Extracting Molecular Binding Relationships from Biomedical Text

Thomas C. RINDFLESCH
National Library of Medicine
8600 Rockville Pike
Bethesda, MD, 20894
tcr@nlm.nih.gov

Jayant V. RAJAN
University of Rochester
Rochester, NY, 14620
Jayant.Rajan@mc.rochester.edu

Lawrence HUNTER
National Cancer Institute
7550 Wisconsin Avenue
Bethesda, MD, 20894
lhunter@nih.gov

Abstract

ARBITER is a Prolog program that extracts assertions about macromolecular binding relationships from biomedical text. We describe the domain knowledge and the underspecified linguistic analyses that support the identification of these predications. After discussing a formal evaluation of ARBITER, we report on its application to 491,000 MEDLINE® abstracts, during which almost 25,000 binding relationships suitable for entry into a database of macromolecular function were extracted.

Introduction

Far more scientific information exists in the literature than in any structured database. Convenient access to this information could significantly benefit research activities in various fields. The emerging technology of information extraction (Appelt and Israel 1997, Hearst 1999) provides a means of gaining access to this information. In this paper we report on a project to extract biomolecular data from biomedical text. We concentrate on molecular binding affinity, which provides a strong indication of macromolecular function and is a core phenomenon in molecular biology. Our ultimate goal is to automatically construct a database of binding relationships asserted in MEDLINE citations.

The National Library of Medicine's MEDLINE textual database is an online repository of more than 10 million citations from the biomedical literature. All citations contain the title of the corresponding article along with other bibliographic information. In addition, a large number of citations contain author-supplied abstracts. Initial studies indicate that there are approximately 500,000 MEDLINE citations relevant to molecular binding affinity.

Our decision to apply information extraction technology to binding relationships was guided not only by the biological importance of this phenomenon but also by the relatively straightforward syntactic cuing of binding predications in text. The inflectional forms of a single verb, *bind,* indicate this relationship in the vast majority of cases, and our initial work is limited to these instances. For example, our goal in this project is to extract the binding predications in (2) from the text in (1).

(1) CC chemokine receptor 1 (CCR1) is expressed in neutrophils, monocytes, lymphocytes, and eosinophils, and binds the leukocyte chemoattractant and hematopoiesis regulator macrophage inflammatory protein (MIP)- 1alpha, as well as several related CC chemokines.

(2) <CC chemokine receptor 1>
BINDS
<leukocyte chemoattractant>
<CC chemokine receptor 1>
BINDS
<hematopoiesis regulator macrophage inflammatory protein-1alpha>
<CC chemokine receptor 1>
BINDS
<related CC chemokine>

Considerable interest in information extraction has concentrated on identifying named entities in text pertaining to current events (for example, Wacholder et al. 1997, Voorhees and Harman 1998, and MUC-7); however, several recent efforts have been directed at biomolecular data (Blaschke et al. 1999, Craven and Kumlien 1999, and Rindflesch et al. 2000, for example). The overall goal is to transform the information

encoded in text into a more readily accessible format, typically a template with slots named for the participants in the scenario of interest. The template for molecular binding can be thought of as a simple predication with predicate "bind" and two arguments which participate (symmetrically) in the relationship: BINDS(<X>, <Y>).

Various strategies, both linguistic and statistical, have been used in information extraction efforts. We introduce a Prolog program called ARBITER (Assess and Retrieve Binding Terminology) that takes advantage of an existing domain knowledge source and relies on syntactic cues provided by a partial parser in order to identify and extract binding relations from text. We discuss the syntactic processing used and then report on a formal evaluation of ARBITER against a test collection of 116 MEDLINE citations in which the binding relations were marked by hand. Finally, we provide a brief overview of the results of applying ARBITER to the 500,000 MEDLINE citations discussing molecular binding affinity.

