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

Fabio Rinaldi, Lenz Furrer, Simon Clematide

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 [6], Gimli [4], iHOP[1] [12, 11], Open Biomedical Annotator [13], AliBaba [16], GOPubMed [8], 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 OntoGene system uses terminology derived from life science databases, storing internally the several possible names for a given term and its unique identifier. All potential terms of relevance in the target collection are automatically annotated with an approximate matching approach. The annotation process not only identifies a string as a potential term of a given type (e.g. a protein), but also associates to it the unique identifier of the term. In case of ambiguos terms, several potential identifiers will be associated to the term. This high-recall approach is then followed by a machine-learning based filtering step which removes false positives and disambiguates ambiguos terms. In the case of the application considered in this paper, the reference database was BioGrid[4] [7] for its good coverage and quality curation. 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.

In order to produce candidate interactions the system first generates all pairwise combinations of the pairs `(term,identifier)` seen in the document, and scores them using information from the original database, as described below. A threshold on the score can then be used to select the best candidate interactions. For each of these interactions, evidence snippets from the text are provided.

BioGrid provides for each protein-protein interaction at least one reference, i.e. a paper where that interaction is mentioned. Our initial collection for the experiment described in this paper consisted of 20,000 PubMed abtracts, selected from the references of BioGrid, in order to insure that at least one protein-protein interaction would be detectable. In addition to the abstract itself, we use also the MeSH keywords and list of Chemical substances which are associated to the abstract in PubMed.

For every annotated term, and for every possible identifier for that term, we compute a *concept score*, which is based on the likelihood of that `(term, identifier)` pair to participate in a curated interaction in the reference database. Additional factors, such as the position in the document (title vs abstracts) are used as weights.

For every pairwise combination of concepts (term identifiers) we also compute a *sentence score*, based on all the sentences which contain annotated terms with the identifiers of the two concepts (excluding cases where the two identifiers happen to refer to the same term). Such sentences (using a distant-learning approach) are assumed to be positives if the pair of identifiers is provided in the reference database as a curated interaction. This probability, estimated with a Naïve Bayes (NB) model having a bag of words as its features, tries to capture

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

the linguistic context in which the interaction can occur. For more technical details about the approach described in this section, please consult [23, 9].

3 Evaluation

The evaluation of the extracted interactions was performed using a 10-fold cross validation. The results are compared against the reference database (BioGrid in this case) taken as a gold standard. Care must be taken in the interpretation of these results, since the reference database does not provide the position of possible interactions, but only the fact that two interacting entities have been identified in a given paper. Therefore any results provided by the system in the form of a triple $(article/id0, entity/id1, entity/id2)$ is considered positive if and only if the interaction $(id1, id2)$ can be found in the database associated with paper $id0$.

Extending this approach to the evaluation of the detection of domain entities, we can consider a detected term as a true positive if it is associated to an identifier which is part of a curated relation for the analyzed article. This approach has the advantage that we can use the manually curated database as a gold standard for evaluation, but it has of course the disadvantage that a correctly annotated term might nonetheless be considered as a false positive simply because it is not part of one of the curated interactions.

Using the approach mentioned above, we evaluated the performance of the entity recognition component using the conventional evaluation measures. The basic system, as described above, reaches a recall of 70.6%, with however only a precision of 11.6% (F-score: 19.9%). After inspecting the results, we introduced a small number of exclusion rules for very frequent false positives, which lead to an improvement of precision (14.6%) at a small cost for recall (70.0%), leading to a clearly better F-score (24.2%).

There are several reasons for the relatively low level of these results. The low level of precision can be partly explained by the indirect evaluation mechanism, i.e. an annotated term is only considered a positive if it is part of an interaction, which results in several false positives which could actually be perfectly correct entity annotations. The recall is affected by the fact that while database curators have access to the full articles, the system analyzes only abstracts.

In any case, our main concern in this experiment is the retrieval of interactions, therefore we aim for for a relatively high recall at the entity level, even at the cost of a low precision. The metric used for the evaluation of the detected interactions is *Threshold Average Precision* (TAP-k) [5], which is a measure of ranking quality. TAP-k can be described in informal terms as "precision after having seen k false positives". The main reason for the choice of this metric is that while it unrealistic to expect the system to produce high levels of precision and recall, we aim at producing the best possible ranking of the results, since we expect the users to inspect only the top results produced by the system.

