Negation of protein–protein interactions: analysis and extraction

Olivia Sanchez-Graillet\(^1\) and Massimo Poesio\(^1,2,*\)

\(^1\)Department of Computer Science, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, UK and \(^2\)DIT and Center for Mind/Brain Sciences, University of Trento, Via Sommarive 14 I-38050 POVO (TN), Italy

ABSTRACT

Motivation: Negative information about protein–protein interactions—from uncertainty about the occurrence of an interaction to knowledge that it did not occur—is often of great use to biologists and could lead to important discoveries. Yet, to our knowledge, no proposals focusing on extracting such information have been proposed in the text mining literature.

Results: In this work, we present an analysis of the types of negative information that is reported, and a heuristic-based system using a full dependency parser to extract such information. We performed a preliminary evaluation study that shows encouraging results of our system. Finally, we have obtained an initial corpus of negative protein–protein interactions as basis for the construction of larger ones.

Availability: The corpus is available by request from the authors.

Contact: osanch@essex.ac.uk or poesio@essex.ac.uk

1 INTRODUCTION AND BACKGROUND

Most text mining research focuses on positive sentences about protein–protein interactions (PPIs), like ‘P1 interacts with P2’. However, negative sentences may also contain evidence of use to biologists. For example, knowing that ‘Rux does not interact with either Drosophila CDK’ in the two-hybrid assay’ tells the biologist that it might not be necessary to carry out an experiment for testing whether Rux interacts with Drosophila CDK, provided that the study supporting the negative information is convincing and the result is a certain conclusion: i.e. negative results may help to avoid the repetition of similar experiments. Furthermore, negative cases contribute to the refinement of protein pathways and can help to avoid bias in form of the publication of positive results only (Knight, 2003).

Some previous research on identifying negative protein–protein interactions (N-PPIs) has been carried out in the medical field; this includes the development of systems such as NegExt (Chapman et al., 2001b) and Lexer (Mutalik et al., 2001). However, this previous research has been very limited in scope, focusing on non-affixal negations expressed with determiner ‘no’ and adverbial ‘not’. The methods used include cascades finite state automata (FSA) (Leroy et al., 2003), information extraction templates (Leroy and Chen, 2002) and regular expressions (Chapman et al., 2001a). In the protein interactions field Kim et al. (2006), have worked on the extraction of contrastive relations (e.g. ‘but not’) and created an online database containing such relations.

Because not much previous research exists, no large corpus of N-PPIs is available, which prevents the use of machine learning methods. For this first implementation we used therefore a heuristic approach, identifying the main cases using general patterns over the output of a functional dependencies grammar (FDG) approach. Because our method matches patterns against a predicate argument structure, the heuristics are able to cover different representations of the same structure.

The structure of the article is as follows. We first present an analysis of negative sentences about protein–protein interactions (henceforth, NSPPIs). We then introduce our semantic representation for PPIs and our system to extract N-PPIs. Finally, we present a preliminary evaluation of our system.

2 AN ANALYSIS OF NEGATED INTERACTIONS IN BIOLOGICAL TEXTS

2.1 Types of negation

Negation can be expressed in a variety of ways, ranging from the use of explicit negative particles both verbal (‘not’) and nominal (determiner ‘no’), to the use of affixes, to the use of inherently negative words like ‘inhibit’. Whereas explicit adverbial negations have been studied in the past, treatments of affixal nominal and inherent negation are less common. In this work, we are addressing all types of negation.

In affixal negation, the negation is expressed by an affix of the word. In biomedical texts, affixal negation is frequently used with verbs: e.g. activate, inactivate. Note that there is a distinction between ‘not activate’ and ‘inactivate’. ‘Not activate’ indicates that there is not an interaction at all between two proteins, whereas ‘inactivate’ indicates that there is such an interaction, but that it is of an inhibitory nature.

Noun phrase negation, also called ‘emphatic’ (Givon, 2001) is expressed syntactically, by using a negative determiner (as in, e.g. ‘No interaction was identified’ or ‘Nothing was found’).’

Finally, inherent negation (Tottie, 1991) is expressed by words with an inherently negative meaning even in their ‘positive’ form (e.g. absent, fail, lack, forget and exclude).

