Journal of Biomedical Informatics 43 (2010) 200–207

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Measuring prediction capacity of individual verbs for the identification of protein interactions

Dietrich Rebholz-Schuhmann *, Antonio Jimeno-Yepes, Miguel Arregui, Harald Kirsch

European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

ARTICLE INFO

Article history:
Received 7 October 2008
Available online 8 October 2009

Keywords:
Text mining
Information extraction
Protein-protein interaction
Semantic relations

ABSTRACT

Motivation: The identification of events such as protein–protein interactions (PPIs) from the scientific literature is a complex task. One of the reasons is that there is no formal syntax to denote such relations in the scientific literature. Nonetheless, it is important to understand such relational event representations to improve information extraction solutions (e.g., for gene regulatory events).

In this study, we analyze publicly available protein interaction corpora (AIMed, BioInfer, BioCreAtIve II) to determine the scope of verbs used to denote protein interactions and to measure their predictive capacity for the identification of PPI events. Our analysis is based on syntactical language patterns. This restriction has the advantage that the verb mention is used as the independent variable in the experiments enabling comparability of results in the usage of the verbs. The initial selection of verbs has been generated from a systematic analysis of the scientific literature and existing corpora for PPIs.

We distinguish modifying interactions (MIs) such as posttranslational modifications (PTMs) from non-modifying interactions (NMIs) and assumed that MIs have a higher predictive capacity due to stronger scientific evidence proving the interaction. We found that MIs are less frequent in the corpus but can be extracted at the same precision levels as PPIs. A significant portion of correct PPI reportings in the BioCreAtIve II corpus use the verb "*associate*", which semantically does not prove a relation.

The performance of every monitored verb is listed and allows the selection of specific verbs to improve the performance of PPI extraction solutions. Programmatic access to the text processing modules is available online (www.ebi.ac.uk/webservices/whatizit/info.jsf) and the full analysis of Medline abstracts will be made through the Web pages of the Rebholz group.

1. Introduction

Since the innovative approach of Blaschke et al. [1], a number of solutions for the identification of binary relations such as protein–protein interactions (PPIs) have been proposed. Until today, no solution is yet publicly available that at the same time identifies from the scientific literature the protein and gene names (PGNs), links them to the concept id (CID) in the biomedical data resources (e.g., to the accession number in UniProtKb) and reads out the relation between two PGNs at a high precision rate (precision = # correctly identified results/all identified results). Several solutions have been proposed (see Section 1.1), including the one that is best-known and called iHOP [2], but none of them offers a comprehensive approach.

Molecular biologists invest a significant portion of their research work to achieve better understanding of the molecular mechanisms of PPIs and use their experimental approaches to identify the type of interactions. This research leads to results that can be described in different ways and that need appropriate text mining methods to deliver results according to their needs. In this research work we explore on the use of language in the scientific literature, in particular in annotated corpora for PPIs to better understand the use of verbs in this context. We follow the hypothesis that PPIs from separate conceptual categories have different linguistic representations: (a) interactions with chemical modifications to one interaction partner ("modifying interaction", MI) and (b) interactions without such changes ("non-modifying interactions", NMI). The distinction between these types is motivated by the assumption that strong experimental proof for the MIs leads to explicit statements in the scientific literature reporting on the interaction (e.g., explicit mention of the interaction partners) and as a consequence, information extraction techniques will achieve better performances in identifying the interactions.

Modifying interactions. The evidence for the MIs is any reporting of chemical changes linked to the interaction partners of the PPI, such as methylation and demethylation and similarly phosphorylation and dephosphorylation as well as other types of chemical

* Corresponding author. Fax: +44 0 1223 494468.
E-mail address: rebholz@ebi.ac.uk (D. Rebholz-Schuhmann).

1532-0464/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2009.09.007

Table 1
Use of verbs in the extraction of binary relations for PPIs that have mentions in at least three scientific publications for use in different information extraction solutions (verbs denoting modifying interactions are represented in boldface). Tk [4], Hg [13], Fm [5], Rz [14], Pk [11], Bk [1] and Sz [7].

	Tk	Hg	Fm	Rz	Pk	Bk	Sz	Total
Activate	1	1		1	1	1	1	6
Bind	1	1	1			1	1	5
Interact	1	1	1	1		1	1	6
Regulate	1	1	1			1	1	5
Phorphorylate	1	1	1	1		1		5
Inhibit	1	1			1	1		4
Down-regulate	1	1	1		1			4
Express	1	1	1	1				4
Suppress	1	1	1			1		4
Up-regulate	1	1	1		1			4
Associate	1	1				1		3
Block	1	1			1			3
Contain	1		1	1				3
Dephosphorylate	1		1	1				3
Inactivate	1	1		1				3
Induce	1	1			1			3
Methylate	1		1	1				3
Modify	1	1	1					3
Overexpress	1		1	1				3
Promote	1	1		1				3
Stimulate	1	1			1			3
Substitute	1		1	1				3
Total	22	17	14	11	7	8	4	

changes (e.g., acetylation, biotinylation; see Table 1). These modifications can be subsumed as posttranslational modifications (PTMs), which are a subcategory of PPIs. Saric et al. [3] have considered this type of interactions into their work. Since the experimental evidence for the interaction shows chemical modifications, it is clear that the two proteins are interacting.

Non-modifying interactions. The second group of reported PPIs forms the largest set and has been commonly used for the identification of PPIs [1,4,5]. This group contains all reported results, where for example one protein activates or binds another protein. This set of interactions is relevant to molecular biologists searching for clues to reconstruct regulatory and signaling pathways in the cell.

