BioNoculars: Extracting Protein-Protein Interactions from Biomedical Text

Amgad Madkour, *Kareem Darwish, Hany Hassan, Ahmed Hassan, Ossama Emam
Human Language Technologies Group
IBM Cairo Technology Development Center
P.O.Box 166 El-Ahram, Giza, Egypt
{amadkour,hanyh,hasanah,emam}@eg.ibm.com, ⋆kareem@darwish.org

Abstract

The vast number of published medical documents is considered a vital source for relationship discovery. This paper presents a statistical unsupervised system, called BioNoculars, for extracting protein-protein interactions from biomedical text. BioNoculars uses graph-based mutual reinforcement to make use of redundancy in data to construct extraction patterns in a domain independent fashion. The system was tested using MEDLINE abstract for which the protein-protein interactions that they contain are listed in the database of interacting proteins and protein-protein interactions (DIPPPI). The system reports an F-Measure of 0.55 on test MEDLINE abstracts.

1 Introduction

With the ever-increasing number of published biomedical research articles and the dependency of new research and previously published research, medical researchers and practitioners are faced with the daunting prospect of reading through hundreds or possibly thousands of research articles to survey advances in areas of interest. Much work has been done to ease access and discovery of articles that match the interest of researchers via the use of search engines such as PubMed, which provides search capabilities over MEDLINE, a collection of more than 15 million journal paper abstracts maintained by the National Library of Medicine (NLM). However, with the addition of abstracts from more than 5,000 medical journals to MEDLINE every year, the number of articles containing information that is pertinent to users needs has grown considerably. These 5,000 journals constitute only a subset of the published biomedical research. Further, medical articles often contain redundant information and only subsections of articles are typically of direct interest to researchers. More advanced information extraction tools have been developed to effectively distill medical articles to produce key pieces of information from articles while attempting to eliminate redundancy. These tools have focused on areas such as protein-protein interaction, gene-disease relationship, and chemical-protein interaction (Chun et al., 2006). Many of these tools have been used to extract key pieces of information from MEDLINE. Most of the reported information extraction approaches use sets of handcrafted rules in conjunction with manually curated dictionaries and ontologies.

This paper presents a fully unsupervised statistical technique to discover protein-protein interaction based on automatically discoverable repeating patterns in text that describe relationships. The paper is organized as follows: section 2 surveys related work; section 3 describes BioNoculars; Section 4 describes the employed experimental setup; section 5 reports and comments on experimental results; and section 6 concludes the paper.

2 Background

The background will focus primarily on the tagging of Biomedical Named Entities (BNE), such genes, gene-products, proteins, and chemicals and the Ex-

BioNLP 2007: Biological, translational, and clinical language processing, pages 89–96,
Prague, June 2007.

traction of protein-protein interactions from text.

2.1 BNE Tagging

Concerning BNE tagging, the most common approaches are based on hand-crafted rules, statistical classifiers, or a hybrid of both (usually in conjunction with dictionaries of BNE). Rule-based systems (Fukuda et al., 1998; Hanisch et al., 2003; Yamamoto et al., 2003) that use dictionaries tend to exhibit high precision in tagging named entities but generally with lower tagging recall. They tend to lag the latest published research and are sensitive to the expression of the named entities. Dictionaries of BNE are typically laborious and expensive to build, and they are dependant on nomenclatures and specific species. Statistical approaches (Collier et al., 2000; Kazama et al., 2002; Settles, 2004) typically improve recall at the expense of precision, but are more readily retargetable for new nomenclatures and organisms. Hybrid systems (Tanabe and Wilbur, 2002; Mika and Rost, 2004) attempt to take advantage of both approaches. Although these approaches tend to generate acceptable recognition, they are heavily dependent on the type of data on which they are trained.

(Fukuda et al., 1998) proposed a rule-based protein name extraction system called PROPER (PROtein Proper-noun phrase Extracting Rules) system, which utilizes a set of rules based on the surface form of text in conjunction with a Part-Of-Speech (POS) tagging to identify what looks like a protein without referring to any specific BNE dictionary. They reported a 94.7% precision and a 98.84% recall for the identification of BNEs. The results that they achieved seem to be too specific to their training and test sets.

