BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 3 2007, pages 365–371
doi:10.1093/bioinformatics/btl616

Data and text mining

RelEx—Relation extraction using dependency parse trees

Katrin Fundel*, Robert Küffner and Ralf Zimmer

Institut für Informatik, Ludwig-Maximilians-Universität München, Amalienstrasse 17, 80333 München, Germany

Received on August 31, 2006; revised on November 8, 2006; accepted on November 28, 2006
Advance Access publication December 1, 2006
Associate Editor: Satoru Miyano

Downloaded from https://academic.oup.com/bioinformatics/article/23/3/365/236564 by guest on 16 August 2022

ABSTRACT

Motivation: The discovery of regulatory pathways, signal cascades, metabolic processes or disease models requires knowledge on individual relations like e.g. physical or regulatory interactions between genes and proteins. Most interactions mentioned in the free text of biomedical publications are not yet contained in structured databases.

Results: We developed RelEx, an approach for relation extraction from free text. It is based on natural language preprocessing producing dependency parse trees and applying a small number of simple rules to these trees. We applied RelEx on a comprehensive set of one million MEDLINE abstracts dealing with gene and protein relations and extracted ~150 000 relations with an estimated perfomance of both 80% precision and 80% recall.

Availability: The used natural language preprocessing tools are free for use for academic research. Test sets and relation term lists are available from our website (http://www.bio.ifi.lmu.de/publications/RelEx/).

Contact: katrin.fundel@bio.ifi.lmu.de

1 INTRODUCTION

Most biological facts are available only in the free text of scientific articles. For information integration or combination with other types of data, these facts have to be extracted from the scientific literature. Information on relations or interactions between genes and proteins is of interest for generating network models of regulatory or metabolic pathways. Various approaches for relation extraction have been applied to the biomedical domain. The simplest approach is the detection of co-occurrences of entities from within sentences or abstracts (Ding *et al*., 2002; Jelier *et al*., 2005; Jenssen *et al*., 2001). It relies on the hypothesis that entities which are repeatedly mentioned together are somehow related. Extracted relations exhibit high sensitivity but very low specificity. Generally, the type and direction of the relation cannot be determined. Pattern-based extraction approaches (Blaschke *et al*., 1999; Blaschke and Valencia, 2001; Leroy and Chen, 2002; Ono *et al*., 2001) were set up to increase specificity, yet they achieve significantly lower recall. Other approaches analyze the underlying sentences in more detail and apply natural language processing (NLP), i.e. analysis of sentence syntax and semantics, typically implemented in complex proprietary software systems. Relation extraction algorithms can also be classified by the way the extraction rules are obtained, they can be manually defined (Divoli and Attwood, 2005; Saric *et al*., 2006; Thomas *et al*., 2000; Yakushiji *et al*., 2001) or learned from large annotated training corpora (Hakenberg *et al*., 2005; Huang *et al*., 2004).

Besides performance criteria, approaches might also be evaluated whether they (1) are available or simple enough so that they can be reproduced, (2) fully disclose the validation procedures and datasets, (3) are able to process publication abstracts in the order of millions in reasonable time, (4) can deal with the human/mammal domain, characterized by complex gene and protein names and complex sentences, (5) annotate genes/proteins involved in interactions with database identifiers so that external information/data can be mapped and (6) cover a broad spectrum of relation types. Most of the existing approaches for relation extraction violate one or more of these criteria.

We developed RelEx, which conforms to all of the above criteria. It shows very good performance despite its simplicity. It uses a small set of simple rules, building upon publicly available tools applied for part-of-speech-tagging, noun-phrase-chunking and dependency.

As an extension to standard relation extraction pipelines, we propose the use of dependency parse trees (Klein and Manning, 2002, 2003; Mel'cuk, 1988) as a means for biomedical relation extraction. Dependency parse trees reveal non-local dependencies within sentences, i.e. between words that are far apart in a sentence. Sentences of biomedical texts tend to be long and complicated and frequently mention a number of possible effectors and effectees. Dependency parse trees provide a useful structure for the sentences by annotating edges with dependency types, e.g. subject, auxiliary, modifier.

Although our approach is not restricted to particular kinds of interactions, we currently focus on physical, genetic and regulatory relations between genes and proteins.

2 METHODS

The RelEx work-flow (Figure 1) extracts directed qualified relations starting from free-text sentences. RelEx requires a synonym dictionary (Fundel and Zimmer, 2006) containing gene and protein names, and a list of restriction-terms[1] that are used to describe relations of interest.

2.1 Text preprocessing

Sentences are part-of-speech (POS)-tagged by MedPost[2] (Smith *et al*., 2004) and noun-phrase chunks are identified by fnTBL[3] (Ngai and Florian, 2001). The POS-tagged sentences are submitted to the Stanford Lexicalized Parser[4] (Version 1.5) (Klein and Manning, 2002, 2003) which generates a dependency parse tree (Figure 2, upper panel) for each sentence and assigns word

*To whom correspondence should be addressed.