The proposed categorization meets the demands from members of curation teams at the EBI that require integration of different interaction types (modifying and non-modifying interactions) into public services (Protein Corral, unpublished). These services will now be properly assessed, after an appropriate evaluation corpus has been made available: the evaluation corpus for PPIs as part of the BioCreAtIve II challenge [6].

Altogether, few results have been presented that report on the use of verbs (and their nominalizations) in the scientific literature as part of information extraction solutions for PPIs. In particular, the distribution of verbal forms for PPIs has not been properly assessed, for example through the analysis of an annotated corpus. Furthermore, we have discovered that only few research teams have integrated verbal forms that represent modifying interactions into their information extraction solutions. As a result, we investigated into the question whether statements on PPIs reporting modifying interactions can be extracted with higher precision than the set of verbal forms that are classified as non-modifying interactions.

In the following sections we will report on the approaches used from other research teams and their choice of verbal forms (Section 1.1). We explain our applied techniques (Section 2) and our experiments and findings (Section 3). Then we discuss our findings (Section 4) and the last section reports on the access to the presented solutions (Section 6).

1.1. Related work

Several researchers have reported on their information extraction solutions, some of which make reference to the verb forms that are integrated into their approach. The earliest work is from Sekimizu et al. [7] who identify frequently used verbs in Medline to parse relations amongst genes from the literature, but they do not provide a list of verbs denoting protein relations. Blaschke et al. [1] applied a text mining solution to Medline abstracts that identifies keywords in conjunction with a selection of verbs (11 verbs and their inflections and three additional language patterns). No evaluation is given in the publication. The selected verbs are listed in Table 1. A similar solution has been proposed by Ono et al. [8] that focuses on relations defined by "*interact*", "*bind*", "*associate*" and "*complex*" and that have been extracted with regular expression for syntactical patterns. Precision is around 94% and recall is around 85% (82.5–86.8%, recall = # correctly identified results/# all known correct results; see Table 5). iHOP also makes use of a rule-based system, but neither the list of verbs nor the evaluation has been disclosed [2].

Several solutions have been proposed that match language patterns in form of finite state automata (FSA) to the scientific literature. Pustejovsky et al. [9] analyzed syntactical language patterns for inhibitory events. Their system performed at recall of 57% and at precision of 90%. They did not offer any solution for the PGN normalization. Part of their solution is the processing of subordinate clauses, sentential coordination and anaphoric resolutions. Leroy et al. [10] also applied cascaded FSAs to extract PPIs from Medline abstracts. They reported 90% precision, but again did not apply any PGN normalization. Saric et al. [3] extracted regulatory gene/protein networks from Medline with cascaded FSAs. They do not state the set of verbs that were used to identify the relations. They suggest that a given verb (e.g., *activate*) can be used to "express different types of relations" (called semantic variation). For their evaluation they opted for semantic correctness in contrast to grammatical correctness and claim to have achieved 83–90% accuracy for expression relations and 86–95% accuracy for phosphorylation relations.

Park et al. [11] applied combinatory categorial grammar in conjunction of seven verbs (including their inflections and noun phrases) denoting a positive regulatory effect (e.g., *activate*, *stimulate*) and five verbs denoting a negative regulatory effect (*inhibit*, *down-regulate*). They consider solving coordination, appositions and anaphoric expressions. They claim 48% recall and 80% precision measured on a selection of 492 sentences.

Friedman et al. [5] use a system that parses text based on grammar rules (semantic patterns, MedLee). The grammar makes use of 22 terms denoting verbs and also nouns (e.g., apoptosis, myogenesis) that are categorized into 14 classes representing actions and processes. All inflectional forms and the nominalizations have been considered for all verbs. NER for PGNs is based on BLAST. [4] applied context-free grammar for the processing of Medline abstracts. They integrated 49 verb forms, their inflectional forms and nominalizations and achieved 63.9% recall at 70.2% precision. Temkin thus reports the biggest coverage of proposed verbs. Finally, Daraselia et al. [12] use context-free grammar to identify PPIs and report 91% precision based on Medline abstracts (recall rate 21%).

Huang et al. [13] did invest effort into an automated analysis of the scientific literature to identify syntactical patterns denoting PPIs. They applied alignment algorithms to the text, filtered out syntactical patterns and identified 30 verbs that are relevant for the identification of PPIs. The list of relevant verbs comprises 91 entries including specific verbs denoting modifying interactions (e.g., *ubiquitinate*), non-modifying interactions (e.g., *interact*, *bind*) and undefined interactions (e.g., *hasten*, *function*) with regards to

protein-related events. The authors use the patterns to extract "interacting" and "binding" relations and report around and above 80% recall and 80% precision. The PGNs are not normalized to CIDs.

Apart from the sheer numbers for precision and recall of the different systems, it is difficult to assess the IE systems against each other. This is mainly due to the fact that the performances have not been measured on a shared corpus. Certainly the approaches based on regular expressions and FSAs scale with the number of language patterns that have been integrated. The number of patterns increases the recall. On the other side, every pattern has to be well crafted to not over-generalize and to reduce the precision of the IE system in an intolerable way. The proposed systems make use of part-of-speech tagging and basic syntactical structures (e.g., identification of noun phrases) to comply with standards in natural language processing approaches.

Those solutions that base on syntactic parsing techniques (combinatorial categorical grammar (CCG) by Park, context-free grammar (CFG) by Daraselia and Temkin) achieve to cover a larger set of verbs and approach more sophisticated language phenomena, such as anaphora resolution. In this respect, we have to give credit to Pustejovsky, too.

