Mining Molecular Binding Terminology from Biomedical Text

Thomas C. Rindflesch, PhD,[1] Lawrence Hunter, PhD,[2] and Alan R. Aronson, PhD[1]
National Library of Medicine[1]
National Cancer Institute[2]
National Institutes of Health, Bethesda, Maryland

Automatic access to information regarding macromolecular binding relationships would provide a valuable resource to the biomedical community. We report on a pilot project to mine such information from the molecular biology literature. The program being developed takes advantage of natural language processing techniques and is supported by two repositories of biomolecular knowledge. A formative evaluation has been conducted on a subset of MEDLINE® abstracts.

INTRODUCTION

There is widespread demand among researchers for factual data about biomolecular function. However, the creation and maintenance of databases to satisfy this demand is, in general, expensive and labor intensive. In this paper we discuss a pilot project aimed at creating a repository of biomolecular function information using data mining techniques to automatically extract information about molecular binding affinities from MEDLINE abstracts.

Currently, there is no adequately detailed and comprehensive source of information regarding biomolecular function. Due to the immense manual effort required, the one database of molecular function which is intended to be comprehensive (the Enzyme Commission database) has quite limited coverage, providing information about a bit over one thousand of the more than 50,000 enzymes in the human genome [1]. The other databases of protein function are intentionally specialized (generally to the well understood enzymes constituting the pathways of intermediary metabolism) and have even lower coverage.

Binding affinity is a central determinant of macromolecular function, and the biomedical literature is replete with references to this molecular relationship. Our goal is to populate a database with assertions of binding affinity, automatically extracted from MEDLINE abstracts. Although such an automatically generated database would have high coverage, the assertions in it would be necessarily less reliable than those in manually generated functional databases, particularly since it is beyond current abilities to capture important contextual information, such as the biochemical environment, cofactors, etc. However, this shortcoming is mitigated by the fact that each assertion of binding affinity in such a database would be associated with specific MEDLINE citations which justify it.

We consider the discovery of molecular binding relations in biomedical free text as a special case of semantic interpretation and thus draw on the natural language processing techniques being developed in the SPECIALIST™ system [2] at the National Library of Medicine (NLM). We further rely on the resources of the UMLS® (Unified Medical Language System®) Metathesaurus® [3] and NCBI (National Center for Biomedical Information) GenBank® [4].

In processing biomedical free text, syntactic predicational structure is used as a proxy for semantic propositions referring to binding relationships. The syntactic structure is defined on the interaction of noun phrases and verbs within a sentence, while semantic type constraints on argument identification complement the syntactic constraints. The arguments of semantic propositions referring to binding relationships identified in the text of abstracts are interpreted, whenever possible, in terms of the UMLS Metathesaurus and NCBI Genbank entries. Each binding relationship with identifiable arguments provides an assertion that can be added to a database of biomolecular function.

BACKGROUND

Our study was conducted on a small set of MEDLINE abstracts which contain sentences asserting molecular binding relationships (most commonly indicated by some form of the verb *bind*). A search limited to entry date June, 1997 and issued for the text words *bind, binds, binding, bound,* and *proteins* produced 116 citations containing 1,141 sentences, 346 of which contained a form of *bind*. This sample reflects 66,250 such citations occurring in MEDLINE over the past two years, and 445,544 from 1975 through mid-1999. The potential arguments of binding relations in the sentences in our test set were marked by hand to be used later as a gold standard against which to evaluate this project.

Although semantic propositions referring to binding relations are normally encoded by a single verb, associated syntactic complexity, as shown in (1) underlies an array of challenges to automatic interpretation.

(1) Cooperatively responsive sequence constructs bound both STAT1alpha and NF-kappaB in nuclear extracts prepared from IFNgamma- and/or TNFalpha-stimulated fibroblasts, although binding of individual factors was not cooperative.