[1] http://www.bio.ifi.lmu.de/publications/RelEx/

[2] ftp://ftp.ncbi.nlm.nih.gov/pub/lsmith/MedPost/medpost.tar.gz

[3] http://nlp.cs.jhu.edu/~rflorian/fntbl/

[4] http://nlp.cs.jhu.edu/~rflorian/fntbl/

© The Author 2006. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

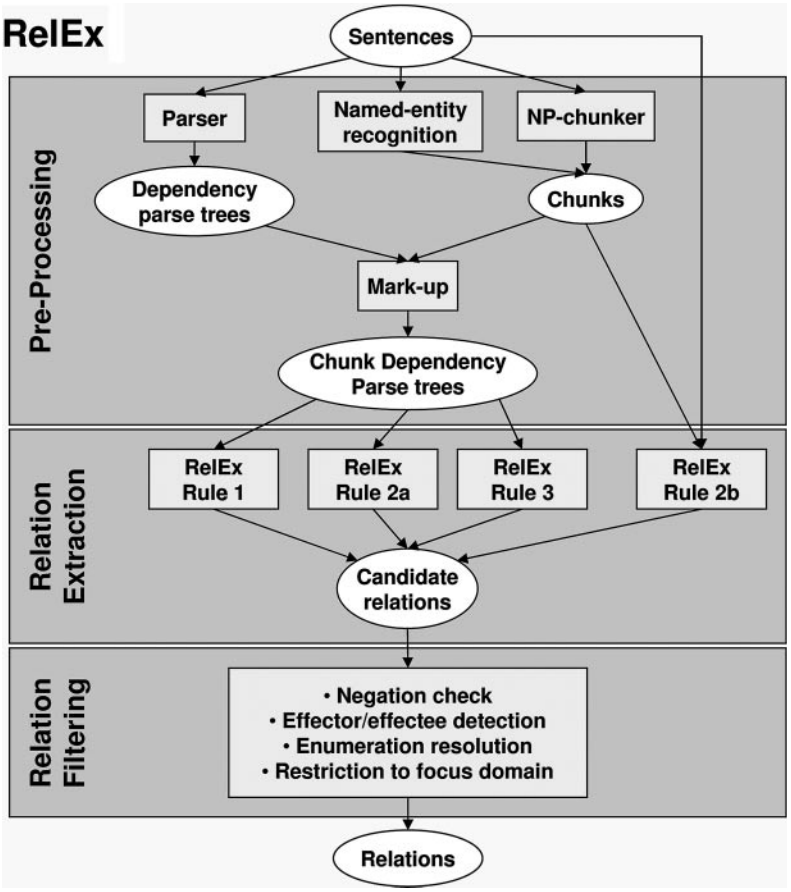


Fig. 1. The work-flow of RelEx is subdivided into preprocessing, relation extraction and relation filtering leading from the original free-text sentences to directed, qualified relations. Preprocessing is based on publicly available tools and named entity identification. Candidate relations are extracted according to rules applied on chunk dependency trees and original sentences, and subjected to filtering steps.

positions to each word. Gene and protein names are identified by ProMiner (Hanisch *et al*., 2005) based on matching to a synonym dictionary (Fundel and Zimmer, 2006). If a noun-phrase chunk contains only a part of a multi-word gene or protein name, the chunk is expanded to contain the complete name. For each chunk, the corresponding nodes in the dependency tree are combined into a chunk-node returning a simplified chunk dependency tree (Figure 2, lower panel).

2.2 Relation extraction

RelEx creates candidate relations by extracting paths connecting pairs of proteins from dependency parse trees. These paths should contain just the relevant terms describing the relation between the given pair of proteins. Currently, we use three rules that reflect the constructs that are most frequently used in English language for describing relations, namely:

(1) effector-relation-effectee ('A activates B')
(2) relation-of-effectee-by-effector ('Activation of A by B')
(3) relation-between-effector-and-effectee ('Interaction between A and B').

Rule 1 (e.g. in Figure 2) extracts paths in the chunk dependency tree that lead from a start-point (generally the effector) to an end-point (generally the effectee). If the chunk dependency tree contains one or more subject-dependencies (*nsubj* or *nsubjpass*), the tree is split so that the parent of each subject-dependency becomes the root of a partial tree, i.e. each resulting partial tree has exactly one subject-dependency. The chunks with an incoming edge labeled as subject-dependency are marked as potential start-points. Starting from these, RelEx constructs paths towards the other gene/protein-containing chunks (potential end-points). If the dependency tree does not include any subject-dependencies all pairs of gene names containing noun-phrase chunks are potential start- and end-points and thus candidate interaction pairs. For each potential start and end-point, the path connecting these two noun phrase chunks is extracted from the chunk dependency tree.

Some of the paths generated by rule 1 are not valid or need to be revised, which is automatically detected and accomplished as follows. A path is invalid if it contains a term occurring after the noun phrase chunk of the end point in the sentence, unless the respective term is contained in the least common ancestor node of the start and end chunk or is part of an enumeration (see below) with the end chunk. This restriction has been found to reduce the number of false paths, especially for long and complex sentences. It reflects that verbs and modifying terms usually occur before the object they refer to.

