What kind of problems do protein interactions raise for anaphora resolution? A preliminary analysis

Olivia Sanchez* Massimo Poesio* Mijail A. Kabadjov* Roman Tesar§

*University of Essex
Wivenhoe Park, Colchester CO4 3SQ,
United Kingdom
{osanch,poesio,malexa}@essex.ac.uk

§University of West Bohemia,
Univerzitni 22, Pilsen 30614,
Czech Republic
romant@kiv.zcu.cz

Abstract

In this preliminary study, we analyzed the kind of anaphoric expressions that occur in expressions describing protein interactions found in biological text. We also studied the impact of anaphora resolution on protein interaction extraction, when an off-the-shelf anaphoric resolver (i.e., not one specially developed for this domain) is used, and looking at full texts as well as abstracts. Our results suggest that about 5% of the descriptions of protein-protein interactions contain anaphoric expressions when full texts are considered. These anaphoric expressions are primarily pronouns, even though most anaphoric expressions are full NPs. The use of our anaphoric resolver gives a small improvement over our baseline system.

1 Introduction and Motivations

Evidence of protein-protein interactions (PPI) is crucial for biologists since it leads to the discovery of more complex structures like protein pathways. Text mining techniques are increasingly used to accelerate this very time consuming process. Currently, recall is one of the most serious problems that such systems have. State-of-the-art systems for mining protein-protein interactions like Blaschke et al. (1999) and Akane et al. (2004) can typically achieve a precision above 70%, but recall is generally of 50% / 60%.

Systems that report high recall (97%) like Cooper and Kershenbaum (2005) have identified coreference as one of the reasons for decreasing recall. The problem arises when one or both of the proteins involved in the interaction is expressed with an anaphoric expression such as a pronoun, as shown in the following text excerpts taken from JBC[1]:

(1) a. The same result was obtained for the **dm-rabphilin**: *it specifically interacted with **dm-Rab27***, not with dm-Rab3 or dm-Rab8. This was in strong contrast to **mouse rabphilin**, because *it interacted with **Rab3A**, **Rab8A**, and **Rab27A***.

b. **Ang-2** blocks the ability of **Ang-1** to activate **Tie2** in ECs, but *it activates **Tie2** expressed in hemangioblast...*

In our research we address two main problems with previous work on anaphora resolution in biological texts (Castaño et al., 2002, Lin and Liang, 2004 and Vlachos et al., 2006). First of all, previous work does not specifically focus on anaphoric expressions involved in protein interactions. Furthermore, all of the mentioned studies have been done on abstracts. We also investigate the effect of anaphora resolution on full text articles.

A further difference is that we tried using a publically available anaphoric resolver not specifically developed for this purpose—which is likely to be the normal case for developers of extraction systems as custom anaphoric resolvers are often unavailable. We use an anaphoric resolver called GUITAR (Poesio and Kabadjov, 2004) designed to be of practical use for the developers of NL applications in that (i) it works as a component in a standard XML-in / XML-out pipeline (ii) it works with a variety of preprocessing components, from basic POS-taggers to chunkers to full parsers (iii) it is highly modular in that, e.g., it is possible to use different named entity coreference components for different applications.

The goal of the work discussed in this paper is exploring the kind of problems raised by using

[1] The Journal of Biological Chemistry: http://www.jbc.org/

an anaphora resolution system in support of a system mining protein-protein interactions. We analyse the cases of interactions where one of the interactors has anaphoric antecedents. Although pronoun and sortal anaphora are the most common cases found in Biological texts (Castaño et al., 2002), in particular in protein interactions, we focus on pronouns because they are those which our NE recogniser finds harder to process.

2 Protein Interaction Extractor (PIE)

We used as baseline system PIE (Poesio and Sanchez, 2005). PIE is a pattern matching system that recognises protein interactions based on verb and noun patterns. The system can resolve cases such as the following examples:

(2) a. *Delta-catenin* interacted with a fragment of *PS1*.
b. Interaction of *Dbs* with *Cdc42*.

2.1 Collecting documents

PIE finds full text articles freely available on the internet by using a web crawler. The files are locally saved in HTML format. Then, the articles are converted to flat text.

2.2 Preprocessing

Stop words are removed. Then the Genia POS tagger (Tsuruoka et al., 2005) is used. Once the text is POS tagged, it is chunked with a noun phrase chunker[2] and converted to XML format in order to be easily manipulated through the use of the DOM model. Then, name entity recognition is performed using the biomedical entity recogniser ABNER (Settles, 2004) configured to recognise only protein names.

