Collection-Wide Extraction of Protein-Protein Interactions

Lenz Furrer* lenz.furrer@uzh.ch
Simon Clematide* siclemat@cl.uzh.ch
Hernani Marques* hernani.marquesmadeira@uzh.ch
Raul Rodriguez-Esteban† raul.rodriguez-esteban@roche.com
Martin Romacker† martin.romacker@roche.com
Fabio Rinaldi* fabio.rinaldi@uzh.ch

Abstract

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 initial 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 goal of extracting protein-protein interactions from a large pool of supporting papers.

1 Introduction

The OntoGene group (http://www.ontogene.org/) 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 (Rinaldi et al., 2008; Rinaldi et al., 2010b; Rinaldi et al., 2013).

The Data Science group at Hoffmann-La Roche supports the development of projects in research and early development with the analysis, management and visualization of biological and chemical data using its expertise in chemoinformatics, text mining, data mining, information science, competitor information, pathway analysis and bioinformatics.

Recently the two groups initiated a collaboration aimed at the development of a system which is capable of automatically processing an input corpus of scientific articles. The system should be able to detect evidence for specific protein interactions described in the input documents. Given an input gene or protein, the system will locate 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 should be 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 should take into consideration not only the local evidence in each paper, but also the global evidence across the collection. In this paper we present the preliminary results of the collaboration.

2 Methods

The OntoGene pipeline is used to provide all the basic text mining capabilities that are needed for the successful realization of the project. The typical OntoGene approach is based on sourcing named entities (terms and identifiers) from one or several reference databases, and use them to annotate the target collection. In a second phase, candidate interactions among the detected entities are generated and they are scored against a target database.

Therefore the 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. In this

*University of Zurich, Institute of Computational Linguistics, Binzmühlestr. 14, 8050 Zürich, Switzerland
†pREDi (Pharma Research and Early Development Informatics), F. Hoffmann-La Roche Ltd, Basel, Switzerland

project, our choice has fallen upon BioGrid[1] (Dolinski et al., 2013) for its good coverage and quality curation. We used release 3.2.115 (June 25, 2014) which contains 747,514 relations and 55,495 interactors (protein entities). However, according to the specific requirements of the industrial partner, and due also to the limited duration of the project, we focused only on relations derived from physical experiments (no genetic), and considered only the human species. Besides, in order to avoid papers describing high-throughput experiments (which typically contain large number of interactions in table format), we selected only articles containing at most 12 curated relations. We were left with 50,784 relations, composed of 9,151 interactors, coming from 20,928 PubMed abstracts.

2.1 Entity recognition and generation of candidate interactions

The terminology derived from the entire BioGrid database is stored as an OntoGene internal lexical resource, and used to annotate the target documents. OntoGene will efficiently recognize all entity names from this resource, and associate them to their database identifiers. OntoGene 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, which can be inspected with a suitable interface, such as ODIN (OntoGene Document Inspector) (Rinaldi et al., 2010a). ODIN has been used in assisted curation experiments in collaboration with major databases, such as PharmGKB (Rinaldi et al., 2012), and RegulonDB (Gama-Castro et al., 2014).

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).

The specific collection that has been used for our experiments is a set 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.

Additionally we use a smaller collection of full text articles (877 documents), which is a random stratified sample of the set above, as we aim at achieving a similar distribution in the number of protein-protein interactions per article.

2.2 Ranking of candidate interactions

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 the OntoGene 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 individ-

[1] http://thebiogrid.org/

ual 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 $P(\text{gold}(A, c))$, i. e. the probability of concept c being part of a relevant relation in article A. The concept score σ_c^{ME} is computed as follows:

$$\sigma_c^{\text{ME}} = \sum_{t \in A | \langle c,t \rangle \in L} f^{\text{b}}(t, c) \times P(\text{gold}(A, c) | c, t, f^{\text{cap}}(t, c)) \tag{1}$$

where $f^{\text{b}}(t, c)$ is the boosted (location-sensitive) frequency of a term t which can be mapped to concept c, as defined by the term lexicon L, and $f^{\text{cap}}(t, c)$ is the capped frequency of t mapped to c. The details of this computations are given in (Clematide and Rinaldi, 2012).

