Rinaldi *et al. BMC Bioinformatics* 2014, **15**(Suppl 14):S6
http://www.biomedcentral.com/1471-2105/15/S14/S6

BMC Bioinformatics

RESEARCH Open Access

OntoGene web services for biomedical text mining

Fabio Rinaldi[1*], Simon Clematide[1], Hernani Marques[1], Tilia Ellendorff[1], Martin Romacker[2], Raul Rodriguez-Esteban[2]

From NETTAB 2013: 13th Network Tools and Applications in Biology Workshop on Semantic, Social and Mobile Applications for Bioinformatics and Biomedical Literature
Venice, Italy. 16-18 October 2013

Abstract

Text mining services are rapidly becoming a crucial component of various knowledge management pipelines, for example in the process of database curation, or for exploration and enrichment of biomedical data within the pharmaceutical industry. Traditional architectures, based on monolithic applications, do not offer sufficient flexibility for a wide range of use case scenarios, and therefore open architectures, as provided by web services, are attracting increased interest.

We present an approach towards providing advanced text mining capabilities through web services, using a recently proposed standard for textual data interchange (BioC). The web services leverage a state-of-the-art platform for text mining (OntoGene) which has been tested in several community-organized evaluation challenges, with top ranked results in several of them.

Background

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. Examples of well known biomedical text mining tools are MetaMap [1], MedEvi [2], WhatIzIt [3], Gimli [4], iHOP [5], cTAKES [6], Open Biomedical Annotator [7].

The biomedical text mining community regularly verifies the progress of the field through competitive evaluations, such as BioCreative [8-10], BioNLP [11,12], i2b2 [13], CALBC [14], CLEF-ER [15], DDI [16], BioASQ [17], etc. Each of these competitions targets different aspects of the problem, sometimes with several subtasks, such as detection of mentions of specific entities (e.g. gene and chemicals), detection of protein interactions, assignment of Gene Ontology tags (BioCreative), detection of structured events (BioNLP), information extraction from clinical text (i2b2), large-scale entity detection (CALBC), multilingual entity detection (CLEF-ER), drug-drug interactions (DDI), question answering in biology (BioASQ).

There are numerous institutional attempts to structure some of the knowledge derived from the scientific literature into a more easily accessible format, such as that represented by life science databases. These are typically high-quality resources, where the primary data is accurately (and expensively) extracted from the scientific literature through a process of (mostly manual) curation. There are hundreds of disparate life sciences databases, each of which aims at representing as accurately as possible a particular subdomain. Examples of well-known biomedical databases are UniProt (proteins) [18], Entrez Gene (genes) [19], NCBI Taxonomy (species) [20], IntAct (protein interactions) [21], BioGrid (protein and genetic interactions) [22,23], PharmGKB (drug-gene-disease relations) [24], CTD (chemical-gene-disease relations) [25], and RegulonDB (regulatory interactions in *E. coli*) [26].

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,

* Correspondence: fabio.rinaldi@uzh.ch
[1]Institute of Computational Linguistics, University of Zurich, Binzmuhlestr. 14, 8050 Zurich, Switzerland
Full list of author information is available at the end of the article

diseases, chemicals. OntoGene derives its lexical resources from life sciences databases, thus allowing a deeper connection between the unstructured information contained in the literature and the structured information contained in databases. The quality of the system has been tested several times through participation in some of the community-organized evaluation campaigns, where it often obtained top-ranked results. One of the goals of the OntoGene group is to develop tools which support the process of curation of the biomedical literature, and promote a move towards *assisted curation*. By assisted curation we mean a combination of text mining approaches and the work of an expert curator, aimed at leveraging the power of text mining systems, while retaining the high quality associated with human expertise. We have implemented a platform for assisted curation called ODIN (OntoGene Document INspector) which aims at serving the needs of the curation community. The usage of ODIN as a tool for assisted curation has been tested within the scope of collaborations with curation groups, including PharmGKB [27], CTD [28], RegulonDB [29]. We believe that it is possible to gradually automate much of the most repetitive activities of the curation process, and therefore free up the creative resources of the curators for more challenging tasks, in order to enable a much more efficient and comprehensive curation process.

Assisted curation is also of utility in the process of pharmaceutical drug discovery. Many text mining tasks in drug discovery require both high precision and high recall, due to the importance of comprehensiveness and quality of the output. Text mining algorithms, however, cannot often achieve both high precision and high recall, sacrificing one for the other. Assisted curation can be paired with text mining algorithms which have high recall and moderate precision to produce results that are amenable to answer pharmaceutical problems with only a reasonable effort being allocated to curation.

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 have set up web services which 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, they can be inspected via a flexible web interface. There are strong drivers also in the pharmaceutical industry (for example at Hoffmann-La Roche) for the usage of web services for the annotation of free text. In particular, there is presently a strong trend to dissociate basic functionalities from siloed applications. It is a major advantage to have an open and modularized architecture where services can be combined into larger work-flows. The annotation web services provided by the OntoGene system fit exactly in this philosophy.

Methods

The text mining pipeline which constitutes the core of the OntoGene system has been described previously in a number of publications [30-32]. We will only briefly describe the core text mining technologies, and instead focus mainly on the novel web services which allows remote access to the OntoGene text mining capabilities. One major recent modification, described in this paper, is the integration of a recently proposed standard for textual data interchange (BioC), which will be discussed later in this section.

The OntoGene text mining system contains modules for entity recognition and relation extraction, based on rule-based approaches (e.g. lexical lookup with variants) as well as machine-learning approaches (e.g. maximum entropy techniques). The first step in order to process a collection of biomedical literature consists in the annotation of names of relevant domain entities (currently the system 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.

One of the problems with sourcing resources from several databases is the possible inconsistencies among them. The fact that domain knowledge is scattered across dozens of data sources, occasionally also with some incompatibilities among them, is a severe problem in the life sciences. Ideally these resources should be integrated in a single repository, as some projects are attempting to do (e.g. OpenPhacts [33]), allowing querying within an unified platform. However, a deep integration of the information provided by the scientific literature and the content of the databases is still missing.

We train our system using the knowledge provided by life sciences databases as our gold standard, instead of hand-labeled corpora, since we believe that the scope and size of manually annotated corpora, however much effort has been invested in creating them, is not sufficient to capture the wide variety of linguistic phenomena that can be encountered in the full corpus of biomedical literature, let alone other types of documents, such as internal scientific reports in the pharma industry, which are not represented at all in annotated corpora. For example, PubMed currently contains more than 23 million records, while the entire set of all annotated publications probably barely reaches a few thousand, most of them sparsely annotated for very specific purposes.

