Procesamiento del Lenguaje Natural, Revista nº 40, marzo de 2008, pp. 137-143 recibido 12-02-08, aceptado 03-03-08

PPIEs: Protein-Protein Interaction Information Extraction system[*]

PPIEs: Sistema de Extracción de Información sobre interacciones entre proteínas

Roxana Danger Paolo Rosso Ferran Pla Antonio Molina
Technical University of Valencia
Cam. Vera, s/n 46022 (Spain)
(rdanger; prosso; fpla; amolina)@dsic.upv.es

Abstract: More than three millions research articles have been written about proteins and Protein-Protein Interactions (PPI). The present work describes a plausible architecture and some preliminary experiments of our Protein-Protein Interaction Information Extraction system, PPIEs. The promising results obtained suggest that the approach deserves further efforts. Some important aspects that need to be improved in the future have been identified: entity recognition; lexical data storage and searching (in particular, controlled vocabularies); knowledge discovery for ontology enrichment.
Keywords: Information Extraction, Protein-Protein Interaction.

Resumen: En la literatura aparecen más de tres millones de artículos acerca de las proteínas y sus interacciones (PPI). En este trabajo se expone una arquitectura plausible y algunos experimentos preliminares de nuestro sistema de extracción de información sobre interacciones entre proteínas, PPIEs. Los resultados obtenidos son muy prometedores, por lo que el trabajo merece ulteriores desarrollos. Este estudio ha permitido, además, identificar algunos aspectos a mejorar en el futuro: el reconocimiento de entidades y el almacenaje y búsqueda de datos léxicos (en particular, los vocabularios controlados) y el descubrimiento de conocimiento para el enriquecimiento de ontologías.
Palabras clave: Extracción de información, Interacción entre proteínas.

1 *Introduction*

The goal of Information Extraction Systems (IES) is the enrichment of knowledge bases with information from texts. None of the different methodologies used to solve this problem has clearly demonstrated its superiority (Reeve and Han, 2005). On the one hand, many of them are based on learning processes. In such cases, the quality of Information Extraction (IE) depends on the representativity of the training data, and the ability for generalization of the systems. On the other hand, the majority of IES uses a complete syntactic and semantic analysis. The quality here is affected by possible errors during Natural Language Processing (NLP).

Background knowledge is an essential element for IES. If the interesting concepts for the task are known, as well as others semantically related concepts (such as their synonyms, antonyms, meronyms, etc.,), its identification could be used for an effective IE. The methods for instance extraction should be based on the own nature of the data to be extracted.

This kind of IES guided by knowledge - or, more formally, by ontology- has demonstrated to be effective when the domain knowledge is enclosed and specific enough. For example, in (Danger, 2007) is described IES to populate an archeology ontology from text collection of archeology site memories. The system has considered both the ontological entities and the complex instances related them, and obtained a 92% of precision and 84% of recall for the archeology ontology with more than 500 concepts and relations.

Our goal is to propose a general architecture for IES guided by ontologies, which allows to enrich both the domain knowledge of ontologies and their instances. This study

[*] This work has been funded by the projects TIN2006-15265-C06-04 and "Juan de la Cierva" of the Ministry of Education and Science of Spain.