All noun phrases in (1) refer to entities which can potentially enter into a binding relationship: *cooperatively responsive sequence constructs*, *STAT1alpha*, *NF-kappaB*, *IFNgamma- and/or TNFalpha-stimulated fibroblasts*, and *individual factors*. Morphological and syntactic phenomena inherent in these phrases which must be addressed in order to effectively extract the binding relations from this sentence include both acronyms (*STAT* for *Signal Transducer and Activator of Transcription*) and coordination (*IFNgamma- and/or TNFalpha-stimulated fibroblasts* for *IFNgamma-stimulated fibroblasts* and *TNFalpha-stimulated fibroblasts*) as abbreviatory devices. A further challenge is presented in determining whether the relative clause *in nuclear extracts prepared from IFNgamma- and/or TNFalpha-stimulated fibroblasts* modifies only *NF-kappaB* or *STAT1alpha* as well. Ideally, the anaphoric reference inherent in *cooperatively responsive sequence constructs* would be resolved with more specific terms appearing elsewhere in the abstract. Effective processing of this sentence must also determine that verb form *bound* does in fact encode a semantic binding proposition in this sentence, while the form *binding* toward the end of the sentence does not.

In order to address the complexity associated with this project, we have divided the processing which supports semantic interpretation into two phases: a) identification and, when possible, interpretation of the noun phrases referring to binding entities and b) final semantic processing which recognizes just those noun phrases which enter into a particular binding relation asserted in the text.

We are developing a Prolog program called ARBITER (Assess and Retrieve BInding TERms) which implements these notions, drawing heavily on existing resources at NLM. For the remainder of the discussion, we focus on the way ARBITER identifies binding terms in free text in preparation for final semantic interpretation. A binding term is a noun phrase referring to a "binding entity," which can be a molecule, a genomic structure, a cell or cell component, or some topographic aspect of a molecule, cell or cell component. In this first phase of the project, for example, ARBITER determines that in processing (2), the noun phrases in bold are binding terms, while those in italic are not.

(2) *These results* lead to *a model* of *rhoGDI function* in which **the carboxy-terminal binding domain** targets **the amino-terminal inhibitory region** to **GTPases**, resulting in *membrane extraction* and *inhibition* of *nucleotide cycling*.

METHODS

The algorithm for determining whether a noun phrase is a binding term depends on the UMLS Metathesaurus and SPECIALIST Lexicon, NCBI's GenBank, and further ARBITER-specific processing which takes advantage of information local to the molecular biology domain. The highly ambiguous character of many of the important words for identifying binding terms indicates the importance of applying ARBITER only to text in the molecular biology domain. Currently this is assured by processing MEDLINE abstracts with appropriate MeSH indexing terms. In order to generalize the processing to text which has not been indexed we are exploring the use of automatic methods of determining the biomedical domain of free text ([5], for example).

Previous research directed at recognizing terminology in free text has often applied to all terms occurring in text and not just those meeting specified criteria. Processing, whether based on a shallow [6] or rather extensive [7] linguistic analysis, normally does not provide an interpretation. Our strategy shares significant features with [8], which, however, concentrates on protein names and does not appeal to an existing knowledge source. Related research focuses on recovering the relationships asserted in text for a given list of proteins [9].

Before submitting input text to natural language processing techniques, ARBITER takes advantage of a program (Aronson) which operates on text as strings in order to expand author-defined acronyms. The enhanced text with locally-defined acronyms expanded is then submitted to ARBITER for further processing.

Existing natural language processing tools including a stochastic tagger [10], the SPECIALIST lexicon and associated tools [11], as well as the SPECIALIST minimal commitment parser and MetaMap [12] provide the input which sets the scene for the binding-specific processing pursued by ARBITER. Input text (3), for example, provides output (4), which includes noun phrases identified with Metathesaurus concept (and semantic type) provided by MetaMap when possible.

(3) The three mutants bind erythrocytes at wild-type toxin levels...

(4) [the three mutants]
[erythrocytes] - UMLS Metathesaurus: Erythrocytes (Cell)
[wild-type toxin levels]

ARBITER uses the information in (4) as the basis for a series of steps which determine the status of noun phrases as binding terms, and, where possible, identify the referent of these terms.