1 Extracting Binding Relationships from Text

Our strategy for extracting binding relationships from text divides the task into two phases: During the first phase we identify all potential binding arguments, and then in the second phase we extract just those binding terms which are asserted in the text as participating in a particular binding predication. In support of this processing, we rely on the linguistic and domain knowledge contained in the National Library of Medicine's Unified Medical Language System® (UMLS®) as well an existing tool, the SPECIALIST minimal commitment parser (Aronson et al. 1994).

The UMLS (Humphreys et al. 1998) consists of several knowledge sources applicable in the biomedical domain: the Metathesaurus, Semantic Network, and SPECIALIST Lexicon (McCray et al. 1994). The Metathesaurus was constructed from more than forty controlled vocabularies and contains more than 620,000 biomedical concepts. The characteristic of the Metathesaurus most relevant for this project is that each concept is associated with a semantic type that categorizes the concept into subareas of biology or medicine. Examples pertinent to binding terminology include the semantic types 'Amino Acid, Peptide, or Protein' and 'Nucleotide Sequence'. The SPECIALIST Lexicon (with associated lexical access tools) supplies syntactic information for a large compilation of biomedical and general English terms.

The SPECIALIST minimal commitment parser relies on the SPECIALIST Lexicon as well as the Xerox stochastic tagger (Cutting et al. 1992). The output produced is in the tradition of partial parsing (Hindle 1983, McDonald 1992, Weischedel et al. 1993) and concentrates on the simple noun phrase, what Weischedel et al. (1993) call the "core noun phrase," that is a noun phrase with no modification to the right of the head. Several approaches provide similar output based on statistics (Church 1988, Zhai 1997, for example), a finite-state machine (Ait-Mokhtar and Chanod 1997), or a hybrid approach combining statistics and linguistic rules (Voutilainen and Padro 1997).

The SPECIALIST parser is based on the notion of barrier words (Tersmette et al. 1988), which indicate boundaries between phrases. After lexical look-up and resolution of category label ambiguity by the Xerox tagger, complementizers, conjunctions, modals, prepositions, and verbs are marked as boundaries. Subsequently, boundaries are considered to open a new phrase (and close the preceding phrase). Any phrase containing a noun is considered to be a (simple) noun phrase, and in such a phrase, the right-most noun is labeled as the head, and all other items (other than determiners) are labeled as modifiers. An example of the output from the SPECIALIST parser is given below in (4). The partial parse produced serves as the basis for the first phase of extraction of binding relationships, namely the identification of those simple noun phrases acting as potential binding arguments (referred to as "binding terms").

1.1 Identifying binding terminology

In order to identify binding terminology in text we rely on the approach discussed in (Rindflesch et al. 1999). Text with locally-defined acronyms expanded is submitted to the Xerox tagger and the SPECIALIST parser. Subsequent processing concentrates on the heads of simple noun

phrases and proceeds in a series of cascaded steps that depend on existing domain knowledge as well as several small, special-purpose resources in order to determine whether each noun phrase encountered is to be considered a binding term.

As the first step in the process, an existing program, MetaMap, (Aronson et al. 1994) attempts to map each simple noun phrase to a concept in the UMLS Metathesaurus. The semantic type for concepts corresponding to successfully mapped noun phrases is then checked against a small subset of UMLS semantic types referring to bindable entities, such as 'Amino Acid, Peptide, or Protein', 'Nucleotide Sequence', 'Carbohydrate', 'Cell', and 'Virus'. For concepts with a semantic type in this set, the corresponding noun phrase is considered to be a binding term.

The heads of noun phrases that do not map to a concept in the Metathesaurus are tested against a small set of general "binding words," which often indicate that the noun phrase in which they appear is a binding term. The set of binding words includes such nouns as *cleft, groove, membrane, ligand, motif, receptor, domain, element,* and *molecule.*

The head of a noun phrase that did not submit to the preceding steps is examined to see whether it adheres to the morphologic shape of a normal English word. In this context such a word is often an acronym not defined locally and indicates the presence of a binding term (Fukuda et al. 1998). A normal English word has at least one vowel and no digits, and a text token that contains at least one letter and is not a normal English word functions as a binding word in this context.

