Using Large Biomedical Databases as Gold Annotations for Automatic Relation Extraction

Tilia Ellendorff, Fabio Rinaldi, Simon Clematide
University of Zurich
Institute of Computational Linguistics, Binzmühlestrasse 14, 8050 Zürich, Switzerland
ellendorff@cl.uzh.ch, fabio.rinaldi@uzh.ch, simon.clematide@uzh.ch

Abstract

We show how to use large biomedical databases in order to obtain a gold standard for training a machine learning system over a corpus of biomedical text. As an example we use the Comparative Toxicogenomics Database (CTD) and describe by means of a short case study how the obtained data can be applied. We explain how we exploit the structure of the database for compiling training material and a testset. Using a Naive Bayes document classification approach based on words, stem bigrams and MeSH descriptors we achieve a macro-average F-score of 61% on a subset of 8 action terms. This outperforms a baseline system based on a lookup of stemmed keywords by more than 20%. Furthermore, we present directions of future work, taking the described system as a vantage point. Future work will be aiming towards a weakly supervised system capable of discovering complete biomedical interactions and events.

Keywords: Biomedical Textmining, Biomedical Databases, Relation Extraction, Distant Supervision

1. Introduction

In order to use approaches of supervised machine learning within the area of biomedical textmining, annotated language resources are necessary. However, the amount of annotated corpora in this area is still limited and not easy to obtain: the human effort involved in linguistic annotations of biomedical text is expensive and not time-efficient. Furthermore, an expert trained to perform such a task needs not only to possess the necessary knowledge of linguistics but also a biomedical background. For this reason, it makes sense to investigate the exploitation of alternative resources. A wide range of manually created scientific databases containing relational data between biomedical entities are already available and most academic publications exist in electronic format and can be accessed relatively easily.

The structured data of scientific databases in combination with the unstructured textual data of publications can be exploited as language resources for constructing and optimizing text mining applications. Methods using the data of a database in connection with textual data are commonly referred to as weakly or distant supervised methods, as there is no direct annotation as there would be with a manually annotated corpus. On the other hand, the system is also not completely unsupervised as a database is used for providing additional information that is linked to the text in a more indirect way.

One very large database which is especially useful for biomedical relation extraction is the Comparative Toxicogenomics Database (CTD) (Davis et al., 2009). It contains data about gene-chemical interactions from so far almost 50,000 publications and is growing continuously. The objective of this paper is to show by means of a case study how this database can be used for building a system to discover relations between biomedical entities.

2. Biomedical Text Mining and Relation Extraction

For promoting research in the field of biomedicine and life science, it is indispensable for scientists to get fast and easily to the information relevant to their research. For a long time, scientific findings contained in biomedical publications have been extracted, organized and stored into databases solely by human manual work. This so-called curation process is performed by human biocurators who are experts in their field and whose task is to expand and maintain the knowledge repository. However, the number of publications in the biomedical domain is continuously increasing and thus, text mining tools are more and more required to speed up the process and support the curation of knowledge contained in scientific publications.

Nowadays most biomedical databases, such as PharmGKB[1] (Klein et al., 2001), IntAct[2] (Hermjakob et al., 2004) and Uniprot[3] (UniProt Consortium, 2007), to name just a few, contain relational data. Many biomedical interactions take place on a molecular level and a huge amount of different entities are involved. Focus and categorization of entities differ between relational databases. The entities which are described in CTD can be categorized into three general groups: genes/proteins, chemical and diseases. Automatic relation extraction represents the current approach for extracting knowledge specific to interactions and associations between different entities of the biomedical domain. As biomedical entities are very diverse and highly ambiguous, relation extraction in the biomedical domain normally has to rely on powerful tools for named entity extraction. This is not always easy, as entities are not always expressed explicitly and unambiguously in the text. But even if the entities have been discovered, one more big challenge of biomedical text mining is finding the relation between these entities. Relations between two or more entities are not al-