Noun phrases which map to UMLS Metathesaurus concepts having any of the semantic types listed in (5) are considered to be binding terms (with referent identified).

(5) 'Amino Acid, Peptide, or Protein', 'Nucleic Acid, Nucleoside, or Nucleotide', 'Gene or Genome', 'Nucleotide Sequence', 'Immunologic Factor', 'Carbohydrate', 'Lipid', 'Organic Chemical', 'Cell', 'Cell Component', 'Virus'

For example, the two noun phrases *arrestin* and *rhodopsin* are identified as binding terms in (6) because they both match Metathesaurus concepts with semantic type 'Amino Acid, Peptide, or Protein'.

(6) ... play important but distinct roles in promoting the binding of **arrestin** to **rhodopsin**.

Noun phrases which match one of the proteins and genes listed in NCBI's GenBank are also identified as binding terms. ARBITER has access to a statistically-based trigraph matching program which provides either exact matches or partial matches with a confidence score. For example, the noun phrase *p53* in (7) matches exactly to the GenBank entry "p53" [DE]. Such a noun phrase constitutes a binding term with referent identified as the GenBank entry.

(7) Binding and modulation of **p53** by p300/CBP coactivators.

Incomplete matches such as *MAPKK Pbs2p* in partially mapping to "pbs2" [GN] are deemed to be binding terms, but no referent can be given.

Those noun phrases which are not found in the Metathesaurus by MetaMap and which do not constitute an exact match to an entry in GenBank through the trigraph program cannot be given a referent. ARBITER attempts to identify binding terms without a referent nevertheless. Such terms are likely to provide valuable information upon human review regarding the nature of the text in which they appear.

The first step the program takes in identifying a noun phrase as a binding term which does not have a referent is to determine whether the head of the noun phrase occurs in a constrained set of words generally designating bindable entities. Examples of words serving as the heads of such noun phrases include those concerned with various biomolecular phenomena (*box, chain, sequence, subunit, repeat, ligand, motif, mutant, terminus, strand*), molecular or cellular topography (*spike, cleft, groove, pit, pocket, surface, membrane*), and general terms referring to various characteristics of bindable entities (*element, receptor, site, complex, component, domain, target*). (8) is a sentence from our test set illustrating these binding words serving as the heads of noun phrases which are binding terms.

(8) ... structure of [this **domain**] shows [a beta-sandwich **motif**] with [a narrow hydrophobic **cleft**] that binds isoprenes, and an [exposed **surface**]..

A further measure to identify binding terms is to consider whether the words constituting a noun phrase exhibit the normal morphological characteristics of English words, namely that they contain at least one vowel and no digits. Text tokens which do not exhibit such characteristics are tagged as being potential binding terms (often acronyms not defined locally). *S343E,* for example, is returned as a binding term by ARBITER on the basis of this evidence.

Some words, such as the components of the phrase *sry delta,* exhibit normal English characteristics and thus do not indicate a binding term by the previous criterion. Hence this phrase, which does not occur in the Metathesaurus or GenBank, and does not contain a general binding word would be missed as a binding argument in (9).

(9) Further, several lines of evidence strongly suggest that **sry delta** binds to DNA as a dimer.

In such instances ARBITER takes advantage of contextual information local to the current abstract. It has been determined that, without exception, any term which occurs immediately to the left of the text token *binding* anywhere in an abstract qualifies as a binding term elsewhere in that same abstract. For example, in the abstract in which (9) occurs, the following sentence also appears.

(10) ... reflecting the cooperativity of **sry delta** *binding* to DNA.

Before ARBITER processes the sentences containing a form of the verb *bind*, it scans the entire abstract looking for terms locally defined by virtue of appearing immediately to the left of *binding.* Such terms are considered to be binding terms for that abstract only.

As a final step, ARBITER joins contiguous simple binding terms meeting certain specified conditions

into a single complex binding term. For example the individual binding terms recognized in (11), namely *coiled-coil domain*, *C terminus*, *PKD1 gene product*, and *polycystin* are combined into the complete binding term highlighted in (11).