Our approach is related to other information extraction research. [34] uses the large semantic knowledge database Freebase for a *distant supervision* approach to optimize information extraction patterns. [35] used the FlyBase database and referenced PubMed abstracts to improve the recognition of gene names by machine learning. [36] presents early work on exploiting biomedical databases as weakly labeled training data to improve relation extraction. Similar techniques have been used very early in the OntoGene system. One example is a version of the system used for the participation of the OntoGene group in the 2006 BioCreative competitive evaluation of text mining systems [37]. For this competition, the OntoGene group generated a training set of positive and negative sentences using techniques of distant supervision. This training set was used for training a classifier able to distinguish between 'background' and 'novel' statements, i.e. sentences reporting previous work as opposed to sentences reporting the actual results generated by the experiment described in the paper. The results of the classifier were used to filter the output of the system and produce only interactions coming from sentences classified as "novel", since the setting of the challenge required the participants to deliver only the most relevant interactions mentioned in the paper. The applied method becomes clear in the following citation:*"A sentence is considered positive if it contains at least one pair of proteins belonging to one of the gold standard interactions for the abstract to which the sentence belongs"* [30]. Furthermore, a similar approach was also applied successfully in the version of the OntoGene system used for participation in the 2009 BioCreative competition [38], which obtained the best results among all participants in the extraction of protein-protein interactions from scientific literature [31].

The specific data source used for the process of training the system through distant supervision depends on the application for which OntoGene/ODIN was customized. For the BioCreative 2009 PPI challenge, the data was sourced from the IntAct database, and was used at various stages in the process: entity recognition, validation of the novel/background statement detection, validation of final PPI. In the more recent applications for the CTD database, CTD entities were used.

In OntoGene, 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. Our normalization rules are similar to the rules reported in [39-41]. This is followed by a disambiguation step which resolves the ambiguity of the matched terms [42]. This process is also used to generate the weakly labeled data for all subsequent training steps based on the distant supervision approach. A marked-up term can be ambiguous for three main reasons. First, a technical term can also be a word of the common language (e.g. a gene called folD, or a chemical called SEX: sodium ethyl xanthate). Although the capitalization in most cases should reveal that it is a special term, and not a common word, there are cases when proper capitalization is not respected. Second, the term can be assigned an ID from different term types. When for example GFP (*Green Fluorescent Protein*) is mentioned in text it is tempting to simply annotate it as an instance of a protein. In fact, most likely it is being mentioned because it is used as a component of an experimental method (fluorescence microscopy) and not for its biological role as a protein. This type of ambiguity is not very common, and we approach it using simple context-based rules that decide on the correct type assignment (e.g. experimental method rather than protein in the case of GFP), in a way similar to what was done in [43]. Third, the term can be assigned several IDs from a single type. This problem is widespread in particular for genes and proteins, which have very ambiguous names, due to the fact that the same gene or protein occurs in many different species. One way to disambiguate such protein names is to apply knowledge about the organisms that are most likely to be the focus of the experiments described in the articles. In [44], we have described an approach to create a ranked list of 'focus' organisms. We use such a list in the disambiguation process by removing all the IDs that do not correspond to one of the organisms detected as relevant for the paper. Additionally, the scores provided for each organism can be used in ranking the candidate IDs for each entity. Such a 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 computing a confidence value for any other derived information, such as interactions where the given entity participates.

Candidate interactions are generated by simple co-occurence of entities within the same syntactic units. However, in order to increase precision, we parse the sentences with our state-of-the-art dependency parser [45], which generates a syntactic representation of the sentence. We have used an existing manually annotated corpus [46] as training corpus for the interaction detection task, based on an approach described in [47]. The corpus was parsed with our dependency parser which has been adapted to and evaluated on the biomedical domain [48,49]. The paths that are extracted from the corpus can directly be used for interaction detection. For example, in the sentence shown in Figure 1, a pattern with the decision 'yes' exists for the relation between *Tim18* and *Tim12*, i.e. the pattern with top node *coimmunoprecipitate*, left path [subj] and right path [pobj]. The details of the algorithm are presented in [31]. The information delivered by the syntactic analysis is used as a factor in order to score and filter candidate interactions based on the syntactic fragment which connects the two participating entities. All available lexical and syntactic information is used in order to

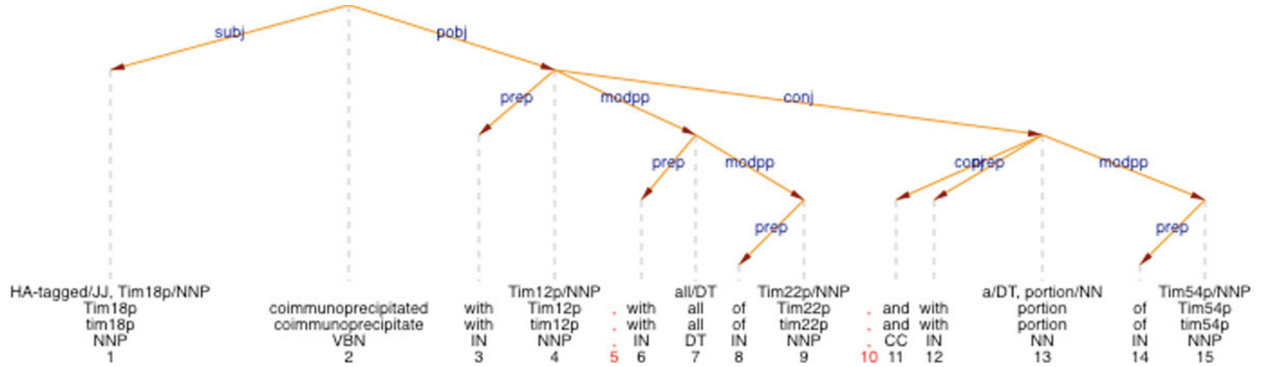


Figure 1 Dependency parser output example. Complete analysis of the sentence "*HA-tagged Tim18p coimmunoprecipitated with Tim12p, with all of Tim22p and with a portion of Tim54p*". The figure shows the linguistic processing performed by OntoGene: tokenization (each word is separated and assigned a unique identifier), part of speech tagging (assignment of grammatical category to each word, e.g. NN for noun), lemmatization (detection of grammatical root form, e.g. *coimmunoprecipitate*), and dependency-based syntactic analysis, represented in the figure by the tree-like orange arrows, which depict the syntactic structure of the sentences, with, for example, Tim18p as subject.

provide an optimized ranking for candidate interactions. The ranking of relation candidates is further optimized by a supervised machine learning method [50]. Since the term recognizer aims at high recall, it introduces several noisy concepts, which we want to automatically identify in order to penalize them. The goal is to identify some global preferences 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 services have been implemented using the RESTful approach [51]. They accept simple XML files as input, based on the BioC specification. The output of the system is generated in the same format. For example, a query aiming at retrieving chemicals and diseases from PubMed abstract 20130422 would generate the output shown in Figure 2.