A path needs to be revised if it contains two nodes tagged as verbs between the least common ancestor and the end node, which are directly linked to each other by a *and*, *but* or *whereas* dependency. In this case the first verb is removed from the path, as it is frequently not relevant for the given path but refers to another child node. This applies for instance to 'Protein A binds B and inhibits C' where 'binds' is not relevant for the interaction between 'A' and 'C'.

Downloaded from https://academic.oup.com/bioinformatics/article/23/3/365/236564 by guest on 16 August 2022

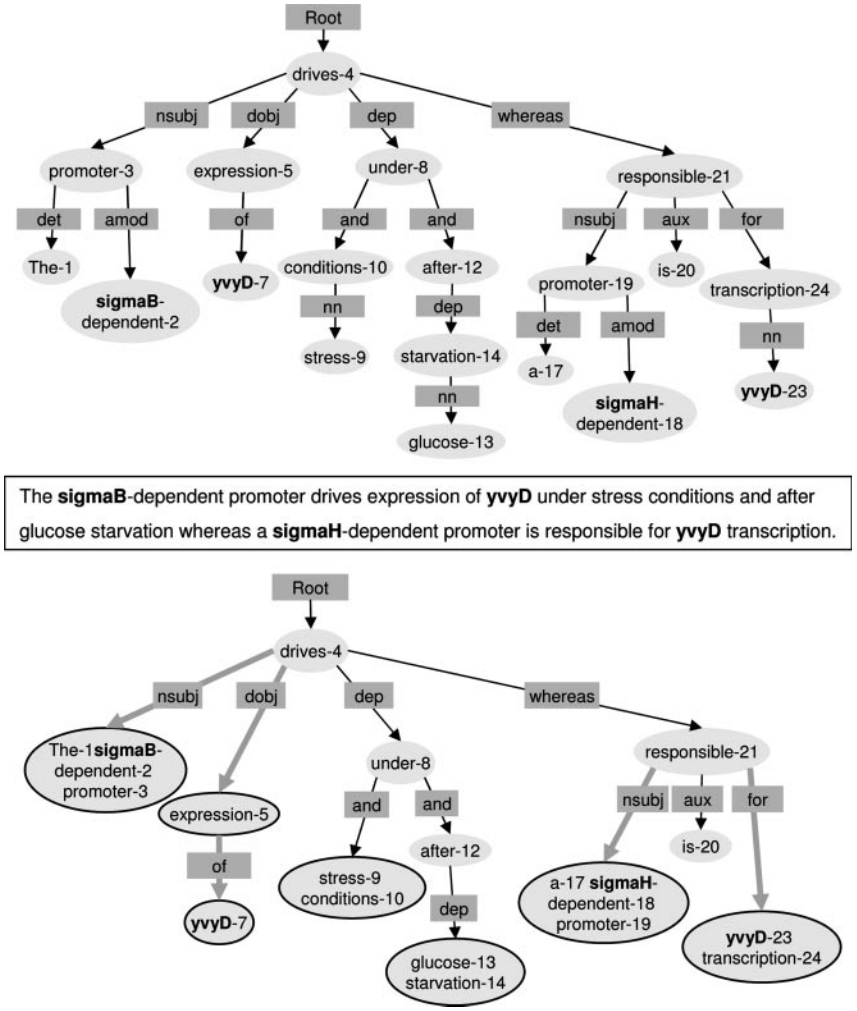


Fig. 2. Upper panel: Dependency parse tree as derived from the Stanford Lexicalized Parser, showing words (ellipses) assigned with word positions (numbers appended to words), dependencies (edges pointing from the head of a dependency to the dependent word), dependency types (rectangles) and the head of the sentence (Root). Lower panel: Corresponding chunk dependency tree that groups the words into noun phrase chunks (framed ellipses). Words marked in bold indicate gene/protein names, thick grey edges indicate paths that are extracted by Rule 1.

Rule 1 applied on the sentence 'This indicates that *the yvyD gene product*, being a member of both the sigmaB and sigmaH regulons, might *negatively regulate the activity of the sigmaL regulon*.' extracts the parts marked in italics as candidate relation.

Rule 2*a* extracts the longest paths through the tree that contain only noun phrase chunks as nodes and dependencies of the types *of*, *by*, *to*, *on*, *for*, *in*, *through*, *with*. The paths containing at least one of these dependencies between two protein containing chunks are retained as candidate relations (e.g. Figure 3, left panel).

Rule 2*b* is similar to Rule 2a, but is applied directly on the chunked sentences. The longest sequences of chunks that are connected by the terms *of*, *by*, *to*, *on*, *for*, *in*, *through*, *with* is extracted. A sequence is retained as candidate relation if it contains at least two of these terms and at least one between two chunks each containing at least one protein. Rule 2 extracts relations described like '*Dephosphorylation of SpoIIAA-P by SpoIIE*' or '*sigmaK-dependent transcription of gerE*'.