Since relations are pairwise combinations of detected entities, an upper boundary for the recall of relation extraction can be estimated from the entity recog-

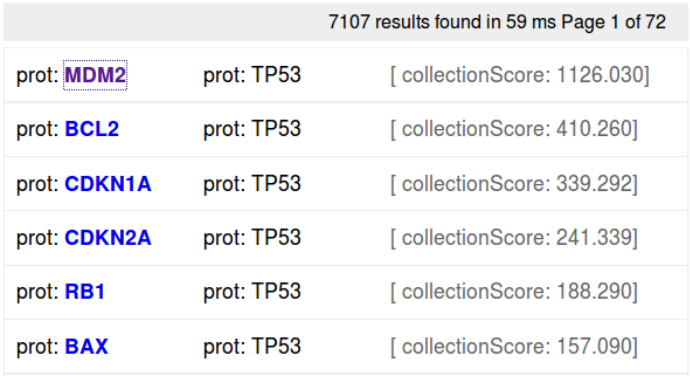

Fig. 1. Example showing best ranked interactors for the protein TP53

nition recall. Given the current recall value of 0.7 for entity recognition, relation extraction is not expected to exceed 0.5 (0.7×0.7). Precision in relation extraction, just like in entity extractions, is limited by the type of evaluation methodology. Relations which have not been curated in the reference databases will be considered as false positives, even though they might be mentioned in the text as legitimate interactions. 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 of extracted relations.

4 Applications

In the previous section we described an approach towards semi-automated semantic annotation of PubMed abstracts (or full papers, when available), using unique identifiers from reference databases. 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 application of the methods described in the previous section, we have analyzed the whole of PubMed, 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 allows examination of the results. The user can enter an arbitrary protein name, and the system will

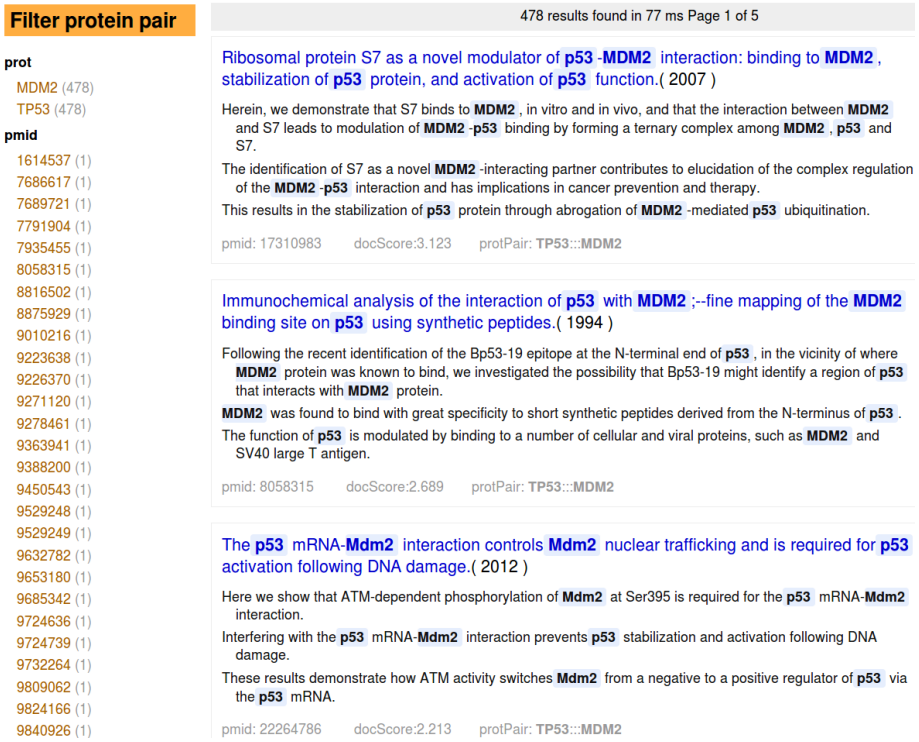

Fig. 2. Example showing top-ranked snippets for the interaction TP53 - MDM2

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), by selecting the highest ranked interaction interaction candidates which are not yet known to the domain expert.