(Hanisch et al., 2003) proposed a rule-based protein and gene name extraction system called ProMiner, which is based on the construction of a general-purpose dictionary along with different dictionaries of synonyms and an automatic curation procedure based on a simple token model of protein names. Results showed that their system achieved a 0.80 F-measure score in the name extraction task on the BioCreative test set (BioCreative).

(Yamamoto et al., 2003) proposed the use of morphological analysis to improve protein name tagging. Their approach tags proteins based on morpheme chunking to properly determine protein name boundary. They used the GENIA corpus for training and testing and obtained an F-measure score of 0.70 for protein name tagging.

(Collier et al., 2000) used a machine learning approach to protein name extraction based on a linear interpolation Hidden Markov Model (HMM) trained using bi-grams. They focused on finding the most likely protein sequence classes (C) for a given sequence of words (W), by maximizing the probability of C given W, P(C—W). Unlike traditional dictionary based methods, the approach uses no manually crafted patterns. However, their approach may misidentify term boundaries for phrases containing potentially ambiguous local structures such as coordination and parenthesis. They reported an F-measure score of 0.73 for different mixtures of models tested on 20 abstracts.

(Kazama et al., 2002) proposed a machine learning approach to BNE tagging based on support vector machines (SVM), which was trained on the GENIA corpus. Their preliminary results of the system showed that the SVM with the polynomial kernel function outperforms techniques of Maximum Entropy based systems.

Yet another BNE tagging system is ABNER (Settles, 2005), which utilizes machine learning, namely conditional random fields, with a variation of orthographic and contextual features and no semantic or syntactic features. ABNER achieves an F-measure score of 0.71 on the NLPA 2004 shared task dataset corpus and 0.70 on the BioCreative corpus.and scored an F1-measure of 51.8set.

(Tanabe and Wilbur, 2002) used a combination of statistical and knowledge-based strategies, which utilized automatically generated rules from transformation based POS tagging and other generated rules from morphological clues, low frequency trigrams, and indicator terms. A key step in their method is the extraction of multi-word gene and protein names that are dominant in the corpus but inaccessible to the POS tagger. The advantage of such an approach is that it is independent of any biomedical domain. However, it can miss single word gene names that do not occur in contextual gene theme terms. It can also incorrectly tag compound gene names, plasmids, and phages.

(Mika and Rost, 2004) developed NLProt, which combines the use of dictionaries, rules-based filtering, and machine learning based on an SVM classifier to tag protein names in MEDLINE. The NLProt system used rules for pre-filtering and the SVM for classification, and it achieved a precision of 75% and recall 76%.

2.2 Relationship Extraction

As for the extraction of interactions, most efforts in extraction of biomedical interactions between entities from text have focused on using rule-based approaches due to the familiarity of medical terms that tend to describe interactions. These approaches have proven to be successful with notably good results. In these approaches, most researchers attempted to define an accurate set of rules to describe relationship types and patterns and to build ontologies and dictionaries to be consulted in the extraction process. These rules, ontologies, and dictionaries are typically domain specific and are often not generalizable to other problems.

(Blaschke et al., 1999) reported a domain specific approach for extracting protein-protein interactions from biomedical text based on a set of predefined patterns and words describing interactions. Later work attempted to automatically extract interactions, which are referenced in the database of interacting proteins (Xenarios et al., 2000), from the text mentioning the interactions (Blaschke and Valencia, 2001). They achieved surprisingly low recall (25%), which they attributed to problems in properly identifying protein names in the text.

(Koike et al., 2005) developed a system called PRIME, which was used to extract biological functions of genes, proteins, and their families. Their system used a shallow parser and sentence structure analyzer. They extracted so-called ACTOR-OBJECT relationships from the shallow parsed sentences using rule based sentence structure analysis. The identification of BNEs was done by consulting the GENA gene name dictionary and family name dictionary. In extracting the biological functions of genes and proteins, their system reported a recall of 64% and a precision of 94%.

Saric et al. developed a system to extract gene expression regulatory information in yeast as well as other regulatory mechanisms such phosphorylation (Saric et al., 2004; Saric et al., 2006). They used a rule based named entity recognition module, which recognizes named entities via cascading finite state automata. They reported a precision of 83-90% and 86-95% for the extraction of gene expression and phosphorylation regulatory information respectively.