The listed verbs (see Table 1) from previous publications will all be considered in our analysis, i.e. use of verbs in the protein interaction corpora and performance measurements on BioCreAtIve II corpus.

2. Methods

The identification of PPIs from the literature is a complex task, which is composed of named entity recognition for proteins, protein name normalization (i.e. identification of the correct CID) and the extraction of the relation between both entities. For the evaluation we relied on the BioCreAtIve II corpus for the IPS task (347,749 sentences from 740 full-text documents), on the AIMed corpus (1942 sentences from 255 abstracts) and on BioInfer (1100 sentences from full text) [6,15,16]. Only the BioCreAtIve corpus delivers a set of CID pairs for every contained document where the CID pair represents a PPI.

2.1. Named entity recognition for proteins/genes

The identification of PGNs has been studied extensively [17–19]. The identification of gene mentions has been solved to a precision close to 90% whereas the gene normalization is still ongoing work. In the presented work, we used two methods that both deliver CIDs as part of the NER task. The first method (SP-tagger) is part of several TM solutions at the EBI (EbiMed, PCorral, MedEvi; [20]). It incorporates all protein names from UniProtKb/SwissProt and their synonyms. Named entity recognition is mainly done by dictionary lookup under consideration of morphological variability [21]. After the first identification step, additional features are considered, for example resolution of acronyms and term frequencies from the British National Corpus to increase the precision of the NER module [22,23]; for SOAP Web services access see [24]. The performance of this tagger is 76% F-measure (precision 95%, recall 64%).

The second protein-tagger again uses dictionary lookup in combination with contextual information to disambiguate gene mentions and to identify the correct boundaries (BL-Tagger [25]). The performance of this method showed 75% F-measure (precision 94%, recall 63%). The underlying term repository is a publicly available lexical resource for biological terms (http://www.ebi.ac.uk/Rebholz-srv/BootStrep/bootstrep.html). The term repository incorporates the BioThesaurus and other terminological resources [26].

Both taggers have been used in this study, since they are based on different design principles. The SP-tagger shows better performance, but the BL-tagger has the advantage that the underlying lexical resource[1] is publicly available. The SP-Tagger is species independent, whereas the BL-Tagger delivers only annotations for human proteins.

2.2. Identification of PPIs

The identification of PPIs from the text is based on the modules of the Whatizit infrastructure [24]. Public access is granted to all modules that are used in this study. Most modules are implemented as finite state automata [27]. The basic NLP modules of the infrastructure comprise the sentenciser and a part-of-speech (PoS) tagger. The PoS tagger was trained on the British National Corpus and incorporates a large-scale biomedical terminological resource to improve the performance on biomedical scientific literature. Noun phrases (NPs) are identified with syntax patterns equivalent to "**DET (ADJ|ADV)+ N+**".

2.2.1. PPI identification based on tri-cooccurrence (3-CO)

For our study we assessed 3-CO against syntactical patterns denoting a PPI (SynP). Both approaches restrict the scope of syntactical expressions that are accepted as representations of PPIs, but have the advantage that they do not rely on more complex solutions to filter out the syntactical structure for the relations from the dependency structure or the predicate-argument structure of a parser. This leads to the result, that the usage and the performance of the verbs can be monitored under restricted but standardized conditions.

3-CO is performed on the stretch of a sentence. Any triplet of two proteins in combination with a verb mention in the following combinations is accepted: (1) "**PGN VP PGN**", (2) "**nomVP PGN PGN**", and (3) "**PGN PGN nomVP**", where VP is the verb phrase that represents all the conjugational forms of a form and nomVP is the nominalized form of a verb phrase. Only the pre-specified verbs are counted and in the case of coordination of two such verbs, both are counted.

The module that identifies and highlights PPIs searches for phrases that contain a verb or a nominal form describing an interaction like bind or dimerization. The first set comprises all verbal forms that report on chemical modifications of a protein: *acetylate*, *acylate*, *amidate*, *brominate*, *biotinylate*, *carboxylate*, *cysteinylate*, *farnesylate*, *formylate*, "*hydrox[iy]late*", *methylate*, *demethylate*, "*myristo?ylate*", "*palmito?ylate*", *phosphorylate*, *dephosphorylate*, *pyruvate*, *nitrosylate*, *sumoylate*, "*ubiquitin(yl)?ate*". The second set of verbs consists of forms that report on interaction and regulation events: *associate*, *dissociate*, *assemble*, *attach*, *bind*, *complex*, *contact*, *couple*, "*(multi|di)meri[zs]e*", *link*, *interact*, *precipitate*, *regulate*, *inhibit*, *activate*, "*down[-]regulate*", *express*, *suppress*, "*up[-]regulate*", *block*, *contain*, *inactivate*, *induce*, *modify*, *overexpress*, *promote*, *stimulate*, *substitute*, *catalyze*, *cleave*, *conjugate*, *disassemble*, *discharge*, *mediate*, *modulate*, *repress*, *transactivate*. "Associate" does not denote any specific binding or transformation event.

If two different verbs have been identified in the context of a CID pair, then both occurrences have been counted. This is also the case for CID pairs that have been identified with syntactical patterns (see below), but then only takes place at a low frequency.

