Large-scale information extraction for assisted curation of the biomedical literature

Fabio Rinaldi
fabio.rinaldi@uzh.ch

Institute of Computational Linguistics,
University of Zurich, Switzerland

Abstract. PubMed, the main literature repository for the life sciences, contains more than 23 million publication references. In average nearly two publications per minute are added. There is a wealth of knowledge hidden in unstructed format in these publications that needs to be structured, linked, and semantically annotated so that it becomes actionable knowledge.

We present an approach towards large-scale processing of biomedical literature in order to extract domain entities and semantic relationships among them. We describe some practical applications of the resulting knowledge base.

1 Introduction

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. Recent comprehensive reviews of the field are [25, 15]. Biomedical text mining involves different levels of document processing: document classification, document structure recognition (zoning), domain entity recognition and disambiguation detection of relations, to name just a few.

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 [3], MedEvi [14], WhatIzIt [18], ChilliBot [7], Gimli [5], iHOP[1] [12, 11], Open Biomedical Annotator [13], AliBaba [16], GOPubMed [9], GeneView[2] [30]. Some of the most commonly used frameworks for the development of text mining systems include IBM LanguageWare, the Natural Language Toolkit (NLTK), the GATE system (General Architecture for Text Engineering) and IBM's UIMA (Unstructured Information Management Architecture).

The biomedical text mining community regularly verifies the progress of the field through competitive evaluations, such as BioCreative [2], BioNLP [17], i2b2 [29], CALBC [20], CLEF-ER [19], DDI [27], BioASQ [1], 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. genes and chemicals),

[1] http://ws.bioinfo.cnio.es/iHOP/

[2] http://bc3.informatik.hu-berlin.de/

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).

Evidence in support of relationships among biomedical entities, such as protein-protein interactions, can be gathered from a multiplicity of sources. The larger the pool of evidence, the more likely a given interaction can be considered to be. In the context of biomedical text mining, this elementary observation can be translated into an approach that seeks to find in the literature all available evidence for a given interaction, and thus provides a reliable means to assign it a likelihood score before delivering the results to an end user.

In this paper we present the results of an on-going collaborative project between a major pharmaceutical company and an academic group with extensive expertise in biomedical text mining, with the initial goal of extracting protein-protein interactions from a large pool of supporting papers, later to be extended to different entity relationships.

The OntoGene group[3] at the University of Zurich (UZH) specializes in mining the scientific literature for evidence of interactions among entities of relevance for biomedical research (genes, proteins, drugs, diseases, chemicals). The quality of the text mining tools developed by the group is demonstrated by top-ranked results achieved at several community-organized text mining competitions [22, 24, 21]. As part of a project funded by a large pharmaceutical company, the OntoGene group recently adapted their text mining, with the goal of detecting evidence for specific protein interactions described in the input documents. Given an input gene or protein, the system locates all interactions of that gene/protein and present them as a ranked list, with evidence coming from all papers where they are mentioned. The interface is structured in a way that allows easy inspection of the original evidence from the publications for any candidate interaction suggested by the system. The ranking computed by the system takes into consideration not only the local evidence in each paper, but also the global evidence across the collection. In summary, the system has the following capabilities:

1. identify all interactions in which a given protein is involved
2. rank them based on evidence in the literature
3. enable curation by an end user through a user-friendly interface

In the rest of this paper we describe the methods that were used in the development of the system (sec. 2), then we briefly report the results of an evaluation (sec. 3), and finally we focus specifically on some applications (sec. 4).

2 Methods

The OntoGene group developed in recent projects an advanced text mining pipeline which is used to provide all the basic text mining capabilities that are

[3] `http://www.ontogene.org/`

needed for the successful realization of the activities described in this paper. Our text mining system has been evaluated in several community-organized competitive evaluation tasks and always shown to perform at state-of-the-art levels, for example obtaining best results in the annotation of experimental methods [22], in the annotation of protein-protein interactions [24], and in the discovery of several other entity types [21].

The system is based on sourcing named entities (terms and identifiers) from one or several reference databases, and use an approximate matching approach to annotate the target collection, typically with high recall. In a second step, a machine learning approach is used to disambiguate the annotations and remove some of the false positives. In a third step, candidate interactions among the detected entities are generated and scored against a target database, using a machine learning approach in order to provide an optimized ranking.

In order to provide a specific application of the system, a preliminary decision to be taken is which reference database (or databases) should be used for sourcing the terminology, and as a reference in training the interaction scoring module. For the work described in this paper we have selected the BioGrid database[4] [8] for its good coverage and quality curation.

