Large-scale Extraction of Gene Regulation for Model Organisms in an ontological context

Jasmin Šarić
European Media Laboratory GmbH
Heidelberg, D-69118
Germany
saric@eml-r.org

Lars J. Jensen
European Molecular Biology Laboratory
Heidelberg, D-69117
Germany
jensen@embl.de

Isabel Rojas
European Media Laboratory GmbH
Heidelberg, D-69118
Germany
rojas@eml-r.org

Abstract

This paper presents an approach using syntacto-semantic rules for the extraction of relational information from biomedical abstracts. The results show that by overcoming the hurdle of technical terminology, high precision results can be achieved. From abstracts related to baker's yeast, we manage to extract a regulatory network comprised of 441 pairwise relations from 58,664 abstracts with an accuracy of 83–90%. To achieve this, we made use of a resource of gene/protein names considerably larger than those used in most other biology related information extraction approaches. This list of names was included in the lexicon of our retrained part-of-speech tagger for use on molecular biology abstracts. For the domain in question an accuracy of 93.6–97.7% was attained on Part-of-speech-tags. The method can be easily adapted to other organisms than yeast, allowing us to extract many more biologically relevant relations. The main reason for the comparable precision rates is the ontological model that was built beforehand and served as a guiding force for the manual coding of the syntacto-semantic rules.

Preliminary results on journal articles from PubMed Central suggest that our rule set performs with equal accuracy when applied to full text rather than abstracts.

1 Introduction and related work

A massive amount of biological information is buried in scientific publications (more than 500,000 publications per year) or comment lines in biological databases. Therefore, the need to extract information in the life sciences is drastically increasing. Most of the ongoing work in the field of computational linguistics is being dedicated to deal with PubMed[1] abstracts. In the field of text mining the technical terminology of biomedicine presents the main challenge of applying IE (information extraction) to such a corpus (Hobbs, 2003).

The goal of our work is to extract from biological abstracts information on which *proteins* are responsible for regulating the expression (*i.e.* transcription or translation) of which *genes* on a general organism independent level. In contrast to the BioCreAtIvE competition tasks[2] that aimed at classifying entities, we thus focus on extracting a specific type of relations between biological entities.

Most NLP (Natural Language Processing) based studies tend to have been focused on extraction of events involving one particular verb, *e.g. bind* (Thomas *et al.*, 2000) or *inhibit* (Pustejovsky *et al.*, 2002). From a biological point of view, the problems of such an approach are two-fold: 1) the meaning of the extracted events will depend strongly on the selectional restrictions and 2) the the same meaning can be expressed using a number of different verbs. In contrast to this and comparable to (Friedman *et al.*, 2001), we instead aim at extracting events related to a specific biological problem only, but attempt to do so for all syntactic variations.

The variety in the biological terminology used to describe regulation of gene expression presents a major hurdle to an IE approach; in many cases the information is buried to such an extent that even a human reader is unable to extract it unless having a scientific background in biology. In this paper we will show that by overcoming the terminological barrier, high precision extraction of entity relations can be achieved within the field of molecular biology. We furthermore show that a rule based system developed for dealing with a particular organism, in our case baker's yeast (details on this system are described in (Šarić *et al.*, 2004)), can be easily adapted to other organisms with no loss of accuracy. We present as well preliminary results from applying our method to full text articles. Finally we close

[1] PubMed is a bibliographic database covering life sciences with a focus on biomedicine, comprising around 12×10^6 articles, roughly half of them including abstract (http://www.ncbi.nlm.nih.gov/PubMed/).

[2] Critical Assessment of Information Extraction systems in Biology, http://www.mitre.org/public/biocreative/

with a discussion of the ontological issues that have to be met when automatising the extraction system.

2 The biological task and our approach

To extract relations, the named entities[3] involved must first be recognised. This is particularly difficult in molecular biology where many forms of variation frequently occur. Synonymy is very frequent due to lack of standardisation of gene names; **BYP1**, **CIF1**, **FDP1**, **GGS1**, **GLC6**, **TPS1**, **TSS1**, and **YBR126C** are all synonyms for the same gene/protein. Additionally, these names are subject to orthographic variation originating from differences in capitalisation and hyphenation as well as syntactic variation of multiword terms (*e.g. riboflavin synthetase beta chain = beta chain of riboflavin synthetase*). Homonymy is frequent too since a gene and its gene product are usually named identically, causing cross-over of terms between semantic classes. Finally, paragrammatical variations are more frequent in life science publications than in common English due to the large number of publications by non-native speakers (Netzel *et al.*, 2003). Other difficulties related to ontological issues like coordination are addressed in a separate section (see section 5).

