RC23179 (W0404-068) April 12, 2004
Computer Science

IBM Research Report

Discovery of Protein-Protein Interactions Using a Combination of Linguistic and Statistical Information

James W. Cooper
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

Discovery of Protein-Protein Interactions Using a Combination of Linguistic and Statistical Information

James W. COOPER

Unstructured Information Management,
IBM T J Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598
jwcnmr@watson.ibm.com

Abstract

The rapid publication of important research in the biomedical literature makes it increasingly difficult for researchers to keep current with significant work in their area of interest. This paper reports a scalable method for the discovery of protein-protein interactions in Medline abstracts, using a combination of text analytics, statistical analysis and a set of easily implemented rules. Using a collection of 564 abstracts describing protein interactions, a precision of 0.92 and a recall on 0.84 were obtained. Applying similar techniques to 12,300 abstracts, a precision of 0.61 and a recall of 0.90 were obtained, (f = 0.72) and when allowing for secondary relations, the precision could be extended to 0.92.

1. Introduction

Scientists in molecular biology find that a significant technique for studying protein function is through the study of protein-protein interactions. While the actual experimental study of such interactions remains the most important manner of obtaining these data, the number of protein-protein interactions reported in the literature is substantial and growing rapidly. There are a number of tabulations of these interactions, such as that provided by the Munich Institute for Protein Sequence (MIPS), these tabulations are of necessity incomplete.

To address this problem, we have been developing a group of biology-specific annotators that work in conjunction with our group's text analytic software, for the discovery of protein-protein relations in text.

In this paper, we undertook a study that utilizes a combination of computational linguistics, statistics and domain-specific rules to detect protein-protein interactions in a set of Medline abstracts.

The system we describe here is particularly appealing because it can be used both to find known interactions and to find interactions not yet tabulated. According to the National Library of Medicine, Medline contains over 11 million abstracts, with about 40,000 being added each month. Thus, having a scalable, robust system for protein interaction discovery provides a major information tool for molecular biologists.

A number of workers have tackled portions of this problem previously with some partial success. The SUISEKI system (Blaschke and Valencia, 2001) recognizes various grammatical frames which may describe protein interactions. They reported high precision (68%) for the shorter patterns and lower precision (21%) for the longer ones.

In a more narrowly focused experiment, Pustejovsky *et. al* (Pustejovsky, 2002) described a computational linguistic system for detecting *inhibit* relations, with 90% precision and recall of 57%.

Recently Leroy (Leroy, *et. al*. 2003) described Genescene, a software package for detecting relations between genes. They used both rule-based detection and co-occurrence based methods, finding that rule-based relations were 95% correct and co-occurrence based relations 60% correct.

Researchers at Ariadne Genomics (Daraselio, *et. al.*) have quite recently described a system called MedScan, which they report as having 91% precision and 21% recall on human protein-protein interactions.

We (Cooper and Byrd, 1997) have previously described methods for detecting relations between noun phrases and methods for displaying them (Cooper and Byrd, 1998). In this paper we propose using these techniques along with a combination of statistical and rule-based approaches to identify protein interactions in Medline abstract text.

Ideally one would imagine constructing a protein interaction network much like the network that allowed discovery of the relationship between “fish oil” and “Reynaud’s disease” (Swanson, 1986). The relations extracted in this paper can be used to form just such a network.

This paper discusses the text analytic tools used, and then describes preliminary experiments against a gold standard of protein relations. Finally the results of mining relations across a large set of Medline abstracts are described.

2. Comparing Approaches

The SUISEKI system and the MedScan system both do fairly deep parses of each sentence in the abstract and align these results with patterns or frames. The Genescene system uses a combination of a very simple parser and a set of rules, as well as a distance co-occurrence measure.

The SUISEKI and Genescene systems attempt to find protein or gene names from patterns and syntax, while the Medscan system uses a compiled dictionary of protein names and synonyms.

In this work, the approach is to use a tagger and shallow parser primarily for sentence boundary recognition, and use a dictionary derived from public sources to recognize the protein names. The goal of this approach is to be fast and scalable as well as to improve precision and recall over other methods.

3. Text Analytic Tools

The system used in these experiments is constructed using the TALENT (Text Analysis and Language Engineering Tools) text mining system (Neff and Byrd, 2003). The current version of this system is called TafTalent and operates in the Unstructured Information Management (UIMA) environment (Ferrucci and Lally, 2003). It consists of a series of document-level annotators that perform preliminary part-of-speech lookup, tag each word for part of speech, perform a shallow parse of each sentence, and annotate yeast proteins in a manner described below. Each of these annotators leaves its results in an annotation repository called the Common Annotation System (CAS).