The *sentence score* of a candidate pair (c_1, c_2) accounts for the linguistic context in which the terms of c_1 and c_2 are found. For each sentence s containing two or more terms, we computed its probability to express a gold interaction $P(\text{gold}(s))$. 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.

In order to ensure a certain confidence in the predicted probability value, $P(\text{gold}(s))$ is set to 0 if it does not equal or exceed a threshold t:

$$P^{\text{conf}}(\text{gold}(s)) = \begin{cases} P(\text{gold}(s)) & \text{if } P(\text{gold}(s)) \geq t \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

The sentence score σ^{NB} for a pair of protein concepts (c_1, c_2) is the sum of the confidence probability $P^{\text{conf}}(\text{gold}(s))$ for each shared sentence s. If c_1 and c_2 never appear in the same sentence, the score results in the back-off value σ_0^{NB}, which we currently set to 0:

$$\sigma_{c_1,c_2}^{\text{NB}} = \sigma_0^{\text{NB}} + \sum_{s \in \mathbb{S}_{c_1,c_2}} P^{\text{conf}}(\text{gold}(s)) \tag{3}$$

$\mathbb{S}_{c_1,c_2}$ is the set of sentences in A which have two distinct terms t_1 and t_2 that can be mapped to c_1 and c_2, respectively:

$$\begin{aligned} \mathbb{S}_{c_1,c_2} = \{ s \in A | \exists t_1, t_2 \in s : \\ \langle c_1, t_1 \rangle \in L \wedge \langle c_2, t_2 \rangle \in L \wedge t_1 \neq t_2 \} \end{aligned} \tag{4}$$

The two different scores are then combined into a *relation score*. For each combination (c_1, c_2), we compute a relation score based on the harmonic mean of the two concept scores $\sigma_{c_1}^{\text{ME}}$ and $\sigma_{c_2}^{\text{ME}}$ and the sentence score $\sigma_{c_1,c_2}^{\text{NB}}$ for all occurrences of c_1 and c_2:

$$\sigma_{c_1,c_2}^{\text{REL}} = (1 - \lambda) \frac{2\sigma_{c_1}^{\text{ME}} \sigma_{c_2}^{\text{ME}}}{\sigma_{c_1}^{\text{ME}} + \sigma_{c_2}^{\text{ME}}} + \lambda \sigma_{c_1,c_2}^{\text{NB}} \tag{5}$$

where λ is a value $0 \leq \lambda \leq 1$, which was used in the experiments to give different weights to the concept score and sentence score (with $\lambda = 0$ indicating that only the concept score should be used, and $\lambda = 1$ selecting only the sentence score).

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 article A, protein c_1 interacts with protein c_2. 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.

3.1 Entity recognition quality

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

	OntoGene	OntoGene 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.

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.

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 (Campos et al., 2013), 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.

3.2 Interaction recognition quality

For evaluating the quality of the interaction recognition, we used *Threshold Average Precision* (TAP-k) (Carroll et al., 2010), 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 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.

λ	mean	min	max	sd
0.0	0.229	0.224	0.234	0.010
0.1	0.228	0.223	0.233	0.010
0.2	0.226	0.222	0.230	0.009
0.3	0.225	0.221	0.229	0.008
0.4	0.224	0.220	0.228	0.008
0.5	0.223	0.219	0.227	0.008
0.6	0.222	0.218	0.225	0.008
0.7	0.221	0.217	0.225	0.008
0.8	0.219	0.215	0.223	0.008
0.9	0.217	0.213	0.221	0.008
1.0	0.041	0.039	0.042	0.002

Table 2: Average TAP-10 values for different settings of the λ weight. All results used a sentence confidence threshold $t = 0.9$.

Table 2 shows the average values of TAP-10 (mean, minimum, maximum and standard deviation of the 10 cross-validation folds) for different values of the parameter λ, which indicates the relative weight of the concept score and the sentence score. The text collection used in this experiment is comprised of the selection of 20,928 articles mentioned above, whereof 877 are full text. The values shown are averaged across the TAP-10 values for each document.