The terminology derived from the entire BioGrid database is stored as an internal lexical resource, and used to annotate the target documents. The system will efficiently recognize all entity names from this resource, and associate them to their database identifiers. It also takes into consideration several possible variants of the input terms (e. g. removal of hyphenation), but if a completely different form is used in a paper, not derivable from any term seen in the database, this will be missed. Applying our term mapping strategy we only have a minor loss in recall. The result of the entity annotation phase is a richly annotated version of the original document, in an XML format which can be inspected with a customized interface, which we describe later in this paper.

Additionally, the pipeline will produce a list of all `(term, identifier)` pairs seen in the document, and it is this list that will be used to generate candidate interactions, by initially considering all possible combinations of the identifiers, and scoring them using information from the original database. Once a score is produced, the best candidate interactions are selected, and for every established relation, we extract the n best text snippets that represent evidence for this protein-protein interaction (PPI).

Initially we worked with a collection of 20,928 PubMed abstracts, selected from those containing interactions that satisfy the conditions described above (must contain 1 to 12 curated relations with human proteins with physical experiments in BioGrid). This collection is then split into a training and test set using a 10-fold validation approach. The information that we use from each abstract is: title, abstract text, MeSH terms and Chemical Substance list. The abstracts have an average length of approximately 300 words. Later we extended the experiment to include all PubMed abstracts that contain mentions of proteins (about 3.5 million).

[4] `http://thebiogrid.org/`

We use a distant-learning approach to train and evaluate the ranking of extracted protein relations per article. Given an article, we expect our system to find and rank highest all pairings of two entities that are part of a curated interaction for this article.

Among the recognized protein concepts in a document, we look at all combinatorial pairs and assign each of them a score, which expresses the likelihood that it is a relevant protein-protein interaction. By ranking the protein pairs by this score, we produce a list of candidate interactions with decreasing confidence.

Since the entity recognition phase of our system is recall-oriented, it introduces numerous false positives that need to be weeded out in a later phase. This is even aggravated when moving to relation extraction, since the error is squared: For example, if 90 % of the extracted proteins are accurate, only 81 % of all protein pairs potentially represent a relevant protein interaction. Therefore, a powerful ranking method is essential for relation extraction, so that the best protein pairs are brought to the top.

The score for each candidate interaction is computed from the scores of the individual proteins, and from a context-based score for the shared sentences in which these proteins are mentioned. The individual score for each protein concept (henceforth *concept score*) expresses the probability of a concept to participate in an interaction, given its surface form and its position inside the document (e. g. in the title). Using a Maximum Entropy (ME) model, we estimate the probability the probability of concept being part of a relevant relation in article. We additionally compute a *sentence score* for each candidate pair of concepts, which accounts for the linguistic context in which the terms are found. For each sentence containing two or more terms, we computed its probability to express a gold interaction. This probability was estimated with a Naïve Bayes (NB) model, having a bag of words as its features. The estimations were calculated in a distant-learning manner, as the training labels were defined as follows: for every sentence with two or more terms, expand every term to all of its possible concepts; if any combination of two concepts (originating from different terms) match a curated relation for this document, the label is "true", "false" otherwise.

The sentence score for a pair of protein concepts is the sum of the confidence probability for each shared sentence. If the two concepts never appear in the same sentence, the score results in a back-off value. The two different scores are then combined into a *relation score*, based on the harmonic mean of the concept scores and the sentence score for all occurrences of the two concepts. For more technical details about the approach described in this section, please consult [23].

3 Evaluation

Using 10-fold cross-validation, we evaluated the ranked PPI lists produced by our system against the curated interactions in BioGrid. While the manually curated relations in BioGrid are undoubtedly of good quality, using BioGrid as a gold standard still needs some caution. Since the mentions of the proteins involved in an interaction are not annotated in BioGrid, all we know is the fact that in a given article, two specific proteins are mentioned as interacting. In our evaluation, we interpreted this as follows: if the system is able to establish a triple (A, c_1, c_2), where c_1 refers to a different textual mention than c_2, and if (A, c_1, c_2) or (A, c_2, c_1) is found in BioGrid, then we grade this as a true positive. Triples found in the system's output, but not in BioGrid, are considered false positives. And finally, triples missing in the output, but present in the subset of BioGrid we focused on, are seen as false negatives.

