A protocol for constructing a domain-specific WordNet ontology for use in lexical-chaining analysis of biomedical texts

Author 1 and **Author 2**
Address 1

Author 3
Address 2

Author 4 and **Author 5***
Address 3

Abstract

Biomedical information extraction is becoming an increasingly important focus in Computational Linguistics research. To perform more semantics-based information extraction, we require specialized domain models, but creating such models can be very difficult and time-consuming. We have developed a hybrid methodology for constructing a domain-specific ontology, "PPIWordNet", which integrates key concepts about protein-protein interactions with the Gene Ontology. In addition, we present a method for using our PPIWordNet ontology in discourse-based information extraction to analyze full-text articles on protein interactions. Our discourse-analysis approach uses "lexical chaining" to extract strings of semantically related words that represent the topic structure of the text. We show that the domain-specific PPIWordNet ontology significantly improves the performance of the lexical-chaining analysis. As well, the topic structure as represented by the lexical chains contains important information about protein interactions which we propose may be useful in evaluating the biological validity of these interactions.

Introduction

Natural Language Processing (NLP) techniques are now widely used in biomedical information extraction (IE). Current approaches typically involve identification of simple syntactic features (e.g., parts-of-speech) together with some form of 'shallow' parsing to identify basic syntactic patterns ('templates') or elemental grammatical 'chunks' (e.g., noun phrases, verb phrases) within the sentence (e.g., (Thomas *et al*., 2000) (Pustejovsky and Castaño, 2002)). Potentially a great deal of additional knowledge could be extracted from scientific articles if we were able to derive detailed linguistic information such as lexical meanings, syntactic structure, semantic content, and discourse structure. However, to perform deeper, more semantics-based, information extraction, we require specialized domain models, but creating such models can be very difficult and time-consuming.

Our subject domain is the automated detection and analysis of protein-protein interactions in full-text articles. Knowledge about the identities and functions of interacting proteins contributes significantly to the understanding of biological processes. Our particular interest is the validation of protein-protein interactions—although a large volume of protein-protein interactions has been identified using biomedical information extraction methods, and this information is now readily available in online databases such as BIND (Bader *et al*., 2001), it may be the case that some interactions mined from the literature are not biologically valid, i.e., do not actually occur in the living cell.

We have developed a hybrid methodology for constructing a domain-specific ontology, "PPIWordNet", which integrates key concepts about protein-protein interactions with the Gene Ontology (Gene Ontology Consortium, 2004). In addition, we present a method for using our PPIWordNet ontology in discourse-based information extraction to analyze full-text articles on protein interactions. Our discourse-analysis approach uses "lexical chaining " (Morris and Hirst, 1991) to extract strings of semantically related words that represent the topic structure of the text. We show that the PPIWordNet ontology significantly improves the performance of the lexical-chaining analysis. As well, the topic structure as represented by the lexical chains contains information about protein interactions which we suggest may be useful in evaluating the validity of these interactions.

Background

Our project (First Author and Second Author, 2005) (First Author, 2007) is developing Natural Language systems for extracting information about protein-protein interactions from online biomedical literature using both discourse-based and Machine Learning methods. Our specific aim is to extract meaningful information that can help to evaluate the biological validity of protein-protein interactions contained in online databases. We base our approach on the inherent biological characteristic of protein-protein relationships, namely that interacting proteins will tend to have similar biological functions:

> . . . Although proteins from different groups of biological functions can still interact with each other, it has been shown that the degree to which interacting proteins are annotated with the same functional category is a measure of quality for the predicted interactions (von Mering *et al.* 2002).

*Author 4 and Author 5 are Joint Author 4.

We may reasonably expect then to find biological terms in the context surrounding a protein interaction that indicate the common functions of these proteins. Our methodology is to determine such terms by an automated method of discourse analysis—"lexical chaining"—to provide an additional means of discovering evidence in the literature that an interaction is indeed biologically valid.