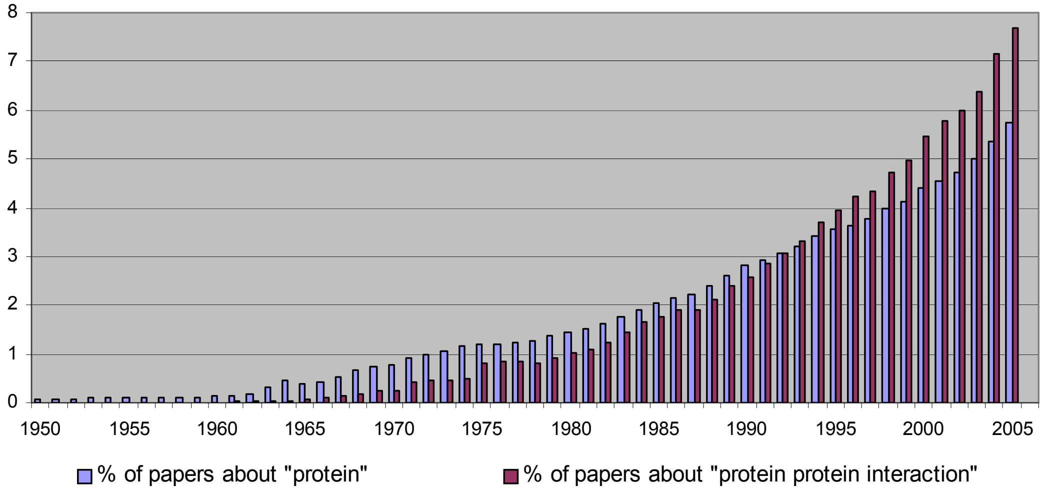

Figure 1: Increasing interest of the biomedical community in PPI research. Data source: http://dan.corlan.net/medline-trend.html.

is part of a research project for the specific biomedical domain[1]. The availability of huge data in text format, the growing interest in the fascinating world of proteins as well as the necessity for biochemistry researchers to arrange all discovered protein features in databases made us decide to carry out some experiments in the Protein-Protein Interaction (PPI) domain. The present work summarizes the available resources which make plausible our proposal and shows some preliminary results of the simplest IES guided by ontology we conceive for the PPI domain.

Section 2 introduces the role of proteins for life, and the importance of PPI. In Section 3 the available resources as well as our first PPIEs (Protein-Protein Interaction Information Extraction system) are described.The results of some preliminary experiments carried out using our PPIEs, are discussed in Section 5. Finally, conclusions and future works are drawn in Section 6.

2 Proteins and Protein-Protein Interaction

Heredity and variation in living organisms are the subject study of Genetics. The discoveries obtained from the pioneer studies of Mendel in 1880 up to have made possible to understand a little but exciting part of the biochemical mechanisms of the living bodies. A very short and shallow summary of genetic discoveries is given below.

Each cell (the human body has about 100 billion of cells) contains *DNA* (Deoxyribonucleic acid) molecules, which are sequences of nucleotides that "describe" hereditary information, contained in a set of chromosomes (23 pairs for humans). DNA fragments containing this hereditary information are *genes*; other fragments are involved in the structural definition or in the regulation processes of the cells. At the beginning of a gene there is a *promoter* which controls its activity, and the coding and non-coding of a sequence. Non-coding sequences regulate the conditions necessary for *gene expression* (the process of converting a gene into a useful form for the cell). The products of gene expression, determined by the coding sequences, are in the majority proteins.

Proteins are linear polymers built from 20 aminoacids. The majority of chemical reactions occurring inside the cell are produced thanks to the protein capability of binding other molecules. Bindings between the same molecule form fibers (structural function). If a protein is associated with other ones, an interaction between proteins is observed.

Protein-protein interactions allow catalyzing chemical reactions (enzymatic function), controlling the cell cycle (control function) and assembling protein complexes (complex functions) which, in turn, are involved in cell signing or in signal transduction functions.

The importance of PPI in living bodies

[1] MIDES: Métodos de aprendizaje para la minería de textos en dominios específicos.
http://gplsi.dlsi.ua.es/text-mess/index.php

has motivated an increasing interest in their study. Figure 1 shows the proportional increasing of the published papers about proteins and PPI since the middle of the last century until nowadays. Up to 2005, more than 3 millions papers about proteins have been published, and at least 5% of them were related specifically to PPI. In the figure, it may be noticed the growing interest of the biomedical community in protein research, and it is clear the faster behaviour of the published papers regarding to PPI.

