Extracting Regulatory Gene Expression Networks from PubMed

Jasmin Šarić
EML Research gGmbH
Heidelberg, Germany
saric@eml-r.org

Lars J. Jensen
EMBL
Heidelberg, Germany
jensen@embl.de

Rossitza Ouzounova
EMBL
Heidelberg, Germany
ouzounov@embl.de

Isabel Rojas
EML Research gGmbH
Heidelberg, Germany
rojas@eml-r.org

Peer Bork
EMBL
Heidelberg, Germany
bork@embl.de

Abstract

We present 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 POS-tags. The method is easily adapted to other organisms than yeast, allowing us to extract many more biologically relevant relations.

1 Introduction and related work

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

The goal of our work is to extract from biological abstracts which *proteins* are responsible for regulating the expression (*i.e.* transcription or translation) of which *genes*. This means to extract a specific type of pairwise relations between biological entities. This differs from the BioCreAtIvE competition tasks[2] that aimed at classifying entities (gene products) into classes based on Gene Ontology (Ashburner et al., 2000).

A task closely related to ours, which has received some attention over the past five years, is the extraction of protein–protein interactions from abstracts. This problem has mainly been addressed by statistical "bag of words" approaches (Marcotte et al., 2001), with the notable exception of Blaschke et al. (1999). All of the approaches differ significantly from ours by only attempting to extract the type of interaction and the participating proteins, disregarding agens and patiens.

Most NLP 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, there are two problems with such approaches: 1) the meaning of the extracted events

[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/

will depend strongly on the selectional restrictions and 2) the same meaning can be expressed using a number of different verbs. In contrast and alike (Friedman et al., 2001), we instead set out to handle only one specific biological problem and, in return, extract the related events with their whole range of 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.

2 The biological task and our approach

To extract relations, one should first recognize the named entities involved. This is particularly difficult in molecular biology where many forms of variation frequently occur. Synonymy is very common 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*). Moreover, many names are homonyms 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 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

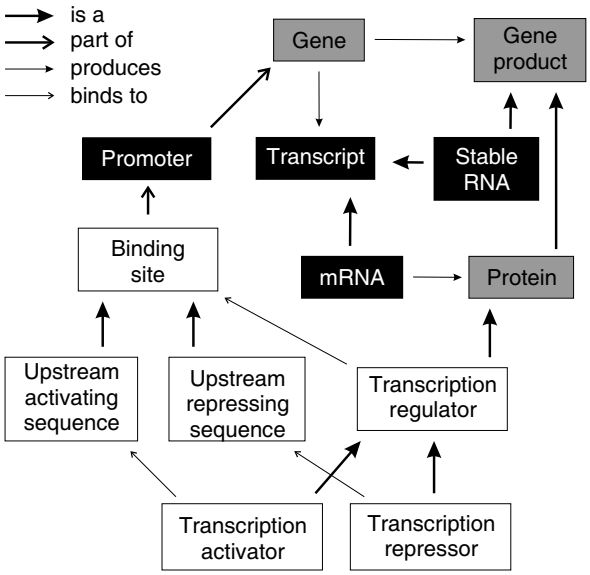

Figure 1: **A simplified ontology for transcription regulation.** The background color used for each term signifies its semantic role in relations: regulator (white), target (black), or either (gray).

one another. An ontology is a great help when writing extraction rules, as it immediately suggests a large number of relevant relations to be extracted. 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 gene transcription or translation or by an indirect mechanism. For this reason, and for simplicity, we decided against trying to extract how the regulation of expression takes place. We do, however, strictly require that the extracted relations provide information about a protein (the regulator, **R**) regulating the expression of a gene (the target, **X**), for which reason three requirements must be fulfilled:

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. Thus, whether the event should be extracted or not 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, leaving the whole relation underspeci-

fied, for which reason it 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 transcription factor activates the expression when binding to the **X** promoter. Linguistically this implies that noun chunks of certain semantic types should be disallowed as agens.

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 patiens 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. The coloring of the entities in Figure 1 helps discern which relations are meaningful and which are not.

The ability to genetically modify an organism in experiments brings about further 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**. The key point is to identify that "*deletion* of **R**" implies that the sentence describes an experiment in which **R** has been removed, but that **R** would normally be present and that the biological impact of **R** is thus the opposite of what the verb *increased* alone would suggest. 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. In this case *mutation* is dealt in a manner similar to *deletion* in the previous example. 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 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 accuracy 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 its 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 wrong.

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.

3 Methods

Table 1 shows an overview of the architecture of our IE system. It is organized in levels such that the output of one level is the input of the next one. The following sections describe each level 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[3]. 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]Medical Subject Headings (MeSH) is a controlled vocabulary for manually annoting PubMed articles.