Extracting that a *protein* regulates the expression of a *gene* is a challenging problem as this fact can be expressed in a variety of ways—possibly mentioning neither the biological process (*expression*) nor any of the two biological entities (*genes* and *proteins*). Figure 1 shows a simplified ontology providing an overview of the biological entities involved in gene expression, their ontological relationships, and how they can interact with one another. An ontology is a great help when writing extraction rules, as it immediately suggests a large number of relevant relations to extract. Examples include "*promoter* contains *upstream activating sequence*" and "*transcription regulator* binds to *promoter*", both of which follow from indirect relationships via *binding site*.

It is often not known whether the regulation takes place at the level of transcription, translation, or by an indirect mechanism. We thus decided against trying to extract how the regulation of expression takes place, however, we strictly require that the extracted relations provide information about a regulatory protein (**R**) regulating the expression of a target gene (**X**):

1. It must be ascertained that the sentence mentions gene expression. "The protein **R** activates **X**" fails this requirement, as **R** might instead activate **X** post-translationally. Whether the event should be extracted or not thus depends on the type of the accusative object **X** (*e.g. gene* or *gene product*). Without a head noun specifying the type, **X** remains ambiguous and thus the whole relation remains underspecified and should not be extracted. It should be noted that two thirds of the gene/protein names mentioned in our corpus are ambiguous for this reason.
2. The identity of the regulator (**R**) must be known. "The **X** promoter activates **X** expression" fails this requirement, as it is not known which protein activates the expression of gene **X**. Linguistically this implies that noun chunks of certain semantic types should be disallowed as agent.
3. The identity of the target (**X**) must be known. "The transcription factor **R** activates **R** dependent expression" fails this requirement, as it is not know which gene's expression is dependent on **R**. The semantic types allowed for theme should thus also be restricted.

The two last requirements are important to avoid extraction from non-informative sentences that—despite them containing no information—occur quite frequently in scientific abstracts.

[3] Named entities can be considered as instances of concepts. *26 Kda heat shock protein* is thus an instance of the concept protein.

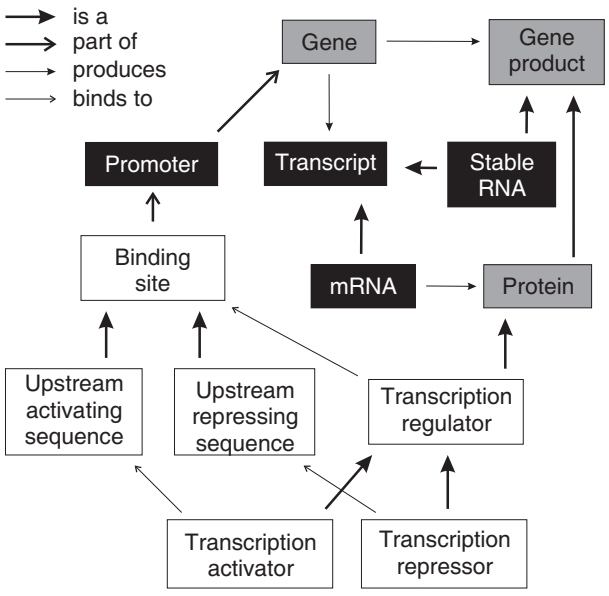

Figure 1: **A simplified ontology for transcription regulation.** The background color used for each term signifies its semantic role in relations: regulator (white), target (black), or either (gray).

The ability to genetically modify an organism brings with it an added complication to IE: biological texts often mention what takes place when an organism is artificially modified in a particular way. In some cases such modification can reverse part of the meaning of the verb: from the sentence "Deletion of **R** increased **X** expression" one can conclude that **R** represses expression of **X**. In other cases the verb will lose part of its meaning: "Mutation of **R** increased **X** expression" implies that **R** regulates expression **X**, but we cannot infer whether **R** is an activator or a repressor. Finally, there are those relations that should be completely avoided as they exist only because they have been artificially introduced through genetic engineering. In our extraction method we address all three cases.