While the underlying TafTalent text analytic system is written in C++, the UIMA framework allows users to write programs in Java that can load the CAS and launch the C++ annotators, and then perform the analysis of the results in Java. This is the approach used in these experiments.

After each Medline abstract is processed by the series of annotators, a CAS consumer program converts these annotations into entries in a DB2 database load file. This file contains all of the salient terms per document, their part of speech and their relative token positions in the document. An additional load file contains the Medline document metadata: dates, titles, authors and ID numbers.

Then it is possible to use a few simple database queries to construct a Terms database table of all the unique terms in the document collection, and compute their frequencies, and the number of documents in which they appear once and more than once. Using these data the salience or IQ (Prager, 1999) of each term can be computed.

4. Computing Relations

This paper explores the idea that the computation of relations between terms that was described earlier by our group (Byrd and Ravin, 1999) can be applied to recognizing protein interactions.

Relations between terms are computed based on their proximity. If two terms occur near each other on several occasions within the collection of documents they have a stronger relation than those that co-occur but once. Since the document number, and token position for each term are stored in the database, it is a simple matter to find terms that co-occur within any specified distance. Further, these relations can be tuned to select only those where one or both of the terms have a salience above a specific value.

The weights of these relations are computed using the mutual information formula

$$m = \log\left(\frac{totalterms \bullet paircount}{freq1 \bullet freq2}\right) \quad (1)$$

where *totalterms* is the total number of unique terms in the collection, *paircount* is the number of documents in which both terms occur, and *freq1* and *freq2* are the frequencies of the two terms in the collection. After computing all the mutual information values *m* for the term pairs, they are scaled to lie between 0 and 100.

In this paper, the co-occurrences are limited to those within a single sentence and no more than 30 words apart.

5. Preliminary Experiments Using MIPS Data

The Munich Institute for Protein Sequences (MIPS) maintains a database of published yeast (*saccharomyces cerevisiae)* protein interactions along with a reference to the Medline abstract of the paper in which the interaction is reported. This table gives 2050 protein names and 2604 pairs of protein interactions and provides links to additional information on each protein. The interaction table was parsed and the protein names and the 564 Medline abstracts downloaded.

An annotator was then developed that compared each lexical token found by TafTalent against the list of proteins and marked those that matched. Then, a simple CAS consumer program was designed to report the location of these proteins within each sentence in each document.

Initially, this was not particularly successful because each protein has a number of possible representations that needed to be matched to a common canonical form. For example, the protein SRV2 can also be represented as Srv2p, SRV2p, CAP and (CAP). Synonyms for most of these proteins are available on pages linked from the original page on the MIPS web site. The dictionary was expanded using these synonyms and the various allowed capitalizations and the analysis rerun, storing all terms and their document positions in a database table.

Even with the expanded protein synonym table, only 388 protein interactions were detected within single sentences that matched those in the MIPS interaction table, and 432 other interactions were detected which did not match those in the MIPS table. This amounted to a precision of 0.47 and a recall of 0.68. Further, there was no particular correlation between the computed strength of the relation (mutual information value) and the likelihood that it agreed with those in the MIPS table.

6. Detecting relations in individual documents

In an effort to improve the accuracy of protein-protein interaction detection, a detailed study of 65 of the abstracts was undertaken to determine what algorithms and approaches would be most effective. In this study, each abstract was examined along with a list of the interactions reported by the MIPS table, including all of the synonyms for each protein. This process led to the following conclusions:

1. Some interactions were not reported in the abstracts, but only in the full papers. In fact some review articles contained no protein names at all in the abstracts. This finding is similar to that previously described (Blaschke and Valencia, 2001).

2. Some interactions were described that were not tabulated by MIPS. For example, the abstract might mention prior work.

3. Protein complexes were frequently mentioned. For example references are made to dimmers such as “Ddc2-Mec1” and trimers such as “Hap2p-Hap3p-Hap5p.” Such complexes do, in fact, represent protein interactions and should also be detected and reported.

4. Proteins were frequently referred to by two synonyms separated by a slash, such as “GIM1/YKE2.”

5. In all but one case, the interactions were described in the same sentence, and thus resolving co-reference issues would add only marginally to the quality of the interaction detection. Thus, the fact that two proteins occurred in the same abstract, but not in the same sentence was not a good metric for the number of relations we should be able to find.