Rule 3 extracts two noun phrase chunks connected by a dependency of the type *between* provided that the successor in the tree contains the word and or has a dependent noun phrase chunk, which is connected via an *and* dependency (e.g. Figure 3, right panel). In the latter case, the dependent noun phrase chunk is included in the candidate relation. This rule extracts relations described like '*the physical association between EGFR and p*185*c-neu*'.

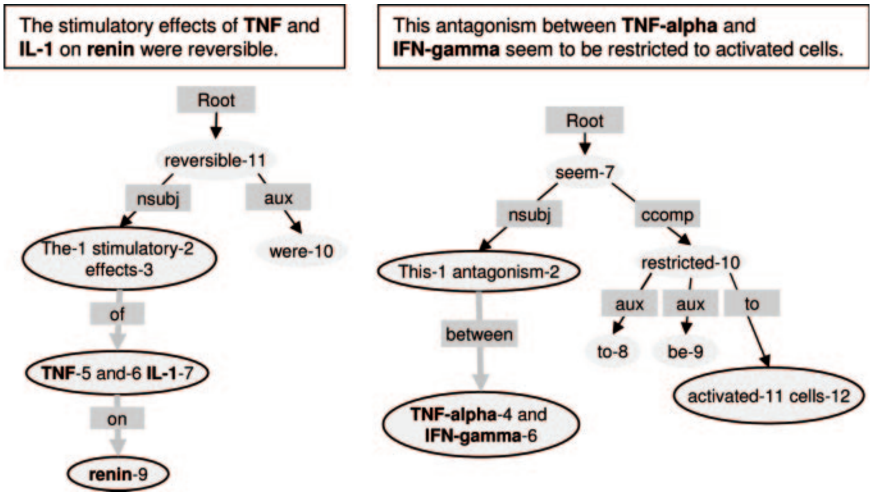
The set of rules can easily be adapted or expanded to extract other types of relations. If, e.g. annotations for individual genes and proteins are sought, the *apposition* dependency is useful as it frequently points from an entity to a description of this entity (e.g. Spo0A-P $\xrightarrow{\text{appos}}$ a major transcription factor).

2.3 Relation filtering and post-processing steps

2.3.1 Negation check A relation is said to be negated if a node in the candidate relation or one of the respective child nodes contains a negation word (*no*, *not*, *nor*, *neither*, *without*, *lack*, *fail(s,ed)*, *unable(s)*, *abrogate(s,d)*, *absen(ce,t)*). Currently, negated relations are excluded from further analysis.

2.3.2 Effector-effectee detection Generally, the named entity appearing first in the extracted relation, i.e. with the smaller sentence position, is assumed to be the effector of the relation while the second named entity is assumed to be the effectee. The roles are switched if some form of passive construct is detected, i.e. if an expression listed in Table 1 matches the

Downloaded from https://academic.oup.com/bioinformatics/article/23/3/365/236564 by guest on 16 August 2022



The stimulatory effects of **TNF** and **IL-1** on **renin** were reversible.

This antagonism between **TNF-alpha** and **IFN-gamma** seem to be restricted to activated cells.

Fig. 3. Dependency parse trees: examples of sentences and chunk dependency parse tree representations for which rules 2 (left panel) or 3 (right panel) extract paths marked by thick gray edges.

Table 1. Effector-effectee detection: terms indicating switched roles, i.e. the named entity with the smaller sentence position is assumed to be the effectee and the named entity with the larger sentence position is assumed to be the effector of the relation

Single words	by, after, with, if, once, require, requires, when, through
Multi-word expressions	due to, in case, provided that, (effect, result, member) of, in response to, (in,under) control of, depend(s,ed,ent) on

relation and is preceded by a verb, noun or adjective ending on *-t*, *-d*, *-ion*, *-ing*. For the word *by* the roles are only switched if *by* is not followed by one of the words *time*, *times*, *fold* or by a verb ending on *-ing*.

2.3.3 Enumeration resolution Noun phrase chunks connected to each other by a *and*, *or*, *nn*, *det*, or *dep* dependency form an enumeration. If a noun phrase chunk contains more than one protein name, these are considered to describe alternative agents/targets. For all candidate relations all gene/protein name containing chunks are analyzed for alternatives from enumerations and chunks containing several protein names. Variants of the candidate relation are generated so that one relation per alternative gene/protein name at each respective position is generated.

2.3.4 Restricting candidate relations to focus domain The words contained in candidate relations are checked against a set of *relation restriction terms*. This list reflects the types of relations that are in the focus of interest, it contains terms that are typically used to describe a relation, most importantly interaction verbs and derived nouns and adjectives. Here, we focus on physical, regulatory and genetic interactions; we compiled a list of 1048 restriction terms with 157 distinct word-stems. A candidate relation is retained if it contains at least one relation term.