(11) ... a previously unrecognized **coiled-coil domain within the C terminus of the PKD1 gene product, polycystin**, and demonstrate...

The conditions which indicate that a term should be incorporated into a larger structure include prepositional modification (*within the C terminus* and *of the PKD1 gene product* in (11)) as well as appositival complementation (*polycystin).*

For evaluation, we submitted the 116 MEDLINE abstracts in our test collection to the various steps of ARBITER processing discussed in the preceding section in order to discover how closely the program matched the determinations made by hand regarding binding term identification.

RESULTS

Of the 2,025 simple noun phrases in the 346 binding sentences in our test collection, 1,179 were marked by hand as being binding terms. Since ARBITER identified 1,064 binding terms, recall as a partial measure of effectiveness was 72%. Of the binding terms retrieved, 845 were correct, and thus precision was 79%.

Somewhat more than two thirds of the unique noun phrases correctly identified as binding terms were assigned a referent in the UMLS Metathesaurus or NCBI's GenBank. Of the terms found in the Metathesaurus; almost two thirds had semantic type 'Amino Acid, Peptide, or Protein'. There was also a certain amount of overlap between terms identified in the Metathesaurus and GenBank (*1-Chloro-3-bromopropene-1 (CBP)* and *signal transducer and activator of transcription 1 (STAT1),* for example); however, we did not formally track this duplication.

Although a third of the binding terms identified could not be given a referent in an available knowledge source, we feel that it is nevertheless valuable to retrieve these terms. Examples include text input *p300* partially matching "p30" [DE] in GenBank, as well as text *GDI* and GenBank "gdi1" [GN]. Another important group of correctly-identified binding terms which were not assigned a specific referent were those with general binding term heads, such as *auxin response elements* and *amino-terminal SH2 domain.* A smaller number of binding terms were identified by failing to conform to normal English morphology (*M-T7* and *2H5* for example) or by virtue of occurring as a locally defined binding term (*Nova-1*).

DISCUSSION

Many of the false negative errors produced by ARBITER while processing the test set occurred because the relevant concept did not appear in any of the knowledge sources to which we appeal, and further, the noun phrase was not subject to the term-specific processing we use. For example, the binding term *trifluoroacetyl* does not occur in either the Metathesaurus or GenBank, nor is it in the set of general binding terms available to ARBITER. Further, this term has the morphological characteristics of a normal English word (and not a binding term acronym) and was not defined by local context elsewhere in the abstract.

Another group of false-negative errors are ultimately due to part-of-speech ambiguity. For example, *sequence* in (12) is listed in the SPECIALIST Lexicon with either verb or noun as part-of-speech.

(12) ... acquires a stem-loop structure and includes a UCU **sequence** that binds to Tat and...

As noted earlier, we employ a stochastic tagger to resolve this type of ambiguity; however, such resolution is not completely accurate. In this instance the part-of-speech of *sequence* was incorrectly determined to be verb. Consequently the parser assigned syntactic structure incorrectly based on this error: *UCU sequence* was analyzed as a noun phrase (*UCU*) followed by the verb *sequence*. This sequence was thus not perceived as a unit qualifying as a binding term.

The majority of the errors produced (both false negative and false positive) are due to a single characteristic of ARBITER, which can be traced to a variety of syntactic phenomena: ARBITER often identifies as a single binding term a phrase which in fact contains several smaller, distinct binding terms. For example in (13), ARBITER identified the entire sequence in bold as a single binding term.

(13) To identify the receptor for TARC, we produced **TARC as a fusion protein with secreted alkaline phosphatase** and used it for specific binding.

The syntactic phenomenon which underlies this error is the fact that in this sentence the preposition *as* functions as a particle for the verb *produce* rather than as an introducer of a noun phrase modifier. The object of *as* is therefore a separate argument of *produce* rather than a part of the noun phrase of which *TARC* is the head.

