# Using biomedical databases as knowledge sources for large-scale text mining

Fabio Rinaldi, Institute of Computational Linguistics,
University of Zurich, Switzerland

**Abstract**

In this paper we discuss how terminological knowledge extracted from biomedical databases can be used effectively in large-scale processing of the biomedical literature. We briefly present an integrated information extraction and text mining environment which is capable of reliably identifying and disambiguating several categories of relevant domain entities, which can then constitute relevant indexing entries in order to allow efficient retrieval of relevant documents and passages. Additionally the system generates ranked lists of candidate interactions among the detected entities, which can be useful for several purposes, from assisted literature curation to question answering systems.

## 1 Introduction

The rapid increase of novel scientific results in the domain of molecular biology renders it necessary to collect this information in structured repositories, so that it becomes easily accessible to the end users. Well-known databases like UniProt, Mint, IntAct, BioGrid, collect information about proteins and their interactions. PharmGKB [4, 12] curates knowledge about the impact of genetic variation on drug response for clinicians and researchers. The Comparative Toxicogenomics Database (CTD) collects interactions between chemicals and genes in order to support the study on the effects of environmental chemicals on health [5]. A significant amount of manual effort is needed in order to extract from the literature the information required to accurately fill those databases (a process referred to as "curation"). Text mining solutions are increasingly requested to support the process of curation of biomedical databases.

The OntoGene project[1] focuses on the improvement of biomedical text mining through the usage of advanced natural language processing techniques. Our approach relies upon information delivered by a pipeline of NLP tools, including sentence splitting, tokenization, part of speech tagging, term recognition, noun and verb phrase chunking, and a dependency-based syntactic analysis of input sentences [11, 8]. The results of the entity detection feed directly into the process of identification of interactions.

[1] http://www.ontogene.org/

Different implementations of the OntoGene system have been used for participation in several well-known text mining shared tasks, such as BioCreative, CALBC and BioNLP, obtaining always competitive results. For example, in the BioCreative 2009 challenge the OntoGene system obtained the best results for protein-protein interactions [10]. More recently, within the scope of the SASEBio project (Semi-Automated Semantic Enrichment of the Biomedical Literature), we have developed a user-friendly interface (ODIN: OntoGene Document INspector) which can be used by database curator to inspect the results of the text mining system. The interface is designed to simplify the interaction of the user with the text mining system, allowing for example modification of incorrect results. The system can then learn based upon this interaction.

In the rest of this short paper we briefly describe the OntoGene pipeline architecture and the ODIN interface for assisted curation.[2]

# 2 Information Extraction

Biomedical terminological resources can be leveraged for construction of large-scale knowledge bases. One example is KaBOB (Knowledge Base of Biology), a large RDF store based upon 17 prominent biomedical daabases. KaBOB contains 5.6-billion RDF-triples [1]. Similar kinds of integrated data networks can be used for knowledge discovery purposes through usage of semantic web technologies (see for example [2]).

In our own work we have used such databases as knowledge sources for the process of semi-automated information extraction. In the rest of this section we describe the OntoGene Text Mining pipeline which is used to (a) provide all basic preprocessing (e.g. tokenization) of the target documents, (b) identify all mentions of domain entities and normalize them to database identifiers, and (c) extract candidate interactions.

## 2.1 Preprocessing and Detection of Domain Entities

Several large-scale terminological resources are used by the OntoGene system in order to detect names of relevant domain entities in biomedical literature (proteins, genes, chemicals, diseases, etc.) and ground them to widely accepted identifiers assigned by the original database, such as UniProt Knowledgebase, National Center for Biotechnology Information (NCBI) Taxonomy, Proteomics Standards Initiative Molecular Interactions Ontology (PSI-MI), Cell Line Knowledge Base (CLKB), etc.

From the original databases we extract preferred names and synonyms for each term, together with its unique identifier. This information is used to annotate the input documents using an efficient lookup procedure. A term normalization step is used to take into account a number of possible surface variations of the terms. The same normalization is applied to the list of known terms at the beginning of the annotation process, when it is read into memory, and to the candidate terms in the input text, so that a matching between variants of the same term becomes possible despite the differences in the surface strings [8]. For more technical details of the OntoGene terminology recognition process, see [7].

