Ontology-Based Protein-Protein Interactions Extraction from Literature using the Hidden Vector State Model

Yulan He
School of Engineering, Computing and Mathematics
University of Exeter
North Park Road, Exeter EX4 4QF, UK
y.he@exeter.ac.uk

Keiichi Nakata and Deyu Zhou
Informatics Research Centre
University of Reading
Whiteknights, Reading RG6 6BX, UK
{k.nakata,d.zhou}@reading.ac.uk

Abstract

This paper proposes a novel framework of incorporating protein-protein interactions (PPI) ontology knowledge into PPI extraction from biomedical literature in order to address the emerging challenges of deep natural language understanding. It is built upon the existing work on relation extraction using the Hidden Vector State (HVS) model. The HVS model belongs to the category of statistical learning methods. It can be trained directly from un-annotated data in a constrained way whilst at the same time being able to capture the underlying named entity relationships. However, it is difficult to incorporate background knowledge or non-local information into the HVS model. This paper proposes to represent the HVS model as a conditionally trained undirected graphical model in which non-local features derived from PPI ontology through inference would be easily incorporated. The seamless fusion of ontology inference with statistical learning produces a new paradigm to information extraction.

1 Introduction

Biomedical literature contains rich information pertaining to genes, proteins, and their role in biological processes. In the past few years, there has been a surge of interest in utilizing text mining techniques to provide in-depth bio-related information services, ranging from identifying gene and protein names within sentences and articles, to trying to establish and predict regulatory networks. Most efforts concerning biomedical literature mining to date focus on automated information extraction which uses natural language processing to search for co-occurrences of names or identifiers of entities along with activation/dependency terms in text. Existing information extraction systems [4, 10, 28] mainly rely on either manually-defined context-free grammars or hand-crafted semantic patterns for the extraction of bio-related information based on strong assumptions about the use of natural language, such as terms typically used to indicate relationships, the ways the named entities are used within sentences, and the typical sentence structures etc. They are not able to extract the inferred relations. As an illustration, consider the following sentence:

In the absence of GCN4, BAS1, and BAS2, the RAP1 protein binds to the HIS4 promoter in vivo but cannot efficiently stimulate HIS4 transcription.

In this sentence, there might be a relationship between GCN4, BAS1, BAS2, RAP1 and the HIS4 promoter. Some additional knowledge is required to reveal the implied relationships.

In biomedical related research, various computational approaches have been proposed to infer protein-protein interactions (PPIs), including those based on genomic information [27], three-dimensional structural information [1] or primary structure of proteins [21], integration of multiple genomic datasets [15], based on evolutionary relationship [22], previously identified domain-domain interactions [16], and protein complex [19], etc. However, inferring PPIs from literature mining is still a less explored area.

The paper explores efficient ways to combine multiple knowledge sources to build a PPI ontology and incorporate the ontology knowledge into relation extraction. A framework of incorporating ontology inference into statistical learning of the Hidden Vector State (HVS) [12] model is comprehensively investigated particularly in response to the following issues:

- Firstly, existing relation extraction approaches mostly focus on the relation itself without concerning about the context where such a relation occurs. For example, relations between biological entities, such as proteins, genes, are conditional and may change when the same entities are considered in a different functional context. As a consequence, every relation between entities

should be linked with the functional context in which the relation was observed. It is crucial to investigate context-dependent relation extraction approaches.

- Secondly, it is important to explore effective ways to incorporate multiple knowledge sources into the process of relation extraction. Some work has been done to combine the ontologies or lexicons etc. with relation extraction. However, it is still a less explored area in integrating multiple knowledge sources into relation extraction.
- Thirdly, since the existing methods depend on the co-occurrence of terms, within a sentence, a phrase, or an abstract, they can only reveal relationships that are already reported in the literature and do not attempt to detect new relations. For example, if there is a report relating protein A to B, and another report relating B to C, it may suggest a possible relation between A and C.

The above issues describe a number of challenging topics of research, leading to an ontology-based relation extraction framework significantly superior to the existing relation extraction approaches; the relation extraction framework would be able to perform inference on hypothesizing new relationships and also be able to present more accurate information. The rest of the paper is organized as follows. Section 2 discusses the related work. Section 3 presents the proposed framework. An example illustrating the feasibility of the proposed approach is given in Section 4. Finally, Section 5 concludes the paper.

