BIOINFORMATICS Bioinformatics Advance Access published July 26, 2005

Extraction of regulatory gene/protein networks from Medline

Jasmin Šarić [a*], *Lars Juhl Jensen* [b*], *Rossitza Ouzounova* [b], *Isabel Rojas* [a], *and Peer Bork* [b]

[a] *EML Research gGmbH, D-69118 Heidelberg, Germany*
[b] *European Molecular Biology Laboratory, D-69117 Heidelberg, Germany*

ABSTRACT

Motivation: We have previously developed a rule based approach for extracting information on the regulation of gene expression in yeast. The biomedical literature, however, contains information on several other equally important regulatory mechanisms, in particular phosphorylation, which we now expanded our rule based system to also extract.

Results: This paper presents new results for extraction of relational information from biomedical text. We have improved our system, STRING-IE, to both capture new types of linguistic constructs as well as new types of biological information (i.e. (de-)phosphorylation). The precision remains stable with a slight increase in recall. From almost one million PubMed abstracts related to four model organisms, we manage to extract regulatory networks and binary phosphorylations comprising **3319** relation chunks. The accuracy is **83–90%** and **86–95%** for gene expression and (de-)phosphorylation relations, respectively. To achieve this, we made use of an organism-specific resource of gene/protein names considerably larger than those used in most other biology related information extraction approaches. These names were included in the lexicon when retraining the part-of-speech tagger on the GENIA corpus. For the domain in question an accuracy of **96.4%** was attained on POS-tags. It should be noted that the rules were developed for yeast and successfully applied to both abstracts and full-text articles related to other organisms with comparable accuracy.

Availability: The revised GENIA corpus, the POS-tagger, the extraction rules, and the full sets of extracted relations are available from `http://www.bork.embl.de/Docu/STRING-IE`.

Contact: saric@eml-r.org

1 INTRODUCTION AND RELATED WORK

More and more scientific discoveries in the life sciences depend on the ability to identify and extract large amounts of data in scientific literature. Several groups have shown that it is possible to apply the engineering techniques from natural language processing to the biomedical domain, where the technical terminology is the major hurdle (Hobbs, 2003). There are two general approaches for extracting information from text: statistical and rule-based approaches. The former have shown good results for the detection of gene names within the BioCreAtIvE[1] or the NLPBA/BioNLP 2004[2] conferences. However, relation extraction is more problematic due to the lack of annotated biomedical corpora. While rule-based approaches usually are considered labour intensive and difficult to adapt to new domains, they are more transparent and thus semantic criteria can more easily be enforced.

In a previous study we developed a rule set for extracting a gene expression network for the yeast *Saccharomyces cerevisiae* (Saric *et al.*, 2004). We here present subsequent changes made to the system in order to 1) improve the recall by capturing linguistic structures previously missed, 2) extend the rule set to extract other types of relations than regulation of gene expression, and 3) allow the system to be applied to other organisms. All of these improvements are illustrated by the example "Lyn, but not Jak2, phosphorylated CrkL". Selective negation in coordinated structures ("A but not B") is one of the new linguistic structures handled by our rule set; we correctly extract that only Lyn phosphorylates the CrkL protein. Moreover, the phrase is concerned with phosphorylation of mouse proteins, meaning that we extract a type relation not previously detected by STRING-IE for a species the system was not developed for. Although the rules were originally developed for *S. cervisiae*, they should be applicable to other model organisms as well, since the only organism-specific part of our system is the list of protein/gene names. Here we show that our rule based system indeed performs equally

*These authors contributed equally

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

[2] The "Joint Workshop on Natural Language Processing in Biomedicine and its Applications" was held at the Coling 2004 conference in Geneva. The proceedings are available through `http://www.genisis.ch/\%7Enatlang/JNLPBA04/`.

well on *Escherichia coli*, *Bacillus subtilis*, and *Mus musculus*. Furthermore, we present preliminary results for a corpus of full text articles, namely PubMed Central.

The goal of our work is to extract from biological abstracts organism specific information on which *proteins* regulate the expression (i.e. transcription or translation) of which *genes* as well as which *proteins* modify which *proteins*.