6. No instances of negation were found.

7. A database query of verbs that lay between two proteins led to the small list shown in Table 1. We note that this list is virtually identical to that used empirically by previous workers. (Blaschke,*et. al.,* 1999)

Table 1 - Verbs Used to Describe Protein Interactions

act
activate
associate
bind
complex
co-precipitate
depend
inhibit

interact
mediate
phosphorylate
stabilize

Accordingly two additional annotators and an extractor to operate on these abstracts were written. One annotator recognized protein complexes: dimers and trimers, and the other recognized protein synonyms in the "slash notation" we illustrated in point 4 above. When the annotator found these synonyms, it only annotated one of the two mentions, to avoid skewing the mention statistics. All protein complexes were treated as reports of interactions and annotated as such.

A CAS consumer was also written to find the verbs or their noun-equivalents in each sentence, if that sentence contained two or more different protein annotations.

7. Evaluation of Revised Annotations

Examination of protein interactions detected in 26 randomly selected documents showed that nearly all of the relations detected by our unnamed relations algorithm actually existed in the document, whether tabulated by MIPS or not, and that of those our algorithm missed, nearly all were not discussed in the abstract at all.

In these 26 documents, MIPS had reported 129 relations. We found that 17 of these were not in the abstracts. We also found an additional 52 interactions by proximity of which only 6 were incorrect. By reporting complexes as protein interactions as well, we found an additional 37 interactions. Overall, the results showed a precision of 0.92 and a recall of 0.84.

While we had anticipated using the protein interaction verbs to filter the excess relations we discovered, we actually found very few cases in this small sample where the verbs provided a meaningful filter.

8. Study of a Larger set of Medline Documents

With these encouraging preliminary results in hand, a study of a larger dataset was undertaken. The query "yeast protein" was submitted against our local indexing of Medline documents through 2002 and a list the top 12,300 documents was obtained. The MIPS protein interaction table was enhanced by one from Stanley Fields (Fields, 2000). These documents were annotated as above using the same series of annotators and database table created of the documents, terms, the proteins found in each of them.

The initial results of this experiment returned 912 relations, but only 133 agreed with the combined gold standard MIPS-Fields table. Considering the large number of abstracts examined, this small number of interactions indicates that the original data referred to by the MIPS table were a serendipitous set which referred specifically to protein-protein interactions. This larger dataset included a number of papers referring to genes which needed to be eliminated from consideration. Modifying the annotator to exclude sentences containing the words "gene," "express," and "encode," improved the accuracy to 110 out of 660.

In this larger set of data, protein names may co-occur in more ways that our initial approach allowed for. To reduce the error rate in these experiments, the annotator was further modified to exclude sentences which did not contain one of the verbs in Table 1, or their nominalizations. This resulted in improving the accuracy to 94 out of 437.

To further explicate the reasons for the remaining 75% apparent false positives, each relation reported was studied in each abstract where it was detected and conservatively rated either true or false. Of the 343 unmatched relations, this resulted in 140 additional relations being discovered which were not in the combined gold standard table but which were definitely reported in the abstracts. This leads to 234 out of 437 relations being discovered correctly.

To further reduce the false positives, sentences containing any negation word (see Table 3) were also excluded from consideration, as were sentences containing the word "allele." It is possible that exclusion of sentences with "not" and the like will also exclude double negatives, but we found only one such case in the entire set of candidate abstracts. This reduced the false positives to 234 out of 381. These results are summarized in Table 2.

Table 2 – Summary of precision in recognizing protein interactions under various conditions.

	Matched relations	All relations	prec
All sentences	133	912	0.14
Exclude genes	110	660	0.17
Require verbs	94	437	0.21
Discovering relations not in MIPS table	234	437	0.53
Exclude negatives, alleles	234	381	0.61
Include secondary	352	381	0.92

9. Study of Secondary Relations

A cursory study of these protein interaction relations leads to the question of whether there are clusters of protein interactions where the non-adjacent nodes can be said to be related indirectly. In the original study of the 564 MIPS Medline documents, secondary connections were frequently found to represent actual, but unobserved, interactions.

In Figure 1, we see a network of term relations around Tip20, composed of both proteins and other noun phrases.

By inspection the relations

Tip20-Ufe1p
Tip20-Sec20p

can be observed. (The figure shows specific rather than canonical protein names.)

But examining the original MIPS data, there are also interactions between

Tip20-SEC22
Sec20p-SEC22
Sec20p-Ufe1p

These additional relations can be observed as "secondary" relations or those one node distant from each other.