2 Related Work

This section presents the existing work in three areas, PPI ontology construction, ontology guided PPI extraction, and incorporating non-local information into relation extraction.

2.1 PPI Ontology Construction

Gene ontology (GO) [2, 17] is the most commonly used biological ontology. It defines a hierarchical vocabulary to describe gene and gene products in any organism. The structure of the ontology consists of biological processes, molecular functions, and cellular components. However, protein-protein interaction domain is not covered by GO.

In recent years, there have been efforts in constructing PPI ontologies. Drabkin et. al [6] did not construct a PPI ontology directly, but they proposed a methodology for integrating and visualizing protein-protein interactions. They constructed protein interaction networks for mouse proteins by utilizing information encoded in the GO annotation. Specifically, they searched for annotations for "protein binding" (GO:0005515) since it is defined as "interacting selectively with any protein or protein complex". These annotations represent an experimentally tested interaction of two proteins. Thus, by searching for GO annotations with "protein binding", a network of protein interactions can be constructed.

He [11] proposed a hybrid approach to build a PPI ontology called PPIWordNet. Key concepts are extracted from source texts and added into the ontology based on two factors on measuring how discriminating a term is for the PPI domain compared to other knowledge domain and how important a term is to the biological evaluation of PPIs. Relationships among concepts are determined manually by knowledge engineers and domain experts. The constructed PPI ontology is then merged into GO.

More recently, Newman *et al.* [20] built a BioMANTA OWL ontology[1] by integrating multiple data sources within a single RDF triple store through a common PPI model. The BioMANTA ontology focused on integrated concepts from PSI-MI[2], BioPAX level 2 [29], Cell Type[3], Gene Ontology [2] and NCBI Taxonomy[3]. The key concepts consist of are different types of observation including experimental, predicted, and inferred; and provenance information including data source, the type of experiment, the cell type, inferencing method, sub-cellular location and observation reference (a BioPAX publication cross reference). A number of protein databases such as UniProt [30], DIP [25], IntAct [13] and MPact [8] were integrated to form a uniform RDF representation.

2.2 Ontology Guided PPI Extraction from Text

The only system that we are aware of incorporating ontology into protein-protein interactions extraction is the PPIEs (Protein-Protein Interaction Information Extraction System) [5]. An ontology in OWL (Ontology Web Language) for protein-protein interactions (PPIs), called PPIO, has been defined which includes interaction and interactor types, biological role of a host in the experiments, cell type on which the experiment was carried out or applied, detection of interaction and identification of the interactor methods in addition to the four essential concepts about the minimum interaction information for PPI, publications, experiments, interactions, and interactors. The PPIO contains 19 concepts and 21 relations. The information extraction system first converts a raw text into a list of words. Then, the words are stemmed and used by ontology entity recognizers which are simply dictionary searches from the *Open*

[1] http://biomanta.sourceforge.net/2007/07/biomanta_extension_02.owl

[2] http://psidev.sourceforge.net/mi/rel2/doc/

[3] http://www.ncbi.nlm.nih.gov/Taxonomy/

Biomedical Ontologies[4]. Those ontology entities to be recognized are defined in forms of concepts and relations of a PPI ontology. The Pellet reasoner[5] has been used to recover, from PPIO, the general descriptions of the concepts and their relations and the lexical information which will be used to generate complex instances describing protein-protein interactions. It is however unclear how the lexical information is gathered in their approach.

2.3 Incorporating Non-Local Information into Relation Extraction

Implied relations do not have direct contextual evidence and thus they require some background knowledge or non-local information in order to be detected. How to incorporate non-local information into information extraction poses a big challenge in the area of natural langauge processing. Traditional approaches that train a probabilistic model use only local features or the constraints imposed by the domain itself. Examples include HMM and its variations, conditional random fields [14] etc. These models can only capture sequential constraints. More recently, rather than being restricted to sequential data, Roth and Yih proposed a linear programming formulation framework [23, 24] to account for constraints supplied by classifiers learned in other contexts and incorporated as background knowledge. Inference could then be modelled as an optimization problem and solved using existing numerical packages. Finkel *et al.* [7] proposed to incorporate non-local structures in a conditional random field based information extraction system with Gibbs sampling, a simple Monte Carlo method used to perform approximate inference.