We have opted for a rule based approach (implemented as cascaded finite state automata) to extract the relations for two reasons. The first is, that a rule based approach allows us to directly ensure that the three requirements stated above are fulfilled for the extracted relations. This is desired to attain high precision on the extracted relations, which is what matters to the biologist. Hence we focus in our evaluation on the semantic correctness of our method rather than on the grammatical correctness. As long as grammatical errors do not result in semantic errors, we do not consider it an error. Conversely, even a grammatically correct extraction is considered an error if it is semantically incorrect.

Our second reason for choosing a rule based approach is that our approach is theory-driven and highly interdisciplinary, involving computational linguists, bioinformaticians, and biologists. The rule based approach allows us to benefit more from the interplay of scientists with different backgrounds, as known biological constraints can be explicitly incorporated in the extraction rules, which is reflected in the ontology. Compared to statistical methods, it is also less prone to being biased by the choice of training data set, allowing the method to better generalise to other corpora, *e.g.* different organisms or full text papers instead of abstracts.

3 Methods

Our IE system is organised in cascaded modules such that the output of one module is the input of the next module. The following sections describe each module in detail.

3.1 The corpus

The PubMed resource was downloaded on January 19, 2004. 58,664 abstracts related to the yeast *Saccharomyces cerevisiae* were extracted by looking for occurrences of the terms "Saccharomyces cerevisiae", "S. cerevisiae", "Baker's yeast", "Brewer's yeast", and "Budding yeast" in the title/abstract or as head of a MeSH term[4]. These abstracts were filtered to obtain the 15,777 that mention at least two names (see section 3.4) and subsequently divided into a training and an evaluation set of 9137 and 6640 abstracts respectively.

3.2 Tokenisation and multiword detection

The process of tokenisation consists of two steps (Grefenstette & Tapanainen, 1994): segmentation of the input text into a sequence of tokens[5] and the detection of sentential boundaries. We use the tokenizer developed by Helmut Schmid at IMS (University of Stuttgart) because it combines a high accuracy (99.56% on the Brown corpus) with unsupervised learning (*i.e.* no manually labelled data is needed) (Schmid, 2000).

The determination of token boundaries in technical or scientific texts is one of the main challenges within information extraction or retrieval. On the one hand, technical terms contain special characters (like brackets, colons, hyphens, slashes, etc.). On the other hand, they often appear as multiword expressions which makes it hard to detect the exact left and right boundaries of the terms. Although a lot of work has been invested in the detection of technical terms within biology related texts (see (Nenadić *et al.*, 2003) or (Yamamoto *et al.*, 2003) for representative results) this task is not yet solved to a satisfying extent. As we are interested in very special terms and high precision results we opted for multiword detection based on semi-automatical acquisition of multiwords (see sections 3.4 and 3.5).

3.3 Part-of-speech tagging

To improve and analyse the improvement of the accuracy of POS-tagging on PubMed abstracts, TreeTagger (Schmid, 1994) was trained on three different corpora. Each training results in a specific parameter file. First, we used the standard english parameter file (Penn Treebank[6] trained) without expanding the lexicon. A second parameter file was generated by training with the GENIA 3.0 corpus (Kim *et al.*, 2003) with an expanded lexicon containing gene names (see section 3.4) and multiwords (see section 3.5). The third version was generated

[4]Medical Subject Headings (MeSH) is a controlled vocabulary for manually annotating PubMed articles.

[5]This is also referred to as detection of word boundaries, where a word unit can refer to a multiword like *upstream activating factor*.

[6]Details on the Penn Treebank can be found in (Marcus *et al.*, 1993).

from a revised version of GENIA 3.0 described in the next paragraph.

The GENIA 3.0 corpus consists of PubMed abstracts and has 466,179 manually annotated tokens. For the last retraining experiment we made three kind of revisions: The first concerns seemingly undecidable cases like *in/or* annotated as `in|cc`, which were split into three tokens: *in*, */*, and *or* each annotated with its own tag. The second set of changes is a refinement of the GENIA tagset concerning auxiliary verbs to distinguish between *be* verbs (`vb...`), *have* verbs (`vh...`) and other verbs `vv...`. The third set of changes removed inconsistencies like `in` annotated as `DT`.

