Use of text mining for Experimental Factor Ontology coverage expansion in the scope of target validation

Şenay Kafkas, Ian Dunham, Helen Parkinson and Jo McEntyre

European Bioinformatics Institute – European Molecular Biology Laboratory (EMBL-EBI), and
Open Targets
Wellcome Genome Campus
Hinxton, CB10 1SD, UK

***Abstract*—Understanding the molecular biology and development of disease plays a key role in drug development. Integrating evidence from different experimental approaches with data available from public resources (such as gene expression level changes and reaction pathways affected by pathogenic mutations) can be a powerful approach for evaluating different aspects of target-disease associations. The application of ontologies is of fundamental importance to effective integration. The Target Validation Platform is a user-friendly interface that integrates such evidences from various resources with the aim of assisting scientists to identify and prioritise drug targets. Currently, the EFO is used as the reference ontology for diseases in the platform, importing terms from existing disease ontologies such as the Human Phenotype Ontology as required. In order to generalize the use of EFO from key target-diseases for wider use, we need to compare the target associated disease coverage in EFO with the scope of other available disease terminology resources. In this study, we address this issue by using text mining and present our initial results.**

Keywords—text mining; ontology; integration; target validation

I. Introduction

Integrating data from *de novo* experiments with data available in public data resources in a user friendly interface to support decision making has been the goal of the Target Validation Platform (https://targetvalidation.org). This platform integrates a variety of evidence for a given target (gene/protein) - disease association, such as reaction pathways that are affected by pathogenic mutations from Reactome [1], and text mined target-disease associations from the Europe PubMed Central (Europe PMC) (http://europepmc.org/) literature database [2]. The application of disease ontologies is critical to integrate such different data types.

The Experimental Factor Ontology (EFO) (http://www.ebi.ac.uk/efo/) is the reference resource for diseases in the platform ("disease" here encompasses both "disease/phenotype" as the disease/phenotype boundary is blurred in both the platform's data sources and ontologically). Therefore, it is important to understand the disease coverage of EFO in the scope of target validation, in comparison to the other available major disease and phenotype resources, in order to expand its disease coverage. In this study, we address this issue by using text mining which is a widely used approach in ontology expansion [3] and target-disease association identification [4], to compare terms available in existing ontologies and present our initial results.

II. Methods

A. Resources Used

We used Europe PMC as the literature database, UniProt for target (gene/protein) names and six major disease terminologies: EFO (V2.69), the Human Phenotype ontology (HP) (access date:31-03-2016) (http://human-phenotype-ontology.github.io/), Orphanet Rare Disease Ontology (ORDO) (V2.1) (http://www.orphadata.org/cgi-bin/inc/ordo_orphanet.inc.php), the Human Disease Ontology (HDO) (06-01-2016 update) (http://disease-ontology.org/), the Mammalian Phenotype Ontology (MP) (access date:31-03-2016)
(http://www.informatics.jax.org/searches/MP_form.shtml), and Unified Medical Language Systems (UMLS) (2014 AB Release) (https://www.nlm.nih.gov/research/umls/).

Europe PMC is one of the largest biomedical literature databases in the World which provides public access to 31 million abstracts and 3.7 million full text articles, covering both PubMed and PubMed Central. In our analyses, we used the latest achieved version of the Open Access full text articles (~1 Million) (http://europepmc.org/ftp/archive/v.2016.03/) from the database.

We generated and refined dictionaries from the human part of the SwissProt Database (the expert annotated part of UniProt) (http://www.uniprot.org/) and disease and phenotype parts of EFO, HP, ORDO, MP, HDO and UMLS before applying text mining. In the refining process, we filtered out the terms that would introduce potentially high numbers of false positives. These are the terms having character length < 3 and the terms that are ambiguous with common English words (e.g. "Large" is a protein name as well). In addition, we generated term variations by replacing the widely used Greek letters in gene/disease names with their symbols (e.g. replacing "alpha" with α). The final target and disease dictionaries consisted of a total of 104,434 Uniprot, 26,617 EFO, 18,332 HP, 20,152 ORDO, 29,800 MP, 21,789 HDO and 75,060 UMLS terms.

B. *Target and disease name identification*

We used the Europe PMC text-mining pipeline, which is based on Whatizit [5] to annotate target and disease names in text with the dictionaries described above. Target and disease name abbreviations can be ambiguous with some other names (e.g. ALS which is “Amyotrophic Lateral Sclerosis”, is ambiguous with “Advanced Life Support”, PMID:26811420). Therefore, we implemented and used abbreviation filters for screening out the potential false positive disease/protein abbreviations introduced during the annotation process. The abbreviation filters operate based on several heuristic rules. For example, text sequences within parentheses (i.e. (XYZ)), appearing in uppercase and having length <6 are identified as a name abbreviation candidate and are retained as an annotation only if any of its long forms from the given disease ontology exists elsewhere in the document.

C. *Target-disease association extraction*

The associations are extracted by identifying the target-disease co-occurrences at the sentence level and applying several filtering rules to reduce noise possibly introduced by the high sensitivity, low specificity co-occurrence approach. The filtering rules utilise heuristic information from a careful manual analysis of the text. They include, filtering out all articles but the “Research” articles (e.g. Reviews, Case Reports), filtering out target-disease associations appearing in certain sections such as “Methods” and “References”, and filtering out target-disease associations that appear only once in the body of a given article but not in the article's title or abstract (see [6] for the details).

