RC23278 (W0407-053) July 23, 2004
Computer Science

IBM Research Report

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

James W. Cooper, Aaron Kershenbaum
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, Statistical and Graphical Information

James W. Cooper and Aaron Kershenbaum

Unstructured Information Management,

IBM T J Watson Research Center

P.O. Box 704

Yorktown Heights, NY 10598

jwcnmr@watson.ibm.com, aaronk@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 two-hop and three-hop relations discovered by graphical analysis, the precision could be extended to 0.82.

Keywords

E.1.d Graphs and networks; H.28.l Text mining; H.3.1.d Linguistic processing; H.3.4.c Information networks; H.2.8.a Bioinformatics (genome or protein) databases.

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 [1] 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* [2] described a computational linguistic system for detecting *inhibit* relations, with 90% precision and recall of 57%.

Recently Leroy [3] 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 [14] have quite recently described a system called MedScan, which they report as having 91% precision and 21% recall on human protein-protein interactions.

We [4] have previously described methods for detecting relations between noun phrases and methods for displaying them [5]. 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” [6]. 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 our 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 [10]. The current version of this system operates in the Unstructured Information Management (UIMA) environment [11]. 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 TALENT 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 database 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 [12] 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 [7] 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, paragraph, sentence and offset 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 a selectable number of tokens apart. We chose a maximum separation value of 30 empirically.

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 reduced to 959 unique relations, and the protein names and the 564 Medline abstracts downloaded.

An annotator was then developed that compared each lexical token found by TALENT 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 [1].

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 dimers 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 13].

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 preliminary experiment 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 [8]. 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. These new 140 relations were added to the "true relations" table in the experiments that follow.

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 239 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	239	381	0.62

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

gene
express
encode
no
not
fail
mRNA
transcription
allele

9. Study of Secondary Relations in the Preliminary Dataset

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 Figure 1, we see a network of term relations around Tip20, composed of both proteins and other noun phrases. The numbers separating the nodes represent the scaled strength of the relation based on the mutual information computation discussed earlier.

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.

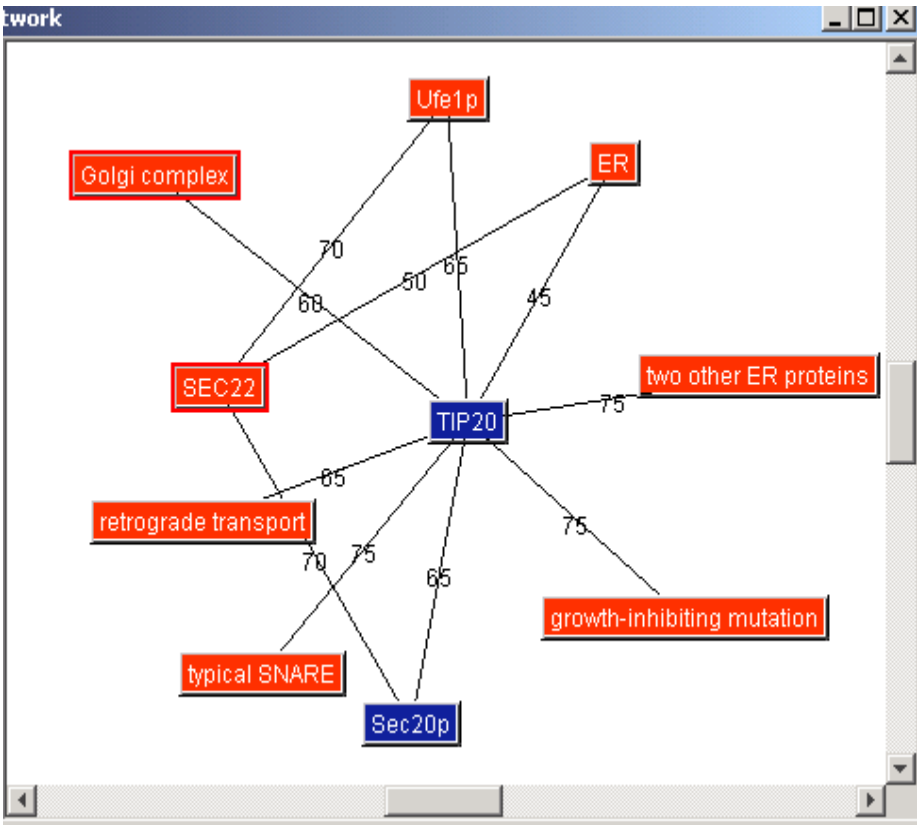

Figure 1 – A network of relations around "Tip20."

10. Graphical Study of Secondary Relations in the Large Dataset

Accordingly, we undertook a study of the graphical relations between proteins, in a similar fashion to that described by Jeong [15]. In this study, we looked at two networks, one of the "true" relations described by the combined table and one described by the network of relations we discovered by text analytic methods. The true relations graph was, of course, larger than the one we mined, and for comparison purposes, we reduced the true relations graph to contain only the nodes found in the experimental data.