3.4 Recognising gene/protein names

To be able to recognise gene/protein names as such, and to associate them with the appropriate database identifiers, a list of synonymous names and identifiers in six eukaryotic model organisms was compiled from several sources (`http://www.bork.embl.de/synonyms/`). For *S. cerevisiae* specifically, 51,640 uniquely resolvable names and identifiers were obtained from Saccharomyces Genome Database (SGD) and SWISS-PROT (Dwight *et al.*, 2002; Boeckmann *et al.*, 2003)[7].

Before matching these names against the POS-tagged corpus, the list of names was expanded to include different orthographic variants of each name. Firstly, the names were allowed to have various combinations of uppercase and lowercase letters: all uppercase, all lowercase, first letter uppercase, and (for multiword names) first letter of each word uppercase. In each of these versions, we allowed whitespace to be replaced by hyphen, and hyphen to be removed or replaced by whitespace. In addition, from each gene name a possible protein name was generated by appending the letter `p`. The resulting list containing all orthographic variations comprises 516,799 entries.

The orthographically expanded name list was fed into the multiword detection, the POS-tagger lexicon, and was subsequently matched against the POS-tagged corpus to retag gene/protein names as such (`nnpg`). To reduce the problem of homonymy, only matches to words tagged as common nouns (`nn`) were accepted.

3.5 Semantic tagging

In addition to the recognition of the gene and protein names, we recognise several other terms and annotate them with semantic tags. This set of semantically relevant terms mainly consists of nouns and verbs, and some few prepositions like `from`, or adjectives like `dependent`. The first main set of nominal terms is classified as follows:

- Nouns representing highly relevant concepts like *gene*, *protein*, *promoter*, *binding site*, *transcription factor*, etc. (153 entries).
- Nouns triggering an experimental or artificial contexts like *mutation*, *deletion*, *fusion*, *defect*, etc. (11 entries).
- Enzyme names like *elongase*, *hexokinase*, etc. (569 entries).
- Species/organism names extracted from the NCBI taxonomy of organisms (Wheeler *et al.*, 2004) (20,746 entries).
- Relational nouns, like nouns of activation (*e.g. derepression* and *positive regulation*), nouns of repression (*e.g. suppression* and *negative regulation*), nouns of regulation (*e.g. affect* and *control*) (69 entries).

The second set of verbal terms contains 50 entries plus inflectional variants. These are crucial for the extraction of relations between entities. The following shows the classification according to their relevance in gene transcription:

- Verbs of activation *e.g. enhance*, *increase*, *induce*, and *positively regulate*.
- Verbs of repression *e.g. block*, *decrease*, *down-regulate*, and *down regulate*.
- Verbs of regulation *e.g. affect* and *control*.
- Other selected verbs like *code* (or *encode*) and *contain* where given their own semantic tags.

Each of the terms consisting of more than one word was utilised for multiword detection.

We also have two additional classes of words to prevent false positive extractions. One class contains words of negation, like *not*, *cannot*, etc. The other contains nouns that are to be distinguished from other common nouns to avoid them being recognised as named entities, *e.g. allele* and *diploid*.

3.6 Extraction of named entities

In the preceding steps we classified relevant nouns according to semantic criteria. This allows us to chunk noun phrases generalising over both POS-tags and semantic tags. This syntacto-semantic chunking was performed to recognise named entities using cascades of finite state rules implemented as a CASS grammar (Abney, 1996). As an example

[7] Operon names were not taken into account, since no complete list of operon names is available.

we recognise gene noun phrases:

[$_{nx\_gene}$
[$_{dt}$ the]
[$_{nnpg}$ CYC1]
[$_{gene}$ gene]
[$_{in}$ in]
[$_{yeast}$ Saccharomyces cerevisiae]]

Other syntactic variants such as "the glucokinase gene **GLK1**" are recognised too. Analogously, we detect at this early level noun chunks denoting other biological entities like proteins, activators, repressors, transcription factors etc.

In subsequent cascades, we recognise more complex noun chunks on the basis of the simpler ones, such as promoters, upstream activating/repressing sequences (UAS/URS), binding sites, etc. At this point it becomes important to distinguish between agent and theme forms of noun chunks. A binding site, for example, is part of a target gene, the name of this gene or by the name of the regulator protein that binds to it. It is thus necessary to discriminate between "binding site of" and "binding site for".