2.2 An analysis of negation in biological texts

We carried out an analysis of negation in biological text. We used 50 articles from the Journal of Biological Chemistry (JBC) which contain a high number of PPI occurrences (Alfarano et al., 2005). The articles were all published in September, 2004.
A program that selects sentences containing potential protein names, verbs denoting interaction and keywords denoting negation (e.g. ‘not’, ‘no’, ‘fail’, etc.) was used to obtain candidate sentences from the JBC files. We found a total of 707 candidate sentences which were manually analysed. The distribution of the negation constructions found is shown in Table 1.

The table above shows the frequency of sentences containing the respective negative constructions and the number of sentences which express N-PPIs.

As the table shows, the most common form of NSPPI are expressions with ‘not’ (49.44%), followed by those containing the construction ‘but not’ (17.98%) and ‘have no effect’ (10.11%).

In the rest of this section, we describe in more detail the most common forms of NSPPI, and the kind of cues that can be used to detect them.

One thing to keep in mind is that, as we will see subsequently, a fundamental property of NSPPI is that they are not only used to express what we will call definite negations of PPIs. Many NSPPIs express uncertainty as to whether a protein interaction does or does not hold: e.g. ‘we failed to detect . . .’, ‘No evidence of’, etc. We highlight those cases in which the existence of a positive or N-PPI is uncertain.

### 2.3 Classes of negation

#### 2.3.1 Adverbial negation

This was the most common type of negation in our data set (49.44%). The negation of the relation expressed by the CUE-VERB (the verb denoting a PPI) is expressed using adverbial ‘not’. Examples include:

1. Co immunoprecipitation studies revealed that Akt does not interact with Grb14.
2. The p46 isoform of JNK was not phosphorylated by ORF36.

#### 2.3.2 Inherent negation with ‘fail’

As said earlier, we have an inherent negation when a verb in its positive form denotes a negative relation. One form of inherent negation is found in constructions with the verb ‘fail’, as in the following example:

(3) More trivially, Y35L might fail to interact with BiP because, as noted earlier, it folds rapidly.

#### 2.3.3 Inability of a protein to interact with another

The fact that a protein does not interact with another can be formulated as an ‘inability’, expressed using a variety of more complex verbal constructions including auxiliaries such as ‘can’, copular constructions with the adjective ‘able’, etc.

**Auxiliary ‘can’**: One way of expressing PPI potential is with auxiliaries ‘can’, ‘could’, etc. in the past or present tense, as in the following example:

(4) In these assays the deacetylase domain of HDAC5 could not interact with MEF2A.

**Unable to interact**: The inability of a protein to interact with another can also be expressed by copular constructions with ‘to be’ together with ‘not’ and an adjective indicating ability (e.g. ‘able’, capable’), or by the positive form of ‘to be’ together with antonyms of ‘able’ (e.g. ‘unable’, ‘incapable’). Some examples are shown below:

(5) First, ePRPP was not able to activate Cys for reaction with glutamine or a glutamine affinity analogue.
(6) In contrast, beta-thrombin was unable to cleave factor V and factor VIII at both Arg372 and Arg1689.

#### 2.3.4 Negative nominals

A negative nominal contains a cue-noun (a noun denoting a type of interaction) either with a negative determiner or as the complement of an inherently negative noun.

**Determiner ‘No’**:

(7) However, TBP 2 Delta, TBP 2 Delta and TBP 2 showed no interaction with Rch1.

‘Lack of’: The cue-noun may occur as a complement of ‘lack’ as in the following example:

(8) . . . and the lack of interaction between GST-PinX1 and the structural RNA U6.

‘Does not exist’: Negations can also be expressed by nominals expressing PPIs in negative ‘there-is’ sentences:

(9) . . . although there does not exist a direct protein–protein interaction between Kv1.5 and caveolin, the channel protein . . .

**Word ‘inability’**: The word ‘inability’ or its synonyms express it directly, like the following example:

(10) The selectivity of this interaction was demonstrated by the inability of hTR to interact with GST . . .

This structure can also occur when the protein is the attribute of ‘inability’ (e.g. ‘the hTR inability to interact with GST’).

#### 2.3.5 Negative coordination (Neither)

The word ‘neither’ can be present either in the subject as well as in the object of
2.3.6 Contrastive structures The construction ‘but not’ is used in contrast structures with a positive and a negative part which are compared. We focus our analysis on the negative part of the contrastive construction, which can either occur in the subject or in the object of the cue-verb. The following examples illustrate this type of contrastive construction.

(13) G Protein beta subunit types differentially interact with a Muscarinic Receptor but not adenylyl cyclase type II or phospholipase C-beta 2/3.