[1] https://www.pharmgkb.org/
[2] http://www.ebi.ac.uk/intact/
[3] http://www.uniprot.org/

```
<ixn id="2892024">
   <taxon id="9606">Homo sapiens</taxon>
   <reference pmid="12035877"/>
   <axn code="exp" degreecode="+" position="1" parentid="2892024">results in increased expression of</axn>
      <actor type="chemical" id="MESH:C081309" parentid="2892024" position="1">irbesartan</actor>
      <actor type="gene" id="GENE:4878" parentid="2892024" form="protein" position="2">NPPA</actor>
</ixn>
```

Figure 1: XML Entry for the abstract with the Pubmed ID 12035877 (Kotridis et al., 2002)

ways expressed within sentence boundaries. Furthermore, relations themselves are also often not always stated in an explicit way.

2.1. Related Work

Our approach of using knowledge database entries and their bibliographic references to unstructured text as a replacement for hand-labeled data is related to other information extraction research. (Mintz et al., 2009) uses the large semantic knowledge database Freebase for a *distant supervision* approach to optimize information extraction patterns. (Morgan et al., 2004) used the FlyBase database and referenced PubMed abstracts to improve the recognition of gene names by machine learning. (Craven and Kumlien, 1999) present early work on exploiting biomedical databases as weakly labeled training data to improve relation extraction. They also use Naive Bayes classifiers for their task.

Similar techniques have already been used in previous applications of 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 (Krallinger et al., 2008). 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 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”* (Rinaldi et al., 2008). Furthermore, a similar approach was also applied successfully in the version of the OntoGene system used for participation in the 2009 BioCreative competition (Leitner et al., 2010), which obtained the best results among all participants in the extraction of protein-protein interactions from scientific literature (Rinaldi et al., 2010).

Axn	**Formulated Action Term**
exp	*results in decreased expression of* *affects the expression of* *results in increased expression of*
rxn	*affects the reaction* *inhibits the reaction* *promotes the reaction*
csy	*results in decreased chemical synthesis of* *affects the chemical synthesis of* *results in increased chemical synthesis of*
rec	*affects the susceptibility to* *results in increased susceptibility to* *results in decreased susceptibility to*
pho	*results in increased phosphorylation of* *results in decreased phosphorylation of* *affects the phosphorylation of*
b	*binds to*
met	*results in decreased metabolism of* *results in increased metabolism of*
w	*co-treated with*
act	*results in decreased activity of* *affects the activity of* *results in increased activity of*

Table 1: Formulated action terms for a given set of action term abbreviations (Axn)

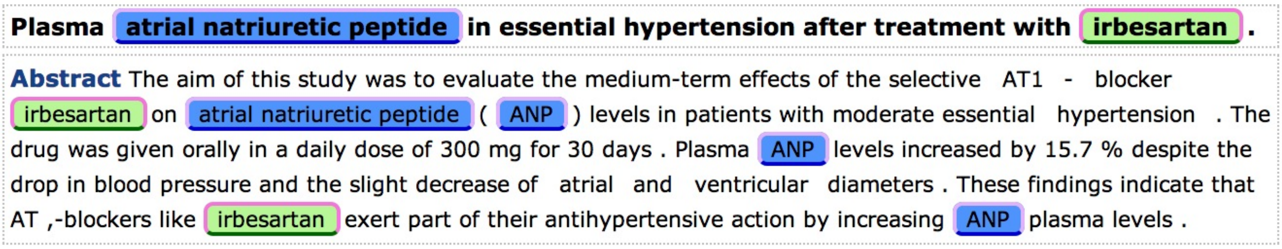
3. The Toxicogenomics Database and its Utility in Language Applications

The Comparative Toxicogenomics Database (CTD) is a continuously growing database which is an important resource and scientific tool for researchers from all biomedical fields. The database was first launched in 2004 and has the aim of promoting *”understanding about the effects of environmental chemicals on human health”* (Davis et al., 2009). It contains manually curated data as well as inferred data, which is added based on logical conclusions drawn from curated data. It presents scientists with a tool that can be used not only for looking up interactions but also for building new hypothesis based on the data that it contains. Apart from dealing with associations between diseases and chemicals as well as proteins, it also contains data about biomolecular interactions between chemicals and genes. The interactions between chemicals and genes are the most interesting part of the database from the point of view of biomedical relation extraction. Associations describe relations that are not as strongly defined as it is the case with real interactions. Furthermore, in CTD associations can be automatically inferred whereas interactions are always manually curated and therefore more reliable. For this reason, we chose the interactions as the focus of our work.