The aforementioned methods only account for non-local information within a particular textual corpus. It is not apparent how external knowledge can be incorporated into the inference procedure. We propose an ontology-based information extraction framework that incorporate ontology knowledge into the relation extraction process in an iterative manner. Details of the framework are presented in the following section.

3 Proposed Framework

The overall process of the proposed framework is shown in Figure 1 which takes the form of four main processes. First, context-dependent information extraction aims to extract enriched PPI information which include PPI attributes, functional context, experimental environment, etc., in additional to the PPIs. The extracted information will be combined with the external knowledge sources such as gene ontology to form a knowledge base. A PPI ontology will then be built and a probability could be calculated to and attached to each protein-protein interaction pair. Automated reasoning can be performed to reveal the implied PPIs. Evidences of PPIs can be fed back to the statistical relation extraction model to extract more accurate information.

3.1 Context-Dependent Information Extraction

Relations between biological entities, such as proteins and genes, are conditional and may change when the same entities are considered in a different functional context. As a consequence, every relation between entities should be linked with the functional context in which the relation was observed. Moreover, without considering the observed context, it is meaningless and impossible to make general statements whether a relation detected by literature mining is a "yes" or a "no" relation. Obviously, to overcome this obstacle, in-depth analysis on sentence or phrase level is requisite.

Existing approaches merely focus on the extraction of the PPIs without retrieving other contextual information such as whether the PPIs were experimentally proved, what are the experimental methods used, and whether the interaction is direct or indirect, etc. Such information is important to the population of the curated PPI databases which is often the ultimate goal of PPI extraction from text. Although there have been some attempts to extract those enriched PPI information [9], none of them gave radical solutions. In fact, a common limitation of existing approaches is that additional information is extracted separately after the PPI relation extraction. It would be useful to incorporate the extraction of such contextual information into the process of relation extraction.

Furthermore, the knowledge extracted from the literature may contradict itself under different environment, conditions, or because of author's errors, experimental errors or other issues. Although the contradictory knowledge may occupy minor part of the whole interaction network, it is worth more attention. To handle this challenge, one way is to capture the contradictory knowledge in a probabilistic reasoning network so that a confidence value could be defined for each PPI extracted and the decision can be made based on these confidence values. The solution can also be applied to handling different parts of an article, such as abstract, introduction, references and so on, which might be assigned different weights.

3.2 Combining Multiple Knowledge Sources

It is possible to perform automatic validation on the relation extraction results from literature using external knowl-

[4] http://www.obofoundry.org/
[5] http://pellet.owldl.com/

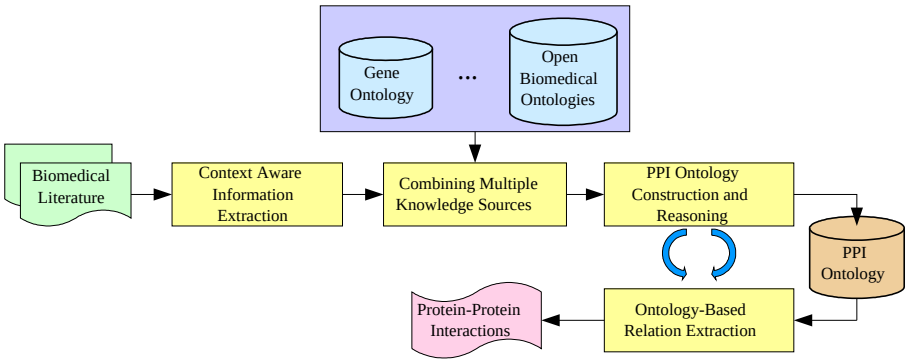

Figure 1. Ontology-based relation extraction for protein-protein interactions.

edge. A novel framework has been proposed in which PPI extraction results are automatically validated using the knowledge mined from gene expression profiles [32]. A probability model has been proposed to score the confidence of protein-protein interactions based on both the PPI extraction results and the gene expression profiles. It is possible to extend the existing work to explore other effective ways to combine multiple knowledge sources.

For example, efforts could be required to focus on linking the knowledge in the databases with text sources available. It is possible to exploit efficiently indirect relationships derived from bibliographic analysis of entities contained in biological databases.