2.3 Extracting verb arguments and verbs

Verb arguments are identified with heuristic methods as follows. The chunker recognises the noun expressions (ne) in each sentence. These noun expressions are then grouped in what will be the arguments of the verbs (denoted here as NP's) by the following regular expression:

$$NP \rightarrow ne\ [\ [\ IN?\ |\ ,?\ |\ CC?\ ]\ ne\ ]^{*}$$

Where the last ne ≠ {PRP, PP, PP$}. Verbal expressions (ve) are formed with the expression:

$$ve \rightarrow MD?\ verb^{+}$$

where verb = {VBZ, VBN, VBD, VBG}

[2] http://www.dcs.shef.ac.uk/~mark/index.html?http://www.dcs.shef.ac.uk/~mark/phd/software/chunker.html

2.4 Protein interaction extraction

The next step is the extraction of protein interactions using patterns. The verb pattern looks for the inflections of special verbs that denote interactions. PIE considers the ones mentioned in Blaschke et al. (1999).

<arg1> verb [IN] <arg2>

Where arg1 and arg2 are protein names. The patterns consider passive and active voice. In that case the order of the arguments is inverted. There can also be a preposition (IN).

In the case of nouns the patterns that are used are similar to those proposed by Plake (2005):

- *Noun* [IN] $prot_1$, $prot_2$ …. [and | IN] $prot_n$
- $prot_1$ *Noun*

Where *Noun* is a noun denoting interaction, like: interaction, association, inhibition, etc.

- $prot_1$ *complex*

Where *complex* includes words like complex, dimmer, heterodimer, homodimer, etc. We consider the lemma of these words.

PIE produces as output both a list with the protein interactions found and the corresponding HTML files to visualise the interactions within the texts. The system can be executed with or without anaphora resolution.

3 GUITAR: A General-Purpose Anaphoric Resolver

For resolving anaphors we used the GUITAR system (Poesio and Kabadjov, 2004) which can resolve definite descriptions, pronoun and proper names. The current version of the system includes an implementation of the MARS pronoun resolution algorithm (Mitkov, 1998) to resolve personal and possessive pronouns, a partial implementation of the algorithm for resolving definite descriptions proposed by Vieira and Poesio (2000) and an implementation of a shallow algorithm for resolving proper names proposed by Bontcheva et al. (2002).

4 Adding anaphora resolution to PIE

GUITAR is invoked before the pattern-matching step. GUITAR adds to the XML file the antecedents and anchors of anaphoric expressions. When PIE identifies a candidate interaction that does not contain a protein name in one of its arguments, it checks its closest antecedent. If the antecedent is a protein then the interaction is annotated. In (3) the candidate interaction contains the pronoun "it" as subject of "phosphorylates". Since GUITAR gives "p70Se kinase" as

the antecedent of "it", the interaction *p70S6 kinase- S6 protein* is annotated:

(3) ***p70S6 kinase*** (p70S6k) is a mitogen-activated protein kinase that plays a central role in the control of mRNA translation. ***It*** physiologically phosphorylates the *S6 protein*…

5 Dataset used

For this study we used the Medstract corpus (Pustejovsky et al., 2002) supplemented by three articles taken from the Journal of Biological Chemistry (JBC)[3].

Medstract contains 32 Medline abstracts. We found 74 explicit anaphoric pairs across all the abstracts. Out of the 74 pairs, 24 are pronominal and 50 are sortal.

We collected the full texts corresponding to the Medline abstracts in Medstract. As some of these articles were not freely available, we could find 20 out of the 32 articles. Therefore, we also collected 3 full texts and their abstracts from JBC. With the Medstracts/JBC documents, we formed two datasets of 23 files each. One of these datasets contains only abstracts and the other full texts. We manually annotated protein interactions and pronominal anaphoric relations in both the abstract and full text dataset.

Pronominal anaphora includes personal, reflexive and possessive pronouns in third person (i.e. they, their, them, themselves, it, its, itself) and sortal considers the cases: both, each, either, this, these, the, those protein(s). The following are the frequencies of PPIs and anaphora relations in the full text corpus. Pronominal and sortal anaphoras are considered.

Total PPIs	402
Total anaphora relations	664
Total PPIs with anaphors	20

Table 1. Frequencies in full text.

Note that only 5% of the PPIs contain anaphors. In the protein interactions there are 18 involving pronominal anaphors and only 2 cases with sortal anaphors. Due to the low occurrence of sortal anaphors, we concentrate therefore on pronominal anaphors.

6 Experiments and Results

First, we compared the number of pronominal anaphors found in both abstracts and full texts against the correct pronominal anaphors given by GUITAR over the same texts.

	Medstract /JBC (Baseline)	GUITAR (correct)	Recall
Abstracts	20	14	70%
Full texts	604	318	52.65%

Table 2. GUITAR results (pronominal anaphors).

Next, we ran GUITAR-PIE over the same texts to compare the number of interactions containing pronominal anaphora. We obtained the following results:

	Medstract/JBC (Baseline)	GUITAR-PIE
Abstracts	1	1
Full texts	18	3

Table 3. GUITAR-PIE results for PPIs with pronominal anaphors.