3 DATASETS

3.1 Learning language in logic (LLL) dataset

The task of the LLL challenge 2005 (Nédellec, 2005) was to extract genic interactions of the types action, regulon, binding and promoter from a set of sentences concerning *Bacillus subtilis* transcription. Participating groups focused on machine learning approaches. The task required identification of genes/proteins that interact and their roles, i.e. agent or target, together with their position within a sentence. The provided data consists of a synonym dictionary for genes/proteins, a training set (55 sentences and 103 interactions) and a test set (80 sentences and 54 interactions). The organizers provided an evaluation script for the training set, and a website for evaluation of the results on the test set.

3.2 Large-scale application

The comprehensive subset of ~1 million MEDLINE abstracts dealing with human gene and protein interactions from 1990 or newer [for details see (Küffner *et al*., 2005)] and a synonym dictionary (Fundel and Zimmer, 2006) containing 338 824 synonyms for 27 141 human genes and proteins were used for large-scale relation extraction.

3.3 Manually annotated subset of large-scale dataset

We randomly selected a subset of 50 abstracts (called hprd50) referenced by the Human Protein Reference Database (HPRD) (Peri *et al*., 2004). Direct physical interactions, regulatory relations, as well as modifications (e.g. phosphorylation) were manually annotated by two annotators with biochemical background (authors K. Fundel and R. Küffner). The consensus contains 138 relation instances (i.e. pairs of genes/proteins with abstract and sentence identifier), corresponding to 92 distinct relations in abstracts (i.e. pairs of genes/proteins with abstract identifier). The inter-annotator agreement was 81% (determined as the intersection of annotated relations divided by the total number of relations) which corresponds to a F-measure of 89% (considering one of the annotations as standard of truth and evaluating the other against it).

3.4 Evaluation criteria

For evaluation, a relation instance *rel* is defined as follows:

rel_{sen}: a pair of interacting proteins/genes in a sentence
rel_{abs}: a pair of interacting proteins/genes in an abstract

Downloaded from https://academic.oup.com/bioinformatics/article/23/3/365/236564 by guest on 16 August 2022

rel_{LLL}: a pair of interacting proteins/genes in a sentence, with defined direction of interaction and sentence position of interactor and interactee.

Results were evaluated in terms of recall R (proportion of known positives identified), precision P (proportion of results known to be true positives), and F-measure F [harmonic mean of precision and recall; $F = 2PR/(P+R)$].

The three definitions of a relation instance correspond to three evaluation criteria. The most generally applied criterion is rel_{sen}. rel_{abs} is useful for comparing manually annotated or RelEx relations against interactions in public databases (e.g. HPRD), which do not provide sentence information. rel_{abs} is less stringent than rel_{sen} as an interaction might be mentioned in several sentences within an abstract. rel_{LLL} is the most stringent criterion as direction and sentence position needs to be defined; this criterion has been used for the LLL-challenge dataset, which is annotated with the required details and only contains directed interactions. The co-occurrence results ($cooc_{sen}$: all pairs of co-occurring genes/proteins identified by ProMiner (Hanisch *et al*., 2005) within a sentence are assumed to interact) indicate the maximum recall that can be achieved by a relation extraction approach working on individual sentences, given the method for gene name identification.

4 RESULTS AND DISCUSSION

4.1 Evaluation on LLL challenge data

4.1.1 Evaluation with LLL-challenge criteria (rel_{LLL}) Evaluation results obtained on the LLL-challenge dataset (Figure 4, F 75%, R 83%, P 68% on the training set; F 72%, R 78%, P 68% for the basic test set) show that RelEx returns relations with significantly higher recall and precision than the approaches previously applied for the LLL-challenge [F 51.8%, R 53.8%, P 50.0% for the basic and F 54.3%, R 53.0%, P 55.6% for the linguistically enriched test set (Nédellec, 2005)].

4.1.2 Evaluation with standard criteria (rel_{sen}) Table 2 shows the evaluation results with standard criteria (i.e. instances of gene/protein pairs in sentences). For comparison, this table also contains precision and recall that would be achieved by co-occurrence extraction. With RelEx, 78–85% of the relations that are found as co-occurrence are extracted as relations. These numbers correspond to inter-annotator agreement for the recognition of gene names and biomedical annotations, which has been shown to be in the range of 69–91% (Colosimo *et al*., 2005) and 70–80% (Wilbur *et al*., 2006). For both datasets (LLL and hprd50) RelEx achieves significantly higher precision and thus F-measure than co-occurrence-search.

4.1.3 Analysis of errors The usage of publicly available preprocessing tools clearly causes RelEx to depend on the quality of the applied tools. The detailed analysis of the results on the hprd50 dataset indicates the most prominent sources of error: out of 28 false positive relations, nine relations were generated by the rules not being specific enough or constructs not being correctly resolved, eight describe undesired types of relations (e.g. homology, part of and similarity), six were generated from sentences where a POS-tagging error occured and four were generated from sentences where the detected gene/protein name actually does not refer to a gene/protein but forms part of a cell name or description of an experimental technique.