The final step in identifying binding terms is to join contiguous simple noun phrases qualifying as binding terms into a single macro-noun phrase. Rindflesch et al. (1999) use the term "macro-noun phrase" to refer to structures that include reduced relative clauses (commonly introduced by prepositions or participles) as well as appositives. Two binding terms joined by a form of *be* are also treated as though they formed a macro-noun phrase, as in *Jel42 is an IgG which binds ...*

The results of identifying binding terms (and thus potential binding arguments) are given in (4) for the sentence in (3). In (4) evidence supporting identification as a binding term is given in braces. Note that in the underspecified syntactic analysis, prepositional phrases are treated as (simple) noun phrases that have a preposition as their first member.

(3) Jel42 is an IgG which binds to the small bacterial protein, HPr and the structure of the complex is known at high resolution.

(4) **[binding_term**([head(Jel42)],
{Morphology Shape Rule}
[aux(is)] ,
[det(an), head(IgG)] {Metathesaurus}
),
[pron(which)],
[verb(binds)],
binding_term([prep(to), det(the),
mod(small), mod(bacterial), head(protein),
punc(,)], {Metathesaurus}
[head(HPr)] {Morphology Shape Rule}
),
[conj(and)],
[det(the), head(structure)],
binding_term([prep(of), det(the),
head(complex)] {General Binding Word}
),
[aux(is)],
[verb(known)],
[prep(at), mod(high), head(resolution),
punc(.)]]

1.2 Identifying binding terms as arguments of relationships

Before addressing the strategy for determining the arguments of binding predications, we discuss the general treatment of macro-noun phrases during the second part of the processing. Although ARBITER attempts to recover complete macro-noun phrases during the first phase, only the most specific (and biologically useful) part of a macro-noun phrase is recovered during the extraction of binding predications. Terms referring to specific molecules are more useful than those referring to general classes of bindable entities, such as *receptor, ligand, protein,* or *molecule.* The syntactic head of a macro-noun phrase (the first simple noun phrase in the list) is not always the most specific or most useful term in the construction.

The Specificity Rule for determining the most specific part of the list of simple binding terms constituting a macro-noun phrase chooses the first simple term in the list which has either of the following two characteristics: a) The head was identified by the Morphology Shape Rule. b) The noun phrase maps to a UMLS concept having one of the following semantic types: 'Amino Acid, Peptide, or Protein', 'Nucleic Acid, Nucleoside, or Nucleotide', 'Nucleotide Sequence', 'Immunologic Factor', or 'Gene or Genome'. For example, in (5), the second simple term, *TNF-alpha promoter*, maps to the Metathesaurus with semantic type 'Nucleotide Sequence' and is thus considered to be the most specific term in this complex-noun phrase.

(5) binding_term(
[transcriptionally active kappaB motifs],
[in the TNF-alpha promoter],
[in normal cells])

In identifying binding terms as arguments of a complete binding predication, as indicated above, we examine only those binding relations cued by some form of the verb *bind* (*bind, binds, bound,* and *binding*). The list of minimal syntactic phrases constituting the partial parse of the input sentence is examined from left to right; for each occurrence of a form of *binds,* the two binding terms serving as arguments are then sought. (During the tagging process, we force *bind, binds,* and *bound* to be labeled as "verb," and *binding* as "noun.")

A partial analysis of negation and coordination is undertaken by ARBITER, but anaphora resolution and a syntactic treatment of relativization are not attempted. With the added constraint that a binding argument must have been identified as a binding term based on the domain knowledge resources used, the partial syntactic analysis available to ARBITER supports the accurate identification of a large number of binding predications asserted in the research literature.