BioC [52] is a novel XML-based standard for data exchange and software integration specifically targeted at biomedical text mining systems. It was developed as an international collaboration of several leading text mining groups worldwide. BioC has the potential to contribute significantly to the advancement of text mining technologies for the biomedical literature, by enabling data sharing, allowing easier integration of different modules, and overcoming various technical barriers that so far have hindered the effective reuse of several tools contributed by the text mining community. BioC is a simple format for text representation which allows a precise but flexible annotation of entities mentioned in text and their relationships. Different implementations in various programming languages have recently been provided by the text mining community [53,54], which can be integrated as libraries in complex text mining applications, and offer tool developers the capability to deal with BioC annotations in a transparent fashion. Additionally, a significant number of corpora have been converted to BioC format, which will provide a critical mass for training future machine-learning based text mining systems.

BioC is a new annotation standard that has been spearheaded by the National Library of Medicine and researchers in the field of biomedical text mining. At present, the main alternatives to BioC in the biomedical realm are the GATE [55] and UIMA frameworks [56]. Each of these frameworks is built around a set of Java libraries that allows the interconnection of text mining components and the incorporation of external plug-ins, using an interoperability layer, to build text mining pipelines. Both UIMA and GATE, however, require a certain level of software expertise for their deployment and maintenance. BioC, on the other hand, provides a light, almost minimalist, alternative to UIMA and GATE by narrowing the scope of the framework to XML annotation standardization. Other XML annotation guidelines have been created for particular annotation types in the context of community challenges, such as BioCreative and the BioNLP Shared Tasks, or in the context of specific annotation tools. BioC has been created as a general standard that allows the flexibility of multiple types of annotations.

The OntoGene web services offer the user some options which 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 in Figure 2). Currently the system uses pre-defined terminology, and only allows the users to decide whether they want to use one of the pre-loaded vocabularies. However, we foresee in the future the possibility to upload user

```
<?xml version=’1.0’ encoding=’UTF-8’?>
<!DOCTYPE collection SYSTEM ’BioC.dtd’>
<collection>
  <source>PUBMED</source>
  <date>20100202</date>
  <key>ctdBCIVLearningDataSet.key</key>
  <document>
    <id>20130422</id>
    <passage>
      <infon key="type">title</infon>
      <offset>0</offset>
      <text>Chemotherapy resistance as a predictor of progression-free
survival in ovarian cancer patients treated with neoadjuvant
chemotherapy and surgical cytoreduction followed by intraperitoneal
chemotherapy: a Southwest Oncology Group Study.</text>
    </passage>
    <passage>
      <infon key="type">abstract</infon>
      <offset>104</offset>
      <text>In vitro testing of the activity of chemotherapeutic agents
has been suggested as 1 method to optimally select drugs for patients
with ovarian cancer. There are limited prospectively obtained data
examining the clinical utility of this approach. We sought to obtain a
preliminary assessment of this strategy in a trial that examined the
administration of neoadjuvant chemotherapy followed by surgical
cytoreduction and intraperitoneal chemotherapy in women with advanced
ovarian cancer. Women with stage III/IV epithelial ovarian carcinoma
that presented with large-volume disease were treated with neoadjuvant
intravenous paclitaxel and carboplatin for three 21-day cycles followed
by cytoreductive surgery. If optimally debulked, patients received
intravenous paclitaxel, intraperitoneal carboplatin and intraperitoneal
paclitaxel for six 28-day cycles. Tumor cloning assay results (Oncotech)
were correlated with progression-free survival. Sixty-two patients (58
eligible) were registered from March 2001 to February 2006. Thirty-six
eligible patients had interval debulking and 26 received
postcytoreduction chemotherapy. Twenty-two patients had tumor cloning
assay results available. The clinical features of this population were
similar to those of the larger group of women who entered this study.
There was no difference in progression-free survival between patients
whose cancers were defined as ’resistant’ or ’nonresistant’ to either
platinum or paclitaxel. While the small patient numbers in this trial do
not permit definitive conclusions, these data fail to provide support
for the argument that prospectively obtained in vitro data regarding
platinum or paclitaxel resistance will be highly predictive of clinical
outcome in advanced ovarian cancer.</text>
      <annotation>
        <infon key="type">chem</infon>
        <infon key="id">MESH_D016190</infon>
        <text>Paclitaxel</text>
      </annotation>
      <annotation>
        <infon key="type">chem</infon>
        <infon key="id">MESH_D017239</infon>
        <text>Carboplatin</text>
      </annotation>
      <annotation>
        <infon key="type">disease</infon>
        <infon key="id">MESH_D010051</infon>
        <text>Ovarian Neoplasms</text>
      </annotation>
    </passage>
  </document>
</collection>
```

Figure 2 Example of output of the OG text mining service in BioC format. The output of the system is generated in the the BioC specification format. This output was generated by querying for chemicals and diseases on PubMed abstract 20130422. Colors added for clarity. Offsets of annotated terms can be obtained with a separate query.

terminologies. Since the OntoGene system does not only deliver the specific terms found in the submitted articles, but also their unique identifiers in the source database(s), it is relatively easy to turn its results into 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

The OntoGene annotator offers an open architecture allowing for a considerable level of customization so that it is possible to plug in in-house terminologies. Besides the considerations at the level of the architecture, Roche has many different use cases that make use of annotation services. Text mining is applied to different text repositories or data feeds on a regular basis. These resources can be categorized into different classes. Open resources like PubMed or full text from PubMed Central are the starting point. Additionally, agreements with some editors permit full text mining from licensed journals [57]. Furthermore, data feeds which are again open (such as clinical trials) or commercial (like Adis R&D Insight, Pharmaprojects or Pharma Partnering) need to be mined. Finally, internal data - both free text and structured data - needs to considered. Given the variety of textual corpora that have to be dealt with, a toolbox of text mining services allowing customization and tailoring of a text mining pipeline to specific use cases is extremely welcome.

Users can submit arbitrary documents to the OntoGene mining services 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 [52]. However, typical usage involves processing 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 user can specify which type of output they would like to obtain: if entities, which entity types, and if relationships, among which entity types.