Another possible direction is to make use of the information from ontologies or terminologies. Ontologies are structured lists of terms and are often used by natural language processing (NLP) technologies to establish the semantic function of a word in a document. Gene Ontology (GO) [2, 17] is a popular ontology in biomedicine. It is possible to semantically annotate biomedical text and actively link it to ontologies. Also, protein interactions occur when two proteins are located in the same cellular component, either a permanent cellular location or a transient complex. Thus, knowledge from GO can be used to verify the extracted PPIs that two interacted proteins should be in the same GO cellular components.

3.3 PPI Ontology Construction and Reasoning

In order to support automatic inference from extracted relations, firstly, there need to be a unified knowledge representation of text on hierarchical semantic relations. Secondly, a learning mechanism should be able to induce such a knowledge representation from raw text. Thirdly, an inference mechanism should be used to infer the implied relations from the representations. Traditionally, knowledge representation and reasoning is often based on the first-order logic (FOL) framework. However, simply resorting to FOL requires the conversion from the semantic representation into FOL languages which is normally too complicated. In recent years, the Semantic Web technologies, OWL and RDF, attracted much interests in representing semantic relations.

For the text based biomedical literature, the entities of interest include genes, proteins, enzymes, diseases, etc. These entities together with the relationships among them will be captured in a PPI ontology.

Knowledge gathered from Section 3.1 and 3.2 such as protein attributes, functional context, experimental environment, etc., will be encoded into the ontology and the dependencies between interactions will be captured. Each interaction between two proteins will be associated with certain confidence value that is obtained from statistical learning of PPIs. Indirect relations would become apparent from the relation path encoded in the ontology. “Interesting” and “emerging” relation patterns could be revealed from the inference performed on the ontology.

An example of the structure of the PPI ontology is given in Figure 2. Terms in rectangles represent classes whilst quoted terms represent respective instances under each of the classes.

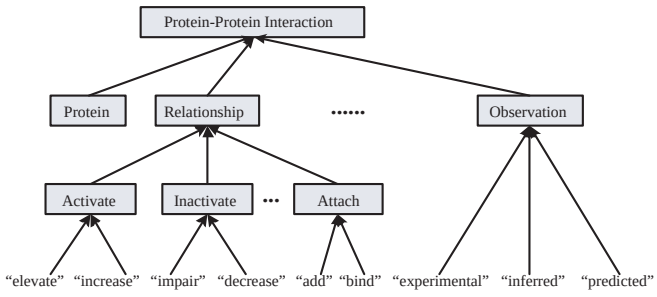

Figure 2. An example of the PPI ontology structure.

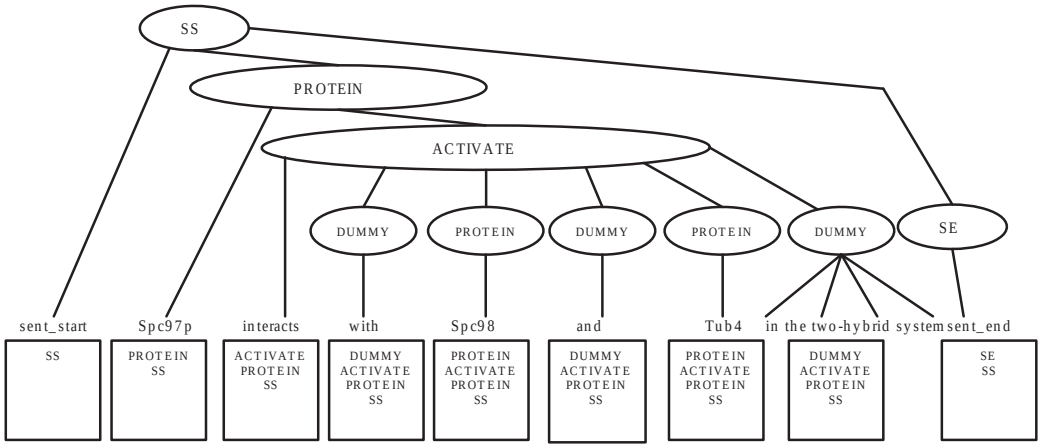

Figure 3. Example of a parse tree and its vector state equivalent.

3.4 Ontology-Based Relation Extraction

We are particularly interested in exploring incorporating PPI ontology knowledge into PPI extraction from biomedical literature based on the Hidden Vector State (HVS) model. The HVS model was originally proposed in [12] and has been successfully applied in biomedical domain for PPI extraction [33].