(Leroy and Chen, 2005) used linguistic parsers and Concept Spaces, which use a generic co-occurrence based technique that extracts relevant medical phrases using a noun chunker. Their system employed UMLS (Humphreys and Lindberg, 1993), GO (Ashburner et al., 2000), and GENA (Koike and Takagi, 2004) to further improve extraction. Their main purpose was entity identification and cross reference to other databases to obtain more knowledge about entities involved in the system.

Other extraction approaches such as the one reported on by (Cooper and Kershenbaum, 2005) utilized a large manually curated dictionary of many possible combinations of gene/protein names and aliases from different databases and ontologies. They annotated their corpus using a dictionary-based longest matching technique. In addition, they used filtering with a maximum entropy based named entity recognizer in order to remove the false positives that were generated from merging databases. The problem with this approach is the resulting inconsistencies from merging databases, which could hurt the effectiveness of the system. They reported a recall of 87.1 % and a precision of 78.5% in the relationship extraction task.

Work by (Mack et al., 2004) used the Munich Information Center for Protein Sequences (MIPS) for entity identification. Their system was integrated in the IBM Unstructured Information Management Architecture (UIMA) framework (Ferrucci and Lally, 2004) for tokenization, identification of entities, and extraction of relations. Their approach was based on a combination of computational linguistics, statistics, and domain specific rules to detect protein interactions. They reported a recall of 61% and a precision of 97%.

(Hao et al., 2005) developed an unsupervised approach, which also uses patterns that were deduced using minimum description lengths. They used pattern optimization techniques to enhance the patterns by introducing most common keywords that tend to describe interactions.

(Jörg et. al., 2005) developed Ali Baba which uses sequence alignments applied to sentences annotated with interactions and part of speech tags.It also uses finite state automata optimized with a genetic algorithm in its approach. It then matches the generated patterns against arbitrary text to extract interactions and their respective partners. The system scored an F1-measure of 51.8% on the LLL'05 evaluation set.

The aforementioned systems used either rule-based approaches, which require manual intervention from domain experts, or statistical approaches, either supervised or semi-supervised, which also require manually curated training data.

3 BioNoculars

BioNoculars is a relationship extraction system that based on a fully unsupervised technique suggested by (Hassan et al., 2006) to automatically extract protein-protein interaction from medical articles. It can be retargeted to different domains such as protein interactions in diseases. The only requirement is to compile domain specific taggers and dictionaries, which would aid the system in performing the required task.

The approach uses an unsupervised graph-based mutual reinforcement, which depends on the construction of generalized extraction patterns that could match instances of relationships (Hassan et al., 2006). Graph-based mutual reinforcement is similar to the idea of hubs and authorities in web pages depicted by the HITS algorithm (Kleinberg, 1998). The basic idea behind the algorithm is that the importance of a page increases when more and more good pages link to it. The duality between patterns and extracted information (tuples) leads to the fact that patterns could express different tuples, and tuples in turn could be expressed by different patterns. Tuple in this context contains three elements, namely two proteins and the type of interaction between them. The proposed approach is composed of two main steps, namely initial pattern construction and then pattern induction.

For pattern construction, the text is POS tagged and BNE tagged. The tags of Noun Phrases or sequences of nouns that constitute a BNE are removed and replaced with a BNE tag. Then, an n-gram language model is built on the tagged text (using tags only) and is used to construct weighted finite state machines. Paths with low cost (high language model probabilities) are chosen to construct the initial set of patterns; the intuition is that paths with low cost (high probability) are frequent and could represent potential candidate patterns. The number of candidate initial patterns could be reduced significantly by specifying the candidate types of entities of interest. In the case of BioNoculars, the focus was on relationships between BNEs of type PROTEIN. The candidate patterns are then applied to the tagged stream to produce in-sentence relationship tuples.

As for pattern induction, due to the duality in the patterns and tuples relation, patterns and tuples are represented by a bipartite graph as illustrated in Figure 1.