As already mentioned, we have annotated a class of nouns that triggers experimental context. On the basis of these we identify noun chunks mentioning for example deletion, mutation, or overexpression of genes. At a fairly late stage we recognise events that occur in nominalisations like "expression of".

3.7 Extraction of relations between entities

This step of processing concerns the recognition of three types of relations between the recognised named entities: up-regulation, down-regulation, and (unspecified) regulation of expression. We combine syntactic properties (subcategorisation restrictions) and semantic properties (selectional restrictions) of the relevant verbs to map them to one of the three relation types.

The following shows a reduced bracketed structure consisting of three parts, a promoter chunk, a verbal complex chunk, and a UAS chunk (theme):

[$_{nx\_prom}$ the ATR1 promoter region]
[$_{contain}$ contains]
[$_{nx\_uas\_pt}$
[$_{dt-a}$ a] [$_{bs}$ binding site] [$_{for}$ for]
[$_{nx\_activator}$ the GCN4 activator protein]].

From this we extract that the **GCN4** protein activates the expression of the **ATR1** gene. We identify passive constructs, too, *e.g.* "**RNR1** expression is reduced by **CLN1** or **CLN2** overexpression". In this case we extract two pairwise relations, namely that both **CLN1** and **CLN2** down-regulate the expression of the **RNR1** gene. We also identify nominalised relations as exemplified by "the binding of **GCN4** protein to the **SER1** promoter in vitro"[8]

4 Results

Using our relation extraction rules, we were able to extract 422 relation chunks from our complete yeast corpus. As one entity chunk can mention several different named entities, these corresponded to a total of 597 extracted pairwise relations. However, as several relation chunks mention the same pairwise relations, this reduces to 441 unique pairwise relations comprised of 126 up-regulations, 90 down-regulations, and 225 regulations of unknown direction.

Figure 2 displays these 441 relations as a regulatory network in which the nodes represent genes or proteins and and the arcs are expression regulation relations. Known transcription factors according to the Saccharomyces Genome Database (SGD) (Dwight *et al.*, 2002) are denoted by black nodes. From a biological point of view, it is reassuring that the proteins serving as regulators in our relations tend to correspond to known *S. cerevisiae* transcription factors.

[8]It should be noted that no subordinate clause information gets extracted.

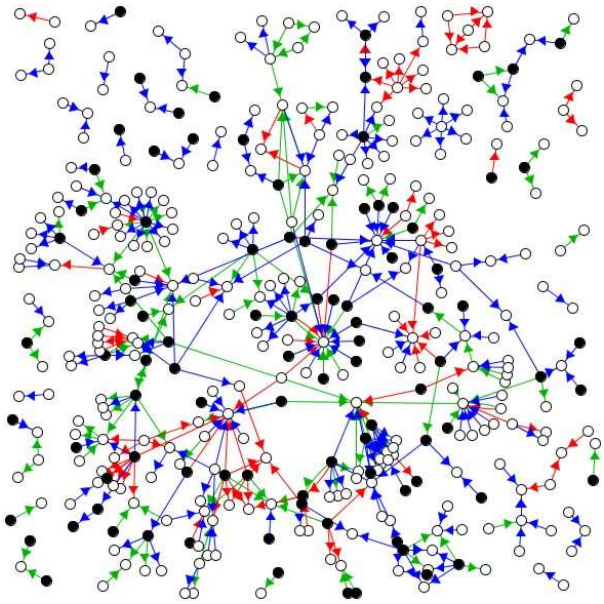

Figure 2: **The extracted network of gene regulation** The extracted relations are shown as a directed graph, in which each node corresponds to a gene or protein and each arc represents a pairwise relation. Arcs point from the regulator to the target and the type of regulation is specified by the color: up-regulation (green), down-regulation (red), and underspecified regulation (blue). Known transcription factors are highlighted as black nodes.

Table 1: Results

Organism	Abstracts	Relations	Prec.
E. coli	195,492	321	85%
B. subtilis	16,270	89	90%
S. cerevisiae	58,664	383	83%
M. musculus	688,937	1636	84%
PubMed central	5,075	158	84%

4.1 Evaluation of relation extraction

To evaluate the accuracy of the extracted relation, we manually inspected all relations extracted from the evaluation corpus using the TIGERSearch visualisation tool (Lezius, 2002). The results are shown in table 1.