In our experimental data, as noted above, we found 385 interactions of which 239 were confirmed by the combined true relations table, while 146 were not, for a precision of 62%. These 385 interactions were among 266 proteins. However, our true relations table contained

only 246 of these proteins. Of the 385 interactions found by our approach, 42 involved one of the 20 proteins not part of the true relations table. If we consider only interactions over the 246 proteins common to both tables, we find that 239 of 343 match and 104 do not, for a precision of 70%.

In examining these two networks, we built a graph corresponding to each interaction found by our approach but not present in MIPS. We then compared the data to find out if relations which were not directly tabulated in the true relations graph but were found in the experimental data could be explained by indirect relations. For example, in Figure 2, there is no direct relationship between Ypt1 and Bet2 in the true relations network. However, our experiments discovered such a relationship, and from examination of Figure 2, it is apparent that there is strong support for this relation. There are relations between Ypt1 and Sec4, Bet2 and Sec4, Bet2 and Mad2 and Mad2 and Sec4. Thus, there is a path of length 2 (Bet1-Sec4-Ypt1) and a path of length 3 (Bet2-Mad2-Sec4-Ypt1) between Ypt1 and Bet2. This lends considerable support for the relationship between Ypt1 and Bet2.

If we then return to our database of computed relations, the document containing this relation is abstract 1903184, and the supporting text for this relation is:

"We propose that Bet2 modifies Ypt1 and Sec4 in an analogous manner."

Thus, our graphical analysis method discovered an actual relation missed by our text mining system. In this case, it was missed because the verb "modifies" was not one of those we selected, as shown in Table 1.

More formally, given a interaction between two proteins, P and Q, we define a neighborhood graph, GN(P,Q), as follows:

1. First, we form 2 graphs, GA and GM, with edges corresponding to interactions found by our algorithm and in MIPS, respectively. The nodes in each graph are the interacting proteins.

2. We merge GA and GM into a single graph, GT. Because we are looking for support for interactions in GA, we use the nodes (proteins) in GA as the nodes of GT. We annotate the nodes to indicate which were present only in GA and which were present in both GA and GM. We include all edges from GA and all edges from GM both of whose endpoints have been included in GT. We annotate the edges to indicate which were present only in GA, which were present only in GM and which were present in both.

3. For each P and Q related by our approach but not present in MIPS, we define a neighborhood graph GN(P,Q) as a subgraph of GT. The nodes of GN(P,Q) are all nodes (proteins) which have relations with P or Q (or both) in GA or GM (or both) and which also are nodes in GT (i.e., which appear in GA). The edges of GN(P,Q) are all edges in GA or GM (or both) who endpoints are included in GN(P,Q).

We then analyze the cohesion of GN(P,Q) for each P and Q and collect statistics on the cohesion. (For a thorough description of graph representation and algorithms for analyzing cohesion, see [16] .)

The cohesion of a graph or subgraph is defined as the ratio of the number of edges present to the possible number of edges. In the case of a single node, n, in an undirected graph, if the degree (number of incident edges) of n is d we define the neighborhood of n as the set of nodes including the endpoints of these *d* edges, and all edges whose endpoints in this node set. Say there are e such edges. The cohesion, *C(n*) is then defined in Equation (2).

$$C(n) = \frac{e}{d(d+1)/2} \quad (2)$$

In this paper, we are analyzing the cohesion of a subgraph defined over the union of the neighborhoods of two nodes, specifically P and Q above. There are also three types of edges in this graph. There are thus many possible definitions of cohesion. For simplicity, we take the conservative approach of only considering 2 and 3 hops paths (i.e., paths between P and Q which contain 2 or 3 edges). This ignores some longer paths which could support the interactions were found, but leads to a clearer picture of what is happening.

In the 104 protein interactions found by our method, but not in the combined true relations table, 32 are related by at least one 2-hop path, 35 are related by at least one 3-hop path and 41 are related by at least one 2-hop or 3-hop path. If we accept the 41 indirectly supported interactions, in addition to the 239 present in our combined true relations table, we find that we have 280 of 343 “correct” interactions, giving a precision of 82%.

Our method found 343 interactions over 246 proteins. The true relations table contained 396 interactions over these same 246 proteins. Thus, of the (246 x 245)/2 = 30,135 pairs of proteins, there are only interactions tabulated between 396/30,315, or 1.2% of them, and only 1286 or 4.2% are connected by 2-hop or 3-hop paths. Also, in most of the 41 cases where the interactions we found were supported by these paths, more than one such path was found. Specifically, in 19 of the 32 cases where 2 hop paths were present, more than one path was present, with the average being 2.3, and similarly, 25 of the 35 occurrences of 3-hop paths were multiple occurrences, with the average being 4.1. Thus, the assumption that these paths support new interactions found by our method seems statistically persuasive. These results are summarized in Table 4.