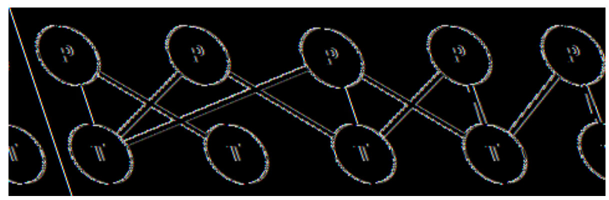

Figure 1: A bipartite graph representing patterns and tuples

Each pattern or tuple is represented by a node in the graph. Edges represent matching between patterns and tuples. The pattern induction problem can be formulated as follows: Given a very large set of data D containing a large set of patterns P, which match a large set of tuples T, the problem is to identify , which is the set of patterns that match the set of the most correct tuples *T*. The intuition is that the tuples matched by many different patterns tend to be correct and the patterns matching many different tuples tend to be good patterns. In other words, BioNoculars attempts to choose from the large space of patterns in the data the most informative, highest confidence patterns that could identify correct tuples; i.e. choosing the most authoritative patterns in analogy with the hub-authority problem. The most authoritative patterns can then be used for extracting relations from free text. The following pattern-tuple pairs show how patterns can match tuples in the corpus:

(protein) (verb) (noun) (prep.) (protein)

Cla4 induces phosphorylation of Cdc24

(protein) (I-protein) (Verb) (prep.) (protein)

NS5A interacts with Cdk1

The proposed approach represents an unsupervised technique for information extraction in general and particularly for relations extraction that requires no seed patterns or examples and achieves significant performance. Given enough domain text, the extracted patterns can support many types of sentences with different styles (such passive and active voice) and orderings (the interaction of X and Y vs. X interacts with Y).

One of the critical prerequisites of the above-mentioned approach is the use of a POS tagger, which is tuned for biomedical text, and a BNE tagger to properly identify BNEs. Both are critical for determining the types of relationships that are of interest. For POS tagging, a decision tree based tagger developed by (Schmid, 1994) was used in combination with a model, which was trained on a corrected/revised GENIA corpus provided by (Saric et al., 2004) and was reported to achieve 96.4% tagging accuracy (Saric et al., 2006). This POS tagger will be referred to as the Schmid tagger. For BNE tagging, ABNER was used. The accuracy of ABNER is approximately state of the art with precision and recall of 74.5% and 65.9% respectively with training done using the BioCreative corpora (BioCreative). Nonetheless we still face entity identification problems such as missed identifications in the text which in turn affects our results considerably. We do believe if we use a better identification method , we would yield better results.

4 Experimental Setup

Experiments aimed at extracting protein-protein interactions for Bakers yeast (Sacharomyces Cerevesiae) to assess BioNoculars (Cherry et al., 1998). The experiments were performed using 109,440 MEDLINE abstracts that contained the varying names of the yeast, namely Sacharomyces cerevisiae, S. Cerevisiae, Bakers yeast, Brewers yeast and Budding yeast. MEDLINE abstracts typically summarize the important aspects of papers possibly including protein-protein interactions if they are of relevance to the article. The goal was to deduce the most appropriate extraction patterns that can be later used to extract relations from any document. All the MEDLINE abstracts were used for pattern extraction except for 70 that were set aside for testing. There were no test documents in the training set. To build ground-truth, the test set was semi-manually POS and BNE tagged. They were also annotated with the interactions that are contained in the text. There was a condition that all the abstracts that are used for testing must have entries in the Database of Interacting Proteins and Protein-Protein Interactions (DIPPPI), which is a subset of the Database of Interacting Proteins (DIP) (Xenarios et al., 2000) restricted to proteins from yeast. DIPPPI lists the known protein-protein interactions in the MEDLINE abstracts. There were 297 protein-protein interactions in the test set of 70 abstracts. One of the disadvantages of DIPPPI is that the presence of interactions is indicated without mentioning their types or from which sentences they were extracted. Although BioNoculars is able to guess the sentence from which an interaction was extracted and the type of interaction, this information was ignored when evaluating against DIPPPI. Unfortunately, there is no standard test set for the proposed task, and most of the evaluation sets are proprietary. The authors hope that others can benefit from their test set, which is freely available.