The downloadable xml data, which is freely available on the official CTD webpage[4], currently contains 611,241 different entries[5]. Each entry within the xml file describes one interaction together with its actors. One example for

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

[5] Number in August 2013



Plasma atrial natriuretic peptide in essential hypertension after treatment with irbesartan .

Abstract The aim of this study was to evaluate the medium-term effects of the selective AT1 - blocker irbesartan on atrial natriuretic peptide (ANP) levels in patients with moderate essential hypertension . The drug was given orally in a daily dose of 300 mg for 30 days . Plasma ANP levels increased by 15.7 % despite the drop in blood pressure and the slight decrease of atrial and ventricular diameters . These findings indicate that AT ,-blockers like irbesartan exert part of their antihypertensive action by increasing ANP plasma levels .

Figure 2: Entity annotation of the abstract with the PubMed ID 12035877, displayed in the ODIN interface

Action Term	Abbreviation	Freq. of PMIDs in CTD	Freq. of PMIDs in Test Set
expression	exp	564434	257
reaction	rxn	93725	140
activity	act	59333	161
cotreatment	w	43778	46
binding	b	33593	82
phosphorylation	pho	19657	22
response to substance	rec	19289	180
secretion	sec	9507	32
methylation	myl	8923	1
abundance	abu	6018	50
localization	loc	5861	23
metabolic processing	met	5309	11
cleavage	clv	4262	8
transport	trt	2444	6
chemical synthesis	csy	2331	13
uptake	upt	2068	5
degradation	deg	1873	6
oxidation	oxd	1171	3
mutagenesis	mut	847	6

Table 2: Number of associated PubMed IDs in CTD and in the Test Set for the most frequent action terms (Sorted by Overall Frequency in CTD)

an entry from the xml file can be seen in Figure 1. Furthermore, each entry in the xml file is connected to at least one reference publication in the format of a PubMed Document Identifier (PubMed ID). This identifier is encoded in the xml data by using the reference tag. With the help of this unique identifier, a scientific abstract can be easily retrieved from the PubMed webpage. Altogether 45,833 different abstracts from PubMed are referenced in the context of gene-chemical interactions.

Within CTD, each interaction is assigned an interaction type which has the format of an action term. Action terms constitute a very central element of each entry in the xml data file. They form a controlled vocabulary used by CTD to describe the nature and type of interactions. An overview of the most frequent action terms is given in Figure 2. There are currently around 50 different action terms in the CTD database. In the xml data, the action term is listed in the axn tags. The action term itself is encoded as its abbreviations as an attribute. The actual content of the tag is represented by a formulation of the action term. In the example in Figure 1 this formulation is *results in increased expression of'*. For each action term, there are one to three different possible formulations in the xml data. These formulations constitute a combination of the action term itself and the "degreecode" attribute, which represents the directionality of the interaction. An overview of action term formulations can be seen in Table 1.

The other main elements of an entry are the actors as arguments of an interaction. Apart from genes and chemicals, other sub-interactions can also have the function of actors within subordinate interaction. In the xml file, genes and chemicals are mostly not encoded in the same format as found in the referred abstract but in a normalized format in order to avoid ambiguities and missunderstandings for the users and also with the aim of facilitating derivation of new interactions. Furthermore, as a reference to other databases, identification numbers are given for chemicals and genes/proteins. It depends very much on the action term if sub-interactions are possible as actors. These so-called nested interactions are most frequent with the action terms "cotreatment" and "reaction" describing the superordinate interaction. Other action terms tend to occur more frequently in the subordinate interactions of nested interaction.

One example for actor entity annotation in CTD can be seen in Figure 1. The gene NPPA and the chemical irbesartan are annotated as actors of an interaction described by the action term expression (abbreviation: "exp"). The referenced abstract for the same PubMed ID can be seen in Figure 2.