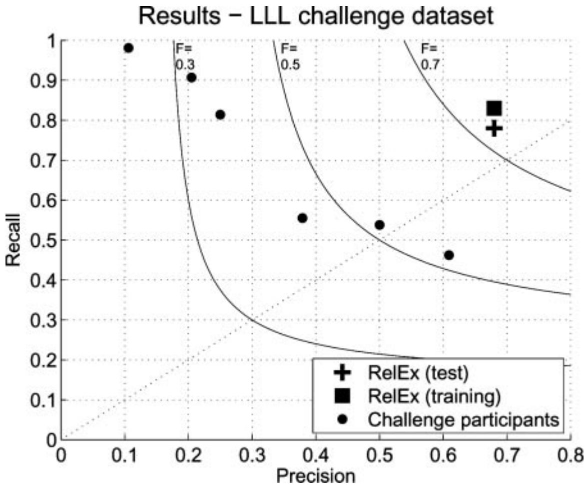


Fig. 4. Evaluation results on the LLL-challenge datasets evaluated with the criteria applied in the challenge (rel_{LLL}).

Table 2. Evaluation of RelEx [rel_{sen}, i.e. instance: pair of genes/proteins with sentence identifier, cooc: sentence co-occurrences]

	LLL		hprd50	
Sentences	55		88	
Co-occurrences ($cooc_{sen}$)	216		294	
Relations (rel_{sen})	97		138	
	cooc	RelEx	cooc	RelEx
Recall (%)	100	85	100	78
Precision (%)	46	79	47	79
F-measure (%)	63	82	64	78

Out of 31 false negative relations, eight are described by a wording that is not covered by the applied rules (e.g. 'a and b are receptors that interact', 'a and b form a complex'), eight relations are described in sentences which contained POS-tagging errors, four false negatives were due to anaphora (e.g. 'which', 'these proteins'), which RelEx currently does not resolve, four relations were not detected due to erroneous subordinate clause attachment produced by the dependency parser, in two cases the relevant relation terms were not contained on the candidate relation paths and in another two cases relations were not extracted due to noun phrase chunks erroneously being split up.

MedPost is a part-of-speech-tagger that has been designed specifically for biomedical texts and generally works very well. The errors mentioned above were due to verbs being annotated as adjectives (in two sentences), verbs being annotated as nouns (in two sentences), and a noun being annotated as verb (one sentence). The dependency parser is sensitive to errors in POS-tagging; tagging errors lead to significantly altered parse trees. As the respective sentences contain several relations, tagging errors lead to several false positive as well as false negative relations.

Downloaded from https://academic.oup.com/bioinformatics/article/23/3/365/236564 by guest on 16 August 2022

The detailed analysis of the effector-effectee detection on the LLL training data showed that in five cases the assigned direction was wrong due to a construct not contained in our list of expressions (Table 1), e.g. 'the bmrUR operon is under sigmaB control'.

4.2 Large-scale application

The large-scale application of RelEx on ∼1 million MEDLINE abstracts yielded a total of 731 432 extracted descriptions of relations between 149 778 distinct pairs of genes or proteins, containing 10 821 distinct genes/proteins.

These relations can be compared against HPRD, which contains interactions that were manually extracted from MEDLINE full-text articles. The comparison provides information with respect to differences and overlaps of the two approaches (Table 3). A large fraction of the HPRD interactions cannot be retrieved from the abstracts. This is demonstrated by the analysis of co-occurrences: only approximately half of the interactions annotated in HPRD can be found in abstract sentences. RelEx extracts a significantly larger number of relations from the abstracts than the number of interactions contained in HPRD.

We analyzed this discrepancy by randomly selecting 50 abstracts annotated in HPRD and annotated these manually (hprd50 dataset).

4.2.1 Comparing RelEx relations with HPRD interactions The hprd50 dataset allows us to estimate the performance based on the abstracts referenced by HPRD (Table 3) and thus to examine the differences between RelEx relations and HPRD interactions. The performance on this data set is slightly lower than on the LLL-challenge dataset. This is in part due to several quite long and complicated sentences. Second, the focus on human genes/proteins represents a more difficult challenge as the multi-word gene and protein names in certain cases impair the construction or analysis of the parse tree.

As shown in Figure 5, many of the HPRD interactions could not be retrieved by RelEx because they were not mentioned in the abstracts at all. We found that a number of additional interactions not annotated by HPRD are contained in the abstracts that exceeds the number of HPRD interactions extracted from the full text articles. Indeed, HPRD and RelEx reported about the same number of valid interactions per paper/abstract.

HPRD is focused on disease-related genes and thus does not yet cover the entire gene/protein space. HPRD makes use of full text articles, yet, abstracts and articles are not necessarily completely annotated, i.e. only a part of the relations mentioned in an abstract or article may be covered. Further differences to our annotation can be explained by the observation that HPRD focuses on direct physical protein–protein interaction data. Gene regulatory relations as well as long-range relations are not covered. Indeed, 17 of the 26 HPRD interactions contained in our manually annotated set were described using just two verbs, 'interact(s/ed/ion)' and 'binds/bound'. The remaining relations contain words like 'cross-link', 'coprecipitated', 'adapter'. This indicates that HPRD uses quite stringent annotation guidelines focused on direct physical interactions; most of them being described with a rather limited set of words and expressions.