The accuracy of the relations was evaluated at the semantic rather than the grammatical level. We thus carried out the evaluation in such a way that relations were counted as correct if they extracted the correct biological conclusion, even if the analysis of the sentence as not as to be desired from a linguistic point of view. Conversely, a relation was counted as an error if the biological conclusion was wrong.

75 of the 90 relation chunks (83%) extracted from the evaluation corpus were entirely correct, meaning that the relation corresponded to expression regulation, the regulator (**R**) and the regulatee (**X**) were correctly identified, and the direction of regulation (up or down) was correct if extracted. A further 6 relation chunks extracted the wrong direction of regulation but were otherwise correct; our accuracy increases to 90% if allowing for this minor type of error. Approximately half of the errors made by our method stem from genetic modifications being overlooked—the relation being extracted is actually mentioned in the sentence, however it is not biologically relevant.

To estimate the coverage of our method, we looked through 100 of the 44,354 yeast sentences that at least two gene/protein names. These contained only 5 relation chunks of the desired type, corresponding to an estimate of 2218 in total. Since 422 of these were successfully extracted by our method, we estimate the coverage of our method to be around 20%. This corresponds to an F-score in the order of 55%, which is respectable by IE standards[9].

A few extraction problems were encountered specifically for *E. coli*, the favoured bacterial species for experiments. Firstly, more errors are made due to artificial constructs since the most common reporter gene, *lacZ*, is itself an *E. coli* gene. Secondly, some abstracts are erroneously associated with *E. coli* hence associating the correct gene names but in the wrong species—this is considered an error since the same gene names is not guaranteed to refer to the same gene in different species.

[9]It should be noted that our approach does not resolve anaphoric relations like “this protein”. In addition if a gene/protein name is missed this sentence is not taken into account for the coverage estimation.

4.2 Entity recognition

For consistency, we have also evaluated our ability to correctly identify named entity at the level of semantic rather than grammatical correctness. Manual inspection of 500 named entities from the evaluation corpus revealed 14 errors, which corresponds to an estimated accuracy of just over 97%. Surprisingly, many of these errors were commited when recognising *proteins*, for which our accuracy was only 95%. Phrases such as “telomerase associated protein” (which got confused with “telomerase protein” itself) were responsible for about half of these errors.

Among the 153 entities involved in relations no errors were detected, which is fewer than should be expected from our estimated accuracy on entity recognition (99% confidence according to hypergeometric test). This suggests that the templates used for relation extraction are unlikely to match those sentence constructs on which the entity recognition goes wrong. False identification of named entities are thus unlikely to have an impact on the accuracy of relation extraction.

4.3 POS-tagging and tokenisation

We compared the POS-tagging accuracy of three parameter files on 24,798 held-out tokens from the GENIA corpus. The best result was achieved using the parameter trained on the revised GENIA corpus, which correctly tagged 96.4% of tokens. This compares favourably to the 93.6% and 85.7% correct tokens achieved using the parameter file for the standard GENIA corpus and the standard English parameter file respectively.

Of 1,068 punctuation marks we recognised 995 correctly as sentences boundaries and all 68 abbreviations correctly, too. This results in an overall precision of 99.5%.

5 Linguistics and ontologies

In the previous sections we have shown that a rule-based approach can be used to extract from biomedical abstracts information on regulation of gene expression. This highly relevant biological problem could be addressed for several model organisms with equal accuracy. The main adaptation required for this was to replace the list of synonymous

gene/protein names to reflect the change of organisms. These high quality results show that the rules that have been used reflect an underlying model, which is independent of the organism. This model is – as already introduced – depicted in Figure Figure 1.

Nonetheless it has to be noted that a major drawback from this rule-based IE approach is that the writing of rules is highly time consuming and scalable only to a limited extent. To reduce the temporal factor and to increase the scalability of the system the development of a a system that allows for (semi-)automatical interaction between ontological information and a rule-based IE system could prove to be highly useful. The process of extracting information from textual data thus should be ontology-driven.