(14) Using NBD peptides, we show that wild type, but not mutant NBD blocks IKK activation and reduces...

In contrast, structures with ‘but not’ and in clause-level parallelism structures as in example (15) (Kim et al., 2006) there are always negative clauses.

(15) Truncated N-terminal mutant huntingtin repressed transcription, whereas the corresponding wild-type fragment did not repress transcription.

There are other ways of expressing contrastive constructions that do not contain negative clauses as in the following example.

(16) ‘...bFGF stimulation led to increased autophosphorylation of Src family members. In contrast, in porcine aortic endothelial cells and lung fibroblasts from Chinese hamster, activation of FGFR caused reduced autophosphorylation of Src and Fyn...’

This structure contains two affirmative sentences. The expression ‘In contrast’ denotes contrast and the senses of the verbs ‘reduce’ and ‘increase’ are contradictory.

2.3.7 ‘No effect’ Some NSPPIs express the fact that a protein (or an event related to a protein) does not have an effect or does not produce any change on another protein (or events of proteins). This NSPPI may contradict or reaffirm existent information of protein interactions. This fact can be expressed by the phrases ‘no effect on’, ‘no effect of’ or ‘not have effect on’ as in the following examples:

(17) Little or no effect of insulin on Cbl phosphorylation at tyrosine 700 was also observed after 5 min...

(18) ING3 did not have effect on Fas ligand expression.

3 SEMANTIC REPRESENTATION

Our system (described in the next section) maps the N-PPIs extracted from text to the semantic structure outlined in Table 2.

<table>
<thead>
<tr>
<th>Table 2. Semantic structure of a PPI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein_name1</td>
</tr>
<tr>
<td>Cue-word</td>
</tr>
<tr>
<td>Polarity</td>
</tr>
<tr>
<td>Certainty</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 3. Examples of semantic relations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Semantic relation</td>
</tr>
<tr>
<td>Activate</td>
</tr>
<tr>
<td>Inactivate</td>
</tr>
<tr>
<td>Create bond</td>
</tr>
<tr>
<td>Break bond</td>
</tr>
</tbody>
</table>

We explained the components of the semantic structure in the next paragraphs.

3.1 Protein_name1 and Protein_name2

These fields are the names of first and second proteins participating in an interaction.

3.2 Cue-word

This is the word expressing a PPI. Cue-words can be verbs or their nominalizations (e.g. interact, interaction, activate, activation, etc.).

3.3 Semantic relation

Semantic relations are the categories in which cue-words are grouped according to their similar effect in an interaction. Table 3 shows some examples of semantic relations and their corresponding cue-words.

Our final list of semantic relations was adapted from Temkin and Gilder (2003) and complemented with categories from Kim et al. (2006) and Friedman et al. (2001).

3.4 Polarity

This field indicates whether the PPI is negated or not.

3.5 Direction

A direction is assigned to the semantic relations according to the effect that proteins cause on other molecules during an interaction. Table 4 shows the directions assigned to each semantic relation.

3.6 Certainty

This field keeps the degree of certainty expressed by the authors. The authors can be completely sure that there is not an
interaction (−) or that there is an interaction (+). However, there are intermediate degrees of certainty as shown in Figure 1.

3.7 Manner
Manner is the adjective or adverb (e.g. directly, weakly, strong, etc.) that affects a cue-word. Manner may reveal levels of interaction or certainty of the interaction, expressed by the authors.

In the present work, we concentrate only on the detection and extraction of the basic fields for N-PPIs. These fields are the names of the interacting proteins (Protein_name1 and Protein_name2), the cue-word and the polarity fields.

4 THE SYSTEM
Our system works by means of patterns that recognize NSPPIs. Candidate NSPPIs are identified by the following procedure:

1. use verbal and nominal cue-words to find potential constructions. Some examples of cue-words are shown in Table 5;
2. use our heuristics for checking that the potential constructions found at step (i) may express a negative PPI, relying on the dependency analysis produced by the FDG parser. Table 6 lists the main dependency relations used by the heuristics;
3. use our term extraction heuristics to extract the arguments of these predicates, again using dependency information, and then
4. verify that these arguments indeed refer to proteins or protein components.