2.2.2. PPI identification based on syntactical patterns

The identification of the syntactical patterns representing PPIs is a more complex process and a computationally intensive task in comparison to the 3-CO analysis. It covers the following

[1] http://www.ebi.ac.uk/Rebholz-srv/BootStrep/bootstrep.html.

components. One module identifies single adjectives (“adj”), combinations of adjectives and adverbs and coordination of adverbs. The second module selects the conjugational forms of “to be”, also in combination with leading, interleaving and trailing adverbs (“beForm”; see Fig. 1). The next module, seeks phrases like “were initially observed” to be combined with “to” and the infinitive of an interaction verb (“shownForm”). In the same sense, modal verbs with optional trailing adverbs, where modal verbs are any of: *can*, *could*, *cannot*, *do*, *may*, *might*, *must*, *need*, *ought*, *shall*, *should*, *would*.

The identification of verb phrases is composed of five parts: **Vsimple** covers the verb itself with only optional leading or trailing adverbs. **Vprep** extends Vsimple by a trailing preposition to catch expressions such as “bound to” or “interact with”. **Vbe** extends both of the above by allowing any of the matches produced by the “beform” stage in front of them and thus captures phrases such as “is regulated” or “are positively regulated by”, **Vshown** allows a match of the Shown stage followed by “to” and a match of Beforms in front of Vsimple and Vprep. This will tag phrases like “have been shown to be phosphorylated”. Finally, **Vmodal** works like Vshown, but uses a modal verb from the “shownForm” stage. It will catch phrases like “may be linked to”.

The identification of noun phrases (NP) selects nouns in combination with adjective modifiers, including coordination of ADJ elements in front of a sequence of nouns. PGNs are treated as nouns. NPs do not include determiners (e.g., “novel orphan receptor TAK1”). Finally the PPI patterns are identified. They are basically combinations of the previously identified information, such as ***NP_P VP det? NP_P*** and ***NP_P VP det? NP of NP_P***, where NP_P is an NP that contains the identified protein.

These construction rules for syntactical patterns lead to the selection of structures that are similar to 3-CO representations but produce results with higher precision. Similar structures have been proposed by Huang et al. [13]. The syntactical patterns apply the same word order as used in the 3-CO extraction method and in addition specify better the verb phrases that are accepted for the extraction of PPIs. The “shownForm” phrase accounts for the hedging used by authors and thus increase the recall of the approach. In the same vein, the use of syntactical patterns denoting nominalizations improve the recall for the identification of PPIs and follow the representation ***VP_NP “(of | with | between | through | from)” det? NP_P “(and | with | within | via | through | by)” det? NP_P***, where VP_NP is the nominalization of the verb form.

2.3. Evaluation

The evaluation was performed using the corpus that has been provided as for the BioCreAtIve II challenge, protein interaction pair sub-task 2 (IPS). The participants of this sub-task had to identify protein interaction pairs from the full-text document. In more detail, the documents in the training and testing corpus were annotated with pairs of protein identifiers, where both proteins of a given pair are known to interact, and the document is known to deliver the evidence to this interaction. The curators of the corpus did not annotate the mentions of the proteins in the documents and therefore it is not possible to measure the performance of the PPI identification methods against individual mentions of interaction pairs in the text. The other available corpora, i.e. AIMed and BioInfer, are small in comparison to the BioCreAtIve IPS corpus and have been deemed less suitable.

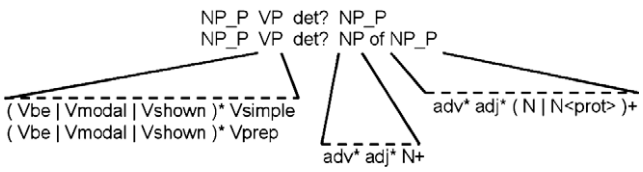

Fig. 1. (Syntactical patterns) The diagram explains the composition of the language patterns. The verb phrase (VP) is composed of several subcomponents that enable the identification of modal verbs (Vmodal), forms of to be (Vbe) and common forms of hedging (Vshown). NP_P is a noun phrase containing a protein mention. For further details refer to Section 2.

3. Results

Identification of verbs denoting PPIs. In the first step we analyzed all three available corpora, i.e. AIMed, BioInfer and BioCreAtIve, and extracted all verbs that cooccur with two mentions of a PGN. This resulted to the identification of 967 verbs for the BioCreAtIve corpus, 165 for AIMed and 162 for BioInfer. 90 were shared in all three corpora. Modal verbs (e.g., do, have) were only considered if they did not appear in combination with other verb forms. Apart from the domain-specific verbs (see Section 2), a large list of general English verbs were extracted: *encode*, *suggest*, *use*, *show*, *test*. They are part of idiomatic phrases such as “we have shown that” or the “encoded protein“. The first type is covered by our syntactical patterns if used as part of the textual protein interaction description.

From the list of NMI verbs 5 were not contained in AIMed (*attach*, *catalyze*, *disassemble*, *modify*, *overexpress*; see Table 6), 5 not in BioInfer (*dimerize*, *down[-]?regulate*, *repress*, *substitute*, *transactivate*) and 3 only in BioCreAtIve (*conjugate*, *multimerize*, *up[-]?regulate*). This shows that the BioCreAtIve corpus has the biggest coverage. It is a small surprise that “*up-regulate*” is not more commonly used.

Regarding the verbs categorized as MI only “*phosphorylate*” appeared in all three corpora and “*acylate*” in two corpora (i.e. not in AIMed). Four verbs appeared only in the BioCreAtIve corpus (*biotinylate*, *dephosphorylate*, *methylate*, *pyruvate*). This leads to the result that MIs are preferably reported in the full-text document and at a low frequency. A complete Medline analysis has lead to the result that only a few verbs for MIs (*biotinylate*, *dophosphorylate*, *hydroxylate*, *methylate*, *phosphorylate*, *pyruvate*) are applied in conjunction with mentions of PGNs, whereas all verbs for NMIs are in use.