For displaying the annotated entities, we use ODIN (Rinaldi et al., 2013b), an interface designed by the Ontogene group for supporting the curation process of human manual curation of biomedical articles. It uses methods of information extraction for displaying named entities and possible relations between them to a curator and thus providing important hints as to where in the text an interaction could be present. However, the suggestions of possible interactions given by ODIN, do not yet include the interaction type.

In the example in Figure 2, the chemical is marked in green colour and the gene in blue. Comparing the xml entry with the text of the abstract, it can be clearly seen, that the surface form in the text does not correspond to the term used in CTD but a normalized form is given. This depicts well how in many cases background knowledge is needed in order to infer the entities from CTD and find them in the actual abstract: the surface form in the text describes a gene product, more precisely a hormon (atrial natriuretic peptide = ANP), while CTD lists the gene from which this gene product is encoded. Furthermore, it also shows how the concrete interaction between the two actors is not mentioned explictly in the text but has to be concluded from the context. If the percentage to the protein (ANP) is observed to be increasing, this means that the chemical ibesartan must have an effect on the gene which is used for encoding the protein.

4. Case Study of using CTD as a Linguistic Resource for biomedical relation extraction

In the following paragraphs we will present a case study for using CTD together with the referenced PubMed abstracts as a language resource for relation extraction. Both sources in combination provide us with annotated gold data for training and evaluating a machine learning system for action term recognition.

The aim of our system, which we developed in the course of the BioCreative 2013 competition[6], and which is part of a bigger system for the detection of CTD entities (Rinaldi et al., 2013a), was to discover action terms describing interactions within PubMed abstracts. Given a PubMed ID, the system is designed to deliver all action terms that describe interactions found in this specific abstract.

4.1. Dataset and Methods

After collecting PubMed IDs from all entries of the CTD download data, we downloaded the abstracts from PubMed and preprocessed them, using sentence splitting and tokenization. By relating the PubMed abstracts to their annotated action terms, we collected the training data used for training the machine learning system.

Our training data consisted of all abstracts which are referenced in CTD . This amounts to a total of 45,836 abstracts. However, for training our system, we took random samples of 4000 abstracts for each action term, of these 2000 positive examples and 2000 negative examples. For action terms for which there were not enough positive examples, we took the highest possible number of abstracts for training.

Our system consisted of several binary classifiers, one for each action term to be classified. These classifiers were built using the Naive Bayes classifiers included in the Natural Language Processing Toolkit (NLTK) (Bird et al., 2009).

We experimented with different feature sets and found that a combination of bag-of-words (the 5000 most frequent words from all abstracts), stem bigrams and MeSH descriptors gave the best results. Medical Subject Headings (MeSH)[7] descriptors are meta-data of PubMed abstracts providing keywords for the main topics of a PubMed abstract.

Furthermore, in order to put our results into a wider context, we built a simple baseline. Our baseline assigned an action term to an abstract if the stem of the action term was present in the abstract. For stemming the action terms as well as the words of each abstract, we used the Lancaster Stemming Algorithm[8], which is also part of NLTK. We decided on using this stemming algorithm for our baseline, as it is more aggressive than, for example, the Porter stemmer and makes sure that the stems of action terms match the stems of corresponding verbs. For multi word action terms, as for example *"response to substance"* (abbreviation: "rec"), we chose a representative word from the action term formulations, in this case "susceptibility", which we stemmed in order to provide an action term stem. The results of our baseline can also be seen in Table 3.

4.2. Evaluation and Results

We evaluated our system on the official gold standard data set from the BioCreative competition. This gold standard consists of a list of 510 PubMed abstracts. For these abstracts we retrieved the correct action terms from the CTD database and compared them to the action terms discovered by our own system. In this way, we were able to determine evaluation scores in terms of overall average precision, recall and F-score for our system but also for each of the classifiers separately (Table 3).

