

# **Corpus Properties of Protein Interaction Descriptions**

**D. Berleant**\*

Department of Electrical and Computer Engineering  
Iowa State University, Ames, Iowa 50011, USA  
berleant@iastate.edu

**J. Ding**

Department of Electrical and Computer Engineering  
Iowa State University, Ames, Iowa 50011, USA

**D. Nettleton**

Department of Statistics  
Iowa State University, Ames, Iowa 50011, USA

\*To whom correspondence should be addressed

**Running head: Corpus Properties of Protein Interactions**

# Corpus Properties of Protein Interaction Descriptions

Daniel Berleant, Jing Ding, and Dan Nettleton

Iowa State University, Ames, Iowa 50011, USA

{berleant, dingjing, dnett}@iastate.edu

## Abstract

*Motivation:* Computer processing of the biomedical “literaturome” is expected to boost efforts to understand and control biological processes. An important aspect of this task is unearthing interactions from biomedical texts. While a body of work exists describing such efforts, current understanding of the evidence that specific sentence attributes provide regarding protein interactions is still meager. Because of the potential value of such knowledge to system builders, we empirically investigated a corpus of MEDLINE abstracts to determine the abilities of several prominent sentence attributes to predict whether protein interactions are described. The investigation also serves as a model that may be useful for obtaining analogous results for other kinds of interactions, such as occur among diseases, drugs, genes, proteins, and other entities. Knowledge of interactions in turn can support a wide range of practical and theoretical goals.

*Results:* This paper contributes to understanding of how passage attributes relate to automatic identification and extraction of protein interactions from texts. The results provide data of potential use in designing biomedical interaction mining systems.

*Availability and supplementary information:* The list of PMIDs of the MEDLINE records in the corpus we used, and the two databases derived from the corpus, in Access format, are downloadable from <http://class.ee.iastate.edu/berleant/s/IEPA.htm>. *Contact:* berleant@iastate.edu.

## 1. Introduction

The emergence of easily accessed, digital texts in large quantities has led to growing interest in automated mining of these texts for useful biological facts. The best-known of the classical work on statistical properties of words and their frequencies is that of George Zipf (1935, 1945a, 1945b), the source of what is now called Zipf’s Law (Fedorowicz 1982; Li 2003). Word presence and frequency are the basis of the *term frequency* (TF) and *inverse document frequency* (IDF) concepts (e.g. Salton and Yang 1973), still widely used in information retrieval. The field of information extraction is a natural conceptual extension to information retrieval and in pure form involves mining texts to fill blanks in fact templates (Rau et al. 1989; Cowie and Lehnert 1996 give a review). In biomedical text mining, techniques have been described for analyzing texts in units ranging in size from individual terms (some recent papers appear in the Pacific Symposium on Bio-computing 2003 proceedings) up to sets of documents (e.g. Shatkay et al. 2000).

A number of reports involve mining of protein-protein interactions from text (Blaschke et al. 1999, Donaldson et al. 2003, Humphreys et al. 2000, Marcotte et al. 2001, Ng and Wong 1999, Ono et al. 2001, Park et al. 2001, Thomas et al. 2000, Wong 2001). Available empirical information on text attributes as evidence that a passage describes a protein-protein interaction is not extensive. However a number of works with related foci describe some findings. Craven and Kumlien (1999) give a list 20 word stems and the ability of each to predict that a sentence describes the subcellular location of a protein, given that it contains the stem, a protein name, and a subcellular location term. However list content and order are noisy due to limited training data. Marcotte et al. (2001) provide a ranked list of the 20 words found most useful in identifying abstracts describing protein interactions. Results were derived from yeast-related abstracts and therefore may be yeast-specific, and the list includes words like *from* and *required* with little comment. Ono et al. (2001) assessed the abilities of four common interaction-indicating terms, each associated with a custom set of templates, to detect descriptions of protein-protein interactions. The quantitative performances of the four are hard to interpret because each used a different template set, but it is interesting that when ordered by precision their order was the same for both the yeast and *E. coli* domains, suggesting domain independence for precision. Thomas et al. (2000) proposed four categories of passages using a rule-based scoring strategy, and gave the IR performance of each category. However the set of rules is vaguely described and apparently complex, making it unclear how the results might be applied by others.