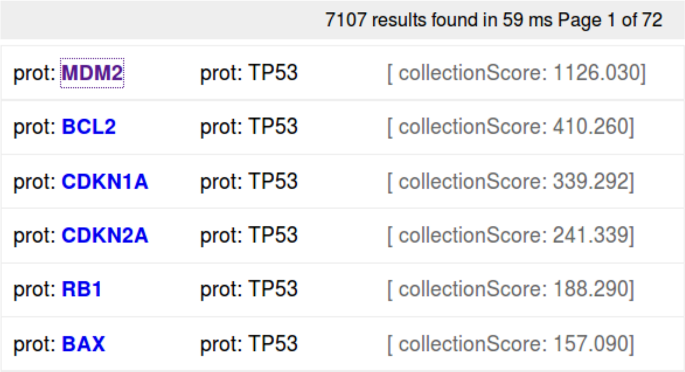
In order to evaluate and optimize the entity recognition pipeline, we further adapted this notion to the intermediate results: the annotated terms. Thus, every concept that was found in the annotations of a document was considered true positive, false positive or false negative based on its presence in the curated relations for this document. This means that a correctly annotated concept is nonetheless counted as a false positive when the protein is only mentioned, but does not participate in a curated protein-protein interaction. Thus, this classification has to be interpreted with respect to the specific usage in relation extraction.

	Our system	Our system optimized	Neji
P	0.116	0.146	0.119
R	0.706	0.700	0.676
F1	0.199	0.242	0.202

Table 1. Entity recognition quality: Precision (P), Recall (R), F-Measure (F1) for different entity recognizers.

Table 1 summarizes the performance of the entity recognition in terms of precision, recall, and their harmonic mean (F1). The first two columns show the effects of optimizing our pipeline. After manual inspection, we added a small number of exclusion rules for very frequent false positives. We also experimented with the pipeline tool Neji [4], which runs out of the box with competitive performance. As we already mentioned, the low level of precision can be explained by the fact that while both tools aim at producing all entities of the selected types, only those participating in interactions will be considered for this specific evaluation, leading to a large number of false positives, which might be perfectly good entities. Please note that the recall of about 70 % is not only due to false negatives of the term recognizer, but also reflects the fact that the text portions we are dealing with (which is the abstract in most cases) do not always mention the proteins of all curated relations.

For evaluating the quality of the interaction recognition, we used *Threshold Average Precision* (TAP-k) [6], which is a measure of ranking quality. While the details are more complicated, it can be roughly described as "precision after having seen k false positives". It is unavoidable that the system will miss some interactions: if a curated interaction is not mentioned in the text portion available to it, there is no chance of finding it. A rough estimate for the upper limit of



7107 results found in 59 ms Page 1 of 72

prot: MDM2	prot: TP53	[collectionScore: 1126.030]
prot: BCL2	prot: TP53	[collectionScore: 410.260]
prot: CDKN1A	prot: TP53	[collectionScore: 339.292]
prot: CDKN2A	prot: TP53	[collectionScore: 241.339]
prot: RB1	prot: TP53	[collectionScore: 188.290]
prot: BAX	prot: TP53	[collectionScore: 157.090]

Fig. 1. Example showing best ranked interactors for the protein TP53

the relation extraction recall can be drawn from the term recognition recall: Assuming a uniform distribution of protein entities in the curated relations, the relation extraction recall will not exceed the square of the term recognition recall, i.e. around 50 % (0.7 × 0.7). This can be verified empirically: The term recognizer is only able to find both protein concepts in 23,191 out of 50,784 gold interactions (45.7 %). Furthermore, the system might detect interactions which are not included in BioGrid and therefore are graded false positive, even though they might be regarded correct by a human expert. Using different variants of the parameter which combines the concept score and the relation score we were able to obtain a best value of 0.229 for the TAP-10 score.