We discovered that including only the 8 best-performing classifiers results in the highest overall F-score of the system. Comparing these 8 classifiers which showed the best performance on their own to the most frequent action terms, it becomes clear that there is a connection between quality of performance of a classifier and the number of abstract for an action term that are referenced by their PubMed ID. Naturally, action terms with more training material available show a better performance. However, some action term classifiers do not show a very good performance even though there is a decent amount of training material available. An example for such an action term is ”cotreatment” (abbreviation: “w”). It can be assumed that for cotreatment, the fact that there is a high frequency of subordinate interactions, taking the role of actors within the interaction, had an impact on the performance of the classifier. On the other hand, classifiers such as for the action term “oxidation”, did

[6] http://www.biocreative.org/tasks/biocreative-iv/

[7] http://www.nlm.nih.gov/mesh/

[8] http://www.comp.lancs.ac.uk/computing/research/stemming/

	System			Baseline		
Action Term	Precision	Recall	F-Score	Precision	Recall	F-Score
exp	0.81	0.77	0.80	0.73	0.66	0.70
rec	0.66	0.72	0.69	0.72	0.13	0.22
act	0.65	0.70	0.67	0.46	0.81	0.59
b	0.50	0.87	0.64	0.29	0.12	0.17
rxn	0.48	0.51	0.50	0.37	0.16	0.22
pho	0.30	0.77	0.44	0.61	0.61	0.61
abu	0.23	0.88	0.37	0.13	0.02	0.03
oxd	0.18	1	0.3	0.04	1	0.07
Macro-Average	0.59	0.64	0.61	0.34	0.44	0.38

Table 3: Evaluation scores of the action term recognition system (System) as compared to the stem baseline (Baseline)

not have a lot of training material available (for example for oxidation only 1171 abstracts) but still show a satisfactory performance. This can be due to the fact that this particular classifier showed a high recall an there were only 3 occurrences of "oxidation" to be found in the test set. Furthermore, for the classifiers for which there is significantly less training material than the default of 2000 abstracts, it can be assumed that the classifier suffers from overfitting since in these cases, the same abstracts that are part of the test set have already been used for training the classifier. This is due to the fact that within the BioCreative competition, the PubMed IDs of the test set have only been released at a point when the classifiers had already been built. However, with the continuous growth of CTD, it is very likely that for most action terms overfitting will not be an issue anymore.

5. Outlook: Future Work

As a next step, we are implementing a system capable of discovering the full interactions, composed of action term and actors. The CTD database, consisting of 328.230 manually curated interactions, constitutes a valuable resource in training such a system. The work presented in this paper represents an important step in this direction.

The classification system that we described can, in a following step, be used for discovering where exactly an interaction is located in the text. After an abstract has been categorized this can be done in two possible ways. Firstly, the system itself can be used for classifying shorter pieces of text, as for example sentences, to find out if an interaction is present. Secondly, the classifier delivers a list of most informative features for a specific action term. These most informative features can also be used separately for identifying the immediate context of an interaction. Together with a powerful tool for named entity extraction in the biomedical domain, such as for example Gimli (Campos et al., 2013), Metamap (Aronson, 2001) or the Ontogene NER module (Kaljurand et al., 2009), our aim is to be able to locate the exact anchorings of the interaction in CTD in the text of an abstract. As soon as the exact anchoring of an interaction from CTD together with all its actors can be found in the text of an abstract, the obtained data can in turn be used as training data for a weakly supervised system. The aim of such a system is to be able to discover complete biomedical interactions from unseen abstracts.

6. Conclusion

As could be seen by the description of CTD, large databases of the biomedical domain can provide a good and applicable alternative to annotated corpora. The big advantages over traditional annotation methods are the efficiency regarding time and money as well as the amount of available information. We showed how the data can be combined relatively easily with the abstracts of the scientific papers from which it was curated, in order to obtain a language resource that can be used in building statistical systems for biomedical text mining. These systems, in turn, can then be used in the future for supporting the process of biocuration.

The system that we presented as a case study has the purpose of recognizing interactions in the form of action terms within scientific abstracts from the biomedical domain. We treated this task as a problem of text classification and found that the resource was very suitable for this approach. The work presented in this paper will be taken as a starting point for finding complete biomedical interaction. Within this future work, CTD will serve as an important resource for providing distant supervision of the interaction detection system.