Another advantage from an ontology-driven approach is related to what Schulze-Kremer calls the *communication problem in molecular biology* [(Schulze-Kremer, 1998)]. Aiming that in a subsequent stage the same ontology guides the integration of the extracted data one can ensure that consistency of data is much more likely than in other approaches.

To shortly explain what is meant by an ontology-driven IE system two layers have to be distinguished[10]. One layer regards the interplay between the ontological concepts and the lexical items (*i.e.* the words) in a text. To give an example, it has to be ensured that an entity recognised as a protein like *GCN4* in "the GCN4 activator protein" (taken from 1) is not linked to the concept *protein* only, but to the even more specific concept *transcription activator protein*. Of course, this concept inherits all properties of *protein*, but it has additional properties specified, *e.g.* that it has the role of activating gene expression.

Even more important, each relational word has to be associated with a relation in the ontology where the arguments are specified. The arguments of the ontological relation *activate* specify at least two possible pairs of semantic types. One, where both are proteins – and thus being part of the protein-protein interactions model – , and a second where one argument is a protein and the second is a gene part – and thus being part of the gene expression model –. For a sentence like example 1, the nominal phrase *the ATR1 promoter region* would be linked to the *promoter* concept, which is a part of the non-coding sequence of the *gene* conceptualisation.

(1) The GCN4 activator protein binds to the ATR1 promoter region.

(2) The ATR1 promoter region contains a binding site for the GCN4 activator protein.

(3) The binding of GCN4 protein to the SER1 promoter in vivo...

The second layer for ontology-driven IE systems that has to be taken into account concerns the clear conceptualisation in combination with the inferential power. Example 2 shows a syntactic variant expressing the same fact as 1. The advantage from a compositional ontology-driven approach is that it allows to treat different syntactic construction in that they are mapped to the same concepts. Of course, one can not judge whether example 3 is about activation or repression of gene expression. A concept *bind* which relates proteins and genes is underspecified with respect to activation or repression. Nonetheless the *binding* relation should be specified between proteins and DNA parts as part of the gene expression model and thus the binding in example 3 would be recognised as part of the gene expression model.

(4) Endotoxin increased NF-kappaB p50/p65 heterodimer binding.

Example 4 illustrates that in some cases combinations of different semantical issues are concerned. One issue concerns coordination and appositions, where the other issues concerns presuppositions. To correctly associate an ontological concept with the term *NF-kappaB p50/p65 heterodimer* a series of processing steps have to be performed. First, *heterodimer* is linked to the appropriate ontological category. This category comprises information that presupposes the existence of two entities *A* and *B*, both of type *protein*, as well as the information that $A \neq B$. Concerning the coordination a common system would try to identify these two entities and associate A with *NF-kappaB* and B with *p50/p65*. Unfortunately in a biological context this can easily fail. To determine the meaning correctly the system needs to know that the slash symbol is used in biological publications to express conjunctions, and, that *NF-kappaB* assigns the type of protein to both *p50* and *65*. The correct answer could then compositionally be computed with $A = p50$ and $B = p65$. This example shows that understanding of coordinated constructions is based on domain knowledge, that has to be fed into a NLP system.

6 Conclusions

We have developed a method that allows us to extract from biomedical abstracts information on reg-

[10] A detailed description of an ontology-driven information extraction system can be found in (Cimiano *et al.*, 2004).

ulation of gene expression. This is a highly relevant biological problem, since much is known about it although this knowledge has yet to be collected in a database. Also, knowledge on how gene expression is regulated is crucial for interpreting the enormous amounts of gene expression data produced by high-throughput methods like spotted microarrays and GeneChips.

Since we developed our method based on an ontological model for gene expression, our method is applicable to several model organisms with comparable accuracy. The main adaptation required for this was to replace the list of synonymous gene/protein names to reflect the change of organism. Furthermore, application of the method to full text journals gave promising preliminary results. We thus intend to systematically apply our rule based method to both abstracts and full text corpora for many more organisms including humans. Additionally, we are working on expanding the rules on a broader ontological model to also extract other, specific types of interactions between biological entities, reusing the many rules responsible for the recognition of named entities.

Acknowledgments

Jasmin Šarić and Isabel Rojas are funded by the Klaus Tschira Foundation gGmbH, Heidelberg (`http://www.kts.villa-bosch.de`). Lars Juhl Jensen is funded by the Bundesministerium für Forschung und Bildung, BMBF-01-GG-9817.

