How Complex are Complex Protein-protein Interactions?

Jari Björne,[1] **Sampo Pyysalo,**[2] **Filip Ginter**[1] and **Tapio Salakoski**[1,2]
[1]Department of IT, University of Turku
[2]Turku Centre for Computer Science (TUCS)
Joukahaisenkatu 3-5
20520 Turku, Finland
firstname.lastname@utu.fi

Abstract

The extraction of protein-protein interactions (PPI) from text requires a formal PPI representation. We use the BioInfer and GENIA corpora to study two such representations: a "binary" interaction model consisting of pairs of proteins and a "complex" model where interactions are defined as a network of proteins and their relations. As both of these formats can be seen as graphs, we contrast them with syntactic dependency graphs, a common tool for PPI extraction. We find that unlike binary interactions, complex interactions closely resemble dependency parses, especially those in the Stanford scheme. We therefore argue that despite appearances, complex interactions might be easier to extract. We also notice the similarity between the independently developed BioInfer and GENIA interaction representations and the Stanford dependency scheme. This suggests an emerging consensus on the representation for complex PPI, supporting the value of these tools and resources for PPI extraction.

1 Introduction

Protein-protein interaction (PPI) extraction is a central, widely studied task in biomedical natural language processing. The simplest model of PPI, used in most corpora and extraction studies, represents each interaction as a pair of protein names. Several systems have been introduced for extracting such binary interactions, but considerable challenges remain (Krallinger et al., 2007).

Recently, two corpora with more detailed interaction annotation have been introduced: the BioInfer (Pyysalo et al., 2007) and GENIA Event corpora (Kim et al., 2008) annotate complex structured relations (Figures 1 and 2). These "complex interactions" differ from binary interactions in that they can have more than two arguments, and allow interactions as arguments, thus enabling annotation of complex nested relations such as in "A causes B to bind C". Complex interactions can also be thought of in terms of semantic frames, with the edges of the complex interaction corresponding to the arguments of a verb frame (Cohen and Hunter, 2006).

For BioInfer, this annotation has also been translated into binary interactions (Heimonen et al., 2008), providing an opportunity to compare complex and binary interactions. In addition to PPI annotation, both BioInfer and GENIA include syntactic annotation that can be accessed in various dependency representations. Dependency has been argued to be well suited for applications such as information extraction, and dependency parsing is both well studied and frequently applied in the biomedical domain (de Marneffe et al., 2006; Clegg and Shepherd, 2007).

We are not aware of methods that would aim to extract the complex interactions annotated in the BioInfer and GENIA Event corpora. Neither has the relationship between simple and complex PPI annotation been studied in detail. Our aim here is to explore this relationship and thus take a first step towards complex PPI extraction.

2 Analysis and Discussion

2.1 Complex vs. Binary Interactions

We first observe that both the "binary" and "complex" representations can be viewed as forms of semantic networks (graphs). In the former case protein nodes are connected by edges expressing interactions, in the latter, both proteins and words

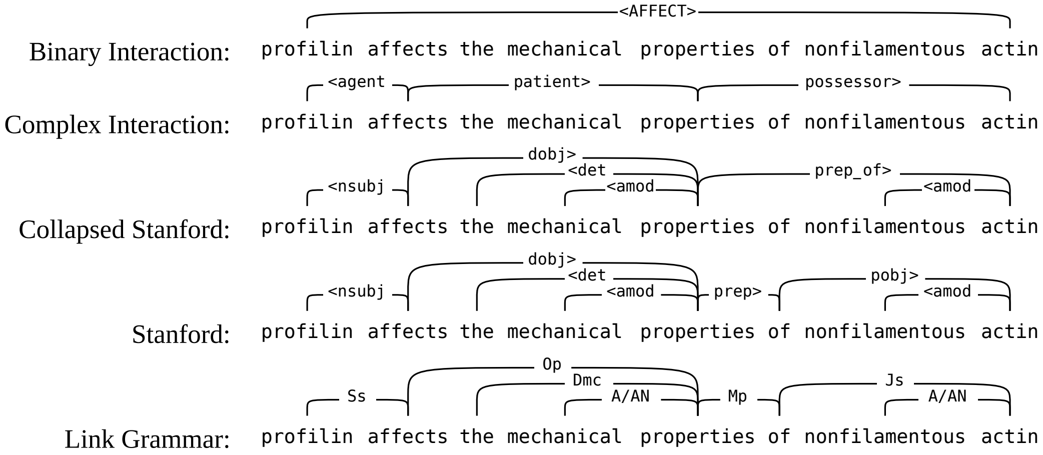
Binary Interaction:
<AFFECT>
profilin affects the mechanical properties of nonfilamentous actin
Complex Interaction:
<agent
patient>
possessor>
profilin affects the mechanical properties of nonfilamentous actin
Collapsed Stanford:
<nsubj
dobj>
<det
<amod
prep_of>
<amod
profilin affects the mechanical properties of nonfilamentous actin
Stanford:
<nsubj
dobj>
<det
<amod
prep>
pobj>
<amod
profilin affects the mechanical properties of nonfilamentous actin
Link Grammar:
Ss
Op
Dmc
A/AN
Mp
Js
A/AN
profilin affects the mechanical properties of nonfilamentous actin