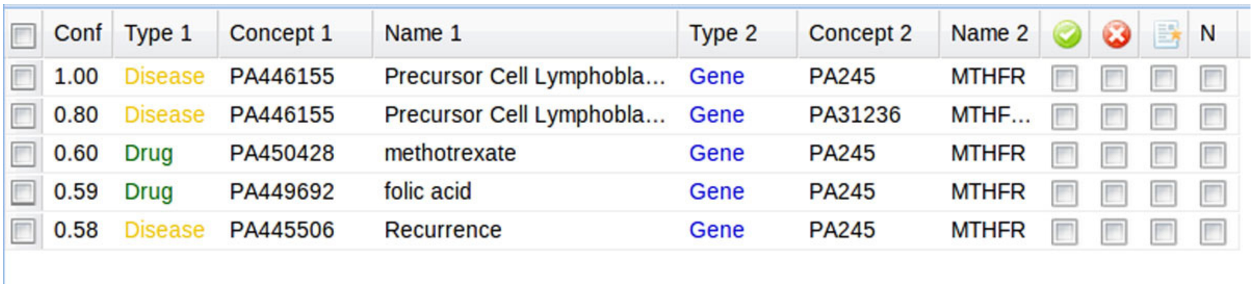
The OntoGene pipeline identifies all relevant entities mentioned in the paper, and their interactions, and reports them back to the user as a ranked list (see Figure 3), where the ranking criteria is the system's own confidence for the specific result. The confidence value is computed taking into account 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 between two interacting entities (for relationships). A detailed description of the factors that contribute to the computation of the confidence score can be found in [31].

The user can choose to either inspect the results, using the ODIN web interface (see Figure 4), or to have them delivered back via the RESTful web services in BioC XML format, for further local processing (see Figure 2). The set of sentences, individually enumerated, contained within those papers can then be viewed through ODIN (OntoGene Document Inspector), a flexible browser-based client application which interfaces with the OntoGene server. The curator can then use the features provided by ODIN to visualize selected annotations, together with the statements from which they were derived, and, if necessary, add, remove or modify them. Once the curator has validated a set of candidate annotations, they can be exported, using a standard format (e.g. CSV, RDF), for further processing by other tools, or for inclusion in a reference database, after a suitable format conversion.

In case of ambiguity, the curator is offered the opportunity to correct the choices made by the system, at any of the different levels of processing: entity identification and disambiguation, organism selection, interaction candidates. The curator can access all the possible readings given by the system and select the most accurate. Candidate interactions are presented in a ranked order, according to the score assigned by the system (see Figure 3). The curator can, for each of them, confirm, reject, or leave undecided. The results of the curation process can be fed back into the system, thus allowing incremental learning.

In order to allow the curators to focus on specific parts of the text that might be most relevant to their curation effort, the OntoGene team recently implemented a new capability within ODIN called "sentence filters", which allows users to specify a simple logical condition that functions as a filter, i.e. only sentences satisfying the specified condition are shown. In a recent experiment in collaboration with RegulonDB [58] such filters proved to be very effective, by allowing the curators to identify, from a subset of sentences constituting in average only 11% of the size of the original articles, all the information about regulatory interactions that they had originally found by reading the entire article set.

As a way to verify the quality of the core text mining functionalities of the OntoGene system, we have participated in a number of text mining evaluation campaigns [59,60,30,61]. Some of our most interesting results include best results in the detection of protein-protein interactions in BioCreative 2009 [31], top-ranked results in several tasks of BioCreative 2010 [62], best results in the triage task of BioCreative 2012 [59]. The usage of ODIN as a curation tool has been tested in collaborations with curation groups, including PharmGKB [63], CTD [27], RegulonDB [64]. Assisted curation is also one of the topics being evaluated at the BioCreative competitions [9,8], where OntoGene/ODIN participated with favorable results. The effectiveness of the web services has been recently evaluated within the scope of one of the BioCreative 2013 shared tasks [65]. Although different implementations can rapidly be produced upon request, a version of the system tailored for the



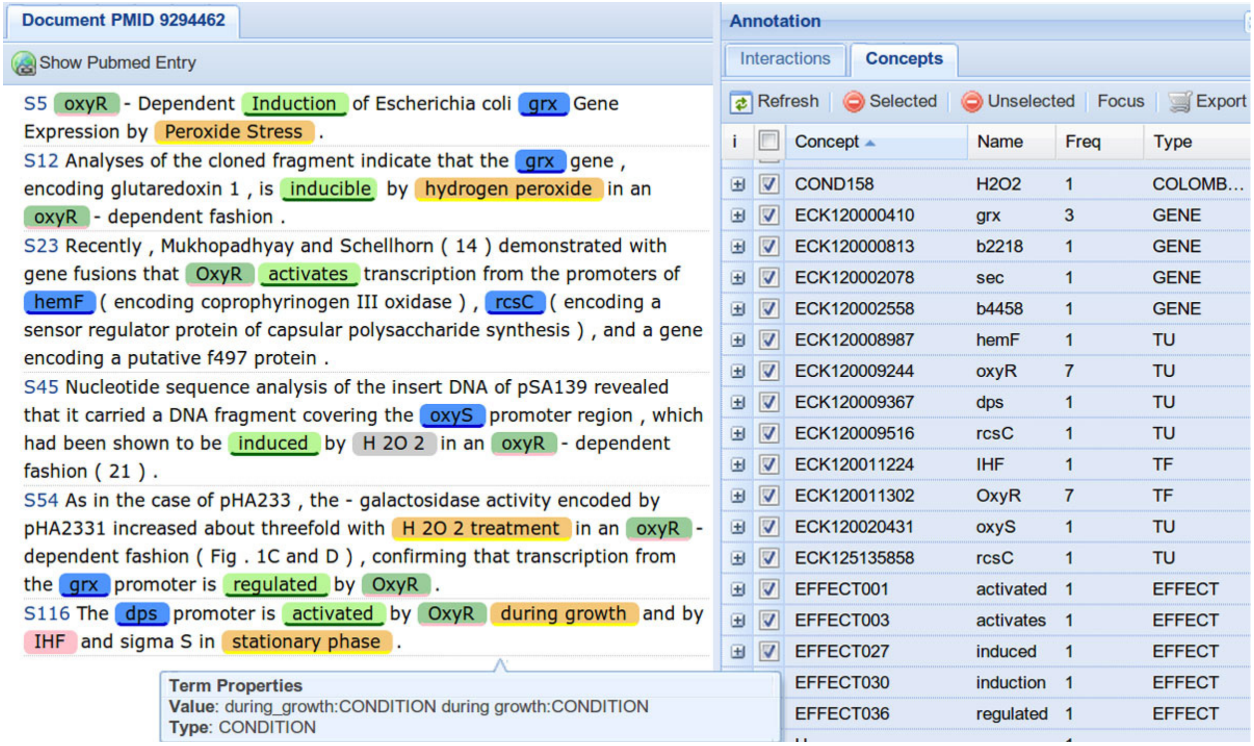
☐	Conf	Type 1	Concept 1	Name 1	Type 2	Concept 2	Name 2	✓	✗		N
☐	1.00	Disease	PA446155	Precursor Cell Lymphobla...	Gene	PA245	MTHFR	☐	☐	☐	☐
☐	0.80	Disease	PA446155	Precursor Cell Lymphobla...	Gene	PA31236	MTHF...	☐	☐	☐	☐
☐	0.60	Drug	PA450428	methotrexate	Gene	PA245	MTHFR	☐	☐	☐	☐
☐	0.59	Drug	PA449692	folic acid	Gene	PA245	MTHFR	☐	☐	☐	☐
☐	0.58	Disease	PA445506	Recurrence	Gene	PA245	MTHFR	☐	☐	☐	☐