The following analysis focuses on the BioCreAtIve corpus only, since it is the largest corpus and the previous figures demonstrate that it provides the largest coverage of relevant verbs.

3.1. Comparison of NER tagging results

3.1.1. Identification of CIDs from abstracts and from full-text documents

In our analysis we used two different Protein-taggers: SP-tagger and BL-tagger. The highest precision was achieved with the SP-tagger on the abstracts (51.4%) and with the BL-tagger on abstracts as well (72.7%). In this evaluation, we considered only the correct identification of the matching CID as the correct solution. Precision was lower, when either one was run in case-insensitive mode (41.3% and 48.6%, respectively). F-measure was rather low in all cases (1.8–6.4%). The same approaches applied to full-text documents showed lower precision (14.8% for SP-tagger and 20.3% for BL-tagger, not shown), but F-measure was higher for the SynP extraction for all methods (6.3–6.8%, not shown). This reflects that a number of PPIs are not reported in the abstract.

In our experiments, we considered the result of the protein-taggers as correct, if the correct CID was contained in the list of attributed CIDs. The resulting number is similar to the frequency of the identified named entities in the text and enables better comparison of results between the different methods (3-CO vs. SynP). Upon this change, the recall on the abstract text and on the full-text

increases significantly (as expected) and on the other side, the precision was lowered to 37.5% and to 44.7% for the case-sensitive SP-tagger and BL-tagger, respectively.

Identification of CID pairs. When analyzing the full-text documents (see Table 2) in comparison to their abstracts, we find that the overall number of identified CID pairs increases (about 10-fold for all methods) as well as the number of correctly identified CID pairs (3- to 5-fold). Precision decreases, recall increases and the F-measure increases to 17.1% and 13.7% for the SP-tagger and the BL-tagger, respectively. The results show, that the full-text documents still contain a significant number of binary relations that make reference to any of the before-mentioned verbs, but that are not relevant interaction pairs.

The evidence extracted with SynP is a true subset of the evidence from the 3-CO method leading to the result that about 50% (49.9–58.8%) of the evidence from 3-CO can be confirmed by the approach using syntactical language patterns. This can be explained by the fact that the predictions are counts of unique CID pairs, which again can be represented by a number of instances in the document. The redundancy in the document counter-balances lower recall of the SynP methods over the 3-CO methods. In the next step we investigated into the distribution of the verb forms that were part of our two approaches.

Identification of CID pairs denoting MIs and NMIs. According to our categorization, we find the following numbers for events representing MIs and NMIs (see Table 3). The most correct predictions are reported in the set of NMIs (325) and the smallest number in the set of MIs (23). Altogether, MIs have a small contribution to all PPIs in the Bio-CreAtIve II corpus. The precision is for both types of events in the same range (18.5% and 17.2%, respectively). Similar results are gained when only processing the abstracts (MI: seven agreements for 18 predictions; NMI: 64 agreements for 241 predictions).

To our surprise, the association of proteins has a significant contribution to the correct identification of relations between proteins. This result is unexpected, since the association of two proteins does not give any clues on the underlying relatedness of the proteins, i.e. a relation based on binding, regulatory or transformational effects.

We further investigated, whether the associated relations have been confirmed by other types of relations. For this analysis we used the case-insensitive BL-tagger that generated higher recall at lower precision than the SP-tagger. From the 64 correct predictions out of 241 total predictions (66 out of 171 for the SP-tagger, see Table 3) 44 were confirmed by NMI interactions (41 correct out of 95 total) and by MI interactions (three correct predictions of five shared predictions). This shows that the authors deliver the scientific evidence to a protein–protein interaction even if the proteins are linked via the verb "associate" in the text.

Identifying the predictive capacity of verbs. We now analyzed the whole set of verb forms used in the corpus to better understand the distribution of contributions in terms of correct predictions (see Table 4). All verbs that are mentioned in Section 2 and that are not listed in Table 4, did not contribute to correct predictions. Amongst these verbs are the following: *upregulate*, *dissociate*, *couple*, *link*, *overexpress*, *repress*, *inactivate*, *cleave* and *acetylate*. When comparing the list of verbs from Table 4 to the proposed verbs from other authors (see Table 1) then we can identify that the verbs "*downregulate*", "*upregulate*", "*inactivate*" and "*stimulate*" do not play an important role, whereas "*associate*" and "*contain*" play an important role for the predictions.

In all our experiments, we found that the integration of nominalizations lead to a decrease in precision of up to 3% in all cases (for 3-CO and SynP, for SP-tagger and BL-tagger), but the F-measure improved up to 2% due to better recall (results not shown). Only in the case of 3-CO tested on full-text documents, the F-measure decreased. This is due to the fact, that 3-CO delivered already a significant portion of recall in the full-text document analysis and the use of nominalizations further decreased the precision.

Altogether our analysis has lead to a prioritized list of verbs that are relevant to the identification of modifying and non-modifying interactions from the scientific literature. The entries in the list can be used to optimize the performance of an information extraction solution, i.e. selection of verbs with a high F-measure to improve the precision/recall ratio of the IE solution and integration of the best performing verbs to improve the overall coverage of the solution. Certainly, more knowledge about the subframe categorizations of the listed verbs will help to further

Table 2
(Processing full-text documents, One-CID) Results for the identification of CID pairs from the BioCreAtIve full text corpus for 3-CO and SynP. SP (SwissProt-tagger), BL (BioLexicon-tagger), cs (case-sensitive), ci (case-insensitive), 3-CO (tri-cooccurrence), SynP (syntactical language patterns for PPIs).