A task closely related to ours, the extraction of protein–protein interactions from abstracts, has received some attention over the past five years but, with the notable exception of (Blaschke *et al.*, 1999), has been mainly addressed by statistical "bag of words" approaches (Marcotte *et al.*, 2001). Our work, on the other hand, is focused on extracting a specific type of relations between biological entities, instead of just classifying those entities as the BioCreAtIvE Project[3] does, and places emphasis on the semantic role of agent and theme. This is done with respect to the biological point of view for two main reasons: 1) the meaning of the extracted event is strongly dependent on the selectional restrictions of the verb and 2) the same meaning can be expressed using a number of different verbs. Unlike some competing approaches that are focused on extraction of events involving one particular verb, e.g. *bind* (Thomas *et al.*, 2000) or *inhibit* (Pustejovsky *et al.*, 2002), and similarly to (Friedman *et al.*, 2001)[4], we aim at extracting events related to a specific biological problem only, but considering all its syntactic variations.

The variety in the biological terminology used to describe the 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, can be easily adapted to other organisms with no loss of accuracy. Finally, we present preliminary results from applying our method to full text articles.

2 CASTING THE BIOLOGICAL TASK TO AN NLP PROBLEM

To extract relations, the named entities involved must first be recognized. This is particularly difficult in molecular biology where many forms of variation occur. Synonymy is very frequent due to lack of standardization 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 capitalization 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).

Extracting the fact that a given *protein* regulates a certain *gene* or *protein* through a particular mechanism is a challenging problem. First, there is the problem of syntactic variation, meaning that the same fact can be expressed in a variety of ways, e.g. active vs. passive voice. Second, the same verb can be used to expressed different types of relations, which is usually referred to as semantic variation. As an example the verb "activate" can equally well refer to regulation of gene expression (e.g. "**A** activates the expression of **B**) or to regulation of protein activity through phosphorylation or de-phosphorylation (e.g. "**A** activates **B** by phosphorylation").

In order for a relation to be extracted, we thus require that the type of regulation can be assigned and that the identity of both the regulatory protein (**R**) and the regulated target gene or protein (**X**) can be determined:

1. It must be ascertained that the sentence mentions either (de-)phosphorylation or regulation of gene expression. "The protein **R** activates **X**" fails this requirement, as there is no information on how **R** activates **X**. Whether the event should be extracted or not thus depends on the semantic types of the agent and theme; without head nouns specifying their types these remains ambiguous. 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. "Phosphorylation of the **X** protein activates **X**" fails this requirement, as it does not give the name of the protein that causes **X** to phosphorylated and hence activated.. Linguistically this implies that noun chunks of certain semantic types should be disallowed as agents.
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 known which gene's expression is dependent on **R**. The theme should thus also be restricted with respect to its semantic type.

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

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

[4] Although implementing a full-sentence parser they are successful on extracting the events that we are interested in, a direct comparison is not possible since only results on protein-protein interactions have been reported so far.

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, e.g. "transcription of the five mutated promoters". 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, because it allows us and to explicitly incorporate known biological constraints and to directly ensure that the three semantic requirements stated above are fulfilled for the extracted relations. Hence we also focus in our evaluation on the semantic correctness of our method rather than on the grammatical correctness. As long as a grammatical error do not result in semantic error, we do not consider it an error. Conversely, even a grammatically correct extraction is considered an error if it is semantically incorrect.

Compared to statistical methods, the rule based approach has the advantage of being able to generalize well to other corpora, as here shown by applying the same rule based extraction system to different organisms and to both abstracts and full text papers. Moreover, we show that by using a modular architecture where several independent relation extraction modules build on top of a common named entity recognition module, the rule based approach can be made highly scalable. New relation types can be added as separate modules, typically requiring only few changes to be made to the named entity recognition module. A modular architecture makes the system much easier to maintain as the system can be expanded without the risk of interference between complex rule sets.

3 METHODS

Our IE system is organized 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. With the notable exception of identification of gene/protein names, none of the modules required changes in order to be applied to other organisms.

3.1 The corpora

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[5]. These abstracts were filtered to obtain the 15,777 that mention at least two names and subsequently divided into a training and an evaluation set of 9137 and 6640 abstracts respectively.

Analogously, corpora were created for *Escherichia coli*, *Bacillus subtilis*, and *Mus musculus*. These were extracted by looking for both the full and abbreviated genus name (e.g. *E. coli*). In the case of *M. musculus* we further checked for occurrences of the words "mouse" and "mice". The size of each corpus can be found in Table 1.