Figure 3 ODIN interactions panel. Example of interactions delivered by OntoGene as shown by ODIN. The confidence value for each candidate interaction is shown on the left. The two actors in the interactions are provided with their unique identifier, type, and reference name. The columns on the right allow the user to inspect and validate each candidate interaction.

Comparative Toxicogenomics Database (CTD) can currently be tested via the ODIN interface at the following URL:

http://www.ontogene.org/webservices/.

Since internally the original database identifiers are used to represent the entities and interactions detected by the system, the annotations can be easily converted into a semantic web format, by using a reference URI for each domain entity, and using RDF statements to express interactions. While it is possible to access the automatically generated annotations for further processing by a reasoner or integrator tool, we strongly believe that at present a process of semi-automated validation is preferable and would lead to better data consistency. A tool such as ODIN, which is currently used by database annotators for practical curation tasks, could also be put in the hands of authors which would be able to confirm or reject candidate annotations suggested by the system, thus allowing the construction of an enhanced document representation, integrating human readable text with machine readable semantic web statements.



S5 oxyR - Dependent Induction of Escherichia coli grx Gene Expression by Peroxide Stress .

S12 Analyses of the cloned fragment indicate that the grx gene , encoding glutaredoxin 1 , is inducible by hydrogen peroxide in an oxyR - dependent fashion .

S23 Recently , Mukhopadhyay and Schellhorn (14) demonstrated with gene fusions that OxyR activates transcription from the promoters of hemF (encoding coprophyrinogen III oxidase) , rcsC (encoding a sensor regulator protein of capsular polysaccharide synthesis) , and a gene encoding a putative f497 protein .

S45 Nucleotide sequence analysis of the insert DNA of pSA139 revealed that it carried a DNA fragment covering the oxyS promoter region , which had been shown to be induced by H 2O 2 in an oxyR - dependent fashion (21) .

S54 As in the case of pHA233 , the - galactosidase activity encoded by pHA2331 increased about threefold with H 2O 2 treatment in an oxyR - dependent fashion (Fig . 1C and D) , confirming that transcription from the grx promoter is regulated by OxyR .

S116 The dps promoter is activated by OxyR during growth and by IHF and sigma S in stationary phase .

Term Properties
Value: during_growth:CONDITION during growth:CONDITION
Type: CONDITION

i	☐	Concept ▲	Name	Freq	Type
⊞	☑	COND158	H2O2	1	COLOMB...
⊞	☑	ECK120000410	grx	3	GENE
⊞	☑	ECK120000813	b2218	1	GENE
⊞	☑	ECK120002078	sec	1	GENE
⊞	☑	ECK120002558	b4458	1	GENE
⊞	☑	ECK120008987	hemF	1	TU
⊞	☑	ECK120009244	oxyR	7	TU
⊞	☑	ECK120009367	dps	1	TU
⊞	☑	ECK120009516	rcsC	1	TU
⊞	☑	ECK120011224	IHF	1	TF
⊞	☑	ECK120011302	OxyR	7	TF
⊞	☑	ECK120020431	oxyS	1	TU
⊞	☑	ECK125135858	rcsC	1	TU
⊞	☑	EFFECT001	activated	1	EFFECT
⊞	☑	EFFECT003	activates	1	EFFECT
⊞	☑	EFFECT027	induced	1	EFFECT
		EFFECT030	induction	1	EFFECT
		EFFECT036	regulated	1	EFFECT

Figure 4 ODIN screenshot. Example of visualization of text mining results using the ODIN interface. The panel on the left shows the document with annotations, the panel on the right the corresponding concepts. The two panels are interconnected by the interface logic: whenever an item is selected in the concept panel, the corresponding terms are highlighted in the document panel.

The documents and the annotations are represented consistently within a single XML file, which also contains a detailed record log of the user interaction, which allows (if requested by the specific application) advanced analysis of user actions. The annotations are selectively presented, in an ergonomic way through CSS formatting, according to different view modalities. While the XML annotations are transparent to the annotator (who therefore does not need to have any specialized knowledge beyond biological expertise), his/her verification activities result in changes at the DOM of the XML document through client-side JavaScript. The use of modern AJAX technology allows for online integration of background information, e.g. information from different term and knowledge bases, or further integration of foreign text mining services.

The presence of the raw XML in the browser document gives the flexibility to compile dynamically tabular grid views of terms and relations including filtering, reordering, and editing the annotations in a spreadsheet-like way (this includes also chart visualizations). To keep the implementation effort feasible, the use of a dedicated JavaScript application framework is crucial. The advantage of a client-side presentation logic is the flexibility for the end user and the data transparency. For text mining applications, it is important to be able to link back curated metainformation to its textual evidence.

Conclusion

We have presented novel web services which aim at making the text mining capabilities of the OntoGene system more easily accessible for a variety of potential users. OntoGene is a mature system which has been shown to perform at state-of-the-art level in a variety of text mining tasks. Open and modular web services, instead of monolithic applications, offer more flexibility to the end users, and are therefore a major advantage. However, such applications need to rely on simple and versatile standards for data exchange. We have discussed one of such standards, BioC, which is used within our text mining services.

Our system relies on dynamic adaptation of terminological resources from databases and ontologies, for named entity recognition and relation extraction. This approach can benefit from the large amount of knowledge present in curated biomedical databases to optimize and/or specialize text mining systems. Additionally, through the ODIN platform, we offer a sophisticated combination of automatic text mining and human curation and validation.