Level	Component
L0	**Tokenization and multiwords** Word and sentence boundaries are detected and multiwords are recognized and recomposed to one token.
L1	**POS-Tagging** A part-of-speech tag is assigned to each word (or multiword) of the tokenized corpus.
L2	**Semantic labeling** A manually built taxonomy is used to assign semantic labels to tokens. The taxonomy consists of gene names, cue words relevant for entity recognition, and classes of verbs for relation extraction.
L3	**Named entity chunking** Based on the POS-tags and the semantic labels, a cascaded chunk grammar recognizes noun chunks relevant for the gene transcription domain, *e.g.* $[_{nxgene}$ The GAL4 gene $]$.
L4	**Relation chunking** Relations between entities are recognized, *e.g.* The expression of the cytochrome genes **CYC1** and **CYC7** is ***controlled*** by **HAP1**.
L5	**Output and visualization** Information is gathered from the recognised patterns and transformed into pre-defined records. From the example in L4 we extract that **HAP1** regulates the expression of **CYC1** and **CYC7**.

Table 1: Overview over the extraction architecture

3.2 Tokenization and multiword detection

The process of tokenization consists of two steps (Grefenstette and Tapanainen, 1994): segmentation of the input text into a sequence of tokens 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 such as brackets, colons, hyphens, slashes, etc. On the other hand, they often appear as multiword expressions which makes it hard to detect the 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 the accuracy of POS-tagging on PubMed abstracts, TreeTagger (Schmid, 1994) was retrained on the GENIA 3.0 corpus (Kim et al., 2003). Furthermore, we expanded the POS-tagger lexicon with entries relevant for our application such as gene names (see section 3.4) and multiwords (see section 3.5). As tag set we use the UPenn tag set (Santorini, 1991) plus some minor extensions for distinguishing auxiliary verbs.

The GENIA 3.0 corpus consists of PubMed abstracts and has 466,179 manually annotated tokens. For our application we made two changes in the annotation. The first one concerns seemingly undecideable cases like *in/or* annotated as `in|cc`. These were split into three tokens: *in*, */*, and *or* each annotated with its own tag. This was done because TreeTagger is not able to annotate two POS-tags for one token. The second set of changes was to adapt the tag set so that `vb...` is used for derivates of *to be*, `vh...` for derivates of *to have*, and `vv...` for all other verbs.

3.4 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 six eukaryotic model organisms was compiled from several sources (available from `http://www.bork.embl.de/synonyms/`). For *S. cerevisiae* specifically, 51,640 uniquely resolvable names and identi-

fiers were obtained from Saccharomyces Genome Database (SGD) and SWISS-PROT (Dwight et al., 2002; Boeckmann et al., 2003).

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`). By accepting only matches to words tagged as common nouns (`nn`), the problem of homonymy was reduced since *e.g.* the name **MAP** can occur as a verb as well.

3.5 Semantic tagging

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

- Relevant concepts in our ontology: *gene*, *protein*, *promoter*, *binding site*, *transcription factor*, etc. (153 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).
- Triggering experimental (artificial) contexts: *mutation*, *deletion*, *fusion*, *defect*, *vector*, *plasmids*, etc. (11 entries).
- Enzymes: *gyrase*, *kinase*, etc. (569 entries).
- Organism names extracted from the NCBI taxonomy of organisms (Wheeler et al., 2004) (20,746 entries).

The second set of terms contains 50 verbs and their inflections. They were classified according to their relevance in gene transcription. These verbs are crucial for the extraction of relations between entities:

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

Each of the terms consisting of more than one word was utilized for multiword recognition.

We also have have two additional classes of words to prevent false positive extractions. The first 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 allowed within named entitities, *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 generalizing over both POS-tags and semantic tags. Syntacto-semantic chunking was performed to recognize named entities using cascades of finite state rules implemented as a CASS grammar (Abney, 1996). As an example we recognize gene noun phrases:

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

Other syntactic variants, as for example "the glucokinase gene **GLK1**" are recognized too. Similarly, we detect at this early level noun chunks de-

noting other biological entities such as proteins, activators, repressors, transcription factors etc.

Subsequently, we recognize more complex noun chunks on the basis of the simpler ones, *e.g.* promoters, upstream activating/repressing sequences (UAS/URS), binding sites. At this point it becomes important to distinguish between agens and patiens forms of certain entities. Since a binding site is part of a target gene, it can be referred to either by 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 trigger experimental context. On the basis of these we identify noun chunks mentioning, as for example deletion, mutation, or overexpression of genes. At a fairly late stage we recognize events that can occur as arguments for verbs like "expression of".

3.7 Extraction of relations between entities

This step of processing concerns the recognition of three types of relations between the recognized named entities: up-regulation, down-regulation, and (underspecified) regulation of expression. We combine syntactic properties (subcategorization 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 consting of three parts, a promoter chunk, a verbal complex chunk, and a UAS chunk in patiens:

$[_{nx\_prom}$ 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 nominalized relations as exemplified by "the binding of **GCN4** protein to the **SER1** promoter in vitro".