1.2.1 Arguments of binding

It is convenient to categorize binding predications into two classes depending on which form of *bind* cues the predication: a) *binding* and b) *bind, binds,* and *bound.* In our test collection (discussed below), about half of the binding relationships asserted in the text are cued by the gerundive or participial form *binding.* In this syntactic predication, the resources available from the underspecified syntactic parse serve quite well as the basis for correctly identifying the arguments of the binding relationship.

The most common argument configuration associated with *binding* is for both arguments to occur to the right, cued by prepositions, most commonly *of* and *to*; however, other frequent patterns are *of-by* and *to-by.* Another method of argument cuing for *binding* is for the subject of the predication to function syntactically as a modifier of the head *binding* in the same simple noun phrase. The object in this instance is then cued by either *of* or *to* (to the right). A few other patterns are seen and some occurrences of *binding* do not cue a complete predication; either the subject is missing or neither argument is explicitly mentioned. However, the examples in (6) fairly represent the interpretation of *binding.*

(6) These results suggest that 2 amino acids, Thr-340 and Ser-343, play important but distinct roles in promoting the **binding** of arrestin to rhodopsin.

<arrestin>
BINDS
<rhodopsin>

Surprisingly, arrestin **binding** to phosphorylated T340E did not increase to the level observed for wild-type rhodopsin.

<arrestin>
BINDS
<phosphorylated t340e>

1.2.2 Arguments of bind

The arguments of forms of *bind* other than *binding* invariably occur on either side of the cuing verb form. The default strategy for identifying both arguments in these instances is to choose the closest binding term on either side of the verb. In the cases we have investigated, this strategy works often enough to be useful for the surface object. However, due to predicate coordination as well as relativization, such a strategy often fails to identify correctly the surface subject of *bind* (*binds* or *bound*) when more than

one binding term precedes the verb. We therefore use the strategy summarized in (7) for recovering the surface subject in such instances.

(7) When more than one binding term precedes a form of *bind* other than *binding*, choose the most specific of these binding terms as the surface subject of the predication.

"Most specific" is determined (recursively) for a series of binding terms in the same way that the most specific part of a complex binding term is determined.

The input text (8) provides an example of a binding predication cued by *binds* in which the arguments appear (immediately) on either side of the cuing verb. The two macro-noun phrases serving as potential arguments are underlined.

(8) A transcription factor, Auxin Response Factor 1, that **binds** to the sequence TGTCTC in auxin response elements was cloned from Arabidopsis by using a yeast one-hybrid system.

(9) <auxin response factor 1>
BINDS
<sequence tgtctc>

In the extracted binding relationship in (9), the Specificity Rule chooses *Auxin Response Factor 1* from the first macro-noun phrase because it maps to the UMLS Metathesaurus with semantic type 'Amino Acid, Peptide, or Protein'. In the second argument, *the sequence TGTCTC* has a head that submits to the Morphology Shape Rule and hence is considered to be more specific than *auxin response elements*.

In (10), the Specificity Rule applies correctly to select the surface subject of the binding predication when multiple binding terms appear to the left of the verb.

(10) Phosphatidylinositol transfer protein has a single lipid-binding site that can reversibly **bind** phosphatidylinositol and phosphatidylcholine and transfer these lipids between membrane compartments in vitro.

<phosphatidylinositol transfer protein>
BINDS
<phosphatidylcholine>
<phosphatidylinositol transfer protein>
BINDS
<phosphatidylinositol>

Both *Phosphatidylinositol transfer protein* and *a single lipid-binding site* occur to the left of *bind* and have been identified as binding terms by the first phase of processing. However, *Phosphatidylinositol transfer protein* maps to the corresponding Metathesaurus concept with semantic type 'Amino Acid, Peptide, or Protein, thus causing it to be more specific than *a single lipid-binding site*. The second predication listed in (10) was correctly extracted due to coordination processing.