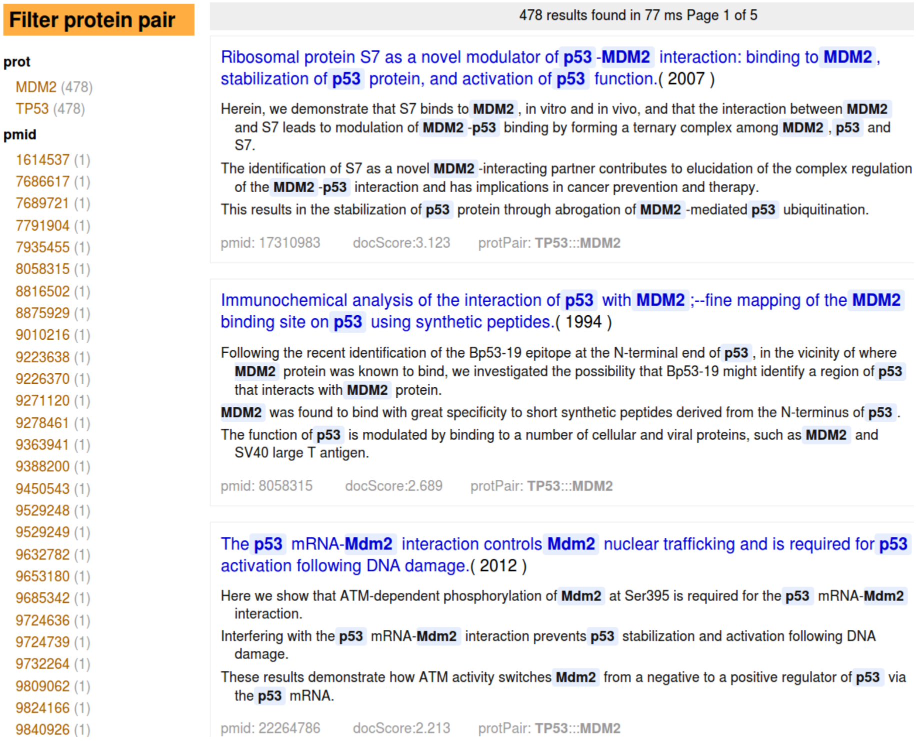
4 Applications

We presented an approach towards semi-automated semantic annotation of PubMed abstracts (or full papers, when available), using unique identifiers from reference daabases. A web-based user interface allows interaction of the expert user with the text mining system in order to achieve an efficient and accurate annotation. The automated annotations are also used in a large-scale application which enables a semantic search for interactions among domain entities.

4.1 Large-scale interaction extraction and interaction validation

As an extension of the work described in the previous section, we have analyzed the whole of PubMed using the approach described above, and produced a database of the extracted protein-protein interactions. Each interaction is characterized by a confidence score (derived from the relation score) which summarizes in a compact form the reliability of the interaction based on the evidence spread across the entire literature.

There are several potential applications for this database. As a demonstration we implemented an interface (using Apache Solr) which allow examination of



Filter protein pair

478 results found in 77 ms Page 1 of 5

prot

MDM2 (478)
TP53 (478)

pmid

1614537 (1)
7686617 (1)
7689721 (1)
7791904 (1)
7935455 (1)
8058315 (1)
8816502 (1)
8875929 (1)
9010216 (1)
9223638 (1)
9226370 (1)
9271120 (1)
9278461 (1)
9363941 (1)
9388200 (1)
9450543 (1)
9529248 (1)
9529249 (1)
9632782 (1)
9653180 (1)
9685342 (1)
9724636 (1)
9724739 (1)
9732264 (1)
9809062 (1)
9824166 (1)
9840926 (1)

Ribosomal protein S7 as a novel modulator of **p53** -**MDM2** interaction: binding to **MDM2** , stabilization of **p53** protein, and activation of **p53** function.(2007)

Herein, we demonstrate that S7 binds to **MDM2** , in vitro and in vivo, and that the interaction between **MDM2** and S7 leads to modulation of **MDM2** -**p53** binding by forming a ternary complex among **MDM2** , **p53** and S7.

The identification of S7 as a novel **MDM2** -interacting partner contributes to elucidation of the complex regulation of the **MDM2** -**p53** interaction and has implications in cancer prevention and therapy.

This results in the stabilization of **p53** protein through abrogation of **MDM2** -mediated **p53** ubiquitination.

pmid: 17310983 docScore:3.123 protPair: **TP53:::MDM2**

Immunochemical analysis of the interaction of **p53** with **MDM2** ;--fine mapping of the **MDM2** binding site on **p53** using synthetic peptides.(1994)

Following the recent identification of the Bp53-19 epitope at the N-terminal end of **p53** , in the vicinity of where **MDM2** protein was known to bind, we investigated the possibility that Bp53-19 might identify a region of **p53** that interacts with **MDM2** protein.

MDM2 was found to bind with great specificity to short synthetic peptides derived from the N-terminus of **p53** .