Three binding terms in (13) should have been identified, namely *TARC, fusion protein*, and *secreted alkaline phosphatase*. The single infelicitous decision noted above resulted in four errors: three false negatives (the smaller components) and one false positive (the longer phrase). Currently, we strictly count these identifications as errors, whereas at least partial credit would not seem unwarranted in these situations.

Coordinate structures pervasively contribute to this particular ARBITER error type as the example in (14) illustrates.

(14) Only LARC but not **five other CC chemokines (MCP-1, RANTES, MIP-1alpha, MIP-1beta,** and TARC) competed with LARC-SEAP for binding to GPR-CY4.

The program identified the entire phrase in bold as a single binding term. Although the opening parenthesis before *MCP-1* contributed to the error, coordination is the primary factor. ARBITER was not able to determine that the components in this phrase are separate noun phrases, each coordinated with *TARC*, rather than being the components of a single (perhaps appositival) term. This deficiency generated five false negatives (*five other CC chemokines, MCP-1, RANTES, MIP-1alpha,* and *MIP-1beta*) and one false positive (the whole phrase). We are currently pursuing a more aggressive approach to coordinate structures to address errors of this type.

CONCLUSION

On the basis of the promising results of this pilot project applying ARBITER to a modest-sized collection of abstracts relevant to molecular binding affinities, we believe that, although deficiencies must be addressed, there are potential benefits to pursuing this approach. We also note the potential to generalize to the creation of many other valuable factual databases from MEDLINE. Although binding relationships were chosen for their immediate significance and relative ease of parsing, databases of many other relationships (for example "catalyze") are feasible with only modest modifications to the general methodology on which ARBITER is based.

Acknowledgements

We are grateful to W. John Wilbur for the use of his trigraph matching program.

References

1. Shah I and Hunter L. Predicting enzyme function from sequence: a systematic appraisal. *Intelligent Systems for Molecular Biology* 1997;5:276-83.

2. McCray AT, Aronson AR, Browne AC, Rindflesch TC, Razi A, and Srinivasan S. UMLS knowledge for biomedical language processing. *Bulletin of the Medical Library Association* 81, 1993, 184-194.

3. Humphreys BL, Lindberg DAB, Schoolman HM, and Barnett GO. The Unified Medical language System: An informatics research collaboration. *Journal of the American Medical Informatics Association* 1998:5(1):1-13.

4. Benson DA, Boguski MS, Lipman DJ, Ostell J, and Ouelette BF. GenBank. *Nucleic Acids Research* 1998;26(1):1-7.

5. Humphrey SN. Automatic indexing of documents from journal descriptors: A preliminary investigation. *Journal of the American Society for Information Science* 1999;50(8):661-674.

6. Tersmette KWF, Scott AF, Moore GW, Matheson NW, and Miller RE. Barrier word method for detecting molecular biology multiple word terms. In Greenes RA (ed.) *Proceedings of the 12th Annual Symposium on Computer Applications in Medical Care*, 1988:207-211.

7. Evans DA and Chengxiang Z. Noun phrase analysis in unrestricted text for information retrieval. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics.* 1996:17-24.

8. Fukuda K, Tsunoda T, Tamura A, and Takagi T. Toward information extraction: Identifying protein names from biological papers. In Altman RB, Dunker AK, Hunter L, and Klein TE (eds.) *Pacific Symposium on Biocomputing '98*, 1998:707-718.

9. Blaschke C, Andrade MA, Ouzounis C, and Valencia A. Automatic extraction of biological information from scientific text: protein-protein interactions. *Intelligent Systems for Molecular Biology 1999 (*to appear*).*

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

11. McCray AT, Srinivasan S, and Browne AC. Lexical methods for managing variation in biomedical terminologies. In Ozbolt JG (ed.) *Proceedings of the 18th Annual Symposium on Computer Applications in Medical Care*, 1994:235-239.

12. Aronson AR, Rindflesch TC, and Browne AC. Exploiting a large thesaurus for information retrieval. *Proceedings of RIAO* 94, 1994:197-216.