Other reports have focused on text properties with the potential for more concrete guidance in system design. Sekimizu et al. (1998) measured the IR performances of 8 interaction-indicating verbs in the context of a shallow parser. The IR capabilities of the verbs could be meaningfully compared, although the extent to which these results would apply to other passage analysis techniques or specifically to protein-protein interactions is not clear. Ding et al. (2002) found that, as vehicles for describing protein-protein interactions, sentences had slightly higher IR effectiveness than phrases despite lower precision, and considerably higher IR effectiveness than whole abstracts. Ding et al. (2003) applied an untuned link grammar parser to sentences containing protein co-occurrences, and found that the presence of a link path as a retrieval criterion raised the IR effectiveness by 7% (5 percentage points).

Research that identifies the abilities of passage attributes to predict whether interactions are described is important because of how it can be used. This includes assisting builders of interaction mining systems, because machine learning as typically done in such systems requires, as input, attributes of passages that are useful in passage classification. It also includes direct application of the knowledge without machine learning to systems that mine the kind of information from the kind of passages to which the knowledge applies. Thirdly, it includes screening and ranking passages for input to human curators of interactions from texts (e.g. Usuzaka et al. 1998, Yeh et al. 2003) for interaction databases (Xenarios et al. 2002, Zanzoni et al. 2002, Ding 2003). Manual curation can be so effort intensive that even modest efficiency improvements are potentially significant (Bader et al. 2003; Donaldson et al. 2003).

## 2. Methods

The IEPA corpus is a body of 303 MEDLINE abstracts chosen because they matched one of ten representative queries to PUBMED. Each query consisted of two protein names, and was elicited from biologists to be typical of the kinds of queries biologists are likely to make. Some further details about the IEPA corpus appear in Ding et al. (2002), and a list of the abstracts in the corpus is downloadable.

Each sentence in the corpus mentioning each term or its synonym in the query that retrieved its containing abstract was analyzed. Its attributes were stored in a database (downloadable – see abstract) to facilitate selecting and tabulating sentences based on their attribute values. Attributes were chosen to illuminate the following questions.

- 1) How much evidence that a sentence describes an interaction between two proteins is provided by the *presence* of verbs or other terms suggesting an interaction?
- 2) Given two protein names and an interaction-suggestive term, how does their collective presence *within* the bounds of a single phrase affect the likelihood that they are described as interacting compared to their presence *across* phrases but within the bounds of a single sentence?
- 3) How does the *order of appearance* of two protein names and an interaction-suggestive term affect the probability that they are described as interacting?
- 4) How does the number of words *intervening* between two protein names affect the probability that they are described as interacting?
- 5) How does whether protein names are in a *subject-object relationship* affect the likelihood that they are described as interacting?

As stated, these questions make intuitive sense but lack the preciseness needed to unambiguously select and tabulate sentences. Therefore we define some terms next to avoid ambiguity later.

*Query term occurrence.* An instance of a query term or its synonym that is not followed by the word “channel,” “receptor,” or “like” and does not have the suffix “-like” is a *query term occurrence*. (Comment: references to the gene for a given protein can count as query term occurrences under this definition.)

*Phrase co-occurrence.* Recall that each abstract in the corpus contains both terms of a two-term query. If a *query term occurrence* of each is present in the same phrase, this is a *phrase co-occurrence*. (Comment: if a phrase has  $m$  query term occurrences of one query term and  $n$  of the other, it has  $n*m$  phrase co-occurrences.)

*Sentence co-occurrence.* If a *query term occurrence* of each query term is present in the same sentence, but not in the same phrase, this is a *sentence co-occurrence*. (Comment: a co-occurrence can be either a phrase co-occurrence or a sentence co-occurrence, but not both.)

*Co-occurrence.* A *phrase co-occurrence* or *sentence co-occurrence*.

*Interaction description.* If a phrase or sentence explicitly or implicitly states that the substance referenced by one *query term occurrence* affects the state of the other, or affects the activity of the other on some chemical or process, an *interaction description* involving the two is present. (Comment: an interaction description is in a phrase or sentence, and a *co-occurrence* is part of an interaction description.)

*Interactor.* A verb or other word used to assert that the terms in a *co-occurrence* do or do not interact is called an *interactor*. (Comment: in the phrase “A is an activator of B,” “activator” is an example of a non-verb interactor.)