The function of **p53** is modulated by binding to a number of cellular and viral proteins, such as **MDM2** and SV40 large T antigen.

pmid: 8058315 docScore:2.689 protPair: **TP53:::MDM2**

The **p53** mRNA-**Mdm2** interaction controls **Mdm2** nuclear trafficking and is required for **p53** activation following DNA damage.(2012)

Here we show that ATM-dependent phosphorylation of **Mdm2** at Ser395 is required for the **p53** mRNA-**Mdm2** interaction.

Interfering with the **p53** mRNA-**Mdm2** interaction prevents **p53** stabilization and activation following DNA damage.

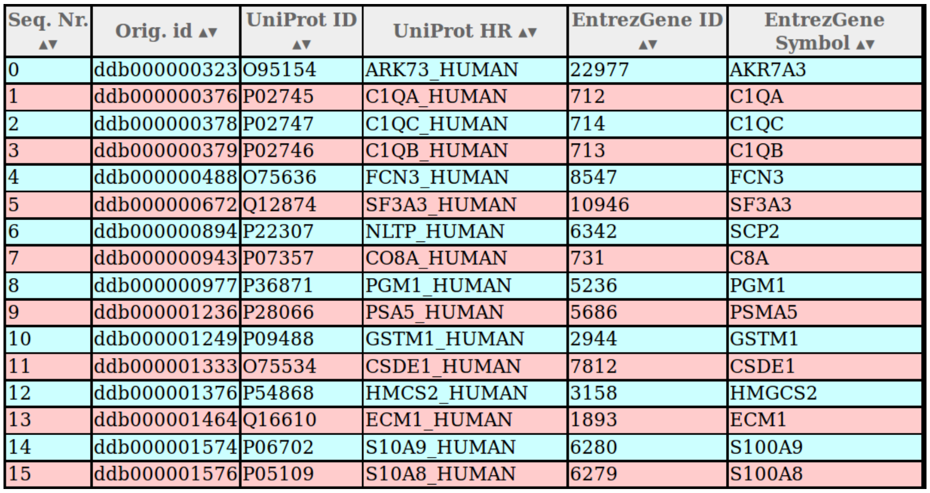
These results demonstrate how ATM activity switches **Mdm2** from a negative to a positive regulator of **p53** via the **p53** mRNA.

pmid: 22264786 docScore:2.213 protPair: **TP53:::MDM2**

Fig. 2. Example showing top-ranked snippets for the interaction TP53 - MDM2

the results. The user can enter an arbitrary protein name, and the system will provide a list of candidate interactors, ranked according to the confidence score (see figure 1). Once the user selects one of these interactions, the system will deliver the textual snippets which are considered to be most relevant for that particular interaction. Figure 2 shows precisely this final step (best evidence for a given interaction) from the current version of the interface.

In another application we have been given by a domain expert a list of several hundred proteins of interest in a particular biological study (see figure 3). The researcher was interested in what are the potential interactions among those proteins. Since the number of potential interactions is quadratic to the number of input proteins, it is useful, before planning an experimental validation, to have some pre-filtering technique that allows to narrow down the space of interactions to be investigated. Using the database described above we were able to reduce considerably this set, removing a huge number of potential interactions for which there is no evidence whatsoever in the literature. Additionally, the remaining set of candidate interactions is ranked according to our confidence score, thus providing a potential way to further narrow down the scope of the experimental investigation (see figure 4).



Seq. Nr. ▴▾	Orig. id ▴▾	UniProt ID ▴▾	UniProt HR ▴▾	EntrezGene ID ▴▾	EntrezGene Symbol ▴▾
0	ddb000000323	O95154	ARK73_HUMAN	22977	AKR7A3
1	ddb000000376	P02745	C1QA_HUMAN	712	C1QA
2	ddb000000378	P02747	C1QC_HUMAN	714	C1QC
3	ddb000000379	P02746	C1QB_HUMAN	713	C1QB
4	ddb000000488	O75636	FCN3_HUMAN	8547	FCN3
5	ddb000000672	Q12874	SF3A3_HUMAN	10946	SF3A3
6	ddb000000894	P22307	NLTP_HUMAN	6342	SCP2
7	ddb000000943	P07357	CO8A_HUMAN	731	C8A
8	ddb000000977	P36871	PGM1_HUMAN	5236	PGM1
9	ddb000001236	P28066	PSA5_HUMAN	5686	PSMA5
10	ddb000001249	P09488	GSTM1_HUMAN	2944	GSTM1
11	ddb000001333	O75534	CSDE1_HUMAN	7812	CSDE1
12	ddb000001376	P54868	HMCS2_HUMAN	3158	HMGCS2
13	ddb000001464	Q16610	ECM1_HUMAN	1893	ECM1
14	ddb000001574	P06702	S10A9_HUMAN	6280	S100A9
15	ddb000001576	P05109	S10A8_HUMAN	6279	S100A8

Fig. 3. Validation of interaction set: example of input proteins.