	Predictions	Correct predictions	Precision (%)	Recall (%)	F-measure (%)
SP-cs, 3-CO	12,771	408	3.2	19.3	5.5
SP-cs, SynP	1539	211	13.7	10.0	11.6
SP-ci, 3-CO	15,823	609	3.8	28.8	6.8
Sp-ci, SynP	2078	358	17.2	17.0	17.1
BL-cs, 3-CO	13,671	479	3.5	22.7	6.1
BL-cs, SynP	1470	239	16.3	11.3	13.3
BL-ci, 3-CO	28,544	659	2.3	31.2	4.3
BL-ci, SynP	3229	367	11.4	17.4	13.7

Table 3
(Processing full-text documents, One-CID, SP-ci, SynP) Predictions from the full-text documents from BioCreAtIve II based on the case-insensitive use of the SP-tagger and SynP. All findings are categorized according to the category of the verb form that has been used in the text in conjunction with the mentioned proteins (see Section 2) (for use of acronyms see Table 2).

	Predictions	Correct predictions	Precision (%)	Recall (%)	F-measure (%)
All, 3-CO	15,823	609	3.8	28.8	6.8
All, SynP	2,078	358	17.2	17.0	17.1
Associate, 3-CO	1,203	180	15.0	8.5	10.9
Associate, SynP	171	66	38.6	3.1	5.8
MI, 3-CO	1,092	71	6.5	3.4	4.4
MI, SynP	124	23	18.5	1.1	2.1
NMI, 3-CO	14,833	596	4.0	28.2	7.0
NMI, SynP	1,893	325	17.2	15.4	16.2

Table 4
(Full-text, One-CID, BL-ci, SynP) All verbs that contributed to a correct prediction of related proteins. They are sorted according to their F-measure. The list can be used to tune an information extraction system for performance (e.g., for precision, recall, speed).

	Predictions	Correct predictions	Precision	Recall	F-measure
Interact	702	136	19.4	6.4	9.7
Bind	562	112	19.9	5.3	8.4
Associate	180	37	20.6	1.8	3.2
Phosphorylate	116	12	10.3	0.6	1.1
Regulate	179	12	6.7	0.6	1.0
Contain	286	12	4.2	0.6	1.0
Inhibit	130	9	6.9	0.4	0.8
Mediate	136	7	5.1	0.3	0.6
Activate	165	7	4.2	0.3	0.6
Modulate	31	5	16.1	0.2	0.5
Precipitate	31	4	12.9	0.2	0.4
Express	218	4	1.8	0.2	0.3
Promote	42	3	7.1	0.1	0.3
Induce	110	3	2.7	0.1	0.3
Modify	6	2	33.3	0.1	0.2
dephosphorylate	8	2	25.0	0.1	0.2
Complex	15	2	13.3	0.1	0.2
Stimulate	41	2	4.9	0.1	0.2
down-regulate	6	2	33.3	0.1	0.2
Methylate	6	1	16.7	0.0	0.1
Substitute	7	1	14.3	0.0	0.1
Assemble	11	1	9.1	0.0	0.1
Block	30	1	3.3	0.0	0.1
Suppress	40	1	2.5	0.0	0.1

Table 5
(Precision and recall of reported solutions) Precision and recall values reported by the authors of the solutions mentioned in Section 1.1. Saric et al. [3] only discloses the range of the accuracy of their system (see text).

First authors	Date of publication	Type of approach	Precision (%)	Recall (%)
Ono	2001	Regular expressions	94	83–87
Pustejovsky	2001	Language patterns	90	57%
Leroy	2003	Cascaded FSAs	90	NA
Saric	2006	Cascaded FSAs	NA	NA
Park	2001	CCG	80	48
Friedman	2001	Grammar rules	NA	NA
Temkin	2003	CFG	70	64
Daresalia	2004	CFG	91	21
Huang	2004	Syntactical patterns	80	80

Table 6
(Overview on the distribution of verbs) All three annotated corpora use full-text documents. In the case of AIMed and BioInfer the protein mentions are annotated in the sentence reporting on the interaction, in the BioCreAtive corpus, however, the document is delivered in conjunction with the protein ids of the interacting proteins in the document. In the case of the Medline analysis, first cooccurrences of proteins were identified to filter out the relevant verb between them or as normalization before them.

All verbs	**AIMed** Full text PGN mention Interactions	**BioInfer** Full text PGN mention Interactions	**BioCreAtive** Full text PGN normalization Documents	**All Medline** Abstracts NA NA
Modifying interactions (MI)				
Acetylate, acylate, amidate, brominate, biotinylate, carboxylate, cysteinylate, farnesylate, formylate, ''hydrox[iy]late'', methylate, demethylate, ''myristo?ylate'', ''palmito?ylate'', phosphorylate, dephosphorylate, pyruvate, nitrosylate, sumoylate, ''ubiquitin(yl)?ate''	Phosphorylate	Phosphorylate, acylate	Acylate, biotinylate, dephosphorylate, methylate, phosphorylate, pyruvate	Biotinylate, dophosphorylate, hydroxylate, methylate, phosphorylate, pyruvate
Non-modifying interactions (NMI)				
Associate, dissociate, assemble, attach, bind, complex, contact, couple, ''(multi|di)meri[zs]e'', link, interact, precipitate, regulate, inhibit, activate, ''down[-]regulate'', express, suppress, ''up[-]regulate'', block, contain, inactivate, induce, modify, overexpress, promote, stimulate, substitute, catalyze, cleave, conjugate, disassemble, discharge, mediate, modulate, repress, transactivate	All except {attach, catalyze, disassemble, modify, overexpress}	All except {dimerize, down[-]?regulate, repress, substitute, transactivate}	All except {conjugate, multimerize, up[-]?regulate}	All

optimize any IE solution and will give contributions to the event identification overall.