4.2 Extraction methods
Our system works by means of patterns that recognize NSPPIs. Candidate NSPPIs are identified by the following procedure:

1. use verbal and nominal cue-words to find potential constructions. Some examples of cue-words are shown in Table 5;
2. use our heuristics for checking that the potential constructions found at step (i) may express a negative PPI, relying on the dependency analysis produced by the FDG parser. Table 6 lists the main dependency relations used by the heuristics;
3. use our term extraction heuristics to extract the arguments of these predicates, again using dependency information, and then
4. verify that these arguments indeed refer to proteins or protein components.

4.2.1 Terms formation
The procedure consists in following the chain of relations between the head noun and its pre and post modifiers. Determiners are not included. This procedure is recursive since a chain of nouns can be part of a noun phrase.

4.2.2 Protein name recognition
In order to recognize protein names, we are using a protein name recognizer, a lexicon and a combination of both.

4.2.3 ABNER
ABNER is a biomedical name recognizer (Settles, 2004) that uses linear-chain conditional random fields (CRFs) with orthographic and contextual features. ABNER was trained with the NLPBA and BioCreative corpora.

Version 1 of our system recognizes protein names in the sentence containing candidate N-PPIs by applying ABNER and the term formation method. The protein names and the terms obtained are compared. If the terms partially or

Table 4. Directions of semantic relations

<table>
<thead>
<tr>
<th>Positive (+)</th>
<th>Negative (−)</th>
<th>Neutral</th>
</tr>
</thead>
<tbody>
<tr>
<td>Activate</td>
<td>Inactivate</td>
<td>Substitute, react</td>
</tr>
<tr>
<td>Create bond</td>
<td>Break bond</td>
<td>Modify, cause</td>
</tr>
<tr>
<td>Generate</td>
<td>Release</td>
<td>Signal, associate</td>
</tr>
</tbody>
</table>

Table 5. Examples of cue-words
activat(e, or, ion), elevat(e, ion), incite, increase, block, decrease, deplet(e, ion), down-regulat(e, ion), demethylat(e, ion), bind, bound, interact(tion), react(tion), express(tion), methylat(e,ation), phosphorylat(e,ation), effect, discharge, mediate, modulat(e,ation), regulat(e,ion), transport(ation)
completely match with the protein names, then the terms are classified as proteins.

4.2.4 Uniprot Uniprot is a database that contains information of protein interactions (Apweiler et al., 2004). We downloaded the database from the Uniprot website and reduced its content to names of proteins, their synonyms and associated genes only.

In version 2 of our system, the candidate term is looked up in the Uniprot file by using the ‘grep’ command of Linux. The words composing the term are separated by spaces or punctuation symbols. Only the first and the last words of the candidate term are looked up.

4.2.5 ABNER–Uniprot Version 3 of our system first carries out protein name recognition by using ABNER. If no name is found, then it looks for the protein name in the Uniprot file.

4.3 Extraction of verb arguments

The arguments of a verbal interaction predicate are usually realized as its subject and object. When extracting these arguments, our algorithm considers verb voices (active or passive), coordination, relative clauses and infinitive verbs (e.g. ‘... shown to activate...’). We detail the steps followed for extracting the respective arguments.

4.3.1 Active voice The heuristics considers relative clauses and coordination of verbs. Figure 2 shows the decision tree for getting the subject and object of a verb in active voice.

4.3.2 Subject

(1) If the cue-verb is coordinated with other verbs, then take the first verb in the coordination chain and get its subject. Otherwise, get the direct connected subject to the cue-verb.

(2) If the subject is a relative pronoun, then get the nearest previous subject.

(3) If the cue-verb has ‘to’ as infinitive clause marker (‘pm’), then the subject is either the node whose postmodifier is the cue-verb or the subject of the verb whose object is the cue-verb.

4.3.3 Relative clause A relative clause can be present in a sentence containing a PPI, as it is shown in the following example.

(19) ‘Chick brain actin depolymerizing factor (ADF) is a 19-kDa protein that severs actin filaments and binds actin monomers’

In this case, the relative pronoun ‘that’ is assigned as the subject of the verbs ‘sever’ and ‘bind’. When the algorithm detects a relative pronoun as subject, it gets the nearest previous subject to the relative pronoun as the subject of the cue-verb. In our example, ‘Chick brain actin depolymerizing factor (ADF)’ is the subject of ‘sever’ and ‘bind’.

4.3.4 Infinitive verbs The verb in infinitive form can be used like a noun phrase expressing the interaction, as in ‘CGB was also shown to interact with CGA’.