The abstracts used for pattern extraction were POS tagged using the Schmid tagger and BNE tagging was done using ABNER. The patterns were restricted to only those with protein names. For extraction of interaction tuples, the test set was POS and BNE tagged using the Schmid tagger and ABNER respectively. A varying number of final patterns were then used to extract tuples from the test set and the average recall and precision were computed. Another setup was used in which the relationships were filtered using preset keywords for relationships such as inhibits, interacts, and activates to properly compare BioNoculars to systems in the literature that use such keywords. The keywords were obtained from the (Hakenberg et al., 2005) and (Temkin and Gilder, 2003). One of the generated pattern-tuple pairs was as follows:

(PROTEIN) (Verb) (Conjunction) (PROTEIN)

NS5A interacts with Cdk1

One consequence of tuple extraction is generation of redundant tuples, which contain the same enti-

Pattern Count	30	59	78	103	147	192	205	217
Recall	0.51	0.70	0.76	0.81	0.84	0.89	0.89	0.93
Precision	0.47	0.42	0.43	0.35	0.30	0.26	0.26	0.16
FMeasure	0.49	0.53	0.55	0.49	0.44	0.40	0.40	0.27

Table 1: Recall, Precision, and F-measure for extraction of tuples using a varying number of top rated patterns

Pattern Count	30	59	78	103	147	192	205	217
Recall	0.31	0.44	0.46	0.48	0.64	0.73	0.74	0.78
Precision	0.31	0.36	0.35	0.34	0.39	0.35	0.35	0.37
FMeasure	0.31	0.40	0.40	0.40	0.48	0.47	0.48	0.50

Table 2: Recall, Precision, and Recall for extraction of tuples using a varying number of top rated patters keyword filtering

ties and relations. Consequently, all protein aliases and full text names were resolved to a unified naming scheme and the unified scheme was used to replace all variations of protein names in patterns. All potential protein-protein interactions that BioNoculars extracted were compared to those in the DIPPPI databases.

5 Results and Discussion

For the first set of experiments, the experimental setup described above was used without modification. Table 1 and Figure 2 report on the resulting recall and precision when taking different number of highest rated patterns. The highest rated 217 patterns were divided on a linear scale into 8 clusters based on their relative weights.

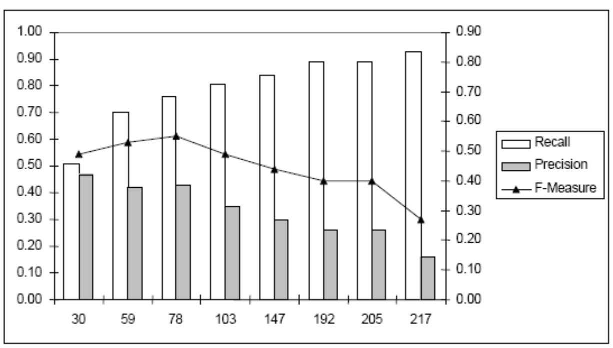

Figure 2: Recall, Precision, and F-measure for tuple extraction using a varying number of top patterns

As expected, Figure 2 clearly shows an inverse relationship between precision and recall. This is because using more extraction patterns yields more tuples thus increasing recall at the expense of precision. The F-measure (with $\beta = 1$) peeks at 78 patterns, which seems to provide the best score given that precision and recall are equally important. However, the technique seems to favor recall, reaching a recall of 93% when using all 217 patterns. The low precision levels warrant thorough investigation.

In the second set of experiments, extracted tuples were filtered using preset keywords indicating interactions. Table 2 and Figure 3 show the results of the experiments.

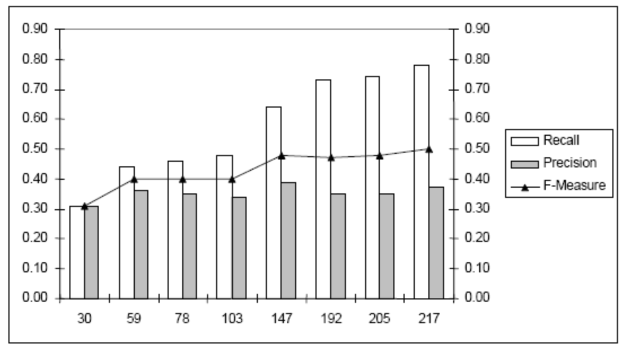

Figure 3: Recall, Precision, and F-measure for tuple extraction using a varying number of top patterns with keyword filtering

The results show that filtering with keywords led to lower recall, but precision remained fairly steady as the number of patterns changed. Nonetheless, the best precision in Figure 3 is lower than the best precision in Figure 2 and the maximum F-measure for this set of experiments is lower than the maximum F-measure when no filtering was used. The BioNoculars system with no filtering can be advantageous for recall oriented applications. The use of no filtering suggests that some interaction may be expressed in more generic forms or patterns. An intermediate solution would be to increase the size of the list of most commonly occurring keywords to filter the extracted tuples further.