4. Discussion

In this work we have systematically analyzed the relevance of different verbs for the identification of PPIs from the scientific literature. We have gathered the verbs from different resources (e.g., scientific publications, annotated corpora) and have estimated the predictive capacity of individual verbs. We have defined the classes of modifying interactions containing all verb forms that report on a chemical transformation of one interaction partner (posttranslational modifications, e.g., methylation, acetylation, phosphorylation), and non-modifying interactions (e.g., interaction, binding, regulatory events). The last class is composed of the undefined interactions (e.g., associations, functions). Much to our surprise the single entry from the class of undefined interactions (''associate'') contributed significantly to the correct predictions in our analysis. A significant portion of the ''association'' of protein pairs could be confirmed by a more informative relation between the proteins from the same document. It needs to be mentioned that all statistical figures with regards to the verbs include the assumption that there is an even distribution of false PGN recognitions over all PPIs.

Friedman et al. [5] proposed a categorization of verbs into semantic classes for actions, process and other relations. It is more fine-grained and distinguishes positive regulation (''*activate*'') from negative regulation (''*inactivate*'') and proposes semantic classes related to bond formation (''*createbond*'', ''*breakbond*'') and general modification actions, reaction actions and others. This approach shows foresight, but could be too detailed to deliver conclusive results from information extraction. Temkin and Gilder [4] propose similar semantic classes. Future experiments will show, which granularity of semantic classes is most suitable for information extraction systems.

Our automatic analysis of the BioCreAtIve II corpus confirmed the top-ranked entries on the list of verbs that have already been used in different information extraction solutions. The verbs ''*phosphorylate*'' and ''*dephosphorylate*'' are the best-ranked protein relations from the list of MIs that have already been extensively analyzed by Saric et al. [3]. Other event denoting verbs have been exploited from a large number of researchers including ''*interact*'', ''*bind*'', ''*regulate*'' and ''*inhibit*''. Another unexpected result is the finding, that up- and downregulation is used at a low frequency in conjunction with protein relations. Verbs that are shared amongst all three corpora but occur at low frequency and that could be relevant for future analyses are: *abolish*, *affect*, *disrupt*, *increase*, *translocate* and *trigger*.

The difference in the performance of our protein-taggers is not very strong, although they apply different techniques for the extraction of the PGNs and one has been tuned for better recall (BL-tagger) whereas the other one is optimized for precision (SP-tagger). This could be due to the fact that full-text documents provide redundant mentions of protein interaction pairs.

The use of syntactic language patterns is a strong restriction. The use of other parsing techniques is work in progress. We assume that the ranking of the verbs according to their predictive capacity will not change when using syntactical parsing techniques, but the precision of the extraction methods will certainly improve. The interpretation of the results requires good knowledge of the underlying parsing techniques.

It is possible that a number of verbs have not been considered in this study, although they are closely related to PGNs and PPIs, if they are only identified with a low reliability. Since the BioCreAtIve corpus is already quite large, we would not expect many of them. Nonetheless, this would be an interesting question to follow up. Furthermore, it has to be kept in mind that our results most likely include a unified distribution of false PGN recognitions for all verbs. This problem could be partly resolved by exploiting the content of reference databases that contain PPIs and the corresponding references to Medline abstracts (similar to the BioCreAtIve IPS corpus). Similar approaches in the past have suffered from inconsistencies in the protein interaction databases (unpublished work).

The presented work contributes to the identification of subcategorization frames (ongoing work in the BootStrep project, www.bootstrep.org). The results help to focus the information extraction task to a selection of verbs that are part of protein interaction events. Certainly more advanced parsing techniques will contribute to improve the performance of the used approach (work in progress).

For the ongoing work in the extraction of gene regulatory events, we will analyze how MI and NMI events contribute to the event extraction. Furthermore, it has to be kept in mind that PPIs are events that report on the relation between two and more entities. Such events are frequently reported in the scientific literature based on syntactical representations that are more complex than binary relation representations between two entities and would even require input from ontological resources to correctly interpret (and possibly parse) the expressed relation. For example the statement ''cpxA gene increases [...] csgA transcription by dephosphorylation of CpxR'' requires domain knowledge to understand that only the product of the cpxA gene can induce via PPIs the activation of cpxR. Future research in the future will give a better understanding of the representation of binary and other relations as well as event representations in the scientific literature and the availability of more advanced ontologies will support the interpretation of such representations [28].

5. Conclusions

The semantic classification of verbs remains to be a challenging task. Our distinction of verbs with regards to their involvement in MI and NMI is meaningful to biologists, but does not lead to higher precision in the identification of MIs as expected. Overall protein–protein interactions are more frequently reported as NMIs than as MIs in the literature and the most complete set of verbs linked to NMIs has already been reported by Temkin and Gilder [4]. Individual verbs for NMIs are well known, e.g., *interact*, *bind*, others show mediocre performance, e.g., *activate*. The verb ''*associate*'' has a strong predictive value, but does not denote an interaction, whereas the verb ''*phosphorylate*'' is the most relevant amongst all verbs used for MI identification. We expect that the reporting on MIs will increase in the future if the screening methods for chemical interactions of proteins will improve.