4.3.5 Object

(1) If the cue-verb is coordinated with other verbs, then take the last verb of the coordination chain and get its object. Otherwise, get the direct connected object to the cue-verb.

(2) If the cue-verb occurs with a preposition, follow the chain formed by the cue-verb’s prepositional complement ‘phr’ (or post-modifier ‘mod’) and its prepositional complement (pcomp) which is the object word.
From here on, when we refer to ‘object’, we assume that the object could have been obtained either by the direct ‘obj’ relation or through a preposition.

4.3.6 Passive voice Figure 3 shows the decision tree for passive voice.

4.3.7 Subject

(1) Look for the by-phrase (‘agt’) or the ‘ha’ relation of the cue-verb or its coordinated verbs in case they exist.
(2) Get the nouns which are the prepositional complement (pcomp) of the ‘agt’ (or ‘ha’) node.

4.3.8 Object The object of a passive verb is obtained by following the procedure to get the subject of a verb in active voice.

4.4 Extraction of the arguments of a cue-noun The verbs denoting PPIs normally take prepositions (e.g. ‘binding to’, ‘interact with’). Therefore, the cue-nouns derived from this kind of verbs also contain prepositions, like in ‘binding of P1 to P2’, ‘interaction of P1 with P2’. An example taken from the literature is shown below.

(20) ‘Specific class I and II histone deacetylases (HDACs) interact in vivo with BCoR’

This sentence can be paraphrased as follows:

(21) ‘Interaction in vivo of the specific class I and II histone deacetylases (HDACs) with BCoR’.

In the texts we have analysed, the cue-nouns indicating interactions usually occur in one of the following constructions:

- Cue-noun followed by the preposition ‘between’ (e.g. ‘interaction between P1 and P2’). In this case the coordination ‘and’ leads to the interactors.
- Cue-noun followed by any other preposition different from ‘between’ (e.g. ‘phosphorylation of P1 by P2’). The first interactor is the noun following the first preposition and the second preposition leads to the second interactor.
- Cue-noun followed by any other preposition different from ‘between’ (e.g. ‘P1 showed interaction with P2’). However, in this case there is not a second preposition that leads to the second interactor. Then, the second interactor is the previous subject to the cue-noun. The trees that are looked up are illustrated in Figure 4.

\[ \text{Fig. 4. Trees followed to get parts of a cue-noun.} \]

The leaf nodes are the items (interactors) to be extracted; the dot line denotes an alternative way in the tree (‘or’ logical operator); and the dots in a line indicate from one to \( n \) lexical items, i.e. coordination (and/or). In this case, coordination is among the head nouns (HNs). The corresponding procedure is described as follows.

(1) Identify a cue-noun in a sentence.
(2) Identify the preposition after the cue-noun.
(3) If the preposition is ‘between’, then the head nouns related to this preposition are obtained.
(4) If the preposition is different of ‘between’, then a second preposition associated to it is looked up.
(5) If the second preposition is found then the head nouns related to each preposition are obtained.
(6) If there is not second preposition then the previous subject to the cue-noun is extracted as well as the head nouns associated to the first preposition.

4.5 Heuristics to extract negative relationships The following heuristics use the modules previously explained to get subject and object of a verb, parts of nouns, formation of terms and protein name recognition, in addition to the particular heuristics for each case of negation which were explained in the previous sections. The heuristics for negative cases are explained in the following paragraphs and the examples are shown with their respective FDG graphs.

4.5.1 Adverbial negation ‘Not’ The basic case of negation is a cue-verb negated by adverbial ‘not’. In the form of FDG encoded by the Connexor parser, these cases are represented as shown in Figure 5. The cue-verb is connected by a verb-chain (v-ch) dependency to the auxiliary node; the negation and subject are also modifiers of this node; the object depends on the cue-verb. The corresponding negative relation for the tree in Figure 5 is: \( \neg \text{interact} \left[\text{Akt}, \text{Grb14}\right] \).

4.5.2 ‘Fail to’ These cases are represented by dependency trees like the one shown in Figure 6. The verb ‘fail’ in positive form has a cue-verb as object; the object is attached to the cue-verb; the subject is attached to the top of the verb-chain ending in ‘fail’.