*Sentence triple.* If a *sentence co-occurrence* is in the same sentence as an *interactor*, then those three terms are called a *sentence triple*. (If  $n$  interactors are present then the co-occurrence will be in  $n$  sentence triples.)

*Phrase triple.* If a *phrase co-occurrence* is in the same phrase as an *interactor*, then those three terms are called a *phrase triple*.

*Triple.* A *sentence triple* or *phrase triple*.

In addition we define sentences, phrases, and passages as follows.

*Sentence.* Either an article title, or a passage beginning and ending with (but not containing) the pattern `<period><whitespace><capital letter>`. In addition, the first non-title sentence (which does not begin with that pattern) and the last sentence (which does not end with that pattern) are each considered a *sentence*.

*Phrase.* A passage that occurs inside a *sentence*, and begins and ends with a bound, where bound ::= `, | ; | : | . | <the beginning of the sentence> | <whitespace>-<whitespace> | ( | )` (exception: a phrase can surround but not include another parenthesized phrase).

*Passage.* A *phrase* or *sentence*.

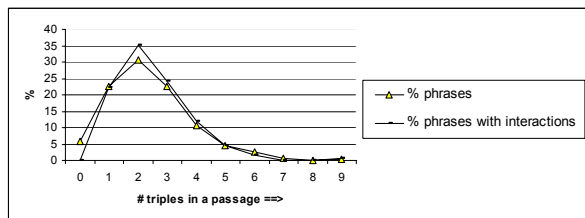
Selected uses of these terms are italicized below (and above) to emphasize their present technical meanings.

### 3. Detailed Data Analyses

This section addresses the questions listed in Section 2.

#### 3.1 Interactors: presence, absence, and quantity

For a given *co-occurrence* there can be multiple *triples* (i.e. multiple *interactors*) present. The great majority of co-occurrences were part of at least one triple (Figure 1). The ones that were not had low precisions (8% for *sentence co-occurrences* and 0% for *phrase co-occurrences*). This was significantly lower than for co-occurrences that were part of at least one triple ( $p < .001$ ,  $\chi^2$  test, for both sentence co-occurrences and phrase co-occurrences). Thus as a source of interactions to be mined, co-occurrences not in triples, i.e. without associated interactors, have comparatively little to offer.



**Figure 1. Percentages of the 285 1-co-occurrence phrases containing 0, 1, 2,... phrase triples. 199 contained interaction descriptions. For sentences (not shown), 76% contained two or more interactors and of those containing an interaction description the figure was 88%.**

**Effect of the number of triples associated with a co-occurrence.** Most passages had only one co-occurrence. Since those passages are likely to be simpler and thus easier to automatically analyze than passages with multiple co-occurrences, they are of particular interest for system design. However, a co-occurrence can be associated with multiple triples (Figure 1), one for each interactor present. Usually at most one of these triples is in an interaction description. Thus the mere presence of an interactor with a co-occurrence is usually not compelling evidence of the type of interaction because other interactors are probably also present.

If recall is important a system will need to mine interaction information from passages in which the co-occurrence of interest is accompanied by multiple interactors, because multiple interactors are so often present. Extracting the type of interaction will generally require determining which interactor applies, which in turn is likely to require deeper analysis. However, the need may be to identify that an interaction is present without identifying its type. This would be the case, for example, in a system that ranks passages for use by human curators.

### Multiple co-occurrences in the same passage

Most phrases containing phrase co-occurrence(s) had only one. The same was true for sentences, though sentences were more likely to have multiple sentence co-occurrences (Table 2). This may be because sentences are longer and have more words of all kinds.

|                  | 1 co-occurrence | 2 co-occurrences | Over 2 co-occurrences |
|------------------|-----------------|------------------|-----------------------|
| <b>Phrases</b>   | 91% (285)       | 8% (24)          | 1.6% (5)              |
| <b>Sentences</b> | 74% (156)       | 18% (39)         | 7.6% (16)             |

Table 2. Data showing that most passages have just one co-occurrence.

