaining to depression published during 2012, we applied BeFree to identify the genes associated to depression. We used the models 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, Table 2). To assess the performance of BeFree, we manually checked a random sample of 100 gene-disease associations. In this sample, we determined the number of true and false associations between depression and any gene, and calculated P and R values for the predictions. In the case of the model trained on the EU-ADR corpus, 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. All in all, the model trained in the EU-ADR corpus, performed a better classification of gene-disease associations in this real case setting (F-score 73.5%).

We then compared the genes identified as related to depression by BeFree with the genes already known to be associated to depression available from DisGeNET (Bauer-Mehren, Rautschka, Sanz, & Furlong, 2010). The BeFree model trained on the EU-ADR corpus identified 168 genes, 41 of them available in DisGeNET, whereas the model trained on the GAD corpus retrieved 104 genes, 37 of them were already reported in DisGeNET (Figure 1). 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 DisGeNET. We analyzed more deeply the set of genes that were predicted by both methods and were not present in DisGeNET (59 genes) by functional enrichment analysis with Gene Ontology terms using DAVID (Dennis et al., 2003). 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 interestingly, some of the genes identified by text mining are putatively involved in RNA regulation, RNA splicing and epigenetic regulation. This is noteworthy since there is an increasing interest in the relationship between the aforementioned processes and the physiopathology of depression (Albert et al., 2012).

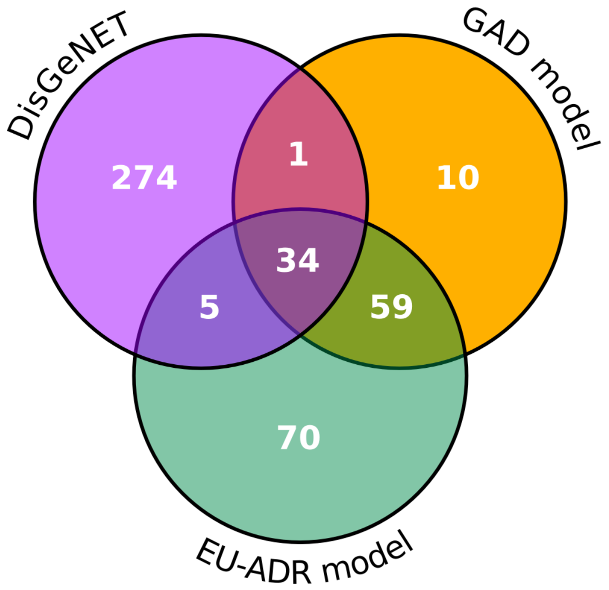
4 Conclusion

We present the BeFree system, a kernel-based approach that using morphosyntactic and dependency information performs competitively (F score at 80 % level) for the identification of drug-disease, drug-target and gene-disease relationships from free text. We show that training BeFree on different corpora (e.g. EU-ADR, GAD) allows the identification of gene-disease associations in a real-case scenario with good performance. Finally, as previously suggested by others (Pakhomov et al., 2012), a corpus developed by semi-automatic annotation is a good resource for developing a RE system in biomedicine. In addition, we evaluated the value of the information extracted by BeFree for a specific case study in translational research. Particularly, BeFree is able to identify genes associated to depression that are not present in other specialized databases. More importantly, some of these genes represent novel aspects of the physiopathology of depression.

Acknowledgements

The research leading to these results has received support from ISCIII-FEDER (PI13/00082 and PI13/00082), the IMI-JU under grants agreements n°115002 (eTOX) and nº115191 (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 ISCIII-FEDER (CP10/00524). The Research Programme on Biomedical Informatics (GRIB) is a node of the Spanish National Institute of Bioinformatics (INB).