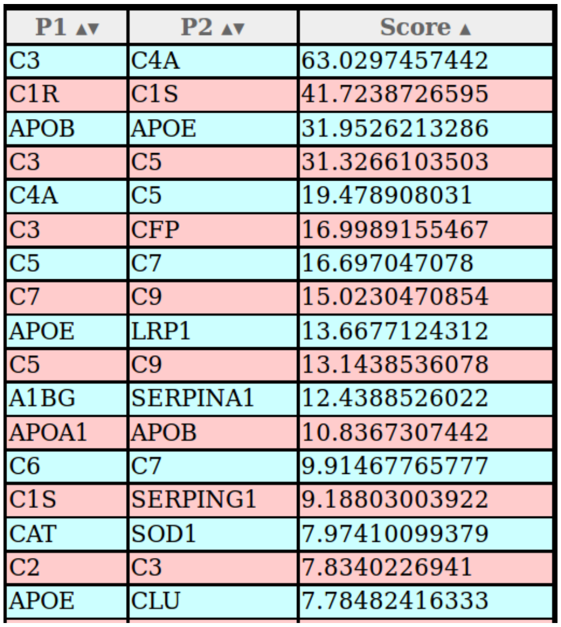
4.2 Assisted curation

Biomedical curators are professionals with a strong background in the life sciences who read the literature in search of particular items of information (e.g. newly detected protein interactions), and store such information in public databases, which can in turn be accessed later by the biologists. For example, UniProt [31] collects information on all known proteins. IntAct [10] is a database collecting protein interactions. PharmGKB [26] collects interactions among genes, drugs, and diseases. BioGrid [28] is a well-known database describing gene and protein interactions.

Most of the information in these databases is derived from the primary literature by a process of manual revision known as “literature curation”. The full scope of curation that has to be done on a single publication is part of ongoing research and leads to the development of new ontologies and to the definition of the most relevant relations that have to be considered.

Despite the significant improvements in the last couple of years, most experts agree that, at least for the time being, it is unrealistic to expect fully automated text mining systems to perform at a level acceptable for tasks that require high accuracy, such as automated database curation. However, existing systems can already achieve results which are sufficiently good to be used in a semi-automated context, where a human expert validates the output of the system. One application where this support is badly needed is biomedical literature curation.

In order to satisfy this need, we have implemented a user-friendly web based interface which interfaces our text mining system and allows a domain expert to inspect the results of the automated annotation process (see Figure 5). The purpose of the system is to enable a human annotator/curator to leverage upon



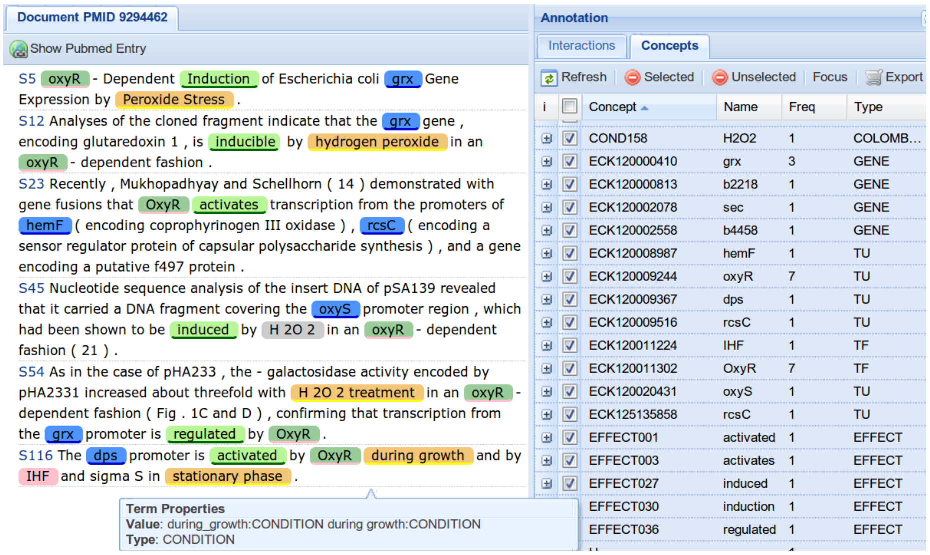
P1 ▴▾	P2 ▴▾	Score ▴
C3	C4A	63.0297457442
C1R	C1S	41.7238726595
APOB	APOE	31.9526213286
C3	C5	31.3266103503
C4A	C5	19.478908031
C3	CFP	16.9989155467
C5	C7	16.697047078
C7	C9	15.0230470854
APOE	LRP1	13.6677124312
C5	C9	13.1438536078
A1BG	SERPINA1	12.4388526022
APOA1	APOB	10.8367307442
C6	C7	9.91467765777
C1S	SERPING1	9.18803003922
CAT	SOD1	7.97410099379
C2	C3	7.8340226941
APOE	CLU	7.78482416333