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 aimed only at facilitating the curation process but not at fully replacing the knowledge of the human experts. Such social application could help address the widening gap between the amount of published literature and the capabilities of curation teams to keep abreast of it.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

FR conceived the work described in this paper and implemented parts of the system. SC implemented the initial versions of ODIN, and was responsible for the machine-learning approach aimed at optimizing ranking of the results. HM implemented the RESTful interface for the OntoGene web services. TE implemented some of the named entity recognition capabilities of the system. MR and RRE provided logistical support, use-case scenarios, and tested ODIN and the OntoGene web services.

Acknowledgements

The OntoGene group (FR, SC, HM, TE) is partially supported by the Swiss National Science Foundation (grant 105315 – 130558/1 to Fabio Rinaldi) and by the Data Science Group at Hoffmann-La Roche, Basel, Switzerland.
This article has been published as part of *BMC Bioinformatics* Volume 15 Supplement 14, 2014: Semantic, social, mobile applications for bioinformatics and biomedical laboratories: selected works from the 13th International Workshop on Network Tools and Applications in Biology (NETTAB 2013). The full contents of the supplement are available online at http://www.biomedcentral.com/bmcbioinformatics/supplements/15/S14.

Authors' details

[1]Institute of Computational Linguistics, University of Zurich, Binzmuhlestr. 14, 8050 Zurich, Switzerland. [2]pREDi (Pharma Research and Early Development Informatics), F. Hoffmann-La Roche Ltd, Grenzacherstrasse 124, 4070 Basel, Switzerland.

Published: 27 November 2014

References

1. Aronson AR, Lang FM: **An overview of MetaMap: historical perspective and recent advances.** *J Am Med Inform Assoc* 2010, **17(3)**:229-236.
2. Kim J, Pezik P, Rebholz-Schuhmann D: **Medevi: Retrieving textual evidence of relations between biomedical concepts from medline.** *Bioinformatics* 2008, **24(11)**:1410-1412.
3. Rebholz-Schuhmann D, Arregui M, Gaudan S, Kirsch H, Jimeno A: **Text processing through Web services: calling Whatizit.** *Bioinformatics* 2008, **24(2)**:296-298.
4. Campos D, Matos S, Oliveira JL: **Gimli: open source and high-performance biomedical name recognition.** *BMC Bioinformatics* 2013, **14**:54.
5. Hoffmann R: **Using the iHOP information resource to mine the biomedical literature on genes, proteins, and chemical compounds.** *Curr Protoc Bioinformatics* 2007, **Chapter 1**:1-16.
6. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, Chute CG: **Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications.** *J Am Med Inform Assoc* 2010, **17(5)**:507-513.
7. Jonquet C, Shah NH, Musen MA: **The open biomedical annotator.** *Summit on Translat Bioinforma* 2009, **2009**:56-60.
8. Arighi C, Roberts P, Agarwal S, Bhattacharya S, Cesareni G, Chatr-aryamontri A, Clematide S, Gaudet P, Giglio M, Harrow I, Huala E, Krallinger M, Leser U, Li D, Liu F, Lu Z, Maltais L, Okazaki N, Perfetto L, Rinaldi F, Saetre R, Salgado D, Srinivasan P, Thomas P, Toldo L, Hirschman L, Wu C: **BioCreative III interactive task: an overview.** *BMC Bioinformatics* 2011, **12(Suppl 8)**:4.
9. Arighi CN, Carterette B, Cohen KB, Krallinger M, Wilbur WJ, Fey P, Dodson R, Cooper L, Van Slyke CE, Dahdul W, Mabee P, Li D, Harris B, Gillespie M, Jimenez S, Roberts P, Matthews L, Becker K, Drabkin H, Bello S, Licata L, Chatr-aryamontri A, Schaeffer ML, Park J, Haendel M, Van Auken K, Li Y,

Chan J, Muller HM, Cui H, Balhoff JP, Chi-Yang Wu J, Lu Z, Wei CH, Tudor CO, Raja K, Subramani S, Natarajan J, Cejuela JM, Dubey P, Wu C: **An overview of the BioCreative 2012 workshop track III: interactive text mining task.** *Database* 2013, **2013.**