The best overall result is achieved with the concept scores only. However, setting λ to a moderate value of 40–70 % still yields reasonable results. This might be explained by the observation that the sentence score has only little influence at the top end of the ranking list: Inspection of the ranked interactions shows that many highly-ranked false positives contain concepts that are false interpretations of the correct surface terms. Since different concepts of the same term mention also point to the same sentences, the sentence score cannot help deciding these cases.

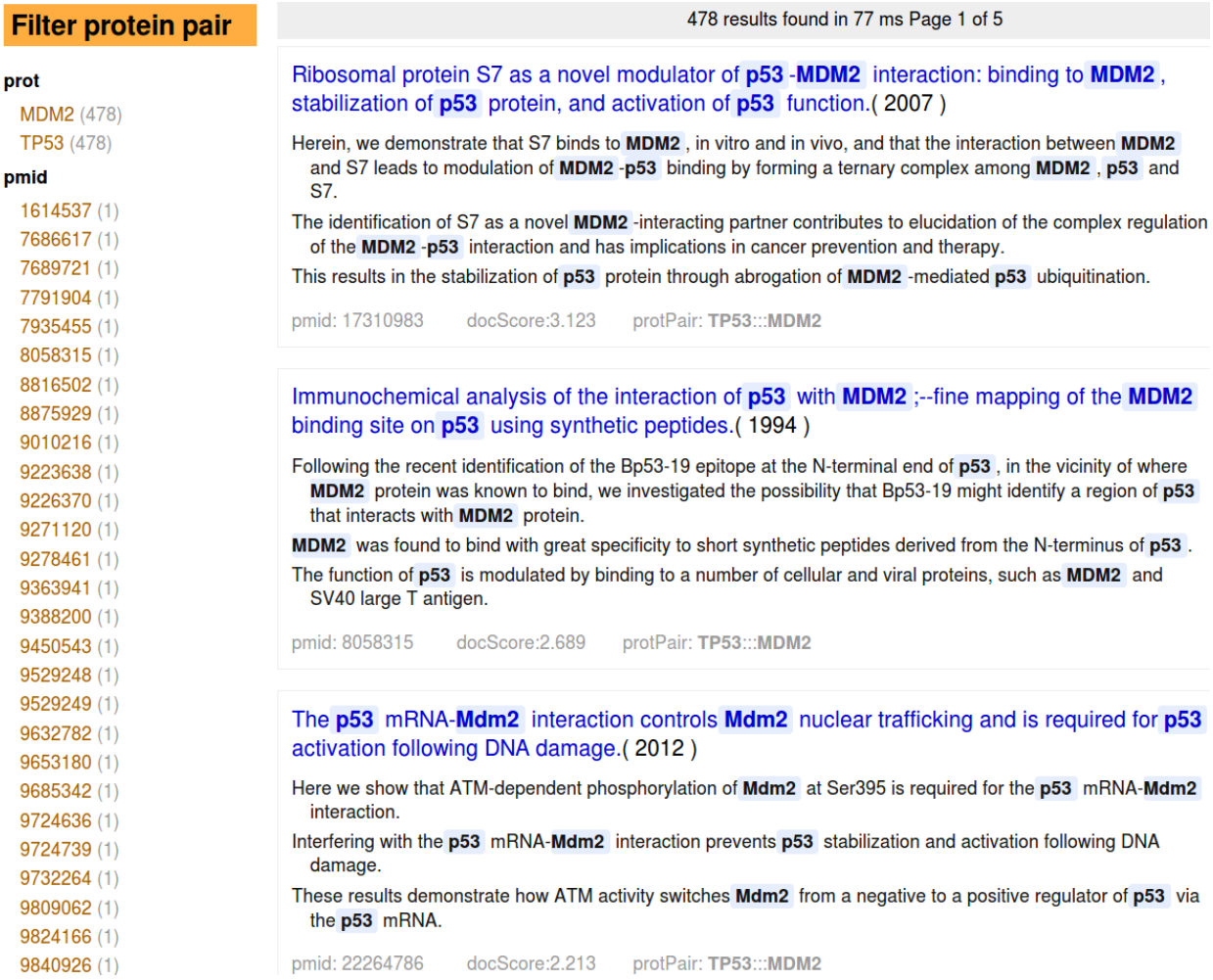

Figure 1: Example showing top-ranked snippets for the interaction TP53 - MDM2

We expect the sentence score to be more helpful in a scenario where documents and/or snippets of textual evidence are ranked, given a pair of protein concepts – rather than ranking interactions, given a document. Such a scenario needs to be evaluated manually by domain experts. While an evaluation by domain experts is planned, at this stage we do not have yet results to report. Figure 1 shows the current version of the interface that will allow examination of the results. The user can enter two arbitrary proteins, and the system will deliver the textual snippets which are considered to be most relevant for that particular interaction.

4 Conclusion

We have presented the preliminary results of an academic-industrial collaboration which is aimed at obtaining support for protein interactions from a very large collection of papers. The project has already developed some innovative approaches to the combination of evidence from large quantities of supporting publications.

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.

References

[Campos et al.2013] David Campos, Sérgio Matos, and José L Oliveira. 2013. A modular framework for biomedical concept recognition. *BMC bioinformatics*, 14(1):281.

[Carroll et al.2010] Hyrum D. Carroll, Maricel G. Kann, Sergey L. Sheetlin, and John L. Spouge. 2010. Threshold average precision (tap-k): a measure of retrieval designed for bioinformatics. *Bioinformatics*, 26(14):1708–1713.