In a similar fashion one might ask if such secondary relations play any part in the discovery of protein interactions. Accordingly, a new database table and query were derived to determine which relations between proteins could be represented as secondary, or separated by a known node, and not reported to interact directly. In the 147 as yet unmatched relations, it was found that 117 were secondary relations. We regard these secondary relations as predictive. It is significant that creating a network of known relations and looking for correlations between these relations and ones discovered in the abstracts, that nearly three quarters if the unmatched relations are indeed just one node apart from each other in this network

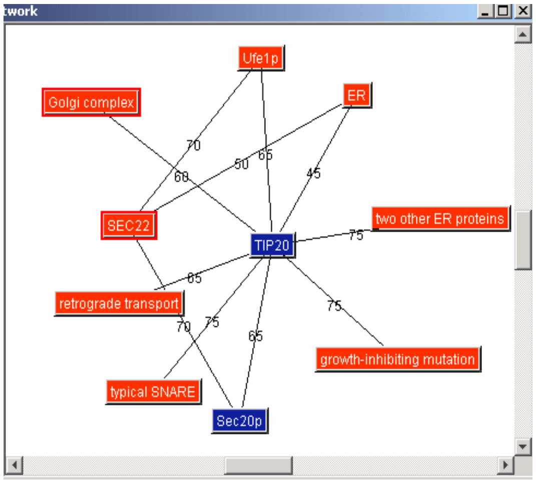

Figure 1 – A network of relations around "Tip20."

While each one of these secondary relations would have to be examined and validated independently, we note that similar assertions have been made by other workers (Blaschke and Valencia, 2001). Thus, this set of interactions separated by one node constitute at least a set of probable interactions. If this set of secondary predicted interactions is added to those tabulated and observed, the total number of correct or likely interactions is 352 out of 381, or a precision of 0.92.

10. Estimation of Recall

Recall, of course, can only be approximated in such a large collection. In the 12,300 document collection, 451 documents were returned as containing one or more of the computed interactions. In reading these documents to validate these interactions, we found only one interaction which was missed by the algorithm because it was referred to across 2 sentences and the co-reference was not resolved by this system.

It is difficult to devise a method for measuring recall when 12,000 documents constitute the sample. Thus, an experiment was devised which would return the most likely candidate documents where protein relations might have been missed. In this experiment, the verb filters (Table 1) were excluded. This approach will return documents containing at least one sentence with two proteins which does not include the word “gene.” The other exclusion terms in Table 3 were not used. This resulted in 581 documents, of which 130 were additional to the original set of 451.

These abstracts were examined in detail for the description of *any* protein interactions anywhere in the abstract, and 12 such interactions were found. Of these, 2 were discovered across sentence boundaries, requiring anaphora resolution and 2 more occurred in sentences containing the word “gene.” This means that 118/130 documents were correctly identified as having no relations, or only 12/130 contained relations, resulting in a recall of at least 90.1%. This allows us to approximate the F-measure as 0.72.

11. Mutual Information and Reliability of Protein Interaction Prediction

At the outset, it was assumed that in a large collection such as the 12,300 Medline documents analyzed in this experiment, the strength of the relation would be predictive of the likelihood that a protein interaction was taking place. Accordingly, a plot of the decile of mutual information value (Eq. 1) versus the percent of relations found to be correct is shown in Figure 2.

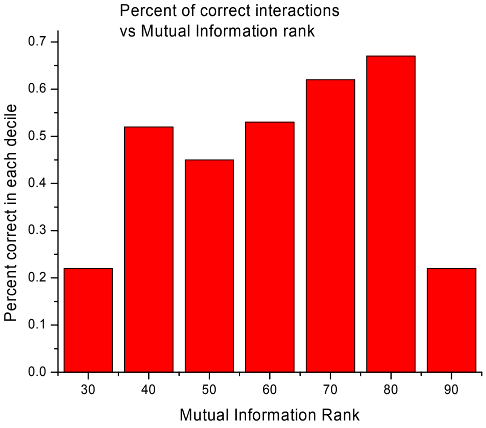

Figure 2 – Plot of mutual information decile versus percent of interactions found to be correct.

There may be no particularly strong correlation between the computed mutual information value and the correctness of the protein interaction, but there is a general upward trend from 0.40 through 0.80, but a downward trend at 0.90 which may only be related to the small umber of relations having this high mutual information value. Over all, this measure appears to be less useful than originally proposed.

gene
express
encode
no
not
fail
mRNA
transcription
allele

Table 3 – Terms that cause a sentence to be excluded from protein interaction discovery.