Fig. 4. Validation of interaction set: best interactions detected by the system.

the result of a advanced text mining system in order to enhance the speed and effectiveness of the annotation process.

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. 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.

The documents and the annotations are represented consistently within a single XML file, which also contains a record of the user interaction, thus allowing advanced logging support. The annotations are selectively presented, in a 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 his 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 methodology 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 advantage of a client-side presentation logic is the flexibility for the



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

Fig. 5. A screenshot of the curation system's interface

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.

In a recently approved NIH-funded project ("High Throughput Literature Curation of Genetic Regulation in Bacterial Models") we intend to leverage the capabilities of the OntoGene/ODIN system in order to improve the efficiency of the curation process of the RegulonDB database. RegulonDB[5] is the primary database on transcriptional regulation in Escherichia coli K-12 containing knowledge manually curated from original scientific publications, complemented with high throughput datasets and comprehensive computational predictions.

5 Conclusion

We have presented an advanced text mining architecture, which is capable of automatically annotating the biomedical literature with domain entities of relevance for specific applications, and to detect interactions among those entities. In particular, we have discussed and evaluated a specific scenario for protein-protein interactions.

Additionally, we discussed an application in assisted curation, and an application for the filtering of potential interactions among a given set of proteins. In order to support a process of assisted curation we provide a user-friendly web-based interface, which is currently being used by life science databases within the scope of large curation projects.

[5] regulondb.ccg.unam.mx

6 Acknowledgments

The OntoGene group at the University of Zurich is partially supported by the Swiss National Science Foundation (grants 100014 − 118396/1 and 105315 − 130558/1) and by F. Hoffmann-La Roche Ltd, Basel, Switzerland. The collaboration with the RegulonDB group at the Mexican National University is sponsored by the NIH grant 1R01GM110597-01A1.

References

1. Androutsopoulos, I.: A challenge on large-scale biomedical semantic indexing and question answering. In: BioNLP workshop (part of the ACL Conference) (08/2013 2013)
2. Arighi, C.N., Carterette, B., Cohen, K.B., Krallinger, M., Wilbur, W.J., Fey, P., Dodson, R., Cooper, L., Van Slyke, C.E., 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, M.L., Park, J., Haendel, M., Van Auken, K., Li, Y., Chan, J., Muller, H.M., Cui, H., Balhoff, J.P., Chi-Yang Wu, J., Lu, Z., Wei, C.H., Tudor, C.O., Raja, K., Subramani, S., Natarajan, J., Cejuela, J.M., Dubey, P., Wu, C.: An overview of the BioCreative 2012 workshop track iii: interactive text mining task. Database 2013 (2013)
3. Aronson, A.R., Lang, F.M.: An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc 17(3), 229–236 (2010)
4. Campos, D., Matos, S., Oliveira, J.L.: A modular framework for biomedical concept recognition. BMC Bioinformatics 14, 281 (2013)
5. Campos, D., Matos, S., Oliveira, J.L.: Gimli: open source and high-performance biomedical name recognition. BMC Bioinformatics 14, 54 (2013)
6. Carroll, H.D., Kann, M.G., Sheetlin, S.L., Spouge, J.L.: Threshold average precision (TAP-k): a measure of retrieval designed for bioinformatics. Bioinformatics 26(14), 1708–1713 (2010)
7. Chen, H., Sharp, B.: Content-rich biological network constructed by mining pubmed abstracts. BMC Bioinformatics 5, 147 (2004)
8. Dolinski, K., Chatr-Aryamontri, A., Tyers, M.: Systematic curation of protein and genetic interaction data for computable biology. BMC Biol. 11, 43 (2013)
9. Doms, A., Schroeder, M.: GoPubMed: exploring PubMed with the Gene Ontology. Nucleic Acids Res. 33(Web Server issue), W783–786 (Jul 2005)
10. 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. 32(suppl 1), D452–455 (2004)
11. Hoffmann, R.: Using the iHOP information resource to mine the biomedical literature on genes, proteins, and chemical compounds. Curr Protoc Bioinformatics Chapter 1, Unit1.16 (Dec 2007)
12. Hoffmann, R., Valencia, A.: A gene network for navigating the literature. Nature Genetics 36, 664 (2004)
13. Jonquet, C., Shah, N.H., Musen, M.A.: The open biomedical annotator. Summit on Translat Bioinforma 2009, 56–60 (2009)