Given a word sequence W, concept vector sequence $\mathbf{C}$ and a sequence of stack pop operations N, the joint probability of $P(W, \mathbf{C}, N)$ can be decomposed as

$$P(W, \mathbf{C}, N) = \prod_{t=1}^{T} P(n_t|\mathbf{c}_{t-1}) P(c_t[1]|c_t[2 \cdots D_t]) P(w_t|\mathbf{c}_t) \quad (1)$$

where $\mathbf{c}_t$, the vector state at word position t, is a vector of D_t semantic concept labels (tags), i.e. $\mathbf{c}_t = [c_t[1], c_t[2], ..c_t[D_t]]$ where $c_t[1]$ is the preterminal concept label and $c_t[D_t]$ is the root concept label (SS in Fig. 3), n_t is the vector stack shift operation at word position t and take values in the range $0, \ldots, D_{t-1}$ and $c_t[1] = c_{w_t}$ is the new preterminal semantic tag assigned to word w_t at word position t.

An example parse tree is illustrated in Fig. 3 which shows the sequence of HVS stack states corresponding to the given parse tree. State transitions are factored into separate stack pop and push operations constrained to give a tractable search space. The result is a model which is complex enough to capture hierarchical structure but which can be trained automatically from only lightly annotated data.

The HVS model computes a hierarchical parse tree for each word string W, and then extracts semantic concepts C from this tree. Each semantic concept consists of a name-value pair where the name is a dotted list of primitive semantic concept labels. For example, the top part of Fig. 3 shows a typical semantic parse tree and the semantic concepts extracted from this parse would be in equation 2.

```
PROTEIN=Spc97
PROTEIN.ACTIVATE=interacts
PROTEIN.ACTIVATE.PROTEIN=Spc98
PROTEIN.ACTIVATE.PROTEIN=Tub4
```
(2)

The original HVS model takes a form of a generative model which makes it difficult to incorporate background knowledge or non-local features. We propose to represent the model as a conditionally trained graphical model similar to the conditional random fields [14]. The HVS model can be viewed as a graphical model as shown in Figure 5. Assuming the vector state stack depth is limited to be 4, that is, there are at most 4 semantic tags (states) relating to each word position. $\mathbf{c}_t$ is the vector state corresponding to the word W_t. S_t is the stack shift operation which consists of popping n_t semantic tags from the previous vector state $\mathbf{c}_{t-1}$ and pushing one pre-terminal semantic tag to the stack and thus producing $\mathbf{c}_t$.

Given a word sequence W, concept vector sequence $\mathbf{C}$ and a sequence of stack pop operations N, the conditional HVS model takes the form

$$\begin{aligned} P_\Theta(\mathbf{C}, N|W) = \frac{1}{Z_w} \exp\Big(&\sum_{t=1}^{T} \sum_k \lambda_k f_k(\mathbf{c_{t-1}}, n_t, W, t) \\ &+ \sum_{t=1}^{T} \sum_k \mu_k g_k(c_t[1], c_t[2 \cdots D_t], W, t) \\ &+ \sum_{t=1}^{T} \sum_k \nu_k h_k(\mathbf{c_t}, W, t)\Big) \end{aligned} \quad (3)$$

where $\Theta = \langle \lambda_1, \lambda_2, ...; \mu_1, \mu_2, ...; \nu_1, \nu_2, ... \rangle$ is the parameter vector of the conditional HVS model. f_k, g_k, h_k are arbitrary feature functions over their respective arguments, and

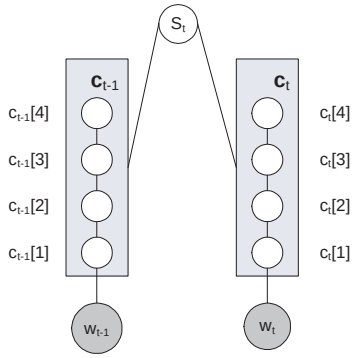

Figure 5. Graphical model representation of the HVS model.

λ_k, μ_k, ν_k are the corresponding learned weights for each feature function.

Inference for the conditional HVS models can be performed efficiently with dynamic programming similar to that described in [14]. Parameter estimation can be performed with standard optimization procedures such as iterative scaling, conjugate gradient descent, or limited memory quasi-Newton method (L-BFGS) [26].