ARBITER pursues limited coordination identification in the spirit of Agarwal and Boggess (1992) and Rindflesch (1995). Only binding terms are considered as candidates for coordination. For each conjunction encountered, the phrase immediately to the right is examined; if it is a binding term, all contiguous binding terms occurring immediately to the left of the conjunct are considered to be coordinate with the right conjunct. Coordination inside the simple noun phrase is not considered, and therefore structures such as *The TCR alpha beta or -gamma delta chains* are not recognized. Nonetheless, as indicated in (11), this limited approach to noun phrase coordination is often effective.

(11) Purified recombinant NC1, **like** authentic NC1, also bound specifically to fibronectin, collagen type I, **and** a laminin 5/6 complex.

<authentic nc1>
BINDS
<laminin 5 / 6 complex>
<authentic nc1>
BINDS
<collagen type i>
<authentic nc1>
BINDS
<fibronectin>
<purified recombinant nc1>
BINDS
<laminin 5 / 6 complex>
<purified recombinant nc1>
BINDS
<collagen type i>
<purified recombinant nc1>
BINDS
<fibronectin>

Although the particular underspecified syntactic analysis used in the identification of binding predications in the biomedical research literature is limited in several important ways, it appears adequate to enable this project with a useful level of effectiveness, and this is supported by evaluation.

2 Evaluation

In order to determine ARBITER's effectiveness, the program was formally evaluated against a gold standard of MEDLINE citations in which the binding predications asserted were marked by hand. A search of MEDLINE limited to one month (June 1997) and based on the text words ((*bind, binds, binding,* or *bound*) and (*protein* or *proteins*)) retrieved 116 citations with 1,141 sentences; of these, 346 contained some form of the verb *bind*. 260 binding predications were identified in the binding sentences. (The binding sentences also contained 2,025 simple noun phrases, 1,179 of which were marked as being binding terms.)

In processing this test collection, ARBITER extracted 181 binding predications, 132 of which were correct. Since ARBITER missed 128 marked binding predications (false negatives) and incorrectly identified 49 such relationships, recall and precision as measures of effectiveness are 51% and 73%, respectively.

In comparing ARBITER's output against that marked in the gold standard, fairly stringent matching criteria were used. A binding predication extracted from a particular sentence by ARBITER had to appear in that same sentence in the gold standard (not just the same citation) in order to be counted as correct. Further, in the gold standard, only the most specific component of a macro-noun phrase was marked as being the correct argument for a particular binding predication. If ARBITER retrieved any other part of a macro-noun phrase in identifying the arguments of that predication, it was assessed as an error.

A large number of ARBITER errors are due to two phenomena: difficulties in correctly identifying binding terms during the first phase of processing and syntactic complexity confounding argument identification during the second phase. An example of the first error type is seen in (12), where the failure to identify *ran* as a binding term caused ARBITER to miss the correct binding predication asserted in this sentence (indicated by "-FN->").

(12) Requirement of guanosine triphosphate-bound ran for signal-mediated nuclear protein export.

-FN-> <guanosine triphosphate>
BINDS
<Ran>
-FP-> < guanosine triphosphate>
BINDS
<signal - mediate nuclear protein export>

This error then led to the false positive error ("-FP->") when the program wrongly interpreted the next noun phrase in the sentence (*signal-mediated nuclear protein export*) as the second argument in this predication.

The interaction of coordination and negation in (13) caused ARBITER to partially misinterpret the binding predications in this sentence.

(13) The nonvisual arrestins, beta-arrestin and arrestin3, but not visual arrestin, bind specifically to a glutathione S-transferase- clathrin terminal domain fusion protein.

<arrestin3>
BINDS
<glutathione s-transferase-clathrin terminal domain fusion protein>
<beta arrestin>
BINDS
<glutathione s-transferase-clathrin terminal domain fusion protein>
<nonvisual arrestin>
BINDS
<glutathione s-transferase-clathrin terminal domain fusion protein>
-FN-> <visual arrestin>
DOES_NOT_BIND
<glutathione s-transferase-clathrin terminal domain fusion protein>

Although some of the coordination in (13) was processed properly, resulting in the relationships listed, the negated coordination associated with the noun phrase *visual arrestin* was not interpreted correctly, and thus ARBITER failed to

identify the predication marked as a false negative.