[2] Readers interested in more details are invited to consult the journal publications available from the OntoGene web site.

The terminological resource obtained as described above is used to annotate biomedical text in a relatively straightforward way. First, in a preprocessing stage, the input text is transformed into a custom XML format, and sentences and tokens boundaries are identified. For this task, we use the LingPipe tokenizer and sentence splitter which have been trained on biomedical corpora. The tokenizer produces a granular set of tokens, e.g. words that contain a hyphen (such as 'Pop2p-Cdc18p') are split into several tokens, revealing the inner structure of such constructs which would allow to discover the interaction mention in "Pop2p-Cdc18p interaction". Tagging of terms is performed by sequentially processing each token in a sentence and, if it can start a term, annotate the longest possible match (partial overlaps are excluded). In the case of success, all the possible IDs (as found in the term list) are assigned to the candidate term.

Ambiguity is a serious problem for several types of entities. For example names of some proteins and genes can refer to several different database identifiers. For example, *hemoglobin* can refer to human hemoglobin or to mouse hemoglobin (or to any other species). Besides, even in humans there are several different types of hemoglobin. Using knowledge about the organisms which are the focus of the experiments described in each paper we can disambiguate to a large extent entities such as proteins and genes. In the OntoGene pipeline we apply an approach which we first described in [3]. We first create a ranked list of 'focus' organisms based on all mentions of proteins, genes, cell lines and organisms in the paper. In the disambiguation process we remove all the IDs that do not correspond to an organism present in the list. Additionally, the scores provided for each organism can be used in ranking the candidate IDs for each entity. Such ranking is useful in a semi-automated curation environment where the curator is expected to take the final decision. However, it can also be used in a fully automated environment as a factor in ranking any other derived information, such as interactions where the given entity participates.

## 2.2 Detection of Interactions

Mentions of relevant domain entities in a given text span are used by the OntoGene system to create candidate interactions. The selected text span can vary from a sentence to a larger observation window. Simple co-occurrence in the selected text span is a low-precision, but high-recall indication of a potential relationship among those entities. In order to obtain better precision the OntoGene system uses the syntactic structure of the sentence, and the global distribution of interactions in the original database. In this section we describe in detail how candidate interactions are ranked by our system, according to their relevance for the original database.

The OntoGene system creates an initial ranking of the candidate relations from the selected text span using only the frequency of the respective entities with the following formula:

$$relscore(e_1, e_2) = (f(e_1) + f(e_2))/f(E)$$

where $f(e_1)$ and $f(e_2)$ are the number of times the entities $e_1$ and $e_2$ are observed in the abstract, while $f(E)$ is the total count of all identifiers in the abstract. An additional zone-based boost might be used in some cases (e.g. for entities mentioned in the title).

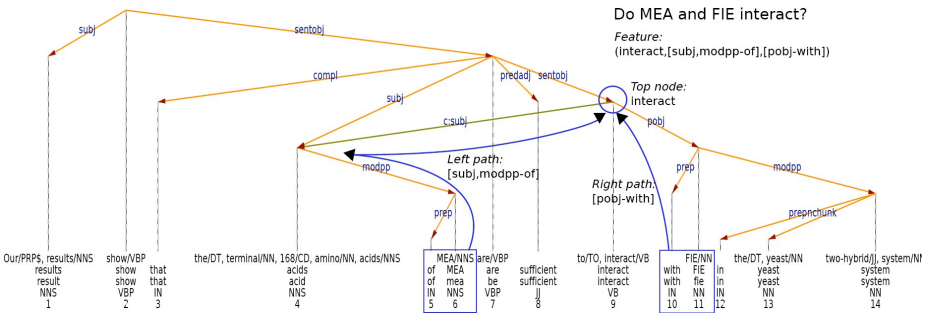


Figure 1: Example of sentence analysis and detection of an interaction.

