Assessment of Modifying versus Non-modifying Protein Interactions

Dietrich Rebholz-Schuhmann[1], Antonio Jimeno[1], Miguel Arregui[1] and Harald Kirsch[1]
[1]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, U.K.

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 definition for the syntactical-semantic representation of the relations with which authors of manuscripts have to comply. In this study, we assess the distribution of verbs denoting binary relations between proteins using different corpora (AIMed, BioInfer, BioCreAtIve II) for protein-protein interactions and measure their performance for the identification of PPI events (in the BioCreAtIve II corpus) based on syntactical patterns. We distinguish modifying interactions (MIs) such as post-translational modifications (PTMs) from non-modifying interactions. We found that MIs are less frequent in the corpus but can be extracted at the same precision levels as PPIs.
Programmatic access to the text processing modules is available online (www.ebi.ac.uk/webservices/whatizit/info.jsf, http://www.ebi.ac.uk/Rebholz-srv/pcorral/).

1 Introduction

Since the innovative approach of (Blaschke et al., 1999), 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 related work), including the one that is best-known and called iHOP (Hoffmann et al., 2005), but none of them offers a comprehensive approach.

In this research work we explore on the use of language in the scientific literature, in particular in annotated corpora for protein-protein interactions to better understand the use of verbs in this context. We follow the hypothesis that language representations for PPIs fall into different categories: (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 thus information extraction techniques will achieve better performances.

The evidence for the modifying interactions is any reporting of chemical changes linked to the interaction partners of the PPI. For example, methylation and demethylation and similarly phosphorylation and dephosphorylation as well as other types of chemical changes (e.g., acetylation, biotinylation) have to be considered here (see table 1). These modifications can be subsumed as posttranslational modifications (PTMs), which are a subcategory of PPIs. (Saric et al., 2006) have integrated these types of interactions into their work. Since the experimental evidence for the reporting of an interaction is linked to chemical changes which require modifying contact between the two proteins, it can be expected that the reported results is a proven protein-protein interaction.

The second group of reported protein-protein interactions forms the largest set and has been commonly used for the identification of PPIs (Temkin et al., 2003; Friedman et al., 2001; Blaschke et al., 1999). 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 protein-protein interactions as part of the BioCreAtIve II challenge (Krallinger et al., 2007).

2 Methods

The identification of protein-protein interactions 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 assessment we relied on the BioCreAtIve II corpus for the IPS task (347,749 sentences from 740 full-text documents), on the AIMed corpus (1,942 sentences from 255 abstracts) and on BioInfer (1,100 sentences from full-text) (Krallinger et al., 2007; Bunescu et al., 2005; Pyysalo et al., 2007). Only the BioCreAtIve corpus delivers a set of CID pairs for every contained document where the CID pair represents a protein-protein interaction.

2.1 Named entity recognition for proteins/genes

The identification of PGNs has been studied extensively (Morgan et al., 2007; Hakenberg et al., 2005; Hirschman et al., 2005). The identification of gene mentions has been solved to a precision close to 90% whereas the gene normalization is still ongoing work. In this work, the applied tagger (SPtagger) delivers CIDs as part of the NER task and is part of several TM solutions at the EBI (EbiMed, PCorral, MedEvi; Rebholz-Schuhmann et al., 2007a). It incorporates all protein names from UniProtKb/SwissProt and named entity recognition is mainly done by dictionary lookup under consideration of morphological variability, acronym resolution and basic disambituation (Tsuruoka et al., 2007; Gaudan et al., 2005; for SOAP Web services access see Rebholz-Schuhmann et al., 2007b).

2.2 Identification of protein-protein interactions

The identification of protein-protein interactions from the text is based on the modules of the Whatizit infrastructure (Rebholz-Schuhmann et al., 2007b) and through Protein Corral. Public access is granted to all modules that are used in this study. Most modules are implemented as Finite state automata (Kirsch et al., 2006). 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, but contains lexicon extensions for the biomedical concepts. Noun phrases (NPs) are identified with syntax patterns equivalent to "**DET (ADJ|ADV) N+**".

For our study we assessed tri-cooccurrence (3-CO) against syntactical patterns denoting a protein-protein interaction (SynP). 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 nomVP is a nominalization of a verb phrase.

The module that identifies and highlights protein-protein interactions searches for phrases that contain a verb or a nominal form describing an interaction like bind or dimerization. The first set comprises all verbal expressions 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.

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 protein-protein interaction patterns (PPI) are identified. They are basi-

cally 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 and VP denotes verbal phrases including modal verbs. These construction rules for syntactical patterns lead to the selection of structures that are similar to tri-cooccurrence representations but generate higher precision. Similar structures have been proposed by (Huang et al., 2004). Nominalizations increase 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.