The notion of lexical chaining was first introduced by Morris and Hirst (1991), and derives from the concept of textual cohesion. There are a number of forms of textual cohesion, including grammatical cohesion (reference, substitution, ellipsis, conjunction) and lexical cohesion (i.e., semantically related words). As an illustration, the following passage shows several types of lexical cohesion.

(1) John has a Jaguar.

(2) He loves the car.

(3) John works in the garage taking care of his Jaguar.

In this passage, the repetition of the word *Jaguar* in sentences (1) and (3) represents a simple form of lexical cohesion; *Jaguar* and *car* form a part-whole semantic relationship; *car* and *garage* have a nonsystematic semantic relationship. Lexical cohesion occurs between two individual terms, but may lead to sequences of related words.

A *lexical chain* may be defined as a sequence of related words in the text, spanning a topical unit of the text which may be of varying length, either short (adjacent words or sentences) or long (entire text). In the passage above, a lexical chain would be {*Jaguar*, *car*, *garage*, *Jaguar*}. In general, a document will contain many lexical chains, each of which forms a portion of the cohesive discourse structure of the document.

In our research, we are using Enss' (2006) lexical-chaining algorithm, a modification of Silber and McCoy's (2002) linear-time algorithm. Silber and McCoy's method uses WordNet (Fellbaum, 1998), an online lexical database, as the knowledge source for the lexical semantic relationships used in constructing the lexical chains. In WordNet, lexical concepts are organized according to various semantic relations. Words (nouns, verbs, adjectives, and adverbs) are organized into 'synonym sets', known as *synsets*, each of which represents the lexical concept underlying a group of words which are synonymic or near-synonymic in a given context. Synsets can be related by various lexical semantic relations: synonymy, antonymy, hyponymy/hypernymy (subclass/superclass, also known as the *IsA* relation), and meronymy (also known as holonymy, representing various types of *part-whole* relationships).

The primary goal of our research is to create a "PPIWordNet", a WordNet-like linguistic ontology for the protein-protein interaction domain. This PPIWordNet will then be used in a lexical-chaining analysis to determines the strings of biologically related terms which we surmise will cluster in protein-interaction contexts.

The Protocol

A Hybrid Strategy

The two main challenges in ontology construction are ontology concept determination (how the concepts in a domain can be discovered and which concepts should go into the ontology) and relationship determination (how the relationships between concepts are determined). We developed a hybrid approach for the construction of a domain-specific ontology, "PPIWordNet", for the protein-protein interaction (PPI) domain. This approach adopted the "middle-out" strategy (Noy and McGuinness, 2001) which first identifies a core of basic domain concepts, and then specifies and generalizes these concepts as necessary. This hybrid strategy combines the semi-automatic extraction of domain concepts and the manual construction of relationships between concepts.

Methodology

The main processes and steps in our hybrid strategy are shown in Figure 1.

Process 1: PPI Ontology Concept Determination The main goal of this process was to identify key concepts and relationships in the protein-protein interaction domain. This process was broken down further into three main steps: corpus selection; extraction of discriminating terms; seed term identification and glossary construction.

Corpus selection. For our corpus of full-text PPI articles, we used the training set provided by the BioCreAtIvE II[1] (Critical Assessment for Information Extraction in Biology) for the interaction extraction task. For a representative corpus of general English usage, we downloaded the current version of Wikipedia[2].

Extraction of discriminating terms. Our basic strategy for extracting the 'PPI discriminating terms' (terms characteristic of the protein-interaction genre) was to compare the PPI articles with Wikipedia texts to filter out terms which were not specific to protein interactions.

Step 1: Prepare articles The original PPI articles were in HTML format rather than the plain-text needed by our lexical-chaining software. JTidy[3], the HTML parser, was used to parse the HTML files to extract the text.

Step 2: Remove stop words Remove words[4] that have no significance.