We propose to incorporate the PPI ontology constructed in Section 3.3 into the HVS-based relation extraction process in an iterative manner. Probabilities associated with PPIs which are learned by accounting for multiple knowledge sources in the ontology could be used as constraints to the training of the HVS model which in turn could identify likely false-positive (and false-negative) interactions and eventually improve the extraction performance. On the other hand, PPIs extracted using the HVS model could be used to expand the PPI ontology knowledge source while at the same time minimize low-confidence inferences.

4 Example Illustration

Our system is still under development, we however give an example shown in Figure 4 to illustrate the feasibility of the proposed approach. Firstly, interacted protein pairs are extracted from the sentence shown in Figure 4(A). One of the extracted protein pairs, *Sentrin* and *UbcH6*, is not valid if checking the extraction result manually. Such an error may be ascribed to the relation extraction model's inability of processing negative sentences.

Identification of the false positive PPIs can be done by employing the knowledge from Gene Ontology (GO) based on the following two observations [31, 18]:

- Interacting proteins often function in the same biological process;
- Physical interactions occur when two proteins are located in the same cellular component, either a permanent cellular location or a transient complex.

Thus, the information about the two proteins is extracted from GO as given in Figure 4(B). Based on the directed acyclic graph (DAG) for cellular component of each protein as shown in Figure 4(C)[6], the strength of the relationship between two proteins can be measured based on the similarity between the paths of them which are constructed from the GO term (for example, cytosol) up to the topmost level of the DAG. The similarity is defined based on the number of common terms between two paths. It can be found that *Sentrin* and *UbcH6* are quite dissimilar. Therefore, it can be inferred that it is unlikely that these two proteins interact with each other. Thus, the false positive result generated by the relation extraction model can be eliminated.

5 Conclusions

The framework proposed in this paper provides an alternative technique in which the extracted information is not limited to the pre-defined semantic units. The surrounding context and the PPI ontology knowledge will also be analyzed to validate the extracted relations. We believe it will advance our technology significantly by addressing the need for exploring semantic aspects for text mining. The extensions to the existing technology would produce immediate improvements to the more constrained task of information extraction.

References

[1] P. Aloy, B. Bttcher, H. Ceulemans, C. Leutwein, C. Mellwig, S. Fischer, A.-C. Gavin, P. Bork, G. Superti-Furga, L. Serrano, and R. B. Russell1. Structure-based assembly of protein complexes in yeast. *Science*, 303(5666):20262029, 2004.

[2] M. Ashburner, C. Ball, J. Blake, and D. Botstein. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–9, 2000.

[3] J. Bard, S. Y. Rhee, and M. Ashburner. An ontology for cell types. *Genome Biology*, 6(2), 2005.

[4] D. P. A. Corney, B. F. Buxton, W. B. Langdon, and D. T. Jones. BioRAT: extracting biological information from full-length papers. *Bioinformatics*, 20(17):3206–3213, 2004.

[5] R. Danger, P. Rosso, F. Pla, and A. Molina. PPIEs: Protein-protein interaction information extraction system. *Procesamiento del lenguaje natural*, 40:137–143, 2008.

[6] H. J. Drabkin, C. Hollenbeck, D. P. Hill, and J. A. Blake. Ontological visualization of protein-protein interactions. *BMC Bioinformatics*, 6(29), 2005.

[6]The DAGs are generated from `http://www.uniprot.org/uniprot/`.

[7] J. R. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370, Morristown, NJ, USA, 2005.

[8] U. Gldener, M. Mnsterktter, M. Oesterheld, P. Pagel, A. Ruepp, H.-W. Mewes, and V. Stmpflen. Mpact: the mips protein interaction resource on yeast. *Nucleic Acids Research*, 34 (Database issue):D436–D441, 2006.

[9] B. Haddow and M. Matthews. The extraction of enriched protein-protein interactions from biomedical text. In *Proceedings of BioNLP 2007*, Prague, Czech Republic, June 2007.

[10] Y. Hao, X. Zhu, M. Huang, and M. Li. Discovering patterns to extract protein-protein interactions from the literature: Part II. *Bioinformatics*, 21(15):3294–3300, 2005.

[11] X. He. A protocol for constructing a domain-specific ontology for use in biomedical information extraction using lexical-chaining analysis. Master's thesis, Computer Science, University of Waterloo, Waterloo, Ontario, Canada, 2006.