Currently, ABNER, which is used by the system, has a precision of 75.4% and a recall of 65.9%. Perhaps improved tagging may improve the extraction effectiveness.

The effectiveness of BioNoculars needs to be thoroughly compared to existing systems via the use of standard test sets, which are not readily available. Most of previously reported work has been tested on proprietary test sets or sets that are not publicly available. The creation of standard publicly available test set can prompt research in this area.

6 Conclusion and Future Work

This paper presented a system for extracting protein-protein interaction from biomedical text call BioNoculars. BioNoculars uses a statistical unsupervised learning algorithm, which is based on graph mutual reinforcement and data redundancy to extract extraction patterns. The system is recall oriented and is able to properly extract 93% of the interaction mentions from test MEDLINE abstracts. Nonetheless, the systems precision remains low. Precision can be enhanced by using keywords that describe interactions to filter to the resulting interaction, but this would be at the expense of recall.

As for future work, more attention should be focused on improving extraction patterns. Currently, the system focuses on extracting interactions between exactly two proteins. Some of the issues that need to be handled include complex relationship (X and Y interact with A and B), linguistic variability (passive vs. active voice; presence of superfluous words such as modifiers, adjectives, and prepositional phrases), protein lists (W interacts with X, Y, and Z), nested interactions (W, which interacts with X, also interacts with Y). Resolving these issues would require an investigation of how patterns can be generalized in automatic or semi-automatic ways. Further, the identification of proteins in the text requires greater attention. Also, the BioNoculars approach can be combined with other rule-based approaches to produce better results.

References

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. 2000. *Gene ontology: tool for the unification of biology.* Nature Genetics,volume 25 pp.25-29.

BioCreative. 2004. [Online].

Blaschke C., M. A. Andrade, C. Ouzounis, and A. Valencia. 1999. *Automatic Extraction of Biological Information from Scientific Text: Protein-Protein Interactions.* ISMB99, pp. 60-67.

Blaschke, C. and A. Valencia. 2001. *Can Bibliographic Pointers for Known Biological Protein Interactions as a Case Study.* Comparative and Functional Genomics,vol. 2: 196-206.

Cherry, J. M., C. Adler, C. Ball, S. A. Chervitz, S. S. Dwight, E. T. Hester, Y. Jia, G. Juvik, T. Roe, M. Schroeder, S. Weng, and D. Botstein. 1998. *SGD: Saccharomyces Genome Database.* Nucleic Acids Research, 26, 73-9.

Chun, H. W., Y. Tsuruka, J. D. Kim, R. Shiba, N. Nagata, T. Hishiki, and J. Tsujii. 2006. *Extraction of Gene-Disease Relations from MEDLINE Using Domain Dictionaries and Machine Learning.* Pacific Symposium on Biocomputing 11:4-15.

Collier, N., C. Nobata, and J. Tsujii. 2000. *Extracting the Names of Genes and Gene Products with a Hidden Markov Model.* COLING, 2000, pp. 201207.

Cooper, J. and A. Kershenbaum. 2005. *Discovery of protein-protein interactions using a combination of linguistic, statistical and graphical information.* BMC Bioinformatics.

DIPPPI http://www2.informatik.hu-berlin.de/ hakenber/corpora. 2006.

Ferrucci, D. and A. Lally. 2004. *UIMA: an architectural approach to unstructured information processing in the corporate research environment.* Natural Language Engineering 10, No. 3-4, 327-348.

Fukuda, K., T. Tsunoda, A. Tamura, and T. Takagi. 1998. *Toward information extraction: identifying protein names from biological papers.* PSB, pages 705716.

Hakenberg, J., C. Plake, U. Leser, H. Kirsch, and D. Rebholz-Schuhmann. 2005. *LLL'05 Challenge: Genic Interaction Extraction with Alignments and Finite State Automata.* Proc Learning Language in Logic Workshop (LLL'05) at ICML 2005, pp. 38-45. Bonn, Germany.