Since the cue-verb ‘interact’ is the object of ‘fail’, ‘Y35L’ is the subject of ‘fail’ and ‘BiP’ is the object of ‘interact’, the resulting protein interaction is \( \neg \text{interact}[\text{Y35L}, \text{BiP}] \).

\[ \text{Fig. 5. Example of negation ‘not’ in active voice.} \]

\[ \text{Fig. 6. Example of ‘fail to’}. \]
4.5.3 Noun negations ‘No’ and ‘lack of’. The system identifies these cases by looking for a cue-noun with ‘no’ as dependent determiner, or for a cue-noun occurring as the prepositional complement (pcomp) of ‘lack of’. The corresponding graphs are shown in Figure 7.

4.5.4 ‘Not exist’. In these constructions, the verb ‘exist’ is negated and a cue-noun is either its subject complement (‘comp’) or subject (‘subj’). This construction can be present before or after a cue-noun. The tree in Figure 8 shows an example.

In the example, the cue-noun ‘interaction’ is complement of ‘exist’. The arguments of ‘interaction’ form the resulting N-PPI: ¬ interact [Kv1.5, caveolin].

4.5.5 Inability to interact. The system looks for structures that contain expressions of inability of the protein to interact. If a cue-verb is post-modifier (‘mod’) of ‘able’ (or synonyms) and in turn ‘able’ is complement (‘comp’) of ‘to be’ in negative form, then the subject of ‘to be’ and the object of the cue-verb are considered as possible arguments. The FDG representation of a construction matching this pattern is shown in Figure 9.

A second matching case is when the cue-verb is a postmodifier (‘mod’) of ‘unable’ (or synonyms) and ‘unable’ is complement (‘comp’) of ‘to be’.

4.5.6 Neither. Can be found in the subject or in the object of a cue-verb.

- In subject: if ‘neither’ is the attributive adverbial (‘ad’) of the subject of a cue-verb, then the subjects and object of the cue-verb are obtained.
- In object: if ‘neither’ has a ‘ha’ relation with a cue-verb, then the subject and objects of the cue-verb are extracted.

As an example, the analysis for ‘... neither K-RasV12 nor N-RasV12 activated PI3K ...’ is shown in Figure 10.

4.5.7 ‘But not’. This contrastive connective can be located in the subject or in the object of a cue-verb. In order to avoid parsing errors due to ambiguity in coordination, we use a combination of pattern matching and functional-dependency relations heuristics. The algorithm looks for the pattern ‘but not’ and gets the closest noun to the left (NL) and the closest noun to the right (NR) of this pattern.

- In subject: if NR is the subject of a cue-verb then NR (and its coordinated nouns) is extracted as well as the object of the cue-verb.
- In object: if NL is the object of a cue-verb, then NR (and its coordinated nouns) is extracted as well as the subject of the cue-verb.

In the example shown in Figure 11, the negated interaction would be ¬ interact [G protein beta, adenylyl cyclase].

4.5.8 ‘No effect on’. The system looks for sentences containing the phrases ‘no effect on/of’ and ‘not have effect on’. When ‘effect’ has the determiner ‘no’ and the preposition ‘on’ as postmodifier (‘mod’), then the complement (‘pcomp’) of ‘on’ and the subject of the word whose object is ‘effect’ are extracted (Fig. 12). If the postmodifier of ‘effect’ is the preposition ‘of’, then the parts of ‘effect’ are extracted as in the case of any other nominal.

If ‘effect’ is object of ‘have’ which is in negative form, then the subject of ‘have’ and the complement (‘pcomp’) of ‘on’ are extracted.

5 PRELIMINARY EVALUATION

We carried out a preliminary evaluation of the performance of the heuristics developed in this work. We created a small corpus as baseline for the evaluation. To create the corpus, we run our searching program over a set of 114 JBC articles. As outlined in Section 2.2, the program looks for sentences containing potential protein names, cue-words expressing interactions, and words denoting negation. Part of the candidate sentences was manually examined in consultation with a biologist to form an annotated corpus. Our corpus consists of 185 sentences of which 90 sentences contain 110 N-PPIs pairwise relations and 95 sentences do not contain any N-PPIs (even though...
they include phrases or words present in N-PPIs). There is no overlap between these sentences and the sentences in the development corpus.

We then measured the performance of the system under the three methods for protein name recognition discussed in Section 4.2 (i.e. using Uniprot, ABNER and both ABNER and Uniprot), as well as using a list with the real names of the proteins contained in the evaluated sentences (i.e. hand-tagged). We then compared the results obtained using different methods with the results obtained using the hand-tagged version of the corpus.