[12] Y. He and S. Young. Semantic processing using the hidden vector state model. *Computer Speech and Language*, 19(1):85–106, 2005.

[13] S. Kerrien et al. Intact–open source resource for molecular interaction data. *Nucleic Acids Research*, 35 (Database issue):D561–D565, 2007.

[14] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *18th International Conference on Machine Learning*, pages 282–289. Morgan Kaufmann, San Francisco, CA, 2001.

[15] N. Lin, B. Wu, R. Jansen, M. Gerstein, and H. Zhao. Information assessment on predicting protein-protein interactions. *BMC Bioinformatics*, 5:154, 2004.

[16] Y. Liu, N. Liu, and H. Zhao. Inferring protein-protein interactions through high-throughput interaction data from diverse organisms. *Bioinformatics*, 21(15):3279 – 3285, 2005.

[17] J. Lomax. Get ready to GO! A biologist's guide to the Gene Ontology. *Briefings in Bioinformatics*, 6(3):298–304, 2005.

[18] M. Mahdavi and Y.-H. Lin. False positive reduction in protein-protein interaction predictions using gene ontology annotations. *BMC Bioinformatics*, 8(1):262–271, 2007.

[19] S. Martin, Z. Mao, L. S. Chan, and S. Rasheed. Inferring protein-protein interaction networks from protein complex data. *International Journal of Bioinformatics Research and Applications*, 3(4):480–492, 2007.

[20] A. Newman, J. Hunter, Y.-F. Li, C. Bouton, and M. Davis. Biomanta ontology: The integration of protein-protein interaction data. In *Interdisciplinary Ontology Conference (InterOntology08 Tokyo)*, Tokyo, Japan, Feb 2008.

[21] S. Pitre, F. Dehne, A. Chan, J. Cheetham, A. Duong, A. Emili, M. Gebbia, J. Greenblatt, M. Jessulat, N. Krogan, X. Luo, and A. Golshani. Pipe: a protein-protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs. *BMC Bioinformatics*, 7(1):365, 2006.

[22] A. Ramani and E. Marcotte. Exploiting the co-evolution of interacting proteins to discover interaction specificity. *Journal of Molecular Biology*, 327:273284, 2003.

[23] D. Roth and W. Yih. A linear programming formulation for global inference in natural language tasks. In H. T. Ng and E. Riloff, editors, *Proc. of the Annual Conference on Computational Natural Language Learning (CoNLL)*, pages 1–8. Association for Computational Linguistics, 2004.

[24] D. Roth and W. Yih. *Introduction to Statistical Relational Learning*, chapter Global Inference for Entity and Relation Identification via a Linear Programming Formulation. MIT Press, 2007.

[25] L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie, and D. Eisenberg. The database of interacting proteins: 2004 update. *Nucleic Acids Research*, 32 (Database issue):D449–D451, 2004.

[26] F. Sha and F. Pereira. Shallow parsing with conditional random fields. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 134–141, 2003.

[27] S. Tsoka and C. A. Ouzounis. Prediction of protein interactions: metabolic enzymes are frequently involved in gene fusion. *Nature Genetics*, 26:141–142, 2000.

[28] H. woo Chun, Y. Tsuruoka, J.-D. Kim, R. Shiba, N. Nagata, T. Hishiki, and J. Tsujii. Extraction of Gene-Disease Relations from MedLine using Domain Dictionaries and Machine Learning. In *The Pacific Symposium on Biocomputing (PSB)*, pages 4–15, 2006.

[29] B. Workgroup. *BioPAX: Biological Pathways Exchange Language (Level 2, Version 0.5 Draft Release) Documentation)*. BioPAX Workgroup, 2004.

[30] C. H. Wu et al. The universal protein resource (uniprot): an expanding universe of protein information. *Nucleic Acids Research*, 34 (Database issue):D187–D191, 2006.

[31] X. Wu, L. Zhu, J. Guo, D.-Y. Zhang, and K. Lin. Prediction of yeast protein-protein interaction network: insights from the Gene Ontology and annotations. *Nucleic Acids Research*, 34(7):2137–2150, 2006.

[32] D. Zhou, Y. He, and C. K. Kwoh. Validating Text Mining Results on Protein-Protein Interactions Using Gene Expression Profiles. In *The International Conference on Biomedical and Pharmaceutical Engineering 2006*, pages 580–585, Singapore, 2006.