Figure 1: Example from the BioInfer corpus, with the interaction annotation and the three parse schemes

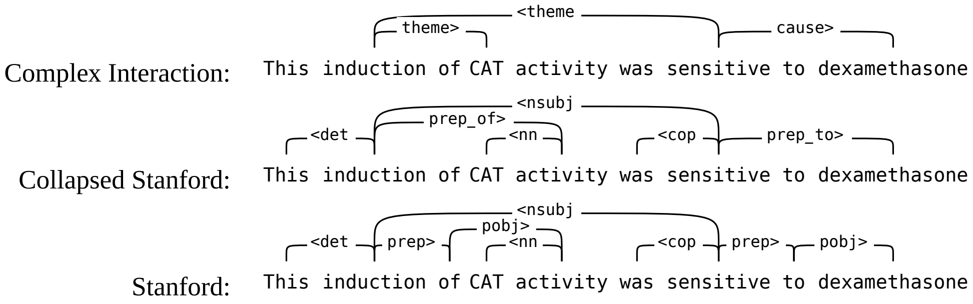
Complex Interaction:
theme>
<theme
cause>
This induction of CAT activity was sensitive to dexamethasone
Collapsed Stanford:
<det
prep_of>
<nsubj
<nn
<cop
prep_to>
This induction of CAT activity was sensitive to dexamethasone
Stanford:
<det
prep>
pobj>
<nsubj
<nn
<cop
prep>
pobj>
This induction of CAT activity was sensitive to dexamethasone

Figure 2: Example from the GENIA corpus, with an annotated interaction and two parse schemes.

stating their relations act as nodes and edges express their roles. Syntactic dependency parses are also graphs where words (nodes) are linked with dependencies (edges). Thus, as syntax and the binary and complex interactions have a graph representation, their relationships can be studied as a mapping between these graphs.

Figures 1 and 2 illustrate sentences from the BioInfer and GENIA corpora with their PPI annotation and dependency syntax, showing all the available graph representations (see Section 2.2 for descriptions of the syntactic annotations). Figure 1 shows a binary interaction between words which are several dependencies away from each other in the syntactic parses. On the other hand the graph of the complex interaction corresponds more closely to the dependency parse. This is typical: the words annotated as expressing interactions frequently fall on the shortest dependency path between the proteins. By providing intermediate nodes along the dependency path that connects the proteins involved in binary interactions, complex interactions subdivide the concept of "interaction" into smaller parts. As these simple relations can correspond better to syntactic features in a sentence, they could be easier to extract than the diverse binary interactions.

To study the feasibility of extracting complex interactions, we compared the dependency parse representations with both binary and complex interactions from the BioInfer corpus, and complex interactions from the GENIA corpus.

