The OntoGene literature mining web service

Fabio Rinaldi

University of Zurich, Switzerland

Received 1 August 2013; **Accepted** 10 September 2013; **Published** 14 October 2013

Competing interests: the authors have declared that no competing interests exist.

Abstract

Text mining technologies are increasingly providing an effective response to the growing demand for faster access to the vast amounts of information hidden in the literature. Several tools are becoming available which offer the capability to mine the literature for specific information, such as for example protein-protein interactions or drug-disease relationships. The biomedical text mining community regularly verifies the progress of such systems through competitive evaluations, such as BioCreative, BioNLP, i2b2, CALBC, CLEF-ER, BioASQ, etc. The OntoGene system is a text mining system which specialises in the detection of entities and relationships from selected categories, such as proteins, genes, drugs, diseases, chemicals. The quality of the system has been tested several times through participation in some of the community-organised evaluation campaigns. In order to make the advanced text mining capabilities of the OntoGene system more widely accessible without the burden of installation of complex software, we are setting up a web service that will allow any remote user to submit arbitrary documents. The results of the mining service (entities and relationships) are then delivered back to the user as XML data, or optionally can be inspected via a flexible web interface.

Motivation and Objectives

Text mining technologies are increasingly providing an effective response to the growing demand for faster access to the vast amounts of information hidden in the literature. Several tools are becoming available which offer the capability to mine the literature for specific information, such as for example protein-protein interactions or drug-disease relationships. The biomedical text mining community regularly verifies the progress of such systems through competitive evaluations, such as BioCreative, BioNLP, i2b2, CALBC, CLEF-ER, BioASQ, etc.

The OntoGene system is a text mining system which specializes in the detection of entities and relationships from selected categories, such as proteins, genes, drugs, diseases, chemicals. The quality of the system has been tested several times through participation in some of the community-organized evaluation campaigns.

In order to make the advanced text mining capabilities of the OntoGene system more widely accessible without the burden of installation of complex software, we are setting up a web service which will allow any remote user to submit arbitrary documents. The results of the mining service (entities and relationships) are then delivered back to the user as XML data, or optionally can be inspected via a flexible web interface.

Methods

The text mining pipeline which constitutes the core of the OntoGene system has been described previously in a number of publications (Rinaldi, 2008; Rinaldi, 2010; Rinaldi, 2012). We will only briefly describe the core text mining technologies, and instead focus mainly on the novel web service which allows remote access to the OntoGene text mining capabilities.

The first step in order to process a collection of biomedical literature consists in the annotation of names of relevant domain entities in biomedical literature (currently the systems considers proteins, genes, species, experimental methods, cell lines, chemicals, drugs and diseases). These names are sourced from reference databases and are associated with their unique identifiers in those databases, thus allowing resolution of synonyms and cross-linking among different resources. A term normalization step is used to match the terms with their actual representation in the text, taking into account a number of possible surface variations. Finally, a disambiguation step resolves the ambiguity of the matched terms.

Candidate interactions are generated by simple co-occurence of terms within the same syntactic units. However, in order to increase precision, we parse the sentences with our state-of-the art dependency parser, which generates a syntactic representation of the sentence. This is in turn used to score and filter candidate interactions based on the syntactic fragment which connects the two participating entities.