Our results indicate, that HPRD, even though being a very large and valuable source for protein interaction data, currently covers only a small part of the human protein-protein relations from very limited relation categories. RelEx provides complementary information.

Table 3. Results of large-scale application of RelEx on a comprehensive set of MEDLINE abstracts (∼1 million abstracts) and comparison against HPRD

	Co-occurrences	RelEx
Instances ($cooc_{sen}/rel_{sen}$)	3.381.602	731.432
Number interacting gene/protein pairs	359.173	149.778
HPRD - Overlap1 (%)	51	40
HPRD - Overlap2 (%)	5	8

Overlaps were determined for pairs of genes/proteins, restricted to the set of genes/proteins common to HPRD and co-occurrence search (5925 genes/proteins), and irrespective of the individual abstract. Overlap1: Proportion of HPRD-relations found by co-occurrence/RelEx; Overlap2: Proportion of co-occurrences/RelEx-relations available in HPRD.

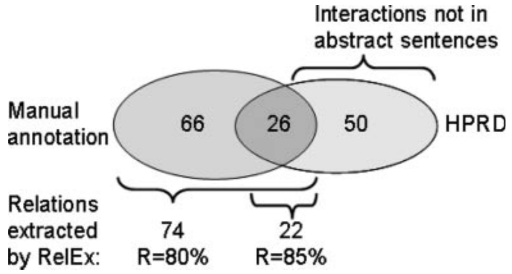

Interactions not in abstract sentences
Manual annotation
66
26
50
HPRD
Relations extracted by RelEx:
74
R=80%
22
R=85%

Fig. 5. Comparison of manually annotated relations, HPRD interactions and relations extracted by RelEx based on the hprd50 dataset (numbers correspond to relations rel_{abs}, R: Recall).

5 CONCLUSIONS

We developed RelEx, a tool for compiling a comprehensive set of causal and physical protein–gene interactions from free text. RelEx is based upon a number of publicly available tools and a simple set of rules. Compared to other approaches it is fairly straight forward to implement and achieves competitive performance.

RelEx is able to cope with different organism domains, which has been validated on publicly available datasets for human and prokaryote interactions. It can be adapted to different kinds of relations by usage of corresponding relation restriction terms and/or entity synonyms.

If RelEx is compared on the rather stringent criteria of the LLL challenge dataset (Nédellec, 2005), performance is significantly higher than previously reported results. Here, the ability is analyzed to specifically extract relations from particular sentences. Most of the published approaches compare themselves regarding the extraction of relations from abstracts, which is considerably relaxed compared to the former criteria. Here, the RelEx performance is in the range of existing approaches (Hu *et al*., 2005; Ono *et al*., 2001; Saric *et al*., 2006). It should be noted, though, that most of the approaches published so far were evaluated on datasets individually created by the authors, generally focused on a very restricted set of interaction types or descriptions (e.g. phosphorylation events). Frequently,

Downloaded from https://academic.oup.com/bioinformatics/article/23/3/365/236564 by guest on 16 August 2022

neither the used benchmark sets nor the software is made available. The LLL challenge dataset can also be considered as rather small, yet, due to its public availability it allows for comparison of methods, and, most importantly, independent evaluation.

In contrast to many other approaches, RelEx can be applied to large corpora. We applied RelEx to ∼1 million abstracts (a comprehensive subset of MEDLINE enriched in human protein–protein interactions) and presented some first results from this large-scale relation extraction. We found ∼150 000 interacting protein pairs and ∼731 000 text passages describing these interactions with an expected recall of 78% and precision of 79%. We showed that the number of valid interactions reported by RelEx per paper is virtually the same compared to large scale annotation approaches like HPRD (Peri *et al*., 2004) even though HPRD manually annotates full text articles instead of just the abstracts. On the other hand, RelEx is able to process far more abstracts and thus yields more interactions. Nevertheless, this requires about a week on a typical Linux cluster (40 Intel Xeon CPUs); the largest part of this time being devoted to dependency parsing. Of course the performance on the whole MEDLINE is difficult to judge as the estimation is based on small hand curated benchmark sets.

Importantly, RelEx not only returns pairs of elements identified to interact, but also assigns public database identifiers to the elements allowing for adding further annotations to texts and objects (Szugat *et al*., 2005). Thus, other data sources can be linked, enabling network-based analysis methods taking experimental data into account (e.g. Küffner *et al*., 2005; Sohler *et al*., 2004). The extracted paths also provide references into abstracts and contexts for the extracted relations. A particular path contains just the relevant subset of terms from a sentence describing a given relation. The paths have already been used to further classify relations as activating/inhibitory, physical/indirect, protein–protein/protein–gene (Küffner *et al*., 2006). Typed relations will help in analyzing pathways and provide a first step in inferring regulatory cascades.