2.2 Processing the Corpora

The BioInfer and GENIA annotation schemes are designed to capture complex biological interactions in detail. The BioInfer format annotates e.g. entities and interactions with predicates appropriate for these tasks. The BioInfer annotation can be converted to several derived formats more suited for different uses. For these experiments, we transformed BioInfer into a semantic network representation in which the entire annotation of a sentence is defined as a directed graph (Heimonen et al., 2008). The edge labels define the semantic roles between entities and relations (e.g. *agent*, *patient*) and between different entities (e.g. *sub/super* for part/whole relations). Predicates not bound to text in the original annotation, such as most occurrences of *EQUAL* (an identity relationship between entities), were converted to edges. To compare the complex interactions with the binary ones, we also used a binarised version of the annotation, where interactions are simple pairs of named entities. BioInfer has manually annotated dependency parses in the Link Grammar (Sleator and Temperley, 1991) and Stanford formats.

For GENIA, we used the recently published event annotation. This annotation has manually annotated complex interactions for 9372 sentences. This was interpreted as a semantic network with the edges labeled *theme* and *cause* as defined in the event annotation; no edges were derived from the entity annotation. From these, we selected the subset of 1968 sentences which had manually annotated parses in the beta version of GENIA Treebank (GTB) (Tateisi et al., 2005). The GTB annotation was converted into dependency which was collapsed with the software introduced by de Marneffe et al. (2006); we refer to this study for a description of the representation. We used the manually annotated gold standard parses for all evaluations.

2.3 Connecting interactions to parses

To compare semantic interaction annotation to dependency parses we have to map the interactions to the sentence text. This is done based on the *text bindings*, which connect the annotations to the words expressing them. However, in both BioInfer and GENIA these text bindings can consist of multiple words. For example when the entity *Acanthamoeba profilin* takes part in an interaction, the edge that links to it connects to this pair of words. By contrast, in a dependency parse, all edges connect to single words. Thus, for comparison with dependency parses, interaction edges connecting to multi-word entities are mapped to a single word. We used the Stanford parse to map these edges to syntactic head tokens.

2.4 Comparing interactions to parses

To see how closely complex interactions resemble a dependency parse, we measured the shortest path in the dependency graph between two tokens connected by an edge in the interaction graph. We compared the lengths of these shortest paths between the available three parses for BioInfer and the two for GENIA (Figure 3). We notice that complex interaction edges most likely have a corresponding dependency in the collapsed Stanford parse. With uncollapsed Stanford and Link Grammar parses, the shortest path more often consists of multiple edges. This supports the design choices of the collapsed Stanford scheme, which was developed to facilitate applications such as information extraction. It is very interesting that the complex interactions of both BioInfer and GENIA correspond so closely to this syntactic representation. The annotators of BioInfer and GENIA were biologists with no formal expertise on the syntactic structure of the sentences they were annotating. Yet the Stanford syntactic parse and the semantic annotations of BioInfer and GENIA, developed independently and with somewhat different aims, result in very similar graph structures.

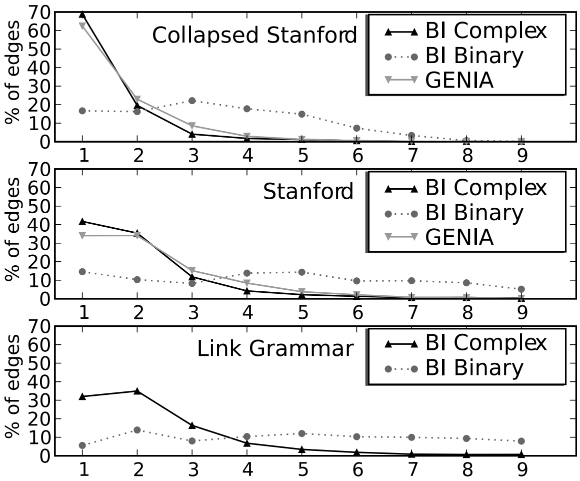

Figure 3: Percentage of interaction edges plotted against the length of the shortest path of dependencies between them. Over 60 % of BioInfer and GENIA complex interaction edges correspond to a single collapsed Stanford dependency. Longer paths are more common for the other parses. The paths for binary interactions are longer than for complex interactions.

For the BioInfer corpus, we also compared the complex interactions to the pairwise binary annotation for the same sentences. The shortest dependency paths corresponding to interaction edges were shorter for the complex interactions than the binary ones. In the case of the collapsed Stanford annotation, over 60 % of complex interaction edges linked neighbouring nodes in the dependency graph. For binary interactions the shortest path most commonly consisted of three dependencies.