Different point of views are emphasized in the studies about proteins: their structural utility, biochemical signals and/or biochemical reactions. All viewpoints have to be combined in order to obtain a general idea of the influence of a determined gene or protein in the organism. Moreover, PPI are important because they may help to discover the functions of other proteins making them interact and observing the successive behaviour. Considering all the above, the current challenge of bioinformatics is to populate biomedical databases with the essential information in order to allow some basic processing, such as searching or general comparison between proteins or their interactions.

Currently, manual and semi-automatic processing are carried out in order to make the recent discoveries available to all biochemical community. The present work aspires to contribute to this process of information diffusion and interchange.

3 PPI resources

The PPI resources which make possible to define an IES are enumerated in the three successive sections. As we explained above, the definition of an ontology to guide the process is essential. In the literature we have found different ontologies regarding PPI. Their study have allowed us to discover the indispensable information needed to be extracted. On the other hand, some biomedical NLP tools have been defined; the understanding of the used methods together with how to improve them is an important issues. Finally, we describe the available data as well as the textual medical databases over which we work.

3.1 PPI ontologies

The biomedical community has been developing a set of ontologies (the OBO, *Open Biomedical Ontologies*[2]) complying with various requirements, including a minimal level of agreement between experts in each domain area. A controlled and consensual vocabulary useful in many tasks may thus be assumed. The most relevant ontologies (structures of databases, in some cases) associated with proteins and their interaction concepts are: *intAct* (Interaction Database), *interPro*, *PO*, *Uniprot/Swiss-Prot*, *MI*, *MGED* and *Tambis*.

All above ontologies share a set of 4 essential concepts, which have been described in (Orchard and et. al., 2007) as the minimal interesting information for PPI:

- *Publications*: a subject research together with its authors, institutions, journal of publication, etc. and the experiments which have been carried out;
- *Experiments*: a description of the experiments which justify the research;
- *Interactions*: a list of interactions occurring in the experiments;
- *Interactors*: a list of interacting molecular elements.

An ontology-driven IES for PPI should consider, in an initial stage, at least the above concepts. In successive stages, other related concepts could be incrementally added.

3.2 Biomedical NLP tools

Recognizing bio-entities (proteins, genes, biological functions, diseases, treatments and others biomedical concepts) is the task in which current developments are focusing on. Given the huge amount of concepts available in the controlled vocabularies which could appear in biomedical texts, some of these recognizers merge Information Retrieval (IR) and IE techniques in order to speed up the recognition process.

Table 1 gives an idea of the quality of protein entity recognizers. Four of the available systems were (trained if necessary and) used to extract proteins from the evaluation sentences provided by BIOCREATIVE'06 challenge[3]. As may be noticed, more than 44% of the proteins remained undetected.

Most of the biomedical recognizers use: rules or dictionary searcher strategies, like in (Hanisch et al., 2005) and (Kou, Cohen, and Murphy, 2005); or machine learning

[2] http://obo.sourceforge.net

[3] http://biocreative.sourceforge.net/biocreative_2.html

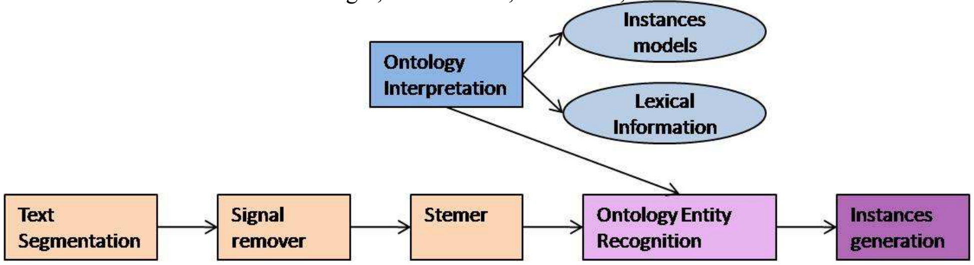

Figure 2: General architecture for a simple IES.