Step 3: Term weighting For each PPI article, the term frequency/inverse document frequency (*tf-idf*) for each

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

[2] http://en.wikipedia.org/wiki/Wikipedia:Database_download

[3] http://jtidy.sourceforge.net

[4] The original list of stop words was taken from: William B. Frakes and Ricardo A. Baeza-Yates (Eds.). *Information Retrieval: Data Structures and Algorithms*. Prentice-Hall. 1992. This list was further modified by Charles Clarke for use in TREC and other information retrieval experiments

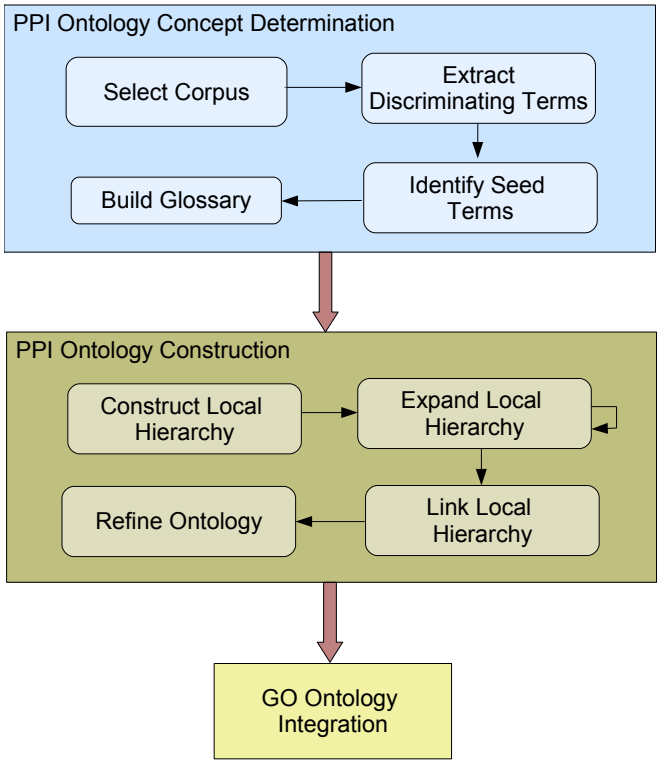

Figure 1: Main processes and steps in our hybrid ontology-construction strategy

term in the articles was calculated. The following formula (Feldman and Dagan, 1995) was used:

(4) $tf - idf(t_k, d_j) = \#(t_k, d_j) * log(|Tr|/\#Tr(t_k))$

in which:

- $\#(t_k, d_j)$ denotes the number of times term t_k occurs in document d_j
- $\#Tr(t_k)$ denotes the number of documents in Tr in which t_k occurs
- $|Tr|$ denotes the number of documents

Step 4: Term matrix For the corpus of PPI articles, a term matrix and a unique list of PPI terms for the entire set of articles were generated. Each row in the term matrix represented an article; each column represented a term appearing in an article and the weight of the term. The frequency of each term in the unique list was calculated by averaging the frequencies of the same term in the set of articles in which this term appeared.

Step 5: Prepare Wikipedia An XML parser was used to extract only the full-text articles from the entire Wikipedia database. Stop words were removed and a Wikipedia term matrix was generated as in the manner described in Step 4. The generated list of unique terms in Wikipedia represented the most common terms in general English usage.

Step 6: Prepare Protein names An XML parser was used to extract the names of the Protein families along with their alternative names from the Human Protein Reference Database (HPRD)[5].

Step 7: Prepare compound names A list of compound names that contain inorganic compounds (compounds without a C-H bond), organic compounds (compounds with a C-H bond), and biomolecules was prepared.

Step 8: Generate discriminating terms The list of protein-protein interaction discriminating terms was generated by removing from the PPI unique list the terms appearing in the Wikipedia, protein, and compounds lists.

Seed term identification and glossary construction. The initial list of 13931 discriminating terms was passed to our two biologist domain experts for manual filtering. Filtering of redundant terms, person names, protein/gene names, etc. reduced the list to 2276 terms with frequent occurrence in the corpora. The domain experts then independently screened the reduced list, scoring terms on a scale of 1 to 5 based on each term's ability to describe generic protein interactions. This resulted in 121 non-redundant terms highly scored by both domain experts. Consultation between the biologists involved removal of terms not having descriptive value for the retrieval of protein interactions from the liter-

[5] http://www.hprd.org

ature, as well as addition of descriptive terms not present in the initial list. This filtering process resulted in a refined list containing 54 'seed terms' from which the PPIWordNet ontology would eventually be created. This process is described more formally as follows.