[Clematide and Rinaldi2012] Simon Clematide and Fabio Rinaldi. 2012. Ranking relations between diseases, drugs and genes for a curation task. *J. Biomedical Semantics*, 3(S-3):S5.

[Dolinski et al.2013] K. Dolinski, A. Chatr-Aryamontri, and M. Tyers. 2013. Systematic curation of protein and genetic interaction data for computable biology. *BMC Biol.*, 11:43.

[Gama-Castro et al.2014] Socorro Gama-Castro, Fabio Rinaldi, Alejandra Lpez-Fuentes, Yalbi Itzel Balderas-Martnez, Simon Clematide, Tilia Renate Ellendorff, Alberto Santos-Zavaleta, Hernani Marques-Madeira, and Julio Collado-Vides. 2014. Assisted curation of regulatory interactions and growth conditions of OxyR in E. coli K-12. *Database: The Journal of Biological Databases and Curation*, bau049.

[Rinaldi et al.2008] Fabio Rinaldi, Thomas Kappeler, Kaarel Kaljurand, Gerold Schneider, Manfred Klenner, Simon Clematide, Michael Hess, Jean-Marc von Allmen, Pierre Parisot, Martin Romacker, and Therese Vachon. 2008. OntoGene in BioCreative II. *Genome Biology*, 9(Suppl 2):S13.

[Rinaldi et al.2010a] Fabio Rinaldi, Simon Clematide, Gerold Schneider, Martin Romacker, and Therese Vachon. 2010a. ODIN: An advanced interface for the curation of biomedical literature. In *Biocuration 2010, the Conference of the International Society for Biocuration and the 4th International Biocuration Conference.*, page 61. Available from Nature Precedings `http://dx.doi.org/10.1038/npre.2010.5169.1`.

[Rinaldi et al.2010b] Fabio Rinaldi, Gerold Schneider, Kaarel Kaljurand, Simon Clematide, Therese Vachon, and Martin Romacker. 2010b. OntoGene in BioCreative II.5. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 7(3):472–480.

[Rinaldi et al.2012] Fabio Rinaldi, Simon Clematide, Yael Garten, Michelle Whirl-Carrillo, Li Gong, Joan M. Hebert, Katrin Sangkuhl, Caroline F. Thorn, Teri E. Klein, and Russ B. Altman. 2012. Using ODIN for a PharmGKB revalidation experiment. *Database: The Journal of Biological Databases and Curation*.

[Rinaldi et al.2013] Fabio Rinaldi, Simon Clematide, Simon Hafner, Gerold Schneider, Gintare Grigonyte, Martin Romacker, and Therese Vachon. 2013. Using the OntoGene pipeline for the triage task of BioCreative 2012. *The Journal of Biological Databases and Curation, Oxford Journals*.