The ranking of relation candidates is further optimized by a supervised machine learning

```
<collection>
<source>PUBMED</source>
<date>20130422</date>
<key>ctdBCIVLearningDataSet.key</key>
<document>
<id>10617681</id>
<passage>
<infon key="type">title</infon>
<offset>0</offset>
<text>
Possible role of valvular serotonin 5-HT(2B) receptors in the cardiopathy associated with
fenfluramine.
</text>
</passage>
<passage>
<infon key="type">abstract</infon>
<offset>104</offset>
<text>
Dexfenfluramine was approved in the United States for long-term use as an appetite suppressant until
it was reported to be associated with valvular heart disease. The valvular changes (myofibroblast
proliferation) are histopathologically indistinguishable from those observed in carcinoid disease
or after long-term exposure to 5-hydroxytryptamine (5-HT)(2)-preferring ergot drugs (ergotamine,
methysergide). 5-HT(2) receptor stimulation is known to cause fibroblast mitogenesis, which could
contribute to this lesion. To elucidate the mechanism of "fen-phen"-associated valvular lesions,
we examined the interaction of fenfluramine and its metabolite norfenfluramine with 5-HT(2)
receptor subtypes and examined the expression of these receptors in human and porcine heart valves.
Fenfluramine binds weakly to 5-HT(2A), 5-HT(2B), and 5-HT(2C) receptors. In contrast, norfenfluramine
exhibited high affinity for 5-HT(2B) and 5-HT(2C) receptors and more moderate affinity for 5-HT(2A)
receptors. In cells expressing recombinant 5-HT(2B) receptors, norfenfluramine potently stimulated the
hydrolysis of inositol phosphates, increased intracellular Ca(2+), and activated the mitogen-activated
protein kinase cascade, the latter of which has been linked to mitogenic actions of the 5-HT(2B)
receptor. The level of 5-HT(2B) and 5-HT(2A) receptor transcripts in heart valves was at least 300-
fold higher than the levels of 5-HT(2C) receptor transcript, which were barely detectable. We propose
that preferential stimulation of valvular 5-HT(2B) receptors by norfenfluramine, ergot drugs, or
5-HT released from carcinoid tumors (with or without accompanying 5-HT(2A) receptor activation) may
contribute to valvular fibroplasia in humans.
</text>
<annotation>
<infon key="type">disease</infon>
<text>HEART VALVE DISEASES</text>
<id>MESH:D006349</id>
</annotation>
</passage>
</document>
</collection>
```

Box 1. The output of the system is generated in the the BioC specification format. This oputput was generated by a query aiming at retrieving the diseases from pubmed abstract 10617681.

method. Since the term recognizer aims at high recall, it introduces several noisy concepts, which we want to automatically identify in order to penalize them. Additionally, we need to adapt to highly-ranked false positive relations which are generated by our frequency based approach. The goal is to identify some global preference or biases which can be found in the reference database. One technique is to weight individual concepts according to their likeliness to appear as an entity in a correct relation, as seen in the target database.

The OntoGene web service has been implemented as a RESTful service (Richardson and Ruby, 2007). It accepts simple XML files as input, based on the BioC specification[1]. The output of the system is generated in the same format. For example, a query aiming at retrieving the diseases from pubmed abstract 10617681 would generate the output presented in Box 1.

Options can be used in the input query to select whether the result should contain in-line annotations (showing where exactly in the text the term was mentioned), or stand-off annotations (as in the example above). Currently the system uses pre-defined terminology, and only allows the users to decide whether they want to use or not to use one of the pre-loaded vocabularies. However we foresee in future the possibility to upload own terminologies.

Since the OntoGene system not only delivers the specific terms found in the submitted articles, but also their unique identifiers in the source

1 http://www.ncbi.nlm.nih.gov/CBBresearch/Dogan/BioC/

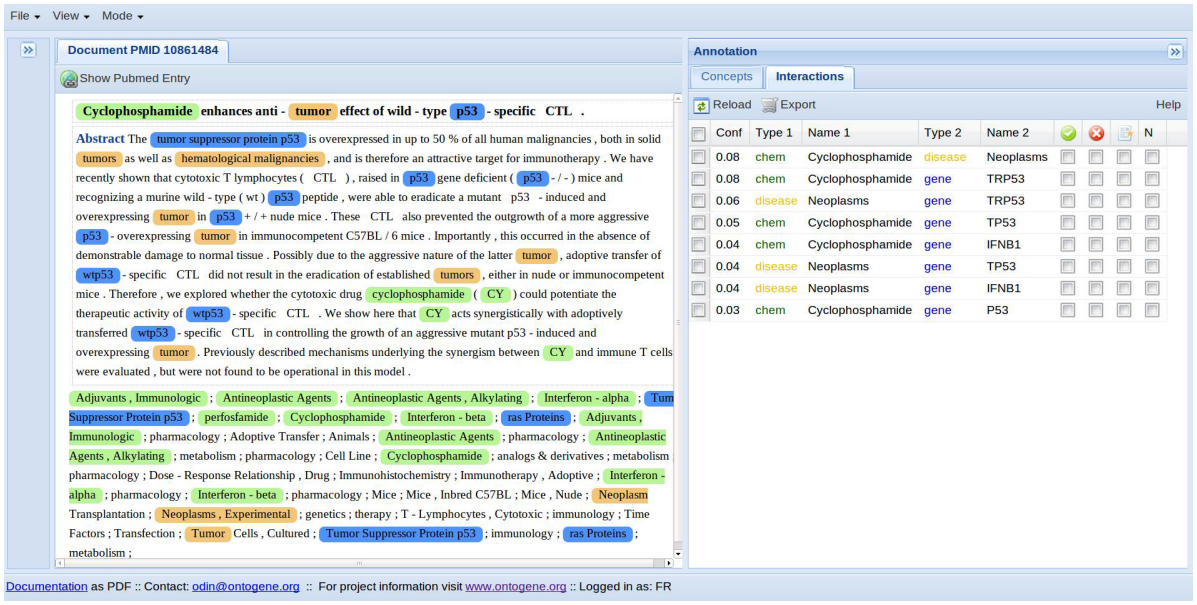

Figure 1. Example of visualization of text mining results using the ODIN interface.