In order to test our extraction rules on full-text articles as well, we downloaded (March 16, 2004) the open-access part of PubMed Central (see Table 1). For the preliminary tests presented here, we did not separate between different parts of a paper although the introduction of a paper tends to list many established facts (in contrast to the results section) (Shah *et al.*, 2003).

3.2 Tokenization and tagging

We extracted abstracts for each of the species listed in Table 1 from PubMed, which we supplemented with the open-access part of PubMed Central to also test our extraction rules on full-text articles.

For segmentation of the input text into a sequence of tokens and detection of sentential boundaries, we use the tokenizer developed by Helmut Schmid, which after training on about 10^6 abstracts attained an overall precision of 99.5% (Saric *et al.*, 2004). Multiwords were acquired semi-automatically to ensure that terms of interest are captured with high accuracy. Three parameter files were tested on 24,798 held-out tokens from the GENIA corpus to optimize the POS-tagging accuracy on PubMed abstracts. The best result was achieved using the parameters trained on a corrected/revised GENIA corpus, which correctly tagged 96.4% of tokens (Saric *et al.*, 2004). This result is comparable to the current state of art (Hahn & Wermter, 2004).

After POS-tagging, we recognize terms of particular interest and re-annotate them with semantic tags. This set of semantically relevant terms mainly consists of nouns (e.g. *gene* or *protein*), verbs (e.g. *activates* or *phosphorylates*), prepositions (e.g. *from*), and adjectives (e.g. *dependent*).

3.3 Recognizing gene/protein names

To be able to recognize gene/protein names as such, and to associate them with the appropriate database identifiers, a list of synonymous names and identifiers in selected model organisms was compiled from several sources (`http://www.bork.embl.de/synonyms/`). For each organism, names and identifiers were obtained from SWISS-PROT (Boeckmann *et al.*, 2003), supplemented by names from Saccharomyces Genome Database (SGD) (Dwight *et al.*, 2002)

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

in the case of *S. cerevisiae*. The name lists were expanded to include orthographic variants of each name before matching them against the POS-tagged corpora (Saric *et al.*, 2004).

The orthographically expanded name lists were included in the lexica used for multiword detection and POS-tagging. Subsequently it was matched against the POS-tagged corpus to retag gene/protein names as such (nnpg).

3.4 Extraction of named entities

In the preceding step we described the recognition of gene/protein names. Although some homonyms can be disambiguated through the POS-tags as previously described (Saric *et al.*, 2004), we still meet two challenges: (i) to disambiguate the gene/protein name when occurring as proper part of a noun phrase, and (ii) the gene/protein names constituting the whole noun phrase.

The first case (i) comprises roughly 50% of the occurrences of gene/protein names in the corpus where they do not occur solely but are modified through adjectives, other nouns or attached prepositional phrases within the same noun phrase, like in "the ArcB sensory kinase in Escherichia coli". To get hold of this problem we built a named entity recognition system to recognize and categorize noun phrases containing gene/protein names on the basis of syntactic information (i.e. generalizing over POS-tag information) augmented with semantic information stemming from a manually curated lexicon.

This approach, which we call syntacto-semantic chunking, recognizes named entities through the use of cascaded finite state automata, which we implemented as a CASS grammar (Abney, 1996). The following simplified example shows how we recognize and semantically categorize the gene noun phrase from the above mentioned example:

$[_{nx\_kinase}$
  $[_{dt}$ the$]$ $[_{nnpg}$ ArcB$]$ $[_{jj}$ sensory $]$ $[_{kinase}$ kinase$]$ $[_{in}$ in$]$
  $[_{org}$ Escherichia coli$]]$

The label *nx_kinase* indicates that this is a noun chunk (*nx*) semantically denoting a *kinase*. Analogously, we detect at this early level noun chunks denoting other biological entities like phosphatases, transcription factors, other proteins, and genes. In subsequent cascades, we recognize more complex (i.e. nested) noun chunks on the basis of the simpler ones, such as gene products, promoters, upstream activating/repressing sequences (UAS/URS), binding sites, etc.:

$[_{nx\_expr}$
  $[_{expr}$ expression$]$ $[_{of}$ of$]$
    $[_{nx\_geneprod}$
      $[_{nx\_gene}$
        $[_{dt}$ the$]$ $[_{nnpg}$ argF$]$ $[_{gene}$ gene$]]$
      $[_{prod}$ product$]]]$

We have implemented rules to distinguish between agent and theme forms of noun chunks as well as a schemes for detecting artificial experimental contexts (Saric *et al.*, 2004) such as gene deletion:

$[_{nx\_del}$
  $[_{vvn}$ targeted$]$ $[_{disr}$ disruption$]$ $[_{of}$ of$]$
    $[_{nx\_gene}$
      $[_{dt}$ the$]$ $[_{nnpg}$ IFN-gamma$]$ $[_{gene}$ gene$]]]$

The second challenge (ii) where gene/protein names constitute the whole noun phrase the disambiguation between these two categories is less straightforward. Generally there exists the possibility, which depends on contextual information (e.g. selectional restrictions imposed by the verb). This is implemented within the following step, *the extraction of relations between entities*, explained in section 3.5. In case there's no rule applicable to disambiguate this gene/protein name it has to be left ambiguous and thus the sentence remains unanalyzed.

3.5 Extraction of relations between entities

This step of processing concerns the recognition of relations between genes and proteins, namely regulation of gene expression and (de-)phosphorylation. To extract these two types of relations we use separate grammar modules, which work on top of the same already introduced named entity recognition module. The gene expression module was based on our original system and extended with additional linguistic structures, whereas the (de-)phosphorylation module was developed from scratch.

To extract both (de-)phosphorylation and gene expression relations we combine syntactic properties (subcategorization restrictions) and semantic properties (selectional restrictions) of the relevant verbs. In order to not write a separate set of rules for each verb we generalize over classes of verbs and relational nouns.

Here we present a series of examples to illustrate how the rules operate and identify the desired information. The combined set of relations extracted from these examples are shown in Figure 1. All examples show a simplified bracketed structure illustrating the major principles of our rules; the internal structure is highly complex and derives from a pass through a number of cascading finite state transducers.

Within the following examples the first line always indicates the type of relation that we extract, which is either phosphorylation, dephosphorylation, or expression regulation. In the latter case, the subtype of expression regulation is detected, i.e. activation, repression, orfc (underspecified) regulation and specified, too. Finally we show whether the relation is verbal—and thus phrased in active or passive voice—or as a nominal relational construct.

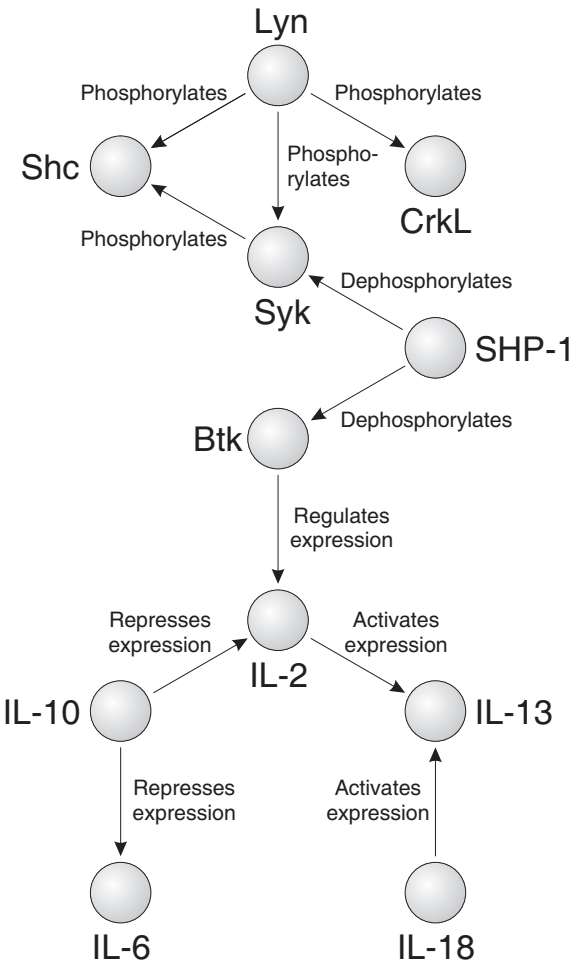

Fig. 1. An example network extracted from the mouse corpus. The network exemplifies the multiple types of relations extracted by our rule based approach; the text from which these relations were extracted are shown in section 3.5.