ACKNOWLEDGEMENTS

This work was funded by projects BEX (Sanofi-Aventis, Frankfurt) and BOA (German ministry for research and education, grant 01GG9824).

Conflict of Interest: none declared.

REFERENCES

Blaschke,C. and Valencia,A. (2001) The potential use of suiseki as a protein interaction discovery tool. *Genome Inform. Ser. Workshop Genome Inform*., **12**, 123–134.

Blaschke,C. *et al.* (1999) Automatic extraction of biological information from scientific text: protein-protein interactions. *Proc. Int. Conf. Intell. Syst. Mol. Biol*., 60–67.

Colosimo,M.E. *et al*. (2005) Data preparation and interannotator agreement: Biocreative task 1b. *BMC Bioinformatics*, **6** (Suppl 1), S12.

Ding,J. *et al*. (2002) Mining medline: abstracts, sentences, or phrases? *Pac. Symp. Biocomput*., **7**, 326–337.

Divoli,A. and Attwood,T.K. (2005) BioIE: extracting informative sentences from the biomedical literature. *Bioinformatics*, **21**, 2138–2139.

Fundel,K. and Zimmer,R. (2006) Gene and protein nomenclature in public databases. *BMC Bioinformatics*, **7**, 372.

Hakenberg,J. *et al*. (2005) LLL'05 challenge: genic interaction extraction—identification of language patterns based on alignment and finite state automata. In *Proceedings of the ICML05 workshop: Learning Language in Logic (LLL05)*. pp. 38–45.

Hanisch,D. *et al*. (2005) Prominer: rule-based protein and gene entity recognition. *BMC Bioinformatics*, **6** (Suppl 1), S14.

Huang,M. *et al*. (2004) Discovering patterns to extract protein–protein interactions from full texts. *Bioinformatics*, **20**, 3604–3612.

Hu,Z.Z. *et al*. (2005) Literature mining and database annotation of protein phosphorylation using a rule-based system. *Bioinformatics*, **21**, 2759–2765.

Jelier,R. *et al*. (2005) Co-occurrence based meta-analysis of scientific texts: retrieving biological relationships between genes. *Bioinformatics*, **21**, 2049–2058.

Jenssen,T.K. *et al*. (2001) A literature network of human genes for high-throughput analysis of gene expression. *Nat. Genet*., **28**, 21–28.

Klein,D. and Manning,C.D. (2002) Fast exact inference with a factored model for natural language parsing. *Adv. Neural Inform. Proc. Syst*., **15**, 3–10.

Klein,D. and Manning,C.D. (2003) Accurate unlexicalized parsing. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics*.

Küffner,R. *et al*. (2005) Expert knowledge without the expert: integrated analysis of gene expression and literature to derive active functional contexts. *Bioinformatics*, **21** (Suppl 2), ii259–ii267.

Küffner,R. *et al*. (2006) Characterization of protein interactions. In *German Conference on Bioinformatics (GCB) 2006: GI-Edition Lecture Notes in Informatics (LNI)*. pp. 64–73.

Leroy,G. and Chen,H. (2002) Filling preposition-based templates to capture information from medical abstracts. *Pac. Symp. Biocomput*., **7**, 350–361.

Mel'cuk,I. (1988) *Dependency Syntax: Theory and Practice*. State University Press of New York, NY.

Nédellec,C. (2005) Learning language in logic - genic interaction extraction challenge. In *Proceedings of the ICML05 workshop: Learning Language in Logic (LLL05)*.

Ngai,G. and Florian,R. (2001) Transformation-based learning in the fast lane. In *Proceedings of the North American Chapter of the Association for Computational Linguistics on Language Technologies 2001 NAACL '01*. pp. 40–47.

Ono,T. *et al*. (2001) Automated extraction of information on protein–protein interactions from the biological literature. *Bioinformatics*, **17**, 155–161.

Peri,S. *et al*. (2004) Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Res*., **32**, D497–D501.

Saric,J. *et al*. (2006) Extraction of regulatory gene/protein networks from medline. *Bioinformatics*, **22**, 645–650.

Smith,L. *et al*. (2004) Medpost: a part-of-speech tagger for biomedical text. *Bioinformatics*, **20**, 2320–2321.

Sohler,F. *et al*. (2004) New methods for joint analysis of biological networks and expression data. *Bioinformatics*, **20**, 1517–1521.

Szugat,M. *et al*. (2005) Web servicing the biological office. *Bioinformatics*, **21** (Suppl 2), ii268–ii269.

Thomas,J. *et al*. (2000) Automatic extraction of protein interactions from scientific abstracts. *Pac. Symp. Biocomput*., **5**, 541–552.

Wilbur,W.J. *et al*. (2006) New directions in biomedical text annotation: definitions, guidelines and corpus construction. *BMC Bioinformatics*, **7**, 356.

Yakushiji,A. *et al*. (2001) Event extraction from biomedical papers using a full parser. *Pac. Symp. Biocomput*., **6**, 408–419.

Downloaded from https://academic.oup.com/bioinformatics/article/23/3/365/236564 by guest on 16 August 2022