Figure 1



References

Albert, P. R., Benkelfat, C., & Descarries, L. (2012). The neurobiology of depression--revisiting the serotonin hypothesis. I. Cellular and molecular mechanisms. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *367*(1601), 2378–81. doi:10.1098/rstb.2012.0190

Bauer-Mehren, A., Rautschka, M., Sanz, F., & Furlong, L. I. (2010). DisGeNET: a Cytoscape plugin to visualize, integrate, search and analyze gene-disease networks. *Bioinformatics*, *26*(22), 2924–2926. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=20861032

Bravo, A., Cases, M., Queralt-Rosinach, N., Sanz, F., & Furlong, L. I. (2014). A knowledge-driven approach to extract disease-related biomarkers from the literature. *BioMed Research International*, *2014*, 253128. doi:10.1155/2014/253128

Bravo, A., Pinero, J., Queralt, N., Rautschka, M., & Furlong, L. I. (2014). *Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research*. *bioRxiv* (p. 007443). Cold Spring Harbor Labs Journals. doi:10.1101/007443

Bundschus, M., Dejori, M., Stetter, M., Tresp, V., & Kriegel, H.-P. (2008, April). Extraction of semantic biomedical relations from text using conditional random fields. Retrieved January 01, 2009, from http://www.biomedcentral.com/1471-2105/9/207/

Buyko, E., Beisswanger, E., & Hahn, U. (2012). The extraction of pharmacogenetic and pharmacogenomic relations--a case study using PharmGKB. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 376–87. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/22174293

Chowdhury, M. F. M., & Lavelli, A. (2012). Combining tree structures, flat features and patterns for biomedical relation extraction. In *EACL '12 Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 420–429).

Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=2380816.2380869

Culotta, A., & Sorensen, J. (2004). Dependency tree kernels for relation extraction. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics - ACL '04* (p. 423–es). Morristown, NJ, USA: Association for Computational Linguistics. doi:10.3115/1218955.1219009

Dennis, G., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C., & Lempicki, R. A. (2003). DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biology*, *4*(5), P3. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3720094&tool=pmcentrez&rendertype=abstract

Giuliano, C., Lavelli, A., & Romano, L. (2006). Exploiting shallow linguistic information for relation extraction from biomedical literature. In *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics (EACL-2006)* (pp. 5–7).

Gurulingappa, H., Mateen-Rajput, A., & Toldo, L. (2012). Extraction of potential adverse drug events from medical case reports. *Journal of Biomedical Semantics*, *3*(1), 15. doi:10.1186/2041-1480-3-15

Hahn, U., Cohen, K. B., Garten, Y., & Shah, N. H. (2012). Mining the pharmacogenomics literature--a survey of the state of the art. *Briefings in Bioinformatics*, *13*(4), 460–94. doi:10.1093/bib/bbs018

Kang, N., Singh, B., Bui, C., Afzal, Z., Mulligen, E. M. van, & Kors, J. A. (2014). Knowledge-based extraction of adverse drug events from biomedical text. *BMC Bioinformatics*, *15*(1), 64. doi:10.1186/1471-2105-15-64

Kim, J.-D., Ohta, T., Pyysalo, S., Kano, Y., & Tsujii, J. (2009). Overview of BioNLP'09 shared task on event extraction. In *BioNLP '09 Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task* (pp. 1–9). Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=1572340.1572342

Kim, S., Yoon, J., & Yang, J. (2008). Kernel approaches for genic interaction extraction. *Bioinformatics (Oxford, England)*, *24*(1), 118–26. doi:10.1093/bioinformatics/btm544

Miwa, M., Saetre, R., Miyao, Y., & Tsujii, J. (2009). Protein-protein interaction extraction by leveraging multiple kernels and parsers. *International Journal of Medical Informatics*, *78*(12), e39–46. doi:10.1016/j.ijmedinf.2009.04.010

Pakhomov, S., McInnes, B. T., Lamba, J., Liu, Y., Melton, G. B., Ghodke, Y., … Birnbaum, A. K. (2012). Using PharmGKB to train text mining approaches for identifying potential gene targets for pharmacogenomic studies. *Journal of Biomedical Informatics*, *45*(5), 862–9. doi:10.1016/j.jbi.2012.04.007

Rebholz-Schuhmann, D., Oellrich, A., & Hoehndorf, R. (2012). Text-mining solutions for biomedical research: enabling integrative biology. *Nat Rev Genet*, *13*(12), 829–839. doi:10.1038/nrg3337

Van Mulligen, E. M., Fourrier-Reglat, A., Gurwitz, D., Molokhia, M., Nieto, A., Trifiro, G., … Furlong, L. I. (2012). The EU-ADR corpus: annotated drugs, diseases, targets, and their relationships. *Journal of Biomedical Informatics*, *45*(5), 879–84. doi:10.1016/j.jbi.2012.04.004