System	Pr	R	F1
ABNER	0.57	0.44	0.50
GAPSCORE (Score ≤ 0.3)	0.67	0.52	0.56
NLPROT	0.57	0.56	0.56
WHATIZIT	0.82	0.54	0.65

Table 1: Comparison of protein recognizers. Pr=Precision, R=Recall.

approaches based on Hidden Markov Models or Conditional Random Fields, like in (Okanohara et al., 2006) and (Sun et al., 2007).

Such bad results are due to the terminology problems observed in bio-entities. Although some molecular names provide useful cues (as the molecular weight, function or the discoverer name), many interactors are described by long, compound, ambiguous, common and jargon English words.

However, in BIOCREATIVE'06 challenge (Wilbur, Smith, and Tanabe, 2007) new proteins recognizers (not freely available) which obtain better results with a highest F1-score of 87.21, have been described. Moreover, combining the results a significant improvement of a 90.66 of F1-score is achieved. This fact reveals us that new bio-entities recognizers, in particular proteins, would be able to reach high quality values by combining different techniques. A similar conclusion was obtained in recent comparison studies (Ponomareva et al., 2007), (Sun et al., 2007).

A representative set of IES for PPI has been met in BIOCREATIVE'06 challenge (Krallinger, Leitner, and Valencia, 2007). The competition was concentrated in detecting pairs of proteins and the kind of interaction between them. The common framework of the systems is to use a complete syntactic and semantic analysis to extract clearly defined interactions. Interactions are extracted considering verb joining two proteins or a set of grammatical rules manually computed. The systems which detected interactions from raw text obtained a F-score of 30, whereas those that used manually interactor annotations reached as much an F-score of 48.

3.3 Public PPI data

The biomedical community publishes various databases in which PPI are described and are constantly updated and supervised by biologists. The most relevant are: *HPRD (Human Protein Reference Database)*, *IntAct (Interaction Database)* and *DIP (Database of Interacting Proteins)*. Each of them provides sophisticated searching capabilities in order to allow users to review, compare and search for particular protein features.

A big amount of researches are public available in various format (pdf, xml, etc.). *Pubmed database*[4] provides access to citations from biomedical literature of many journals and conferences. Moreover, the data available in databases are referred to Pubmed paper identifiers. Therefore, combining both sources of information, sets of texts for training and evaluation purposes may be easily defined.

4 Defining our first PPIEs

The simplest approximation we may conceive for an IES guided by ontologies is represented in Figure 2. It is composed basically by a process which converts a raw text in a list of words (by using a text segmentation, which includes the recognition of simple datatypes such as those that use regular expressions, and a signs remover). Then, the words are stemmed and used by ontology entity recognizers.

Ontology entities to be recognized are defined in form of concepts and relations of a

[4] http://www.ncbi.nlm.nih.gov/PubMed/

Type of entity	Vocabulary Resource
Biological role	psi-mi.obo#biological role
Cell type	cell.obo#cell
Detection method	psi-mi.obo#interaction detection method
Identification method	psi-mi.obo#participant identification method
Interaction type	psi-mi.obo#interaction type
Interactor type	psi-mi.obo#interactor type
Tissue type	http://www.expasy.org/cgi-bin/lists?tisslist.txt
Protein name	Uniprot/Swiss-Prot database[5]

Table 2: PPI controlled vocabulary. *Notation: Ontology name#concept base in the Ontology.*

PPI ontology. We assume that the lexical information to extract them from text is also specified in the ontology. Therefore, a reasoner should be used to: 1) interpret the ontology, that is, the concepts and their relations; and 2) make available lexical information needed for the IE task.

The instance generator makes use of the algorithm proposed in (Danger, 2007). This algorithm defines a set of rules for the complex instance generation which use the ontology interpretation to properly link a list of ontological entities.

The above architecture is useful for a study of the complexity of the problem we are facing. In the following sections we describe, our PPI including how the lexical information has been linked to the appropriated ontological elements and the inference process used to generate the complex instances.