Sentences tend to be syntactically more complex than phrases. Also, it may be harder to unravel the interaction(s) of a pair of biochemicals from passages with multiple co-occurrences compared to those with a single co-occurrence. Thus interaction mining system designers may decide to condition whether and how a system will analyze a passage based on whether it is a phrase or a sentence, and whether it has one or multiple co-occurrences. For such decisions Table 3 may be useful. It shows, for different categories of passage, how good a source of interactions it is. The Table shows that 1-co-occurrence phrases have by far the most interaction descriptions, while 1-co-occurrence sentences are next. System design might also be impacted by the observation that sentences with multiple co-occurrences form the category with the lowest recall, meaning that the category that is presumably the most complicated to analyze is also the category with the least to contribute.

| Interaction descriptions in... | 1 co-occurrence                             | >1 co-occurrence                          | Total   |
|--------------------------------|---|---|---|
| ...phrases                     | 256 in 286 phrases<br>$r=0.62, p=0.90$      | 41 in 29 phrases<br>$r=0.10, p=1.41^*$    | <b>297 in 315 phrases</b><br>$r=0.72, p=0.94$   |
| ...sentences                   | 75 in 167 sentences<br>$r=0.18, p=0.45$     | 38 in 53 sentences<br>$r=0.092, p=0.72$   | <b>113 in 220 sentences</b><br>$r=0.28, p=0.51$ |
| ...both                        | <b>331 in 453 cases</b><br>$r=0.81, p=0.73$ | <b>79 in 82 cases</b><br>$r=0.19, p=0.96$ | <b>410 in 535 cases</b><br>$r=1.0, p=0.77$      |

Table 3. Co-occurrence categories and the recall  $r$  and precision  $p$  of each with respect to protein-protein interactions. \* A passage can contain more than one interaction description.

### 3.2 Co-occurrences: term separation, term order, and subject-object relationship

Every co-occurrence has, among its attributes, the separation of its two query term occurrences, whether an interactor intervenes between them, and whether they are in a syntactic subject-object relationship.

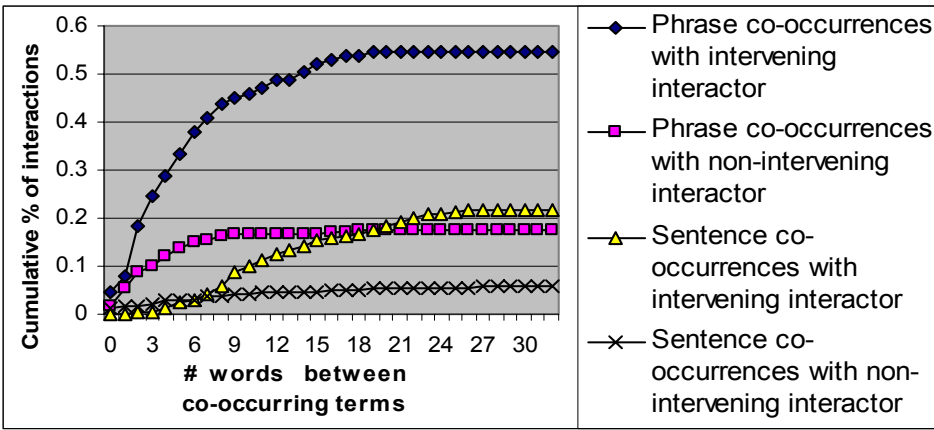
#### Term separation and term order

Co-occurring terms can be separated by any number of intervening words, from zero up. The separation can be zero when the terms are next to each other or when intervening material is hyphen-connected to a term. For example, "...A-induced B..." is considered to have zero full words between A and B. Overall, a median of seven words intervened between terms in a co-occurrence. Although occasionally there were as many as a few dozen intervening words, none of the 334 co-occurrences in an interaction description had over 31. Table 4 shows that whether an interactor intervenes between terms of a co-occurrence or instead is present outside of it has a major effect on both recall and precision. The recall figures suggest the importance for a system design to handle cases where the interactor intervenes, and the precision figures suggest the relative difficulty of extracting interactions when the interactor is not intervening, an attribute also associated with comparatively low recall and hence mining potential.