10. Krallinger M, Vazquez M, Leitner F, Salgado D, Chatr-aryamontri A, Winter A, Perfetto L, Briganti L, Licata L, Iannuccelli M, Castagnoli L, Cesareni G, Tyers M, Schneider G, Rinaldi F, Leaman R, Gonzalez G, Matos S, Kim S, Wilbur W, Rocha L, Shatkay H, Tendulkar A, Agarwal S, Liu F, Wang X, Rak R, Noto K, Elkan C, Lu Z, Dogan R, Fontaine JF, Andrade-Navarro M, Valencia A: **The protein-protein interaction tasks of BioCreative III: classification/ranking of articles and linking bio-ontology concepts to full text.** *BMC Bioinformatics* 2011, **12(Suppl 8)**:3.
11. Kim JD, Pyysalo S, Ohta T, Bossy R, Nguyen N, Tsujii J: **Overview of BioNLP shared task 2011.** *ACL HLT 2011* 2011, 1.
12. Cohen KB, Demner-Fushman D, Ananiadou S, Pestian J, Tsujii J, Webber B: **Proceedings of the BioNLP 2009 Workshop.** Association for Computational Linguistics, Boulder, Colorado; 2009 [http://www.aclweb.org/anthology/W09-13].
13. Sun W, Rumshisky A, Uzuner O: **Evaluating temporal relations in clinical text: 2012 i2b2 Challenge.** *J Am Med Inform Assoc* 2013, **20(5)**:806-813.
14. Rebholz-Schuhmann D, Yepes A, Li C, Kafkas S, Lewin I, Kang N, Corbett P, Milward D, Buyko E, Beisswanger E, Hornbostel K, Kouznetsov A, Witte R, Laurila J, Baker C, Kuo CJ, Clematide S, Rinaldi F, Farkas R, Mora G, Hara K, Furlong LI, Rautschka M, Neves M, Pascual-Montano A, Wei Q, Collier N, Chowdhury M, Lavelli A, Berlanga R, Morante R, Van Asch V, Daelemans W, Marina J, van Mulligen E, Kors J, Hahn U: **Assessment of NER solutions against the first and second CALBC silver standard corpus.** *Journal of Biomedical Semantics* 2011, **2(Suppl 5)**:11.
15. Rebholz-Schuhmann D, Clematide S, Rinaldi F, Kafkas S, van Mulligen EM, Bui C, Hellrich J, Lewin I, Milward D, Poprat M, Jimeno-Yepes A, Hahn U, Kors J: **Entity recognition in parallel multi-lingual biomedical corpora: The CLEF-ER laboratory overview.** In *Information Access Evaluation Multilinguality, Multimodality, and Visualization Lecture Notes in Computer Science.* Springer, Valencia;Forner, P., Mueller, H., Rosso, P., Paredes, R 2013:353-367[http://www.zora.uzh.ch/82216/].
16. Segura-Bedmar I, Martínez P, Sánchez-Cisneros D: **The 1st ddi extraction-2011 challenge task: Extraction of drug-drug interactions from biomedical texts.** *Proc DDI Extraction-2011 Challenge Task* Huelva, Spain; 2011, 1-9.
17. Androutsopoulos I: **A Challenge on Large-scale Biomedical Semantic Indexing and Question Answering.** *BioNLP Workshop (part of the ACL Conference)* 2013 [http://www.bioasq.org/workshop/programme], presentation.pdf.
18. Consortium T: **The universal protein resource (UniProt) in 2010.** *Nucleic Acids Research* 2010, **38(suppl 1)**:142-148.
19. Maglott D, Ostell J, Pruitt KD, Tatusova T: **Entrez Gene: gene-centered information at NCBI.** *Nucleic Acids Res* 2011, **39(Database)**:52-57.
20. Federhen S: **The NCBI Taxonomy database.** *Nucleic Acids Res* 2012, **40(Database)**:136-143.
21. Hermjakob H, Montecchi-Palazzi L, Lewington C, Mudali S, Kerrien S, Orchard S, Vingron M, Roechert B, Roepstorff P, Valencia A, Margalit H, Armstrong J, Bairoch A, Cesareni G, Sherman D, Apweiler R: **IntAct: an open source molecular interaction database.** *Nucl Acids Res* 2004, **32(suppl 1)**:452-455.
22. Dolinski K, Chatr-Aryamontri A, Tyers M: **Systematic curation of protein and genetic interaction data for computable biology.** *BMC Biol* 2013, **11**:43.
23. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M: **Biogrid: A general repository for interaction datasets.** *Nucleic Acids Research* 2006, **34**:535-9.
24. Sangkuhl K, Berlin DS, Altman RB, Klein TE: **PharmGKB: Understanding the effects of individual genetic variants.** *Drug Metabolism Reviews* 2008, **40(4)**:539-551.
25. Davis A, King B, Mockus S, Murphy C, Saraceni-Richards C, Rosenstein M, Wiegers T, Mattingly C: **The comparative toxicogenomics database: update 2011.** *Nucleic Acids Res* 2011, **39(Database)**:1067-72.
26. Gama-Castro S, Salgado H, Peralta-Gil M, Santos-Zavaleta A, Muniz-Rascado L, Solano-Lira H, Jimenez-Jacinto V, Weiss V, Garcia-Sotelo JS, Lopez-Fuentes A, Porron-Sotelo L, Alquicira-Hernandez S, Medina-Rivera A, Martinez-Flores I, Alquicira-Hernandez K, Martinez-Adame R, Bonavides-Martinez C, Miranda-Rios J, Huerta AM, Mendoza-Vargas A, Collado-Torres L, Taboada B, Vega-Alvarado L, Olvera M, Olvera L, Grande R, Morett E, Collado-Vides J: **RegulonDB version 7.0: transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units).** *Nucleic Acids Res* 2011, **39(Database)**:98-105.
27. Rinaldi F, Clematide S, Garten Y, Whirl-Carrillo M, Gong L, Hebert JM, Sangkuhl K, Thorn CF, Klein TE, Altman RB: **Using ODIN for a PharmGKB re-validation experiment.** *Database: The Journal of Biological Databases and Curation* 2012.
28. Rinaldi F, Clematide S, Hafner S: **Ranking of CTD articles and interactions using the OntoGene pipeline.** *Proceedings of the 2012 BioCreative Workshop* Washington D.C; 2012.
29. Gama-Castro S, Rinaldi F, López-Fuentes A, Balderas-Martínez YI, Clematide S, Ellendorff TR, Collado-Vides J: **Assisted curation of growth conditions that affect gene expression in e. coli k-12.** *Proceedings of the Fourth BioCreative Challenge Evaluation Workshop* 2013, **1**:214-218.
30. Rinaldi F, Kappeler T, Kaljurand K, Schneider G, Klenner M, Clematide S, Hess M, von Allmen JM, Parisot P, Romacker M, Vachon T: **OntoGene in BioCreative II.** *Genome Biology* 2008, **9(Suppl 2)**:13.
31. Rinaldi F, Schneider G, Kaljurand K, Clematide S, Vachon T, Romacker M: **OntoGene in BioCreative II.5.** *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 2010, **7(3)**:472-480.
32. Rinaldi F, Schneider G, Clematide S: **Relation mining experiments in the pharmacogenomics domain.** *Journal of Biomedical Informatics* 2012, **45(5)**:851-861.
33. Williams AJ, Harland L, Groth P, Pettifer S, Chichester C, Willighagen EL, Evelo CT, Blomberg N, Ecker G, Goble C, Mons B: **Open PHACTS: semantic interoperability for drug discovery.** *Drug Discovery Today* 2012, **17(21-22)**:1188-1198.
34. Mintz M, Bills S, Snow R, Jurafsky D: **Distant supervision for relation extraction without labeled data.** *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP Volume 2 ACLIJCNLP 09* 2009, **2(2005)**:1003.
35. Morgan AA, Hirschman L, Colosimo M, Yeh AS, Colombe JB: **Gene name identification and normalization using a model organism database.** *Journal of Biomedical Informatics* 2004, **37(6)**:396-410.
36. Craven M, Kumlien J: **Constructing biological knowledge bases by extracting information from text sources.** *Proceedings International Conference on Intelligent Systems for Molecular Biology* 1999, 77-86.
37. Krallinger M, Leitner F, Rodriguez-Penagos C, Valencia A: **Overview of the protein-protein interaction annotation extraction task of BioCreative II.** *Genome Biology* 2008, **9(Suppl 2)**:4.
38. Leitner F, Mardis SA, Krallinger M, Cesareni G, Hirschman LA, Valencia A: **An overview of BioCreative II.5.** *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 2010, **7(3)**:385-399.
39. Hakenberg J: **What's in a gene name? Automated refinement of gene name dictionaries.** *Proceedings of BioNLP 2007: Biological, Translational, and Clinical Language Processing; Prague, Czech Republic* 2007, 153-160 [http://www.aclweb.org/anthology-new/W/W07/W07-1020.pdf].
40. Hakenberg J, Plake C, Royer L, Strobelt H, Leser U, Schroeder M: **Gene mention normalization and interaction extraction with context models and sentence motifs.** *Genome Biol* 2008, **9(Suppl 2)**:14.
41. Wang X, Matthews M: **Distinguishing the species of biomedical named entities for term identification.** *BMC Bioinformatics* 2008, **9(Suppl 11)**:6.
42. Kaljurand K, Rinaldi F, Kappeler T, Schneider G: **Using existing biomedical resources to detect and ground terms in biomedical literature.** *Proceedings of the 12th Conference on Artificial Intelligence in Medicine (AIME09)* 2009, 225-234.
43. Tanabe L, Wilbur W: **Tagging gene and protein names in biomedical text.** *bioinformatics* 2002, **18(8)**:1124-32.
44. Kappeler T, Kaljurand K, Rinaldi F: **TX Task: Automatic Detection of Focus Organisms in Biomedical Publications.** *Proceedings of the BioNLP Workshop* Boulder, Colorado; 2009, 80-88.
45. Schneider G: **Combining shallow and deep processing for a robust, fast, deep-linguistic dependency parser.** In *ESSLLI 2004 Workshop on Combining Shallow and Deep Processingfor NLP.* Nancy, France; Hinrichs, E., Simov, K 2004:41-50.
46. Kim JD, Ohta T, Tateisi Y, Tsujii J: **GENIA corpus-semantically annotated corpus for bio-textmining.** *Bioinformatics* 2003, **19(Suppl 1)**:180-182.
47. Schneider G, Kaljurand K, Kappeler T, Rinaldi F: **Detecting Protein/Protein Interactions using a parser and linguistic resources.** *Proceedings of CICLing*