4.1 PPI ontology

We have defined an ontology in OWL (Ontology Web Language) for PPI, based on the recommendations about the minimal interesting information for PPI (Orchard and et. al., 2007). We include other important and well classified concepts related to this domain knowledge such as: interaction and interactor types, biological role of a host in the experiments, cell type on which the experiment was carried out or applied, detection interaction and identification of the interactors methods.

The ontology we defined, PPIO, contains 19 concepts and 21 relations. Moreover, it has been enriched with lexical information in two annotation properties, *lex* and *lexValue*. Through them the lexical methods for identifying ontological elements (concepts and properties) and properties values are described. In the current implementation *lex* and *lexValue* are limited to list entity examples.

Entity recognizers are simply dictionary searchers. In Table 2 the resources from which the dictionaries have been created are described. Almost all of them are ontologies from the *Open Biomedical Ontologies*[6].

4.2 Ontology Reasoner and instance generation

The Pellet reasoner[7], the most popular reasoner for OWL, has been used to recover, from *PPIO*, the instances models (general descriptions of the concepts and their relations) and the lexical information which will be used to generate complex instances describing protein-protein interactions.

For simplicity, the reader should assume that we obtain, for each concept, the other concepts and relations associated with it, its position in the hierarchy with respect to the others concepts, and how to recognize it in a text. Therefore, using all this information, the ontology entities in texts may be discovered. It is easy to infer the compositions of relations linking two concepts and the semantic distances between them. The two aspects above allow, by using the algorithm introduced in (Danger, 2007), to infer the complex ontological instances described in texts.

5 *Preliminary experiments*

Experiments have been carried out on two resources developed and maintained by EBI[8]. The first resource is *IntAct*, the previously mentioned database, and the second one is a set of 3422 paragraphs extracted from PPI research papers along with the interaction identification number (*Accession number, AC*) in IntAct database which represents the interaction described in the paragraph. Each paragraph represents a complex interaction in-

[6] http://obo.sourceforge.net

[7] http://www.mindswap.org/2003/pellet/

[8] http://www.ebi.ac.uk/

Type of entity	%of Parag.	Precision	Recall
Biological role	100	90	46
Cell type	32	92	69
Detection method	100	70	23
Identification method	100	98	85
Interaction type	100	99	83
Interactor type	100	100	78
Tissue type	9	58	35
Protein name	100	95	78

Table 3: Entities in text paragraphs.

stance: there are 3422 interaction instances which include a total of 87186 relations.

For example, given a typical paragraph such as:

"Co-immunoprecipitation from T-cells of theta PKC and p59fyn.",
ontological entities are recognized using dictionary searchers, as in the example:

<*detect_method*>Co-immunoprecipitation </*detect_method*> from <*tissue_type*> T-cells </*tissue_type*> of <*protein*> theta PKC </*protein*> and <*protein*> p59fyn </*protein*>.

Finally, the corresponding instance is reconstructed using the instance generator as follows. The indentation is used to identify relations with previously defined instances. As it may be noticed, the complex instance is created using the list of recognized entities. The appropriate relations are selected and used to link the corresponding instances. Some instances (such as *experiment*) and data (such as *interaction type*) are inferred using the ontology information.

```
interaction
 has_been_produced_by :: experiment
  found_in_source :: ncbiTaxId=9606
  has_tissue_type :: Peripheral blood T-lym.
  detect_method :: anti bait coimmunoprecipit.
 has_participant :: Concrete_interactor
  name :: Proto-oncog. tyros.-protein kin. Fyn
  interactorType :: protein
 has_participant :: Concrete_interactor
  name :: Protein kinase C theta type
  interactorType :: protein
 has_interaction_type :: physical interaction
```

Table 3 shows for each type of entity mentioned in the paragraphs, the percentage of paragraphs in which it has been found and the precision and recall obtained by the particular ontology entity recognizer.