From the tables above we can notice that the performance is not particularly high. It was mainly caused by incorrect protein tagging and wrong assignment of antecedents, for instance when the correct antecedent is a protein, but the antecedent given is not, like in "*However, **it** has been reported that TRAP1 is also present in the cytoplasm, where **it** interacts with the retinoblastoma protein during mitosis or heat shock"*. The first "*it*" is taken as the antecedent of the second "*it*" instead of *TRAP1*.

Finally, to measure the effect of pronominal anaphora resolution on information extraction, we calculated precision/recall of PIE and GUITAR-PIE, both for abstracts and for full texts.

		Recall	Precision
Abstracts	PIE	38.23	61.90
	GUITAR-PIE	42.42	63.63
Full texts	PIE	57.96	69.96
	GUITAR-PIE	58.70	70.87

Table 4. PPI Extraction with/without pronominal anaphora resolution.

This table shows that even by taking an off-the-shelf anaphoric resolver like GUITAR one can expect a small improvement both in precision

[3] http://www.jbc.org/

and recall. (More texts would of course be needed to measure significance.)

7 Discussion

Our results suggest that even a general-purpose anaphoric resolver may lead to small improvements in PPI extraction. These small effects may nevertheless become significant when large amounts of full texts are processed. In future work, we will examine larger collections and establish if these improvements will prove statistically significant. We will also analyse the impact of anaphora resolution on a perfect protein annotated corpus.

Another interesting finding is that although sortal anaphora cases are more frequent than pronominal anaphora in biological texts, their number in PPI is smaller. This is good as sortal anaphora requires more knowledge to be resolved.

Finally, we found that pleonastic *it* (as in *it has been proved*) is as frequent as referential *it*, suggesting that methods for detecting such expressions would be useful (Mitkov, 2002).

References

Y. Akane, Y. Miyao, Y. Tateisi and J. Tsujii. 2005. Biomedical Information Extraction with Predicate-Argument Structure Patterns. In *Proceedings of the First International Symposium on Semantic Mining in Biomedicine*. pp. 60-69, Hinxton, UK

C. Blaschke, M. A. Andrade, C. Ouzounis, A. Valencia. 1999. Automatic extraction of biological information from scientific text: protein-protein interactions. In *Proceedings of the International Conference on Intelligent Systems for Molecular Biology*. Vol.60-7.

J. Bontcheva, M. Dimitrov, D. Maynard, V.Tablan, and H. Cunningham. 2002. Shallow methods for named entity coreference resolution. In *Chaînes de références et résolveurs d'anaphores*, workshop TALN 2002, Nancy, France.

J. Castaño, J. Zhang, J. Pustejovsky. 2002. Anaphora Resolution in Biomedical Literature. In the *International Symposium on Reference Resolution*, Alicante, Spain.

J. Cooper and A. Kershenbaum. 2005. Discovery of protein-protein interactions using a combination of linguistic, statistical and graphical information. In *BMC Bioinformatics* 6:143 (7 June 2005)

Y. Lin and T. Liang. 2004. Pronominal and Sortal Anaphora Resolution for Biomedical Literature. In *Proceedings of ROCLING XVI*, Taipei, Taiwan, pp. 101-110.

R. Mitkov. 1998. Robust pronoun resolution with limited knowledge. In *Proceedings of COLING*.

R. Mitkov. 2002. Anaphora resolution: Longman.

C. Plake, J. Hakenberg, U. Leser. 2005. Optimizing syntax patterns for discovering protein-protein interactions. *SAC 2005*: pp.195-201.

M. Poesio and M. Kabadjov. 2004. A General-Purpose, off-the-shelf Anaphora Resolution Module: Implementation and Preliminary Evaluation. In *Proceedings of LREC*. Lisbon, Portugal.

M. Poesio, O. Sanchez, 2005. Acquisition of Causal Knowledge from Text: Applications to Bioinformatics. In the *First International Symposium on Semantic Mining in Biomedicine*, Hinxton, UK

J. Pustejovsky, J. Castaño, R. Saurí, A. Rumshisky, J. Zhang, W. Luo., Medstract: Creating Large-scale Information Servers for Biomedical Libraries. *ACL 2002 Workshop on Natural Language Processing in the Biomedical Domain*. Philadelphia, PA.

B. Settles. 2004. Biomedical Named Entity Recognition Using Conditional Random Fields and Rich Feature Sets. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications* (NLPBA), Geneva, Switzerland, pp. 104-107.

Y. Tsuruoka, Y. Tateishi, J. Kim, T. Ohta, J. McNaught, S. Ananiadou and J. Tsujii. 2005. Developing a Robust Part-of-Speech Tagger for Biomedical Text. In *Advances in Informatics-10th Panhellenic Conference on Informatics*, LNCS 3746, pp. 382-392.

R. Vieira and M. Poesio. 2000. An empirically-based system for processing definite descriptions. In *Computational Linguistics*, 26(4).

A. Vlachos, C. Gasperin, I. Lewin, T. Briscoe. 2006. Bootstrapping the recognition and anaphoric linking of named entities in Drosophila articles, to appear in *Proceedings of the Pacific Symposium in* Biocomputing 2006.