database(s), it is relatively easy to turn its results in a semantic representation, as long as the original databases are based on a standardized ontology. Any term annotation can be turned into a monadic ground fact (possibly using a suitable URI), and interactions can be turned into RDF statements, which could then potentially be integrated across a large collection of documents.

Results and Discussion

Users can submit arbitrary documents to the OntoGene mining service by embedding the text to be mined within a simple XML wrapper. Both input and output of the system are defined according to the BioC standard [Comeau et al., 2013]. However typical usages will involve processing of PubMed abstracts or PubMed Central full papers. In this case the user can provide as input simply the PubMed identifier of the article. Optionally the users can specify which type of output they would like to obtain: if entities, which entity types, and if relationships, which combination of types.

The OntoGene pipeline identifies all relevant entities mentioned in the paper, and their interactions, and reports them back to the user as a ranked list, where the ranking criteria is the system own confidence in the specific result. The confidence value is computed taking into account several factors, including the relative frequency of the term in the article, its general frequency in PubMed, the context in which the term is mentioned, and the syntactic configuration among two interacting entities (for relationships). A detailed description of the factors that contribute to the computation of the confidence score can be found in (Rinaldi et al, 2010).

The user can chose to either inspect the results, using the ODIN web interface (see figure 1), or to have them delivered back via the RESTful web service in BioC XML format, for further processing locally. The usage of ODIN as a curation tool has been tested within the scope of collaborations with curation groups, including PharmGKB, CTD, RegulonDB (Rinaldi, 2012).

The effectiveness of the web service has been recently evaluated within the scope of one of the BioCreative 2013 shared tasks. The official results will be made available at the BioCreative workshop (to be held at the NIH, Bethesda, Maryland, 7-9 October 2013), where only two groups have been invited to present their results, thus showing that the OntoGene/ODIN system is among the top achievers, and will be discussed at the NETTAB workshop when this paper is presented. The system can currently be tested via the ODIN interface[2].

As a future development we envisage the possibility that ODIN could be turned into a tool for collaborative curation of the biomedical literature, with input from the text mining system

2 http://kitt.cl.uzh.ch/kitt/ontogene/bc2013-ctd/

aimed only at facilitating the curation process but not at fully replacing the knowledge of the human experts. It is already possible in ODIN for any user to easily add, remove or modify annotations provided by the system. Such social application could help address the widening gap between the amount of published literature and the capabilities of curation teams to keep abreast with it.

Acknowledgements

The OntoGene group is partially supported by the Swiss National Science Foundation (grants 100014-118396/1 and 105315-130558/1). A continuation of this work is planned within the scope of a collaboration with Roche Pharmaceuticals, Basel, Switzerland.

References

Comeau DC, Islamaj Doğan R, *et al.* (2013) BioC: A Minimalist Approach to Interoperability for Biomedical Text Processing, *Database (Oxford)* **2013**, bat064. doi:10.1093/database/bat064

Richardson L and Sam R (2007), *RESTful Web Services*, O'Reilly, ISBN 978-0-596-52926-0.

Rinaldi F, Kappeler T, *et al.* (2008). OntoGene in BioCreative II. *Genome Biol* **9**:S13. doi:10.1186/gb-2008-9-s2-s13

Rinaldi F, Schneider G, et al. (2010) OntoGene in BioCreative II.5 *IEEE/ACM Trans Comput Biol Bioinform* **7**(3), 472-480. doi:10.1109/TCBB.2010.50

Rinaldi F, Clematide S, et al. (2012) Using ODIN for a PharmGKB revalidation experiment. *Database (Oxford)*, bas021; doi:10.1093/database/bas021

Rinaldi F, Schneider G, and Clematide S. (2012) Relation Mining Experiments in the Pharmacogenomics Domain. *J Biomed Inform.* **45**(5), 851-861. doi:10.1016/j.jbi.2012.04.014