2009, 10th International Conference on Intelligent Text Processing and Computational Linguistics Springer, Mexico City, Mexico; 2009, 406-417.

48. Schneider G, Kaljurand K, Rinaldi F, Kuhn T: **Pro3Gres parser in the CoNLL domain adaptation shared task.** *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007* Prague; 2007, 1161-1165 [http://www.aclweb.org/anthology/D07-1128].
49. Haverinen K, Ginter F, Pyysalo S, Salakoski T: **Accurate conversion of dependency parses: targeting the Stanford scheme.** *Proceedings of Third International Symposium on Semantic Mining in Biomedicine (SMBM 2008)* Turku, Finland; 2008.
50. Clematide S, Rinaldi F: **Ranking relations between diseases, drugs and genes for a curation task.** *Journal of Biomedical Semantics* 2012, **3(Suppl 3)**:5.
51. Richardson L, Ruby S: **RESTful Web Services.** O'Reilly, Sebastopol, California; 2007, ISBN 978-0-596-52926-0.
52. Comeau DC, Doğan RI, Ciccarese P, Cohen KB, Krallinger M, Leitner F, Lu Z, Peng Y, Rinaldi F, Torii M, Valencia A, Verspoor K, Wiegers TC, Wu CH, Wilbur WJ: **Bioc: a minimalist approach to interoperability for biomedical text processing.** *The Journal of Biological Databases and Curation bat064* 2013.
53. Liu W, Comeau DC, Doğan RI, Kwon D, Marques H, Rinaldi F, Wilbur WJ: **Bioc implementations in go, perl, python and ruby.** *Database: The Journal of Biological Databases and Curation* 2014, under review.
54. Rinaldi F, Marques H: **PyBioC: a Python implementation of the BioC core.** *Proceedings of the Fourth BioCreative Challenge Evaluation Workshop* 2013, **1**:2-4.
55. Cunningham H, Tablan V, Roberts A, K B: **Getting more out of biomedical documents with gate's full lifecycle open source text analytics.** *PLoS Comput Biol* 2013, **9(2)**:1002854.
56. Ferrucci D, Lally A: **Building an example application with the unstructured information management architecture.** *IBM Systems Journal* 2004, **43(3)**:455-475, 2004.
57. Noorden RV: **Elsevier opens its papers to text-mining.** *Nature* 2014, **506(17)**.
58. Gama-Castro S, Rinaldi F, López-Fuentes A, Balderas-Martínez YI, Clematide S, Ellendorff TR, Santos-Zavaleta A, Marques-Madeira H, Collado-Vides J: **Assisted curation of regulatory interactions and growth conditions of OxyR in E. coli K-12.** *Database: The Journal of Biological Databases and Curation bau049* 2014.
59. Rinaldi F, Clematide S, Hafner S, Schneider G, Grigonyte G, Romacker M, Vachon T: **Using the OntoGene pipeline for the triage task of BioCreative 2012.** *The Journal of Biological Databases and Curation, Oxford Journals* 2013.
60. Clematide S, Rinaldi F, Schneider G: **OntoGene at CALBC II and some thoughts on the need of document-wide harmonization.** *Proceedings of the CALBC II Workshop* EBI, Cambridge, UK; 2011, 16-18, March.
61. Rinaldi F, Kappeler T, Kaljurand K, Schneider G, Klenner M, Hess M, von Allmen JM, Romacker M, Vachon T: **OntoGene in Biocreative II.** *Proceedings of the II Biocreative Workshop* 2007.
62. Schneider G, Clematide S, Rinaldi F: **Detection of interaction articles and experimental methods in biomedical literature.** *BMC Bioinformatics* 2011, **12(Suppl 8)**:13.
63. Rinaldi F, Clematide S, Schneider G, Romacker M, Vachon T: **ODIN: An advanced interface for the curation of biomedical literature.** *Biocuration 2010, the Conference of the International Society for Biocuration and the 4th International Biocuration Conference* 2010, 61, Available from Nature Precedings http://dx.doi.org/10.1038/npre.2010.5169.1.
64. Rinaldi F, Gama-Castro S, López-Fuentes A, Balderas-Martínez Y, Collado-Vides J: **Digital curation experiments for regulondb.** *BioCuration 2013, April 10th* Cambridge, UK; 2013.
65. Rinaldi F, Clematide S, Ellendorff TR, Marques H: **OntoGene: CTD entity and action term recognition.** *Proceedings of the Fourth BioCreative Challenge Evaluation Workshop* 2013, **1**:90-94.

doi:10.1186/1471-2105-15-S14-S6
Cite this article as: Rinaldi *et al.*: OntoGene web services for biomedical text mining. *BMC Bioinformatics* 2014 **15**(Suppl 14):S6.