6. Access to resources

The information extraction pipeline for the PPIs for MIs and NMIs is accessible via the Whatizit SOAP Web services. The information extraction solution described in this publication was used to extract relations between proteins from Medline. The latest compilation delivered the following results for the identified relations: inhibit (71,328), bind (40,407), interact (18,705), regulate (37,755), phosphorylate (8864), link (8399), complex (1542) and dissociate (1565).

Acknowledgments

This research was sponsored by the EC STREP project ''BOOTStrep'' (FP6-028099, www.bootstrep.org) including the development of the term repository (the pre-stage to the BioLexicon).

Whatizit has been funded by the NoE "Semantic Mining" (NoE 507505).

References

[1] Blaschke C, Andrade MA, Ouzounis C, Valencia A. Automatic extraction of biological information from scientific text: protein–protein interaction. Proc Int Conf Intell Syst Mol Biol 1999:60–7.
[2] Hoffmann R, Valencia A. Implementing the iHOP concept for navigation of biomedical literature. Bioinformatics 2005;21(Suppl. 2):ii252–8.
[3] Saric J et al. Extraction of regulatory gene/protein networks from Medline. Bioinformatics 2006;22(6):645–50.
[4] Temkin JM, Gilder MR. Extraction of protein interaction information from unstructured text using a context-free grammar. Bioinformatics 2003;19(16): 2046–53.
[5] Friedman C et al. GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. Bioinformatics 2001;17(Suppl. 1):S74–82.
[6] Krallinger M, Leitner F, Valencia A. Assessment of the second BioCrea-tive PPI task: automatic extraction of protein–protein interactions. In: Proceedings of the second BioCreative challenge evaluation workshop; 2007.
[7] Sekimizu T et al. Identifying the interaction between genes and gene products based on frequently seen verbs in Medline abstracts. Genome Informatics 1998:62–71.
[8] Ono T et al. Automated extraction of information on protein–protein interactions from the biological literature. Bioinformatics 2001;17(2):155–61.
[9] Pustejovsky J et al. Robust relational parsing over biomedical literature: extracting inhibit relations. In: Pac Symp Biocomput; 2001. p. 362–373.
[10] Leroy G et al. A shallow parser based on closed-class words to capture relations in biomedical text. J Biomed Inform 2003;36(3):145–58.
[11] Park JC et al. Bidirectional incremental parsing for automatic pathway identification with combinatory categorial grammar. In: Pac Symp Biocomput; 2001. p. 396–407.
[12] Daraselia N et al. Automatic extraction of gene ontology annotation and its correlation with clusters in protein networks. BMC Bioinformatics 2007;10(8):243.
[13] Huang M et al. Discovering patterns to extract protein–protein interactions from full texts. Bioinformatics 2004;20(18):3604–12.
[14] Rzhetsky A et al. GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. J Biomed Inform 2004;37(1): 43–53.
[15] Bunescu R, Ge R, Kate RJ, Marcotte EM, Mooney RJ, Ramani AK, et al. Comparative experiments on learning information extractors for proteins and their interactions. Artif Intell Med 2005;33(2):139–55.
[16] Pyysalo S, Ginter F, Heimonen J, Björne J, Boberg J, Järvinen J, et al. BioInfer: a corpus for information extraction in the biomedical domain. BMC Bioinformatics 2007;9(8):50.
[17] Morgan A, Hirschman L. Overview of BioCreative II gene normalization. In: Proceedings of the second BioCreative challenge evaluation workshop; 2007.
[18] Hakenberg J et al. Systematic feature evaluation for gene name recognition. BMC Bioinformatics 2005;6(Suppl. 1):S9.
[19] Hirschman L et al. Overview of BioCreAtIvE task 1B: normalized gene lists. BMC Bioinformatics 2005;6(Suppl. 1):S11.
[20] Rebholz-Schuhmann D et al. EBIMed: text crunching to gather facts for proteins from Medline. Bioinformatics 2007;23(2):e237–44.
[21] Tsuruoka Y et al. Learning string similarity measures for gene/protein name dictionary look-up using logistic regression. Bioinformatics 2007;23(20): 2768–74.
[22] Gaudan S, Kirsch H, Rebholz-Schuhmann D. Resolving abbreviations to their senses in Medline. Bioinformatics 2005;21(18):3658–64.
[23] Rebholz-Schuhmann D et al. Annotation and disambiguation of semantic types in biomedical text: a cascaded approach to named entity recognition. In: Workshop on "Multi-Dimensional Markup in NLP", EACL 2006, Trente, Italy; 2006.
[24] Rebholz-Schuhmann D, Arregui M, Gaudan S, Kirsch H, Jimeno A. Text Processing through Web Services: Calling Whatizit. Bioinformatics 2008;24(2):296–8.
[25] Pezik P, Jimeno A, Lee V, Rebholz-Schuhmann D. Static dictionary features for term polysemy identification. In: Proceedings of the language resources and evaluation conference (LREC-2008), Marrakech (Morocco), 28–30 May 2008; 2008.
[26] Liu H, Hu Z, Zhang J, Wu C. BioThesaurus: a web-based thesaurus of protein and gene names. Bioinformatics 2006;22(1):103–5.
[27] Kirsch H et al. Distributed modules for text annotation and IE applied to the biomedical domain. Int J Med Inform 2006;75(6):496–500.
[28] Beisswanger E, Lee V, Kim JJ, Rebholz-Schuhmann D, Splendiani A, Dameron O, et al. Gene regulation ontology (GRO): design principles and use cases. Stud Health Technol Inform 2008;136:9–14.