14. Kim, J., Pezik, P., Rebholz-Schuhmann, D.: Medevi: Retrieving textual evidence of relations between biomedical concepts from medline. Bioinformatics 24(11), 1410–1412 (2008)
15. Lu, Z.: Pubmed and beyond: a survey of web tools for searching biomedical literature. Database 2011 (2011)
16. Plake, C., Schiemann, T., Pankalla, M., Hakenberg, J., Leser, U.: AliBaba: PubMed as a graph. Bioinformatics 22(19), 2444–2445 (Oct 2006)
17. Pyysalo, S., Ohta, T., Miwa, M., Cho, H.C., Tsujii, J., Ananiadou, S.: Event extraction across multiple levels of biological organization. Bioinformatics 28(18), i575–i581 (Sep 2012)
18. Rebholz-Schuhmann, D., Arregui, M., Gaudan, S., Kirsch, H., Jimeno, A.: Text processing through Web services: calling Whatizit. Bioinformatics 24(2), 296–298 (2008)
19. Rebholz-Schuhmann, D., Clematide, S., Rinaldi, F., Kafkas, S., van Mulligen, E.M., 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: Forner, P., Mueller, H., Rosso, P., Paredes, R. (eds.) Information Access Evaluation. Multilinguality, Multimodality, and Visualization, pp. 353–367. Lecture Notes in Computer Science, Springer, Valencia (2013)
20. Rebholz-Schuhmann, et al.: Assessment of ner solutions against the first and second calbc silver standard corpus. Journal of Biomedical Semantics 2(Suppl 5), S11 (2011)
21. 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)
22. Rinaldi, F., Kappeler, T., Kaljurand, K., Schneider, G., Klenner, M., Clematide, S., Hess, M., von Allmen, J.M., Parisot, P., Romacker, M., Vachon, T.: OntoGene in BioCreative II. Genome Biology 9(Suppl 2), S13 (2008)
23. Rinaldi, F., Schneider, G., Clematide, S.: Relation mining experiments in the pharmacogenomics domain. Journal of Biomedical Informatics 45(5), 851–861 (2012)
24. 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 7(3), 472–480 (2010)
25. Rodriguez-Esteban, R.: Biomedical text mining and its applications. PLoS Comput. Biol. 5(12), e1000597 (Dec 2009)
26. Sangkuhl, K., Berlin, D.S., Altman, R.B., Klein, T.E.: PharmGKB: Understanding the effects of individual genetic variants. Drug Metabolism Reviews 40(4), 539–551 (2008), pMID: 18949600
27. Segura-Bedmar, I., Martínez, P., Sánchez-Cisneros, D.: The 1st ddi extraction-2011 challenge task: Extraction of drug-drug interactions from biomedical texts. In: Proc DDI Extraction-2011 challenge task. pp. 1–9. Huelva, Spain (2011)
28. Stark, C., Breitkreutz, B.J., Reguly, T., Boucher, L., Breitkreutz, A., Tyers, M.: Biogrid: A general repository for interaction datasets. Nucleic Acids Research 34, D535–9 (2006)
29. Sun, W., Rumshisky, A., Uzuner, O.: Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. J Am Med Inform Assoc 20(5), 806–813 (2013)
30. Thomas, P., Starlinger, J., Vowinkel, A., Arzt, S., Leser, U.: GeneView: a comprehensive semantic search engine for PubMed. Nucleic Acids Res. 40(Web Server issue), W585–591 (Jul 2012)
31. UniProt Consortium: The universal protein resource (uniprot). Nucleic Acids Research 35, D193–7 (2007)