Step 1: Filter discriminating term list

- Remove redundant terms
- Remove compound, antibiotics, names
- Remove taxonomic and morphologic terms

Step 2: Score terms for relevance

- Score filtered terms based on PPI relevance

Step 3: Identify seed terms

- Select top-scored terms as seed-term list
- Refine seed-term list

Step 4: Glossary building from seed terms

- Provide WordNet-like definitions for seed terms
- Provide synonyms then build WordNet-like synsets

Process 2: PPI Ontology Construction Each of the 54 seed terms from the previous step was subsequently classified by our domain experts into one of three categories based loosely on the Gene Ontology (GO) categorization scheme (interaction detection method, biological function, interaction property). Each domain expert also defined and listed parent and child terms for each term. Using the parent and child relationships, three local hierarchies were created. The process of constructing the complete PPIWordNet ontology was performed progressively, integrating each local hierarchy in turn.

Step 1: Construct local hierarchies

- Classify seed terms into Gene Ontology categories.
- Construct a local hierarchy for each category consisting of the directly related terms for each seed term. Only *IsA* and *PartOf* relationships were considered.

Step 2: Refine local hierarchies

- Use nouns to replace verbs and adjectives
- Introduce any necessary concepts not in discriminating term list needed to make the hierarchy coherent, e.g., more-abstract concepts (superclass), more-specific concepts (subclass), concepts that have the same superclass (siblings) for each seed term.

Step 3: Expand each local hierarchy recursively

- If a newly added term in the previous step appeared in the discriminating list and was considered important, a local hierarchy was built for the new term.

Step 4: Link local hierarchies

- If common terms were found between local hierarchies, the hierarchies were linked.

Step 5: Refine combined ontology of local hierarchies

- Concepts and relationships in the combined ontology were checked for consistency and completeness.

Portions of our ontology of PPI domain concepts are shown in Figure 2 (PPI Method terms) and Figure 3 (PPI Molecular Function terms).

Process 3: Integration with Gene Ontology In the last step, our ontology of PPI domain concepts and the Gene Ontology were integrated to produce the final PPIWordNet ontology. Three integration cases were identified:

Case 1: Unique concept is **used as it is**.

Case 2: Same concept having different subclasses is **expanded** to include all available subclasses.

Case 3: Same concept having different superclasses is **specialized** so the more-detailed relation is adopted.

Evaluation

An Experiment in Lexical Chaining

To evaluate our PPIWordNet ontology, we used it as the basis for a lexical-chaining analysis of protein-protein–interaction full-text articles. The lexical-chaining algorithm was modified to compute lexical chains on a paragraph-by-paragraph basis[6].

We randomly selected 100 articles totalling 2461 paragraphs from the BioCreAtIvE II protein-interaction extraction task's training data as the test set. These articles had been verified as containing detailed information that could be used to identify protein-protein interactions, so would also contain a good sampling of the biological terms likely to occur in protein-protein–interaction contexts.

The PPIWordNet ontology was divided into four components: the original Gene Ontology, PPI Method, PPI Interaction Property, and PPI Molecular Function. The ontology components were added in turn to the lexical-chaining analysis, in the following order:

Step 1: Only Gene Ontology (GO)

Step 2: GO + PPI Method terms

Step 3: GO + PPI Method terms + PPI Interaction Property terms

Step 4: GO + PPI Method terms + PPI Interaction Property terms + PPI Molecular Function terms

In evaluating the results of our experiment, we looked for evidence that the addition of our PPIWordNet ontology had a positive effect on the lexical chains generated, in terms of both quality and quantity.