The first example shows a phosphorylation relation in active voice. The participating proteins are shown in boldfaced letters. The relational word is underlined. The selective negation is also marked by the *negation*-bracket. We extract that **Lyn** phosphorylates **CrkL** from the following example:

[$_{phosphorylation\_active}$
Lyn, [$_{negation}$ but not **Jak2**]
phosphorylated
CrkL]

This active voice phosphorylation construct is detected through the relational noun *phosphorylation* as argument of *participates*. It should be noted that the *phosphorylation* bracket is triggered through the key word *phosphorylation*. We extract that **Lyn** phosphorylates **syk** from:

[$_{phosphorylation\_active}$
Lyn
also participates in
[$_{phosphorylation}$ the tyrosine *phosphorylation*
and *activation* of **syk**]]

The following two examples illustrate nominalisation for phosphorylation. The arguments are attached through the *of* and *by* prepositional phrases, where the latter identifies the agent role:

[$_{phosphorylation\_nominal}$
the phosphorylation of
the adapter *protein* **SHC**
by the Src-related *kinase* **Lyn**]

[$_{phosphorylation\_nominal}$
phosphorylation of **Shc** by
the hematopoietic cell-specific
tyrosine *kinase* **Syk**]

The system is also able to identify dephosphorylation relations, as exemplified by the following nominalisation example, from which we extract that both **Syk** and **Btk** are dephosphorylated by **SHP-1**:

[$_{dephosphorylation\_nominal}$
Dephosphorylation of
Syk and **Btk**
mediated by
SHP-1]

The following examples shows gene expression relations. The first of these illustrates the ability of our system to deal with passive voice. Based on the verb ("induce") and the relational noun ("expression") we conclude that **IL-2** and **IL-18** activate expression of **IL-13**:

[$_{expression\_activation\_passive}$
[$_{expression}$ **IL-13** *expression*]
induced by
IL-2 + **IL-18**]

Repression of gene expression relation be the next example, where one protein (**IL-10**) represses the expression of two other genes (**IL-2** and **IL-6**):

[$_{expression\_repression\_active}$
IL-10
also decreased
[$_{expression}$ mRNA *expression* of
IL-2 and **IL-6** cytokine *receptors*]]

Table 1. Corpus statistics and evaluation of extraction results.

Corpus	Papers	Tokens	Gene/protein matches	Expression Relations	Expression Accuracy	Phosphorylation Relations	Phosphorylation Accuracy
E. coli	195,492	28,568,983	380,362	395	**85%**	19	**89%**
B. subtilis	16,270	2,022,852	67,758	118	**90%**	22	**91%**
S. cerevisiae	58,664	9,447,237	580,654	475	**83%**	106	**95%**
M. musculus	688,937	106,027,447	3,599,912	1862	**84%**	322	**86%**
PubMed Central	5,075	19,199,318	558,941	158	**84%**		

In the final example, the expression regulation is underspecified, that is we can only extract that **Btk** regulates the expression of the **IL-2** gene, not whether it activates or represses it:

$[_{expression\_regulation\_active}$
Btk
regulates
$[_{expression}$ the *transcription* of
the **IL-2** *gene*]]

4 RESULTS

Using our relation extraction rules, we were able to extract 3319 relation chunks for four organisms from PubMed abstracts (Table 1). A network of relations extracted from a small subset of the the mouse corpus, namely the examples shown in section 3.5, is shown in Figure 1.

4.1 Evaluation of relation extraction

To evaluate the accuracy of the extracted relations for yeast, we manually inspected all relations extracted from the evaluation corpus using the TIGERSearch visualization tool (Lezius, 2002). Since the rules written for the yeast training corpus were applied unchanged to other corpora, these were entirely used for evaluation.

The accuracy of the relations was evaluated at the semantic level rather than at 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 was 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. In contrast to what is normally done in IE, this type of evaluation can only be carried out by a biologist.

For yeast, 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. The accuracies obtained for other organisms/corpora are comparable, see Table 1. For (de-)phosphorylation relations, the accuracy appears to be marginally better although this is difficult to say for sure given the smaller number of extracted relations.