[33] D. Zhou, Y. He, and C. K. Kwoh. Extracting Protein-Protein Interactions from the Literature using the Hidden Vector State Model. *International Journal of Bioinformatics Research and Applications*, 4:64–80, 2008.

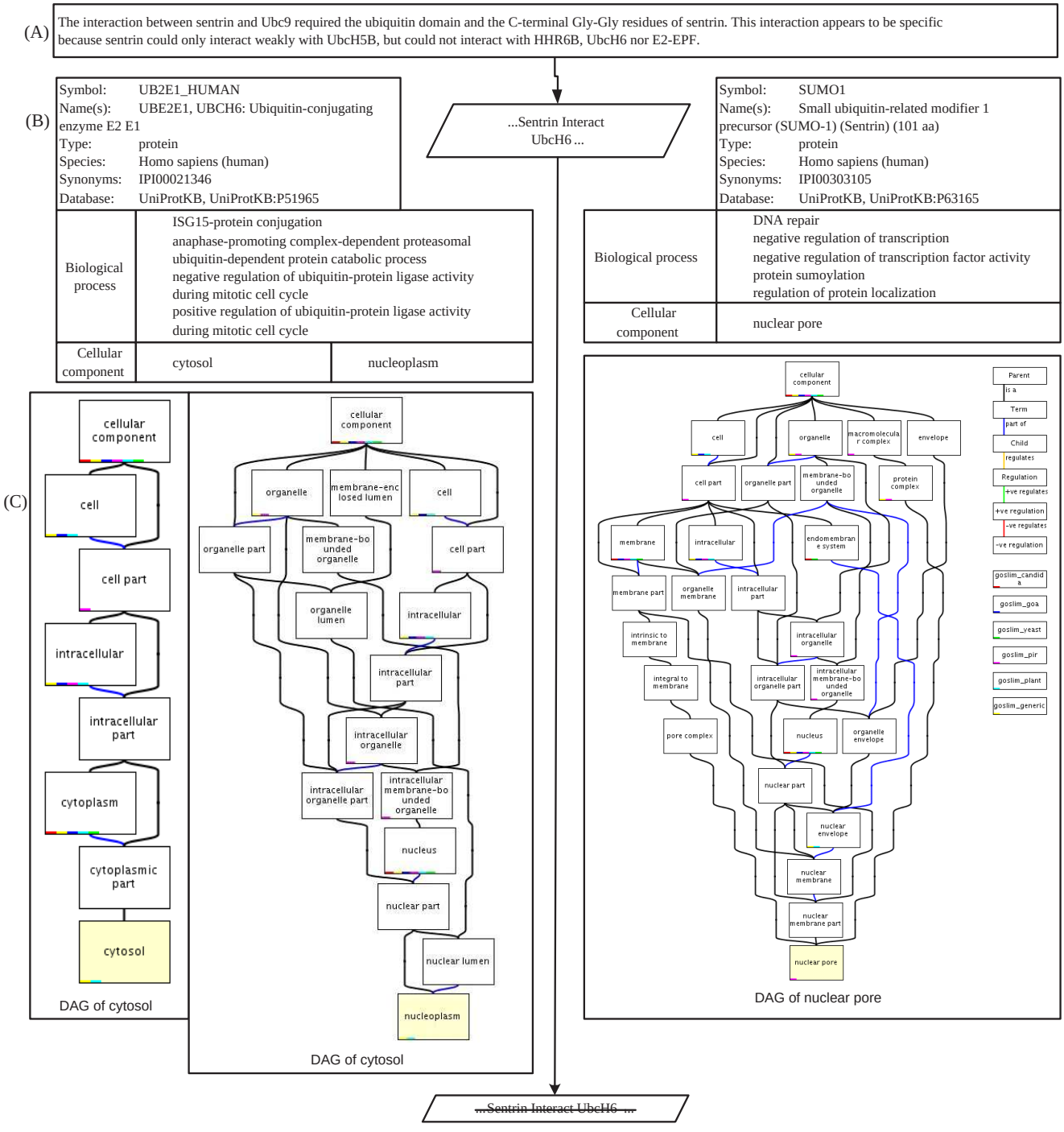

Figure 4. An example of ontology-based PPI extraction.