Results

We performed a statistical analysis on the number and types of lexical chains generated at each step in the analysis. The results of the analysis are shown in Table 1. The results show significant improvements in the quantity of lexical chains, with mild improvements in the quality of the chains. By quantity, we mean the number of lexical chains generated. The size of the ontology between each step increased by an

[6]Descriptions of protein interactions seldom spread across more than one paragraph as observed in our manual lexical-chaining study (First Author and Second Author, 2005)

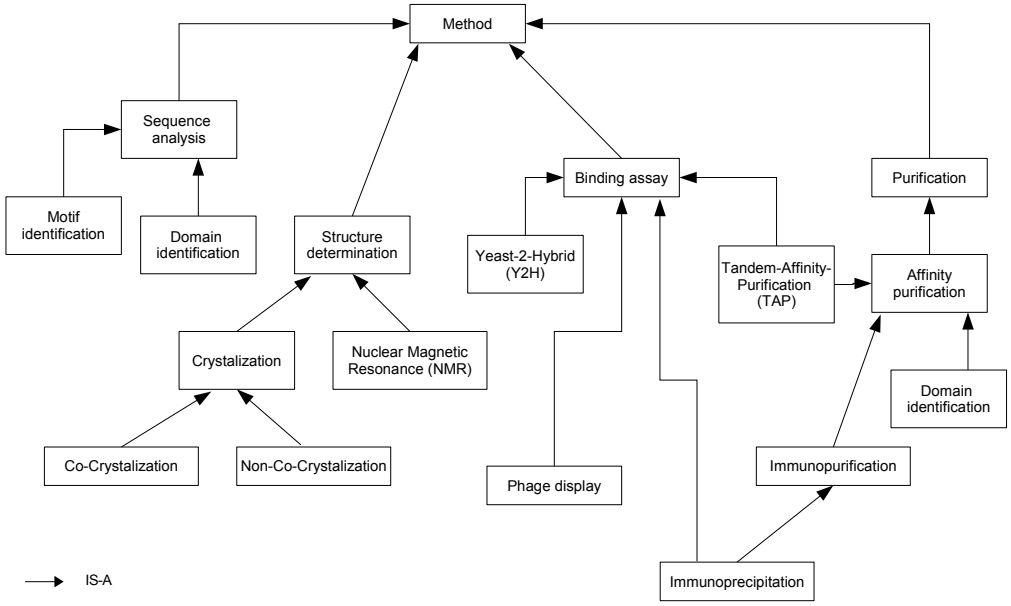

Figure 2: Portion of the PPIWordNet ontology for Method terms

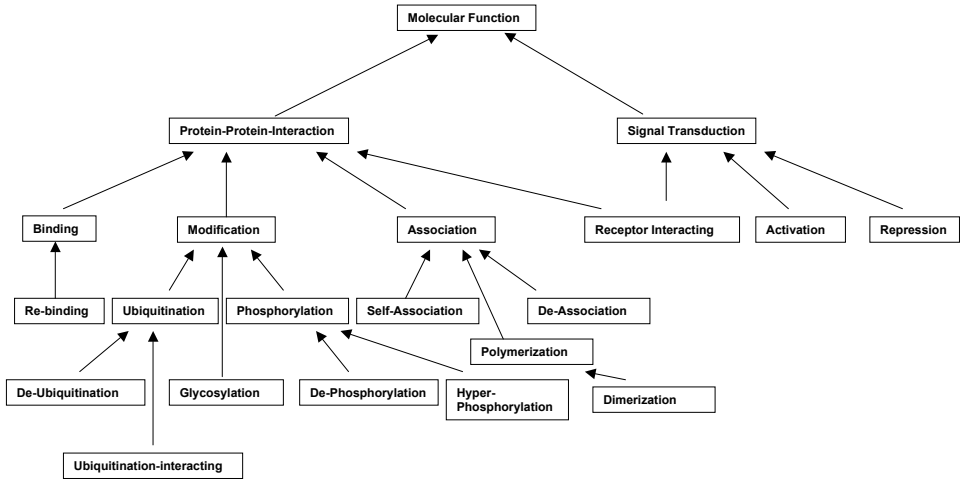

Figure 3: Portion of the PPIWordNet ontology for Molecular Function terms

Measurement	GO	GO + Method	GO + Method + IP	GO + Method + IP + MF
# of terms in ontology	60020	60038	60070	60089
# of chains	4536	5030	5652	5898
average length	5.23	5.16	5.03	5.10
average lemmas	1.19	1.20	1.29	1.40