III. Results and Discussion

Our target-disease extraction system achieves a Mean-Average Precision value of 81% [6]. Figure 1 presents a Venn diagram showing the disease terms found in the corpus that are associated with targets, after application of the target-disease heuristics above, for each of the six different disease resources. There are 3,859 HDO, 3,384 MP, 1,610 ORDO, 4,277 HP and 17,584 UMLS target associated distinct disease terms that are not found by EFO. Possible reasons for the difference in coverage between EFO and the other terminologies are twofold: nonexistence of a given disease name in EFO, the coverage of a given disease with different synonyms and different classification of a given term in EFO. For example, “fetal valproate syndrome” and “Chagas cardiomyopathy” from ORDO are not covered by EFO. “HIV” is classified as “disease and syndrome” in UMLS, indicating “HIV infection”, however, in EFO, it is classified as a virus name. Results suggest that there is some room for improvement in the EFO and this will be explored for future releases of EFO.

IV. Conclusion and Future Work

In this study, we demonstrate the use of text mining for analysing and suggesting approaches to expand the disease/phenotype coverage of EFO within the scope of target validation. We focused on the target-associated disease terms from EFO and five other major disease resources, but there is no reason why this approach could not be applied to other contexts in efforts to integrate across terminologies and ontologies. In future, we will extend our analysis to discover any trends over the resources, to understand the disease/phenotype target space derived from literature and how much of the associations that we find in EFO scope is relevant.

Acknowledgments

This work is funded by the Open Targets.

References

[1] A. Fabregat, K. Sidiropoulos, P.Garapati, M. Gillespie, K. Hausmann et al., “The Reactome pathway Knowledgebase,” Nucleic Acids Res., 44(D1):D481-7, 2016.

[2] Europe PMC Consortium, “Europe PMC: a full-text literature database for the life sciences and platform for innovation,” Nucleic Acids Res., 43(Database issue):D1042-8, 2015.

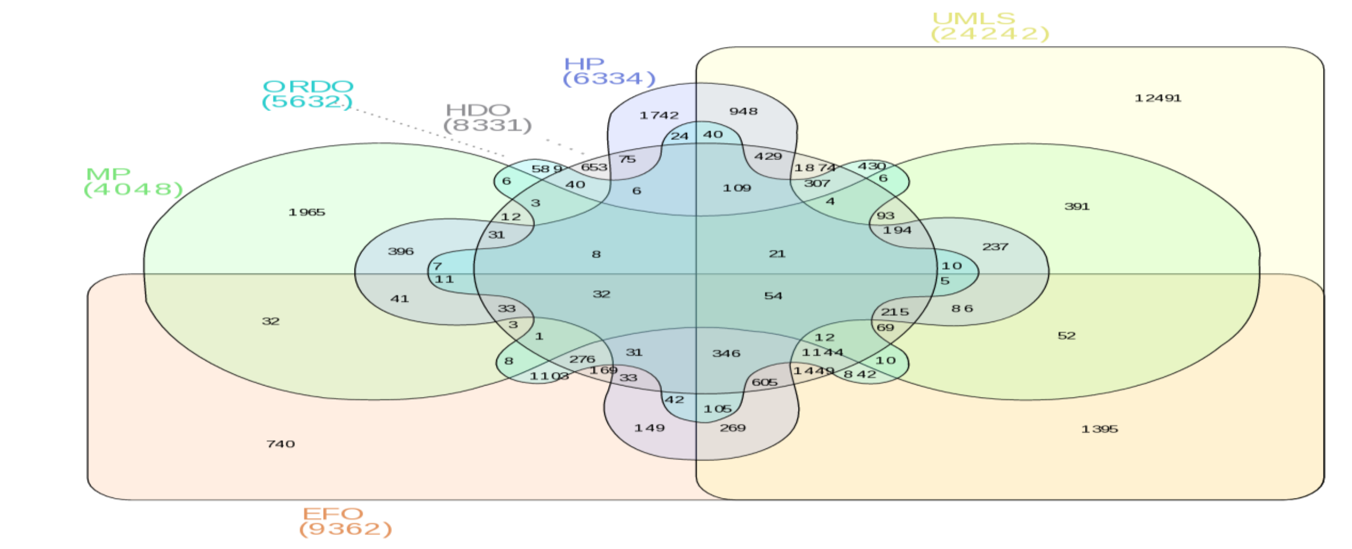
[3] I. Spasic, S. Ananiadou, J. McNaught, A. Kumar, “Text Mining and Ontologies in biomedicine:Making sense of raw text,” Breefings in Bioinformatics, 6(3), pp. 239-251, 2005.

[4] S. Pletscher-Frankild, A. Pallejà, K. Tsafou, J.X. Binder, L.J. Jensena, “DISEASES: Text mining and data integration of disease–gene associations,” Methods, 34, pp.83-89, 2015.

[5] D. Rebholz-Schuhmann, M. Arregui, S. Gaudan, H. Kirsch, A. Jimeno, “Text processing through Web services: calling Whatizit,” Bioinformatics, vol. 24(2), pp.296-8, 2008.

[6] Ş. Kafkas, I. Dunham and J. McEntyre, “Literature Evidence in Open Targets– a target validation platform,” Phenotype Day @ISMB 2016, special session of the Bio-Ontologies SIG, 8-12 July 2016, Orlando, Florida, U.S.

Fig1. Venn diagram showing overlapping target associated disease terms


UMLS
(24242)
HP
(6334)
ORDO
(5632)
HDO
(8331)
MP
(4048)
EFO
(9362)