3 RESULTS

In the first step we analyzed all three available corpora, i.e. AIMed, BioInfer and BioCreAtIve, and extracted all verbs that co-occur 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 method sections), 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), 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 already by the number of provided sentences 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). 4 verbs appeared only in the BioCreAtIve corpus (biotinylate, dephosphorylate, methylate, pyruvate). This leads to the result that MIs are preferrably 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

In our assessment, we considered the result of the protein-tagger as correct, if the right concept id (CIDs) 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).

Table 2. (Processing full-text documents, One-CID) The table shows the results for the identification of CID pairs from the BioCreAtIve full text corpus for 3-CO and SynP.
SP (SwissProt-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	1,539	211	13.7%	10.0%	11.6%
SP-ci, 3-CO	15,823	609	3.8%	**28.8%**	6.8%
Sp-ci, SynP	2,078	358	**17.2%**	17.0%	**17.1%**

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 counterbalances 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.

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 protein-protein interactions in the BioCreAtIve II corpus. The preci-

sion 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: 7 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.

Table 3. (Processing full-text documents, One-CID, SP-ci) The table shows the predictions from the full-text documents from BioCreAtIve II based on the case-insensitive use of the SP-tagger. 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 methods section). (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%**

4 DISCUSSION

In the presented work, we 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.

(Friedman et al., 2001) 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.

For the ongoing work in the extraction of gene regulatory events, we will analyze how MI and NMI events contribute to the event extraction.

Acknowledgments

This research was sponsored by the EC STREP project "BOOTStrep" (FP6-028099, www.bootstrep.org) including the development of the Term Repository (the prestage to the BioLexicon). Whatizit has been funded by the NoE "Semantic Mining" (NoE 507505).

References

Bunescu,R., Ge,R., Kate,R.J., Marcotte,E.M., Mooney,R.J., Ramani,A.K. and Wong,Y.W. (2005) Comparative experiments on learning information extractors for proteins and their interactions. Artificial Intelligence In Medicine. 33 (2): 139-155.

Friedman,C., et al. (2001) GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. Bioinformatics, 17(Suppl 1), S74–82.

Gaudan, S., Kirsch, H., and Rebholz-Schuhmann, D. (2005) Resolving abbreviations to their senses in Medline. Bioinformatics 21(18):3658-64

Hakenberg,J., et al. (2005). Systematic feature evaluation for gene name recognition. BMC Bioinformatics; 6 Suppl 1:S9.

Hirschman,L., et al. (2005) Overview of BioCreAtIvE task 1B: normalized gene lists. BMC Bioinformatics, 6(Suppl 1):S11.

Hoffmann,R. and Valencia,A. (2005) Implementing the iHOP concept for navigation of biomedical literature. Bioinformatics 21 (Suppl 2):ii252-8.

Huang,M., et al. (2004) Discovering patterns to extract protein–protein interactions from full texts. Bioinformatics 20(18):3604-3612

Kirsch,H., et al. (2006) Distributed modules for text annotation and IE applied to the biomedical domain. Int. J. Med. Inform. 75(6):496-500.

Krallinger,M., Leitner,F., and Valencia,A. (2007) Assessment of the Second BioCrea-tive PPI task: Automatic Extraction of Protein-Protein Interactions. Proc Second BioCreative Challenge Evaluation Workshop.

Morgan,A., and Hirschman,L. (2007) Overview of BioCreative II Gene normalization. Proc Second BioCreative Challenge Evaluation Workshop.

Pyysalo,S., Ginter,F., Heimonen,J., Björne,J., Boberg,J., Järvinen,J., and Salakoski,T. (2007) BioInfer: a corpus for information extraction in the biomedical domain. BMC Bioinformatics. 9;8:50.

Rebholz-Schuhmann,D., et al. (2007a) EBIMed: text crunching to gather facts for proteins from Medline. Bioinformatics 23(2):e237-e244.

Rebholz-Schuhmann,D., et al. (2007b) Text processing through Web services: Calling Whatizit. Bioinformatics 2007 Nov 21

Saric,J., et al. (2006) Extraction of regulatory gene/protein networks from Medline. Bioinformatics 22(6):645-650.

Temkin,J.M. and Gilder,M.R. (2003) Extraction of protein interaction information from unstructured text using a context-free grammar. Bioinformatics. 19(16):2046-53.

Tsuruoka,Y., et al. (2007). Learning string similarity measures for gene/protein name dictionary look-up using logistic regression. Bioinformatics. 23(20):2768-74.