References

Abney, S. (1996). Partial parsing via finite-state cascades. Journal of Natural Language Engineering *2*(4), 337–344.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., & Schneider, M. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. *31*, 365–370.

Cimiano, P., Reyle, U., & Saric, J. (2004). Ontology driven discourse analysis for information extraction. Data and Knowledge Engeneering Journal.

Dwight, S. S., Harris, M. A., Dolinski, K., Ball, C. A., Binkley, G., Christie, K. R., Fisk, D. G., Issel-Tarver, L., Schroeder, M., Sherlock, G., Sethuraman, A., Weng, S., Botstein, D., & Cherry, J. M. (2002). Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). Nucleic Acids Res. *30*, 69–72.

Friedman, C., Kra, P., Yu, H., Krauthammer, M., & Rzhetsky, A. (2001). Genies: a natural language processing system for the extraction of molecular pathways from journal articles. Bioinformatics *17 Suppl.1*, 74–82.

Grefenstette, G. & Tapanainen, P. (1994). What is a word, what is a sentence? problems of tokenization. in *The 3rd International Conference on Computational Lexicography* pp. 79–87.

Hobbs, J. R. (2003). Information extraction from biomedical text. J. Biomedical Informatics *35*, 260–264.

Kim, J.-D., Ohta, T., Tateisi, Y., & Tsujii, J. (2003). GENIA corpus—a semantically annotated corpus for bio-textmining. Bioinformatics *19 suppl. 1*, i180–i182.

Lezius, W. (2002). TIGERSearch—ein Suchwerkzeug für Baumbanken. in *Proceedings der 6. Konferenz zur Verarbeitung natürlicher Sprache (KONVENS 2002)* (Busemann, S., ed.) Saarbrücken, Germany.

Marcus, M. P., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. Computational Linguistics, Special Issue on Using Large Corpora *19*(2), 273–290.

Nenadić, G., Rice, S., Spasić, I., & Ananiadou, S. an d Stapley, B. (2003). Selecting text features for gene name classification: from documents to terms. in *Proceedings of the ACL 2003 Workshop on Natural Language Processi ng in Biomedicine* (Ananiadou, S. & Tsujii, J., eds.) pp. 121–128.

Netzel, R., C., P.-I., Bork, P., & Andrade, M. A. (2003). The way we write. EMBO Rep. *4*, 446–451.

Pustejovsky, J., Castaño, J., Zhang, J., Kotecki, M., & Cochran, B. (2002). Robust relational parsing over biomedical literature: Extracting inhibit relations. Pac. Symp. Biocomput. *7*, 362–373.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. in *International Conference on New Methods in Language Processing* Manchester, UK. unknown.

Schmid, H. (2000). Unsupervised learning of period disambiguation for tokenisation Technical report Institut für Maschinelle Sprachverarbeitung, University of Stuttgart.

Schulze-Kremer, S. (1998). Ontologies for molecular biology. in *3rd Pacific Symposium on Biocomputing* pp. 693–704.

Thomas, J., Milward, D., Ouzounis, C., Pulman, S., & Carroll, M. (2000). Automatic extraction of protein interactions from scientific abstracts. Pacific Symposium on Biochemistry , 538–549.

Šarić, J., Jensen, L. J., Ouzounova, R., Rojas, I., & Bork, P. (2004). Extraction of regulatory gene expression networks from pubmed. in *Proceedings of the ACL 2004 Conference* pp. 192–199 Barcelona, Spain.

Wheeler, D. L., Church, D. M., Edgar, R., Federhen, S., Helmberg, W., L., M. T., U., P. J., D., S. G., M., S. L., Sequeira, E., Suzek, T. O., Tatusova, T. . A., & Wagner, L. (2004). Database resources of the national center for biotechnology information: update. Nucleic Acids Res. *32*, D35–40.

Yamamoto, K., Kudo, T., Konagaya, A., & Matsumoto, Y. (2003). Protein name tagging for biomedical annotation in text. in *Proceedings of the ACL 2003 Workshop on Natural Language Proces sing in Biomedicine* (Ananiadou, S. & Tsujii, J., eds.) pp. 65–72.