12. Rules Used in Finding Protein Interactions

This section summarizes the rules and techniques used in finding the protein interactions.

1. Exclude any sentence containing the words in Table 3.
2. Recognize proteins from a dictionary of proteins and their synonyms and variant spellings. Exclude all lowercase spellings, which usually represent mutations.
3. Recognize protein complexes by hyphenation.
4. Recognize protein synonyms when separated by a slash.
5. Require any sentence with two or more proteins to contain one of the verbs in Table 1.
6. Allow any sentence containing “form” and “complex” along with two or more proteins.
7. Recognize secondary interactions based on those computed as secondary from the primary table of correct interactions.

13. Summary and Conclusions

In a small set of abstracts describing protein-protein interactions, it is possible to use shallow parsing along with a dictionary, mutual co-occurrence and dimer recognition to achieve 0.92 precision and 0.84 recall (F-measure = 0.89).

In a larger set of abstracts, the primary task is filtering out sentences in documents which describe genes and other non-protein interactions. Once this is done, 61% precision is possible, and if the predictions of secondary interactions hold true, the precision reaches 92%. Based on reading of the abstracts the recall is estimated to be at least 90% The F measure is 0.72, based on a precision of 0.61. Little correlation was found between the mutual information value and the likelihood of there being a protein interaction.

These experiments result in respectable precision and considerably higher recall than previously reported methods and tend to indicate that a combination of statistical and linguistic methods can give better results than linguistic (frame based) methods alone.

14. Acknowledgements

We thank Bhavani Iyer for writing the XML extractor from our database representation, Eric Brown for the use of his DictMatcher code for detecting dictionary terms, and Bob Mack for numerous helpful discussions. The graphical layout system in Figure 1 was developed by Tunkelang (Tunkelang, Byrd and Cooper, 1997).

15. References

C. Blaschke and A. Valencia. 2001. A potential Use of SUISEKI as a Protein Interaction Discovery Tool. *Genome Informatics* 12: 123-134.

J. Pustejovsky, J. Castado, J. Zhang, M. Kotecki and B. Cochran, 2002. Robust Relational Parsin over Biomedical Literature Extracting Inhibit Relations. *Proceedings of the Pacific Symposium on Biocomputing (PSB)* 2002.

G. Leroy, *et. al.* 2003. Genescene: Biomedical Text and Data Mining. *Joint Conference on Digital Libraries,* Houston, TX, 2003.

J. Cooper and R. Byrd 1997. Lexical Navigation: Visually Prompted Query Refinement. *ACM Digital Libraries Conference*, 1998, Philadelphia, PA.

J. Cooper and R. Byrd. 1998. OBIWAN: A Visual Interface for Prompted Query Refinement. *Hawaii International Conferences on System Sciences,* 1998, Kona, HI.

D. R. Swanson. 1986, Fish oil, Reynaud's syndrome and undiscovered public knowledge. *Perspectives in Biology and Medicine* 30(1), 7-18, 1986.

R. J. Byrd and Y. Ravin, Identifying and Extracting Relations in Text, *Proceedings of NLDB 99,* Klagenfurt, Austria.

S. Fields, 2000. A network of protein-protein interactions in yeast. *Nature Biotechnology*, 18, 1257-1261, 2000.

D. Tunkelang, R.J. Byrd and J. W. Cooper, 1997. Lexical navigation: Using Incremental Graph Drawing for Query Refinement, *Graph Drawing, 1997.*

M. Neff, R.J. Byrd and B. Boguraev. 2003. The Talent System: Textract Architecture and Data Model, *NAACL Workshop on Software Engineering and Architecture of Language technology Systems*, Edmonton, Alberta, Canada, 2003.

D. Ferrucci, and A. Lally. 2003. Accelerating Corporate Research in the Development, Application and Deployment of Human Language Technologies, *NAACL Workshop on Software Engineering and Architecture of Language Technology systems*, Edmonton, Alberta, Canada, 2003.

J. Prager 1999. Linguini: Recognition of Language in Digital Documents, *Proceedings of the 32nd Hawaii International Conference on System Sciences*, Wailea, HI, January, 1999.

C. Blaschke, M. Andrade, C. Ouzounis, and A Valencia, 1999. Automatic extraction of biological information from scientific text. *International Conference on Intelligent Systems for Molecular Biology,* Heidelberg, 1999.

N. Daraselia, A Yuryev, S. Egorov, S. Novichkova, A. Nikitin and I Mazao. 2004. Extracting human protein interactions from Medline using a full-sentence parser. *Bioinformatics* 20(50 604-611, 2004.