4 Results

Using our relation extraction rules, we were able to extract 422 relation chunks from our complete corpus. Since one entity chunk can mention several different named entities, these corresponded to a total of 597 extracted pairwise relations. However, as several relation chunks may 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 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 these tend to correspond to proteins serving as regulators in our relations.

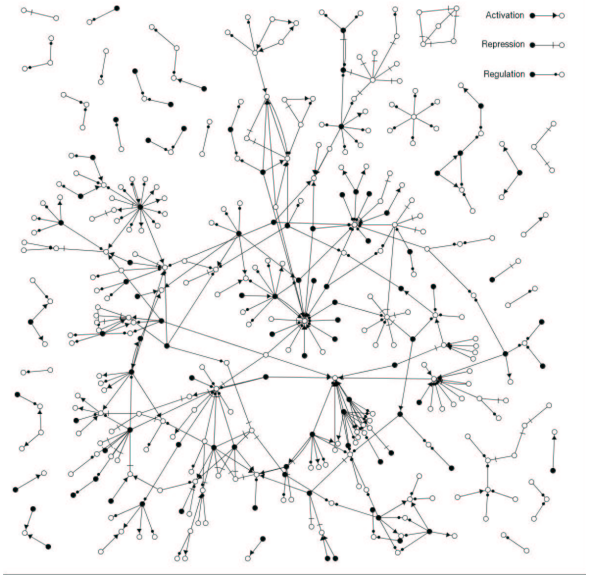

Figure 2: **The extracted network of gene regulation** The extracted relations are shown as a directed graph, in which each node corresponds to a gene or protein and each arc represents a pairwise relation. The arcs point from the regulator to the target and the type of regulation is specified by the type of arrow head. Known transcription factors are highlighted as black nodes.

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 visualization tool (Lezius, 2002).

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 is 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. 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 overlooked genetic modifications—although mentioned in the sentence, the extracted relation is not biologically relevant.

4.2 Entity recognition

For the sake of 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 evaluation corpus revealed 14 errors, which corresponds to an estimated accuracy of just over 97%. Surprisingly, many of these errors were commited 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 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 tokenization

We compared the POS-tagging performance of two parameter files on 55,166 tokens from the GENIA corpus that were not used for retraining. Using the retrained tagger, 93.6% of the tokens were correctly tagged, 4.1% carried questionable tags (*e.g.* confusing proper nouns for common nouns), and 2.3% were clear tagging errors. This compares favourably to the 85.7% correct, 8.5% questionable tags, and 5.8% errors obtained when using the Standard English parameter file. Retraining thus reduced the error rate more than two-fold.

Of 198 sentences evaluated, the correct sentence boundary was detected in all cases. In addition, three abbreviations incorrectly resulted in sentence marker, corresponding to an overall precision of 98.5%.

5 Conclusions

We have developed a method that allows us to extract information on regulation of gene expression from biomedical abstracts. 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.

Although we developed and evaluated our method on abstracts related to baker's yeast only, we have successfully applied the method to other organisms including humans (to be published elsewhere). The main adaptation required was to replace the list of synonymous gene/protein names to reflect the change of organism. Furthermore, we also intend to reuse the recognition of named entities to extract other, specific types of interactions between biological entities.

Acknowledgments

The authors wish to thank Sean Hooper for help with Figure 2. Jasmin Šarić is 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

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

M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. 2000. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29.

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

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

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

C. Friedman, P. Kra, H. Yu, M. Krauthammer, and A. Rzhetsky. 2001. GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics*, 17 Suppl. 1:S74–S82.

G. Grefenstette and P. Tapanainen. 1994. What is a word, what is a sentence? problems of tokenization. In *The 3rd International Conference on Computational Lexicography*, pages 79–87.

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

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

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

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

G. Nenadić, S. Rice, I. Spasić, S. Ananiadou, and B. Stapley. 2003. Selecting text features for gene name classification: from documents to terms. In S. Ananiadou and J. Tsujii, editors, *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine*, pages 121–128.

R. Netzel, Perez-Iratxeta C., P. Bork, and M. A. Andrade. 2003. The way we write. *EMBO Rep.*, 4:446–451.

J. Pustejovsky, J. Castaño, J. Zhang, M. Kotecki, and B. Cochran. 2002. Robust relational parsing over biomedical literature: Extracting inhibit relations. In *Proceedings of the Seventh Pacific Symposium on Biocomputing*, pages 362–373, Hawaii. World Scientific.

B. Santorini. 1991. Part-of-speech tagging guidelines for the penn treebank project. Technical report, University of Pennsylvania.

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

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

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

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

K. Yamamoto, T. Kudo, A. Konagaya, and Y. Matsumoto. 2003. Protein name tagging for biomedical annotation in text. In S. Ananiadou and J. Tsujii, editors, *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine*, pages 65–72.