To estimate the coverage of our method, we looked through 250 of the 44,354 sentences that contain at least two gene/protein names. These contained only 8 relation chunks of the desired type, corresponding to an estimate of 1419 in total. Since 422 of these were successfully extracted by our method, we estimate the coverage of our method to be around 30%. This corresponds to an F-score in the order of 44%, which is respectable by IE standards.

Approximately half of the errors made by our method stem from genetic modifications that are overlooked due to long distance (anaphoric) relationships for example. This problem is particularly frequent for *E. coli*, the favored bacterial species for experiments, because the most commonly used reporter gene, *lacZ*, is itself an *E. coli* gene. Because *E. coli* is often used as an expression system (host) for foreign genes, *E. coli* is often mentioned in abstracts concerned with the expression of genes from other organisms. Our method thus in some cases correctly extracts a relation between two gene names, but erroneously attributes this relation to the *E. coli* genes with the same names.

4.2 Entity recognition

For consistency, we have also evaluated our ability to correctly identify named entities at the level of semantic rather than grammatical correctness. Manual inspection of 500 named entities from the yeast evaluation corpus revealed 14 errors, which corresponds to an estimated accuracy of just over 97%. Surprisingly, many of these errors were committed when recognizing *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 is thus unlikely to have an impact on the accuracy of relation extraction.

5 CONCLUSIONS

We have developed a method that allows us to extract information on gene regulation as well as (de-)phosphorylation from biomedical text. This is a highly relevant problem, since much is known about it although this knowledge has yet to be systematically collected in a database. Also, knowledge on gene expression and phosphorylation is crucial for understanding many important biological processes, e.g. the mitotic cell cycle and signaling cascades.

Although we developed our method on abstracts related to baker's yeast only, we have applied our method to several other 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 organism. Furthermore, application of the method to full text journals gave promising preliminary results. Additionally, we expanded the rules to also extract (de-)phosphorylation relations, reusing the many rules responsible for the recognition of named entities. The relations extracted for 180 organisms will soon be available through the STRING database (von Mering *et al.*, 2005, `http://string.embl.de`).

ACKNOWLEDGMENTS

Jasmin Šarić is funded by the Klaus Tschira Foundation gGmbH, Heidelberg (`http://www.kts.villa-bosch.de`). This work was supported by grants LSH6-CT-2003-503265 and LSH6-CT-2004-503567 from the European Union.

REFERENCES

Abney, S. (1996) Partial parsing via finite-state cascades. In *Proceedings of the ESSLLI '96 Robust Parsing Workshop* pp. 8–15, Prague, Czech Republic.

Blaschke, C., Andrade, M. A., Ouzounis, C. & Valencia, A. (1999) Automatic extraction of biological information from scientific text: protein–protein interactions. In *Proc. of Intelligent Systems for Molecular Biology* vol. 7, pp. 60–67 AAAI Press, Menlo Park, CA.

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.

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**, S74–S82.

Hahn, U. & Wermter, J. (2004) Tagging medical documents with high accuracy. In *PRICAI* pp. 852–861.

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

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

Marcotte, E. M., Xenarios, I. & Eisenberg, D. (2001) Mining literature for protein–protein interactions. *Bioinformatics*, **17**, 359–363.

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. In *Proceedings of the Seventh Pacific Symposium on Biocomputing* pp. 362–373 World Scientific, Hawaii.

Saric, J., Jensen, L. J., Ouzounova, R., Rojas, I. & Bork, P. (2004) Extracting regulatory gene expression networks from pubmed. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*.

Shah, P. K., Perez-Iratxeta, C., Bork, P. & Andrade, M. A. (2003) Information extraction from full text scientific articles: where are the keywords? *BMC Bioinformatics,* **4**, 20.

Thomas, J., Milward, D., Ouzounis, C., Pulman, S. & Carroll, M. (2000) Automatic extraction of protein interactions from scientific abstracts. In *Proceedings of the Fifth Pacific Symposium on Biocomputing* pp. 707–709 World Scientific, Hawaii.

von Mering, C., Jensen, L. J., Snel, B., Hooper, S. D., Krupp, M., Foglierini, M., Jouffre, N., Huynen, M. A. & Bork, P. (2005) STRING: known and predicted protein–protein associations, integrated and transferred across organisms. *Nucleic Acids Res.,* **33**, D433–D437.