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

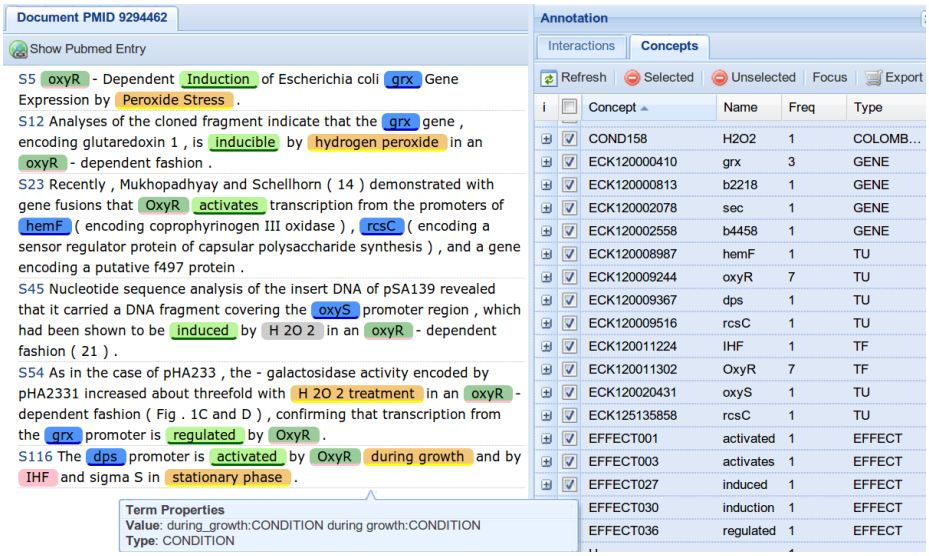

Fig. 5. A screenshot of the curation system's interface

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), http://bioasq.org/sites/default/files/BioNLP_presentation.pdf
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), http://database.oxfordjournals.org/content/2013/bas056.abstract
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.: Gimli: open source and high-performance biomedical name recognition. BMC Bioinformatics 14, 54 (2013)
5. 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)
6. Chen, H., Sharp, B.: Content-rich biological network constructed by mining pubmed abstracts. BMC Bioinformatics 5, 147 (2004)
7. Dolinski, K., Chatr-Aryamontri, A., Tyers, M.: Systematic curation of protein and genetic interaction data for computable biology. BMC Biol. 11, 43 (2013)
8. Doms, A., Schroeder, M.: GoPubMed: exploring PubMed with the Gene Ontology. Nucleic Acids Res. 33(Web Server issue), W783–786 (Jul 2005)
9. Furrer, L., Clematide, S., Marques, H., Rodriguez-Esteban, R., Romacker, M., Rinaldi, F.: Collection-wide extraction of protein-protein interactions. In: Proceedings of The Sixth International Symposium on Semantic Mining in Biomedicine (SMBM), Aveiro, Portugal (October 2014)
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), http://nar.oxfordjournals.org/cgi/content/abstract/32/suppl_1/D452
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), http://database.oxfordjournals.org/content/2011/baq036.abstract
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), http://bioinformatics.oxfordjournals.org/cgi/content/abstract/24/2/296
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), http://www.zora.uzh.ch/82216/
20. 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, C.J., Clematide, S., Rinaldi, F., Farkas, R., Mora, G., Hara, K., Furlong, L.I., 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 2(Suppl 5), S11 (2011), http://www.jbiomedsem.com/content/2/S5/S11
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), http://genomebiology.com/2008/9/S2/S13
23. Rinaldi, F., Schneider, G., Clematide, S.: Relation mining experiments in the pharmacogenomics domain. Journal of Biomedical Informatics 45(5), 851–861 (2012), http://www.sciencedirect.com/science/article/pii/S1532046412000676
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), http://informahealthcare.com/doi/abs/10.1080/03602530802413338, 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)