3 Application

As an initial application of ARBITER we ran the program on 491,356 MEDLINE citations, which were retrieved using the same search strategy responsible for the gold standard. During this run, 331,777 sentences in 192,997 citations produced 419,782 total binding assertions. Extrapolating from the gold standard evaluation, we assume that this is about half of the total binding predications asserted in the citations processed and that somewhat less than three quarters of those extracted are correct.

The initial list of 419,982 binding triples represents what ARBITER determined was asserted in the text being processed. Many of these assertions, such as those in (14), while correct, are too general to be useful.

(14) <receptors>
BINDS
<Peptides>
<Erythrocytes>
BINDS
<Antibodies>

Further processing on ARBITER raw output extracted specific protein names and genomic structures and reduced the number of such binding predications to 345,706. From these more specific binding predication, we began the construction of a database containing binding relations asserted in the literature. More detailed discussion of this database can be found in (Rajan et al. in prep); however, here we give an initial description of its characteristics.

We submitted the 345,706 more specific ARBITER binding predications to a search in GenBank (Benson et al. 1998) and determined that 106,193 referred to a GenBank entry. The number of Genbank entries with at least one binding assertion is 11,617. Preliminary results indicate that the database we are constructing will have some of the following characteristics:

- 10,769 bindings between two distinct Genbank entries (5,569 unique)
- 875 more binding assertions found between an entry and a specific DNA sequence
- 27,345 bindings between a Genbank entry and a UMLS Metathesaurus concept
- 5,569 unique relationships among pairs of entries (involving 11,617 unique entries)

Conclusion

The cooperation of structured domain knowledge and underspecified syntactic analysis enables the extraction of macromolecular binding relationships from the research literature. Although our implementation is domain-specific, the underlying principles are amenable to broader applicability.

ARBITER makes a distinction between first labeling binding terms and then identifying certain of these terms as arguments in a binding predication. The first phase of this processing is dependent on biomedical domain knowledge accessible from the UMLS. Applying the techniques we propose in other areas would require at least a minimum of semantic classification of the concepts involved. General, automated techniques that could supply this requirement are becoming increasingly available (Morin and Jacquemin 1999, for example).

Although we concentrated on the inflectional forms of a single verb, the principles we invoke to support argument identification during the second phase of processing apply generally to English predication encoding strategies (with a minimum of effort necessary to address prepositional cuing of gerundive arguments for specific verbs). The approach to noun phrase coordination also applies generally, so long as hypernymic classification is available for the heads of the potential conjuncts.

Acknowledgements

We are grateful to John Wilbur for assistance with accessing GenBank, to Alan Aronson for modifications to MetaMap, and to James Mork for providing the distributed system that supported the processing of MEDLINE citations.

References

Agarwal R. and Boggess L. (1992) *A simple but useful approach to conjunct identification.* Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics, pp. 15-21.

Ait-Mokhtar S. and Chanod J.-P. (1997) *Incremental finite-state parsing.* Proceedings of the Fifth Conference on Applied Natural Language Processing, pp. 72-9.

Appelt D. E. and Israel D. (1997) *Tutorial on building information extraction systems.* Fifth Conference on Applied Natural Language Processing.

Aronson A. R., Rindflesch T. C., and Browne A. C. (1994) *Exploiting a large thesaurus for information retrieval.* Proceedings of RIAO 94, pp. 197-216.

Benson D. A., Boguski M. S., Lipman D. J., Ostell J., and Ouelette B. F. (1998) *GenBank.* Nucleic Acids Research, 26/1, pp. 1-7.