For paths of length one, we also measured which dependency types correlated best with each interaction graph edge type (Table 1). Certain edge types correspond very strongly to a specific dependency type. For example, an interaction edge of type *EQUAL* has most often a corresponding edge of type *appos* in the collapsed Stanford parse. This is promising for the development of systems for detection of interaction type.

[%]	appos	nn	nsubj	prep_of
EQUAL	**73.45**			
MEMBER	**27.27**	**59.60**		
agent	0.45	2.91	**22.6**	5.82
possessor		**31.96**		**48.45**
sub		**45.52**		2.76
super		**22.35**		**47.06**

Table 1: Selected BioInfer complex interaction edges (vertical) of which over 20 % have a one-to-one correlation to a collapsed Stanford format dependency (horizontal). The percentages are of all interaction edges, including those not corresponding to a single dependency. Values > 20 % are emphasized with bold text.

3 Conclusions and future work

Comparison of the interaction annotation to different parse schemes showed that the complex interactions of both BioInfer and GENIA are closer to the collapsed Stanford parse than to the other considered parse representations, supporting its value in extracting complex interactions.

The independently developed complex interaction formats of BioInfer and GENIA and the collapsed Stanford dependency parse are strikingly similar. We assume this indicates that these schemes succeed in capturing the essential structure and information of the annotated text. Our analysis is the first comparison of the relative complexity of BioInfer and GENIA interactions and our results suggest that they are of roughly similar complexity in this regard.

Comparison of complex and binary interactions indicates that while complex interactions can correspond closely to a syntactic dependency parse, binary interactions often link syntactically distant words. Therefore, despite appearances, complex interactions may prove to be easier to extract than binary ones. As previous studies have shown (Pyysalo et al., 2008), with binary interactions the definition of "interaction" also varies substantially, leading easily to ambiguous data. We hope that complex annotation will allow a more precise definition of the various concepts falling under the term "interaction", allowing both the development of better extraction systems and more consistent evaluation of the results.

These findings will be useful when we attempt to use the studied parses and annotations in the development of an automated system for the extraction of complex interactions. Our preliminary study indicates that the resources we evaluated can provide a consistent basis for future work.

Acknowledgments

This work was supported by the Academy of Finland.

References

A. Clegg and A. Shepherd. 2007. Benchmarking natural-language parsers for biological applications using dependency graphs. *BMC Bioinformatics*, 8(1):24.

K. B. Cohen and L. Hunter. 2006. A critical review of PASBio's argument structures for biomedical verbs. *BMC Bioinformatics*, 7(Suppl 3):S5.

J. Heimonen, S. Pyysalo, F. Ginter, and T. Salakoski. 2008. Complex-to-pairwise mapping of biological relationships using a semantic network representation. In *Proc. of SMBM'08*. To appear.

J-D. Kim, T. Ohta, and Tsujii J. 2008. Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, 9(1):10.

Martin Krallinger, Florian Leitner, and Alfonso Valencia. 2007. Assessment of the second BioCreative PPI task: Automatic extraction of protein-protein interactions. In *Proc. of BioCreative II*, pages 41–54.

M.-C. de Marneffe, B. MacCartney, and C. D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proc. of LREC'06*, pages 449–454.

S. Pyysalo, F. Ginter, J. Heimonen, J. Björne, J. Boberg, J. Järvinen, and T. Salakoski. 2007. BioInfer: A corpus for information extraction in the biomedical domain. *BMC Bioinformatics*, 8(1):50.

S. Pyysalo, A. Airola, J. Heimonen, J. Björne, F. Ginter, and T. Salakoski. 2008. Comparative analysis of five protein-protein interaction corpora. *BMC Bioinformatics*, 9(Suppl 3):S6.

D. D. Sleator and D. Temperley. 1991. Parsing English with a Link Grammar. Technical Report CMU-CS-91-196, Carnegie Mellon University.

Y. Tateisi, A. Yakushiji, T. Ohta, and J. Tsujii. 2005. Syntax annotation for the genia corpus. In *Proc. of the IJCNLP 2005*, pages 222–227.