Table 1: The lexical-chaining experiment results (GO = the Gene Ontology, IP = Interaction Property, MF = Molecular Function, lemmas = # of unique terms in a chain, length = # of terms in a chain)

average of 0.05%, while the increase in the number of chains was 10.9%, 12.3%, and 4.3% respectively, very significant compared to the trivial change of the ontology's size. In terms of quality of lexical chains, we looked at two factors: 'strength' and 'richness' of a lexical chain. The strength of a lexical chain is indicated by the length of the chain, and the richness of a lexical chain is indicated by the lemmas (unique terms in the chain).

The length of a lexical chain represents the degree of importance of a topic (i.e., theme) in the text, that is, an author may emphasize the current topic by repeatedly using closely related terms within a single paragraph. We may reasonably assume that the longer a chain, the more important the theme represented by the overriding sense of the chain. We determined that the average length of a chain decreased by an average of 0.76% between each step.

In terms of richness, the number of lemmas increased by an average of 5.9% between each step, still significant compared to changes in the size of the ontology. We suggest that a larger number of unique terms in a chain will provide more valuable information about the protein interaction in the surrounding context, and will thus be more useful in providing evidence about the biological validity of the interaction.

Conclusions and Future Work

We have outlined a method for biomedical information extraction that makes use of the lexical-chaining discourse structure in scientific articles to determine strings of biologically related words in protein-interaction contexts. Our hypothesis is that lexical chains may provide evidence for the biological significance of the protein interactions occurring in the same context. We have developed a protocol for constructing a domain-specific linguistic ontology to use in lexical-chaining analysis of protein-interaction articles. The results of our study have shown that this specialized ontology signficantly improved the quantity of lexical chains generated, and to some extent the quality as well. Our next task is to investigate methods for improving the quality of the chains, in particular, by enhancing the ontology and lexical-chaining analysis with more-varied lexical semantic relations.

References

First Author. 2007. Master's thesis.

First Author and Second Author. 2005. Conference paper.

Bader, G.D., Donaldson, I., Wolting, C., Ouellette, B.F., Pawson, T., and Hogue, C.W. 2001. BIND—The Biomolecular Interaction Network Database. *Nucleic Acids Res.* 29: 242–245.

Enss, M.J.R. 2006. An Investigation of Word Sense Disambiguation for Improving Lexical Chaining. Master's thesis. University of Waterloo, Waterloo, Ontario.

Feldman, R. and Dagan, I. 1995. Knowledge Discovery in Textual Databases (KDT). In Proceedings of First International Conference on Knowledge Discovery and Data Mining, 112–117.

Fellbaum, C. ed. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press.

Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. 2004. *Nucleic Acids Research*, 32:D258–D261. http://www.geneontology.org.

Morris, J. and Hirst, G. 1991. Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text. *Computational Linguistics* 17: 21–43.

Noy, N.F. and McGuinness, D.L. 2001. Ontology Development 101: A Guide to Creating Your First Ontology. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880.

Pustejovsky, J. and Castaño, J.M. 2002. Robust Relational Parsing over Biomedical Literature: Extracting Inhibit Relations. In Proceedings of the Pacific Symposium on Biocomputing, 362–373.

Silber, H.G. and McCoy, K.F. 2002. Efficiently Computed Lexical Chains as an Intermediate Representation for Automatic Text Summarization. *Computational Linguistics* 28: 487–496.

Thomas, J., Milward, D., Ouzounis, C., Pulman, S., and Carroll, M. 2000. Automatic Extraction of Protein Interactions from Scientific Abstracts. In Proceedings of the 5th Pacific Symposium on Biocomputing, 541–553.

von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G. 2002. Comparative Assessment of Large-Scale Data Sets of Protein-Protein Interactions. *Nature* 417: 399–403.