Blaschke C., Andrade M. A., Ouzounis C., and Valencia A. (1999) *Automatic extraction of biological information from scientific text: protein-protein interactions.* Intelligent Systems for Molecular Biology (ISMB), 7, pp. 60-7.

Church K. W. (1988) *A stochastic parts program and noun phrase parser for unrestricted text.* Proceedings of the Second Conference on Applied Natural Language Processing, pp. 136-143.

Craven M. and Kumlien J. (1999) *Constructing biological knowledge bases by extracting information from text sources.* Intelligent Systems for Molecular Biology (ISMB), 7, pp. 77-86.

Cutting D. R., Kupiec J., Pedersen J. O., and Sibun P. (1992) *A practical part-of-speech tagger.* Proceedings of the Third Conference on Applied Natural Language Processing.

Fukuda F., Tsunoda T., Tamura A., and Takagi T. (1998) *Toward information extraction: Identifying protein names from biological papers.* Pacific Symposium on Biocomputing (PSB), 3, pp. 705-16.

Hearst M. A. (1999) *Untangling text data mining.* Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, pp. 3-10.

Hindle D. (1983) *Deterministic parsing of syntactic non-fluencies.* Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics, pp. 123-8.

Humphreys B. L., Lindberg D. A. B., Schoolman H. M., and Barnett G. O. (1998) *The Unified Medical language System: An informatics research collaboration.* Journal of the American Medical Informatics Association, 5/1, pp. 1-13.

McCray A. T., Srinivasan S., and Browne A. C. (1994) *Lexical methods for managing variation in biomedical terminologies.* Proceedings of the 18th Annual Symposium on Computer Applications in Medical Care, pp. 235-9.

McDonald D. D. (1992) *Robust partial parsing through incremental, multi-algorithm processing.* In "Text-Based Intelligent Systems," P. S. Jacobs, ed., pp. 83-99.

Morin E. and Jacquemin C. *Projecting corpus-based semantic links on a thesaurus.* Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, pp. 389-96.

MUC-7. *Message Understanding Conference Proceedings.* http://www.muc.saic.com.

Rajan J. V., Hunter L., and Rindflesch T. C. (In prep.) *Mining MEDLINE.*

Rindflesch T. C. (1995) *Integrating natural language processing and biomedical domain knowledge for increased information retrieval effectiveness.* Proceedings of the 5th Annual Dual-Use Technologies and Applications Conference, pp. 260-5.

Rindflesch T. C., Hunter L., and Aronson A. R. (1999) *Mining molecular binding terminology from biomedical text.* Proceedings of the AMIA Annual Symposium, pp. 127-131.

Rindflesch T. C., Tanabe L., Weinstein J. N., and Hunter L. (2000) *EDGAR: Extraction of drugs, genes and relations from the biomedical literature.* Pacific Symposium on Biocomputing (PSB), 5, pp. 514-25.

Tersmette K. W. F., Scott A. F., Moore G.W., Matheson N. W., and Miller R. E. (1988) *Barrier word method for detecting molecular biology multiple word terms.* Proceedings of the 12th Annual Symposium on Computer Applications in Medical Care, pp. 207- 11.

Voorhees E. M. and Harman D. K. (1998) *The Seventh Text Retrieval Conference.*

Vourtilainen A. and Padro L. (1997) *Developing a hybrid NP parser.* Proceedings of the Fifth Conference on Applied Natural Language Processing, pp. 80-7.

Wacholder N., Ravin Y., and Choi M. (1997) *Disambiguation of proper names in text.* Proceedings of the Fifth Conference on Applied Natural Language Processing, pp. 202-208.

Weischedel R., Meteer M., Schwartz R., Ramshaw L., and Palmucci J. (1993) *Coping with ambiguity and unknown words through probabilistic models.* Computational Linguistics, 19/2, pp. 359-382.

Zhai C. (1997) *Fast statistical parsing of noun phrases for document indexing.* Proceedings of the Fifth Conference on Applied Natural Language Processing, pp. 312-31.