### 5.1 Quantitative results

The results in terms of precision, recall and F-score are shown in Table 7.

The ‘hand-tagged’ version is the upper bound—i.e. the performance that the system might achieve without protein name recognition errors. The ABNER–Uniprot method performed better than the ABNER and Uniprot methods in isolation. Since the corpus used for this evaluation is small, we expect a reduction in the performance when using a larger corpus. The errors in the system that we have detected are described in the following section.

### 5.2 Error analysis

The most frequent errors detected in the hand-tagged protein names system were caused by cases not considered in the heuristics such as uncertain negation (e.g. ‘Although the direct interaction between RIN2 and Ha-Ras was not observed…’); words denoting interactions in themselves (e.g. ‘C8 oligomerization’); complicated grammatical structures (e.g. ‘SREBP-1a and -2 interacted specifically with p300; however, SREBP-1c did not’). These cases account for 43.17% of the errors.

The second major source of errors was incorrect parsing. Incorrect parsing was provoked by pre-processing, ambiguity in coordination and erroneous part of speech (POS) tagging, since the parser is not trained on biological texts (e.g. ‘monitoring’, ‘binding’, etc.). Incorrect parsing caused around 29.55% of the errors.

Other problems were caused by errors in the heuristics (9.10%); ambiguous cases where the arguments are rather protein events implying more than one protein [e.g. ‘PKA inhibitor does not block (phosphorylation of eNOS-S1179 induced by AktMyr)’]. These cases account for 9.10% of the errors. Finally, problems in the formation of complex terms (9.10%) like in ‘other [proteins of the extracellular matrix such as laminin and fibronectin] did not bind OSM’.

### 6 UNCERTAINTY WITH RESPECT TO THE EXISTENCE OF A PROTEIN INTERACTION

Some statements about PPIs do not express a definite statement about the lack of interaction, but poor confidence about its existence. It is important to consider this kind of uncertain sentences when looking for positive interactions, since they can be mistakenly conceived as real protein-protein interactions (i.e. false positives).

Biologists may encounter difficulties in finding protein-protein interactions (e.g. due to limitations of the method used, because of problems associated with the experimental design, the lack of evidence supporting the interaction, etc). However, from the fact that biologists may face difficulties in identifying PPIs, it does not follow that these interactions do not exist.

The following are examples of NSPPIs that express uncertainty rather than definite negation.

<table>
<thead>
<tr>
<th>Evaluation Method</th>
<th>Recall</th>
<th>Precision</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tagged</td>
<td>67.27</td>
<td>89.15</td>
<td>76.68</td>
</tr>
<tr>
<td>ABNER-Uniprot</td>
<td>61.82</td>
<td>64.15</td>
<td>62.96</td>
</tr>
<tr>
<td>Uniprot</td>
<td>52.72</td>
<td>66.66</td>
<td>58.88</td>
</tr>
<tr>
<td>ABNER</td>
<td>40.00</td>
<td>84.61</td>
<td>54.32</td>
</tr>
</tbody>
</table>

We do not have any evidence that hTERT binds directly to DNA-PKcs…

There is no evidence that nuclear factor kappa B phosphorylates I kappa B.
We also plan to run the system over a larger set of data. Positive and N-PPIs, such as the discovery of contradictions, are challenging biological knowledge discovery that needs both detecting uncertain N-PPIs. Furthermore, we will explore more approaches and evaluations for the construction of larger datasets that may be useful for N-PPI. In this work, we have obtained an initial dataset as basis for 'gold-standard data sets' we need to reach an agreement among biologists to determine when a sentence actually expresses an interaction. The evaluation of biological information extraction systems is generally difficult due to the fact that, in order to create ‘gold-standard data sets’ we need to reach an agreement among biologists to determine when a sentence actually expresses an N-PPI. In this work, we have obtained an initial dataset as basis for the construction of larger datasets that may be useful for other approaches and evaluations.

In future work, we plan to complete the heuristics for detecting uncertain N-PPIs. Furthermore, we will explore more challenging biological knowledge discovery that needs both positive and N-PPIs, such as the discovery of contradictions. We also plan to run the system over a larger set of data.

ACKNOWLEDGEMENTS

We would like to thank Joaquin Zacarias and Andrew Harrison from the Department of Biology of the University of Essex for their comments and advice for this work.

Conflict of Interest: none declared.

REFERENCES