The OntoGene pipeline makes use of an internally developed dependency parser [13] in order to parse all sentences in the input documents. The information derived from the dependency analysis is used to improve on the baseline ranking for candidate interaction. Besides, the syntactic analysis provides useful information for the extraction of the interaction type. Given two terms identified in the same sentence, a collector traverses the tree from each of the two terms upwards to the lowest common parent node, recording all intermediate nodes and dependency paths along the route. An example of such a traversal can be seen in Figure 1. Such traversals have been used in many PPI applications, they are commonly called tree walks or paths.

Each candidate interaction is assigned a score, obtained by combining several features, including: (1) *Syntactic path*, which encodes the information provided by the dependency structure between the two entities in the candidate interaction; (2) *Known interaction*: in order to better distinguish between 'novel' interactions (more important for the curation process) and 'older' interactions (already known, thus less important for the curation process), we penalize interactions that are already reported in the reference databases, in proportion to their 'age' (date at which the interaction was first reported); (3) *Novelty score*: we also use linguistic clues in order to to distinguish between sentences that report the results detected by the authors (e.g. *"Here we report that..."*) from sentences that report background results. Interactions in 'novelty' sentences are scored higher than interactions in 'background' sentences; (4) *Zoning*: different structural zones of the paper have often different levels of relevance. We observed that novel interactions are often mentioned in the abstract and the conclusions, while the introduction and methods section are less likely and therefore get lower scores; (5) *Pair salience*: the frequency of mentions in the paper of each of the entities in the candidate pair is an important indicator of the relevance of that interaction in the paper. Scores from each feature are then combined and normalized to the `[0,1]` range, in order to produce a ranking for the candidate interactions.

The results of the OntoGene text mining system are made accessible through a curation system called **ODIN** ("OntoGene Document INspector") which allows a user to dynamically inspect the results of the text mining pipeline. An experiment in interactive curation has been performed recently in collaboration with the PharmGKB database [4, 12]. The results of this experiment are described in [6]. [9] provides fur-

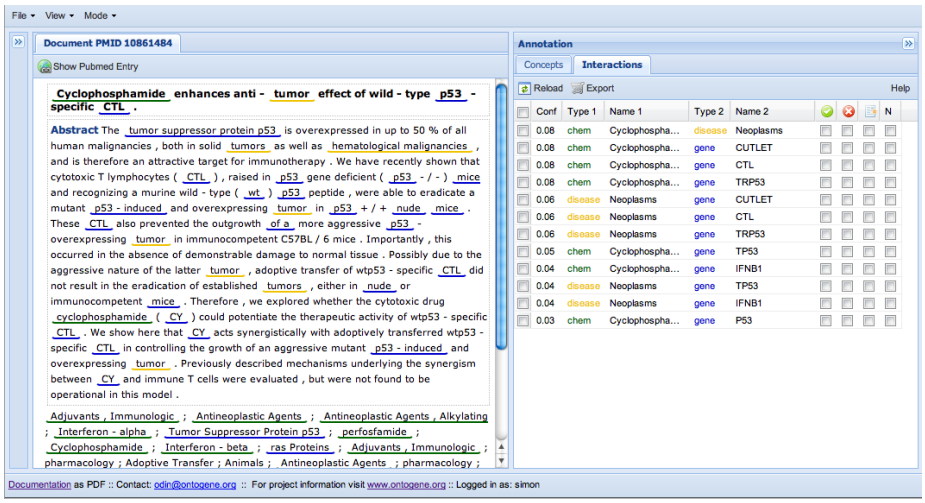


Figure 2: Entity annotations and candidate interactions on a sample PubMed abstract

ther details on the architecture of the system. Figure 2 shows a screenshot of ODIN.

# 3 Conclusion

In this paper we briefly described the OntoGene text mining system, targeted at the extraction of entities and relationships from the biomedical literature. The OntoGene pipeline leverages upon manually curated resources and is capable of reliably identifying entity and relationships which can optionally be delivered using standard semantic-web formats such as RDF or OWL. The long-term vision of the project is a deeper integration of databases and literature.