7. Acknowledgements

The OntoGene group 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.

8. References

Aronson, A. R. (2001). Effective mapping of biomedical text to the umls metathesaurus: the metamap program. *Proc AMIA Symp*, pages 17–21.

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly, Beijing.

Campos, D., Matos, S., and Oliveira, J. L. (2013). Gimli: open source and high-performance biomedical name recognition. *BMC Bioinformatics*, 14(1):54.

Craven, M. and Kumlien, J. (1999). Constructing biological knowledge bases by extracting information from text sources. *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology*, pages 77–86.

Davis, A. P., Murphy, C. G., Saraceni-Richards, C. A., Rosenstein, M. C., Wiegers, T. C., and Mattingly, C. J. (2009). Comparative toxicogenomics database: a knowledgebase and discovery tool for chemical-gene-disease networks. *Nucleic Acids Research*, 37(Database-Issue):786–792.

Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S. E., Vingron, M., Roechert, B., Roepstorff, P., Valencia, A., Margalit, H., Armstrong, J., Bairoch, A., Cesareni, G., Sherman, D. J., and Apweiler, R. (2004). Intact: an open source molecular interaction database. *Nucleic Acids Research*, 32(Database-Issue):452–455.

Kaljurand, K., Rinaldi, F., Kappeler, T., and Schneider, G. (2009). Using existing biomedical resources to detect and ground terms in biomedical literature. In Combi, C., Shahar, Y., and Abu-Hanna, A., editors, *AIME*, volume 5651 of *Lecture Notes in Computer Science*, pages 225–234.

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

Kotridis, P., Kokkas, B., Karamouzis, M., Sakadamis, G., Kanonidis, I., Dadous, G., Karantona, C., Gouli, O., Karadoumanis, J., Papadopoulos, P. C., and Papadopoulos, C. L. (2002). Plasma atrial natriuretic peptide in essential hypertension after treatment with irbesartan. *Blood Press.*, 11(2):91–94.

Krallinger, M., Leitner, F., Rodriguez-Penagos, C., and Valencia, A. (2008). Overview of the protein-protein interaction annotation extraction task of BioCreative II. *Genome Biology*, 9(Suppl 2):S4.

Leitner, F., Mardis, S. A., Krallinger, M., Cesareni, G., Hirschman, L. A., and Valencia, A. (2010). An overview of biocreative ii.5. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 7(3):385–399.

Mintz, M., Bills, S., Snow, R., and Jurafsky, D. (2009). 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*, 2(2005):1003.

Morgan, A. A., Hirschman, L., Colosimo, M., Yeh, A. S., and Colombe, J. B. (2004). Gene name identification and normalization using a model organism database. *Journal of Biomedical Informatics*, 37(6):396–410.

Rinaldi, F., Kappeler, T., Kaljurand, K., Schneider, G., Klenner, M., Clematide, S., Hess, M., von Allmen, J.-M., Parisot, P., Romacker, M., and Vachon, T. (2008). OntoGene in BioCreative II. *Genome Biology*, 9(Suppl 2):S13.

Rinaldi, F., Schneider, G., Kaljurand, K., Clematide, S., Vachon, T., and Romacker, M. (2010). OntoGene in BioCreative II.5. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 7(3):472–480.

Rinaldi, F., Clematide, S., Ellendorff, T. R., and Marques, H. (2013a). Ontogene: Ctd entity and action term recognition. In *Proceedings of the Fourth BioCreative Challenge Evaluation Workshop*, volume 1, pages 90–94.

Rinaldi, F., Davis, A. P., Southan, C., Clematide, S., Ellendorff, T. R., and Schneider, G. (2013b). ODIN: a customizable literature curation tool. In *Proceedings of the Fourth BioCreative Challenge Evaluation Workshop*, volume 1, pages 219–223.

UniProt Consortium. (2007). The universal protein resource (uniprot). *Nucleic Acids Research*, 35:D193–7.