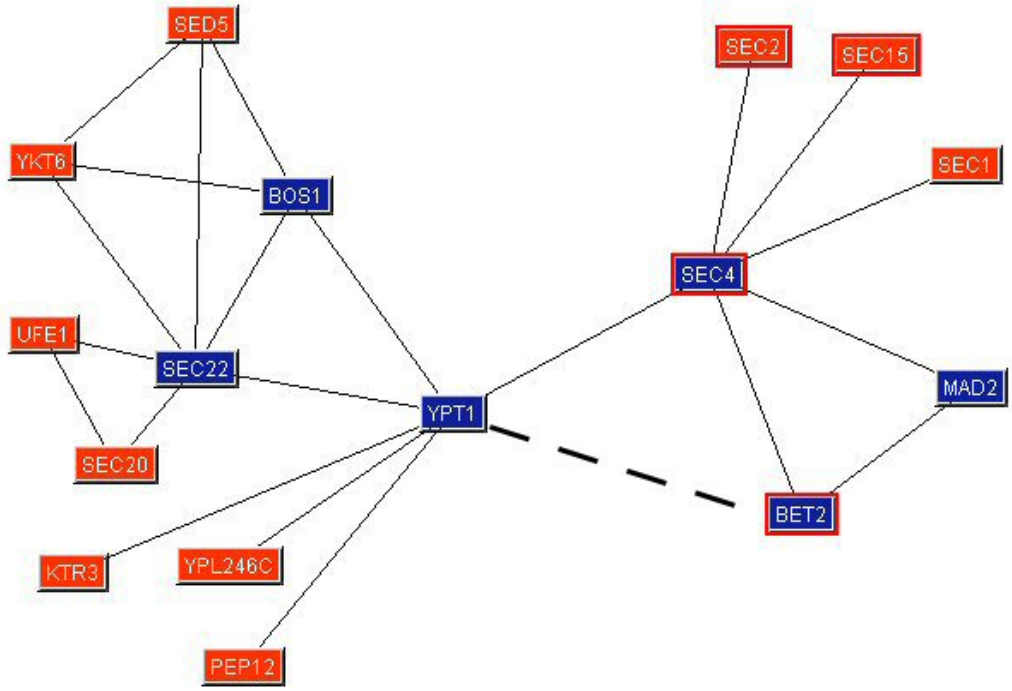

Figure 2 – The true relations network around the discovered YPT1-BET2 relation.

Table 4 – Results of including network analysis data

	Matched relations	All relations	prec
Exclude negatives, alleles (from Table 2)	239	381	0.62
Exclude 20 proteins that are not part of true relations table.	239	343	0.70
Include 2-hop and 3-hop relations.	280	343	0.82

11. Validation of Graphical Computations

Computation of all relations having 2-hop and 3-hop paths which do not have direct reported interactions gave 30 relations deduced from 2-hop paths and 60 relations deduced from 3-hop paths. Of the 2-hop path relations, 15 (50%) of them were found to be true by examination of the text of the abstracts. A similar proportion was found for the 3-hop paths.

12. 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.

We note that is an extremely conservative measure of recall, since the sample selected for detailed analysis is *less* likely to contain correct relations, because we did not carry out all the exclusions noted in Table 3

13. 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 3.

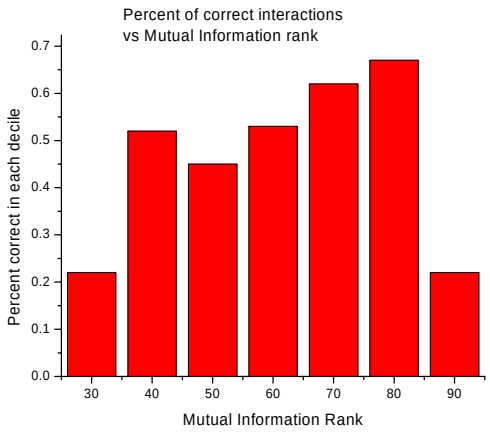

Figure 3 – Plot of mutual information decile versus percent of interactions found to be correct.

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 number of relations having this high mutual information value. Over all, this measure appears to be less useful than originally proposed.

14. 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 found by 2-hop and 3-hop connections in the primary table of correct interactions.

15. 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 82%. 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, and is 0.85 based on the precision of 0.82. There is some possible correlation 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.

Finally, we note that there is apparently no "silver bullet" to improve detection of protein-protein relations. Instead, the process is one of incremental improvement based on rules and filters of data. However, the set of rules we report here appear to have the highest F-measure yet reported.

16. 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).

17. References

1. C. Blaschke and A. Valencia. 2001. A potential Use of SUISEKI as a Protein Interaction Discovery Tool. *Genome Informatics* 12: 123-134.
2. 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.
3. G. Leroy, *et. al.* 2003. Genescene: Biomedical Text and Data Mining. *Joint Conference on Digital Libraries*, Houston, TX, 2003.

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

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

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

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

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

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

10. 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.

11. 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.

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

13. 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.

14. 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(5) 604-611, 2004.

15. H. Jeong, S. Mason, A. Barabási and Z. Oltvai, Centrality and Lethality of Protein Networks, *Nature* 411, 41-42, 2001.

16. *Introduction to Algorithms*, T. Cormen, C. Leiserson and R. Rivest, MIT Press, 1990