**Acknowledgments**

This research is partially funded by the Swiss National Science Foundation (grant 100014-118396/1) and Novartis Pharma AG, NIBR-IT, Text Mining Services, CH-4002, Basel, Switzerland.

# References

[1] Michael Bada, Kevin Livingston, and Lawrence Hunter. An ontological representation of biomedical data sources and records. *Bio-Ontologies*, 2011.

[2] Huajun Chen, Li Ding, Zhaohui Wu, Tong Yu, Lavanya Dhanapalan, and Jake Y. Chen. Semantic web for integrated network analysis in biomedicine. *Briefings in Bioinformatics*, 10(2):177–192, 2009.

[3] Thomas Kappeler, Kaarel Kaljurand, and Fabio Rinaldi. TX Task: Automatic Detection of Focus Organisms in Biomedical Publications. In *Proceedings of the BioNLP workshop, Boulder, Colorado*, pages 80–88, 2009.

[4] T.E. Klein, J.T. Chang, M.K. Cho, K.L. Easton, R. Fergerson, M. Hewett, Z. Lin, Y. Liu, S. Liu, D.E. Oliver, D.L. Rubin, F. Shafa, J.M. Stuart, and R.B. Altman. Integrating genotype and phenotype information: An overview of the PharmGKB project. *The Pharmacogenomics Journal*, 1:167–170, 2001.

[5] C.J. Mattingly, M.C. Rosenstein, G.T. Colby, J.N. Forrest Jr, and J.L. Boyer. The Comparative Toxicogenomics Database (CTD): a resource for comparative toxicological studies. *Journal of Experimental Zoology Part A: Comparative Experimental Biology*, 305A(9):689–692, 2006.

[6] Fabio Rinaldi, Simon Clematide, Yael Garten, Michelle Whirl-Carrillo, Li Gong, Joan M. Hebert, Katrin Sangkuhl, Caroline F. Thorn, Teri E. Klein, and Russ B. Altman. Using ODIN for a PharmGKB re-validation experiment. *Database: The Journal of Biological Databases and Curation*, 2012.

[7] Fabio Rinaldi, Kaarel Kaljurand, and Rune Saetre. Terminological resources for text mining over biomedical scientific literature. *Journal of Artificial Intelligence in Medicine*, 52(2):107–114, June 2011.

[8] Fabio Rinaldi, Thomas Kappeler, Kaarel Kaljurand, Gerold Schneider, Manfred Klenner, Simon Clematide, Michael Hess, Jean-Marc von Allmen, Pierre Parisot, Martin Romacker, and Therese Vachon. OntoGene in BioCreative II. *Genome Biology*, 9(Suppl 2):S13, 2008.

[9] Fabio Rinaldi, Gerold Schneider, and Simon Clematide. Relation mining experiments in the pharmacogenomics domain. *Journal of Biomedical Informatics*, 2012. *doi:10.1016/j.jbi.2012.04.014*.

[10] Fabio Rinaldi, Gerold Schneider, Kaarel Kaljurand, Simon Clematide, Therese Vachon, and Martin Romacker. OntoGene in BioCreative II.5. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 7(3):472–480, 2010.

[11] Fabio Rinaldi, Gerold Schneider, Kaarel Kaljurand, Michael Hess, and Martin Romacker. An Environment for Relation Mining over Richly Annotated Corpora: the case of GENIA. *BMC Bioinformatics*, 7(Suppl 3):S3, 2006.

[12] Katrin Sangkuhl, Dorit S. Berlin, Russ B. Altman, and Teri E. Klein. PharmGKB: Understanding the effects of individual genetic variants. *Drug Metabolism Reviews*, 40(4):539–551, 2008. PMID: 18949600.

[13] Gerold Schneider. *Hybrid Long-Distance Functional Dependency Parsing*. Doctoral Thesis, Institute of Computational Linguistics, University of Zurich, 2008.