High recall values were obtained for *proteins*, but these results are due to the completeness of the protein dictionary, which also includes protein synonyms. In the future, we should use a molecular (protein) recognizer based on morpho-syntactic features of protein names, and protein synonyms should be discovered and matched to the corresponding most common protein names. We limit the analysis to *protein interactor types*: therefore, the precision is of 100% and the recall coincides with the recall of *protein name.*

Other entities have different behaviours. The *interaction type*, *identification method* and *cell type* concepts are well recognized due to the stability of their vocabulary, whereas a low proportion of *detection method*, and *tissue type* are recognized. We plan to perform a thorough study of the dynamism of biomedical terminology in order to recognize new terms, as well as to improve the entity disambiguation mechanism. Also, a process for identifying typing errors will be included, because we notice a high frequency of such mistakes in the processed text.

With respect to the instance generation process, a precision of 72% and a recall of 67% were obtained considering all paragraphs. We consider that an instance is well recognized if it is referred to the correct concept and all its relations are well formed.

In spite of the rather simple linguistics processing, the precision and recall values obtained by the system are satisfactory. We will try to maintain linguistic processing complexity as low as possible in future developments. Moreover, we plan to improve the entity recognition process to make it less dictionary-dependent.Other two issues will be considered in the future. These are the learning of new terms, synonyms, acronyms and metonyms to enrich the controlled vocabulary, and the efficient recognition of such terms in texts. The latter aspect includes the use of efficient indexing strategies for searching terms appearing in texts.

6 Conclusions and further work

In this paper we have introduced an architecture for an information extraction system about protein protein interactions, PPIEs. The most important resources available regarding PPI have been summarized. Such resources have been used in order to perform information extraction in relevant papers. A domain ontology on PPI has been defined which includes lexical information regarding ontological entities. Preliminary experimental results are encouraging. They indicate that the proposed set of tools is suitable for PPI identification, although a more sophisticated mechanism for entity identification should be used in the future. Furthermore, we plan to study the dynamism of the biomedical vocabulary (including the recognition and evolution of new terms, synonyms, acronyms and metonyms), the disambiguation process and the extension of the PPIO ontology.

References

Danger, Roxana. 2007. *Extraction and analysis of information from the Semantic Web perspective (in Spanish: Extracción y análisis de información desde la perspectiva de la Web Semántica).* Ph.D. thesis.

Hanisch, Fundel, Mevissen, Zimmer, and Fluck. 2005. Prominer: rule-based protein and gene entity recognition. *BMC Bioinformatics*, 6 Suppl 1.

Kou, Zhenzhen, William Cohen, and Robert Murphy. 2005. High-recall protein entity recognition using a dictionary. *Bioinformatics*, 21(1):266–273.

Krallinger, Martin, Florian Leitner, and Alfonso Valencia. 2007. Assessment of the second biocreative ppi task: Automatic extraction of protein-protein interactions. In *Proceedings of the Second BioCreative Challenge Evaluation Workshop*, pages 41–54.

Okanohara, Aisuke, Yusuke Miyao, Yoshimasa Tsuruoka, and Junichi Tsujii. 2006. Improving the scalability of semi-markov conditional random fields for named entity recognition. *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL*, pages 465–472.

Orchard, Sandra and et. al. 2007. The minimum information required for reporting a molecular interaction experiment (mimix). *Nature Biotechnology*, 25(8):894–898.

Ponomareva, Natalia, Paolo Rosso, Ferrán Pla, and Antonio Molina. 2007. Conditional random fields vs. hidden markov models in a biomedical named entity recognition task. In *Proc. of Int. Conf. Recent Advances in Natural Language Processing, RANLP*, pages 479–483.

Reeve, Lawrence and Hyoil Han. 2005. Survey of semantic annotation platforms. In *SAC*, pages 1634–1638.

Sun, Chengjie, Yi Guan, Xiaolong Wang, and Lei Lin. 2007. Rich features based conditional random fields for biological named entities recognition. *Computers in Biology and Medicine*, 37(9):1327–1333.

Wilbur, Johm, Larry Smith, and Lorrie Tanabe. 2007. Biocreative 2. gene mention task. In *Proceedings of the Second BioCreative Challenge Evaluation Workshop*, pages 7–16.