Hanisch, D., J. Fluck, HT. Mevissen, and R. Zimmer. 2003. *Playing biologys name game: identifying protein names in scientific text.* PSB, pages 403414.

Hao, Y., X. Zhu, M. Huang, and M. Li. 2005. *Discovering patterns to extract protein-protein interactions from the literature: Part II.* Bioinformatics, Vol. 00 no. 0 2005 pages 1-7.

Hassan, H., A. Hassan, and O. Emam. 2006. *Unsupervised Information Extraction Approach Using Graph Mutual Reinforcement.* Proceedings of Empirical Methods for Natural Language Processing (EMNLP).

Humphreys B. L. and D. A. B. Lindberg. 1993. *The UMLS project: making the conceptual connection between users and the information they need.* Bulletin of the Medical Library Association, 1993; 81(2): 170.

Jörg Hakenberg, Conrad Plake, Ulf Leser. 2005. *Genic Interaction Extraction with Alignments and Finite State Automata.* Proc Learning Language in Logic Workshop (LLL'05) at ICML 2005, pp. 38-45. Bonn, Germany (August 2005)

Kazama, J., T. Makino, Y. Ohta, and J. Tsujii. 2002. *Tuning Support Vector Machines for Biomedical Named Entity Recognition.* ACL Workshop on NLP in Biomedical Domain, pages 18.

Kleinberg, J. 1998. *Authoritative sources in a hyperlinked environment.* In Proc. Ninth Ann. ACM-SIAM Symp. Discrete Algorithms, pages 668-677, ACM Press, New York.

Koike A. and T. Takagi. 2004. *Gene/protein/family name recognition in biomedical literature.* BioLINK 2004: Linking Biological Literature, Ontologies, and Database, pp. 9-16.

Koike, A., Y. Niwa, and T. Takagi 2005. *Automatic extraction of gene/protein biological functions from biomedical text.* Bioinformatics, Vol. 21, No. 7.

Leroy, G. and H. Chen. 2005. *Genescene: An Ontology-enhanced Integration of Linguistic and Co-Occurance based Relations in Biomedical Text.* JASIST Special Issue on Bioinformatics.

Mack, R. L., S. Mukherjea, A. Soffer, N. Uramoto, E. W. Brown, A. Coden, J. W. Cooper, A. Inokuchi, B. Iyer, Y. Mass, H. Matsuzawa, L. V. Subramaniam. 2004. *Text analytics for life science using the Unstructured Information Management Architecture.* IBM Systems Journal 43(3): 490-515.

Mika, S. and B. Rost. 2004. *NLProt: extracting protein names and sequences from papers.* Nucleic Acids Research, 32 (Web Server issue): W634W637.

Saric, J., L. J. Jensen, R. Ouzounova, I. Rojas, and P. Bork. 2004. *Extracting regulatory gene expression networks from PUBMED.* Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, Barcelona, Spain, pp.191-198.

Saric, J., L. J. Jensen, R. Ouzounova, I. Rojas, and P. Bork. 2006. *Extraction of regulatory gene/protein networks from Medline.* Bioinformatics Vol.22 no 6,pp. 645-650.

Schmid, H. 1994. *Probabilistic Part-of-Speech Tagging Using Decision Trees.* In the International Conference on New Methods in Language Processing, Manchester, UK.

Settles, B. 2004. *Biomedical Named Entity Recognition Using Conditional Random Fields and Rich Feature Sets.* In Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA), Geneva, Switzerland, pages 104-107.

Settles, B. 2005. *ABNER: an open source tool for automatically tagging genes, proteins, and other entity names in text.* Bioinformatics, 21(14): 3191-3192.

Tanabe L., and W. J. Wilbur. 2002. *Tagging gene and protein names in biomedical text.* Bioinformatics, 18(8):11241132.

Temkin, J. M. and M. R. Gilder. 2003. *Extraction of protein interaction information from unstructured text using a context-free grammar.* Bioinformatics 19(16):2046-2053.

Xenarios I, Rice DW, Salwinski L, Baron MK, Marcotte EM, Eisenberg D. 2000. *DIP: the Database of Interacting Proteins.* Nucleic Acids Res 28: 289291.

Yamamoto, K., T. Kudo, A. Konagaya, Y. Matsumoto. 2003. *Protein Name Tagging for Biomedical Annotation in Text.* Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine, pp. 65-72.