|                                | Interactor intervening     |                            | Interactor elsewhere       |                            | Interactor anywhere        |                            |
|--------------------------------|----------------------------|----------------------------|----------------------------|----------------------------|----------------------------|----------------------------|
| <b>Phrase co-occurrences</b>   | $r=0.55$                   | $p=0.63$                   | $r=0.18$                   | $p=0.24$                   | <b><math>r=0.72</math></b> | <b><math>p=0.45</math></b> |
| <b>Sentence co-occurrences</b> | $r=0.22$                   | $p=0.30$                   | $r=0.058$                  | $p=0.091$                  | <b><math>r=0.28</math></b> | <b><math>p=0.21</math></b> |
| <b>All co-occurrences</b>      | <b><math>r=0.77</math></b> | <b><math>p=0.48</math></b> | <b><math>r=0.23</math></b> | <b><math>p=0.17</math></b> | <b><math>r=1</math></b>    | <b><math>p=0.34</math></b> |

Table 4. Analysis of (i) the recalls of co-occurrence categories with respect to mining interaction descriptions, representing their potentials as sources for mining, and (ii) the precisions of the categories, representing their richness as sources for mining.

Figure 3 shows percentages of co-occurrences (vertical axis) that have  $x$  words or fewer between the co-occurring terms, for each of four disjoint categories. The two curves for phrase co-occurrences level off sooner than the two for sentence co-occurrences, probably because phrases tend to be shorter than



sentences. The implication for system design is that there is little reason to look for phrase co-occurrences separated by over about 20 words, or sentence co-occurrences separated by much more than that.

Figure 3. Cumulative interactions for four disjoint categories of co-occurrences.

### Precision vs. separation of co-occurring terms

Mining systems could use the number of intervening words to help estimate the likelihood that an interaction description is present, if these separations are associated with different precisions. Thus we plotted the precisions for different separations (Figure 4), for each of the four co-occurrence categories of Figure 3. Because the number of data for any given separation was often small, the graph is noisy. However, a tendency for precision to decrease as the separation increases is evident (the effect is visually more apparent if one ignores plotted precision values of 0 or 1, many of which are based on a single datum and therefore forced to be 0 or 1.)

We used the logistic regression model to describe precision  $p$  as a function of separation  $x$ . This implies the equation  $\log(p/(1-p))=a+bx$  where  $a$  and  $b$  are unknown parameters estimated separately for each co-occurrence category. For a given separation, the number of co-occurrences in a given category was assumed to follow a binomial distribution with success probability equal to its precision  $p$ . Maximum likelihood methods (Agresti 1990 pp. 112-117) were used to obtain estimates and approximate standard errors. The estimated curves along with error bars representing plus and minus one standard error on the  $\log(p/(1-p))$  scale appear in Figure 4. Many of the points displayed in Figure 4 are based on only one or a few co-occurrences, especially for larger separations, so there is substantial uncertainty in the estimated logistic regression curves. For instance, the estimated curve for phrase co-occurrences without an intervening interactor rises, but its slope coefficient  $b$  is not significantly different from zero, and the true curve could actually be decreasing. The other three curves could be used as sources of evidence by an interaction mining system, while further investigation would enable using this statistical approach to estimate all four curves as accurately as desired.

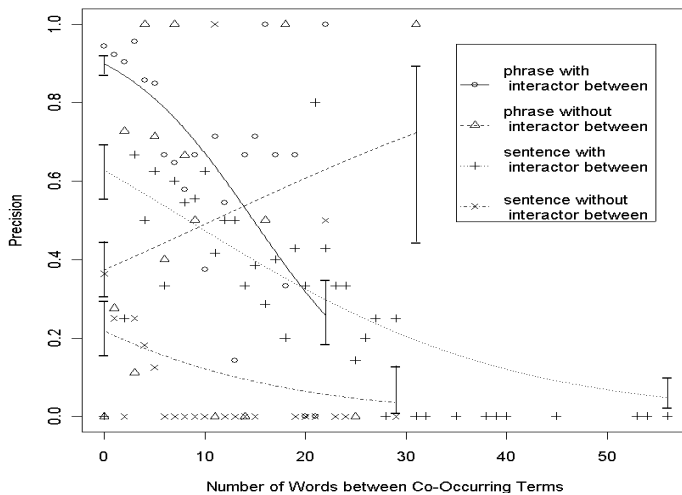


Figure 4. Precision data for four categories of co-occurrences and various separations.

## Subject-object vs. other syntactic relationships

System designs that perform syntactic analyses of passages, as systems reported in the most recent literature usually do, may benefit from data on the kinds of syntactic structures used to describe interactions. Table 5 provides some information of that type (see also Ding et al. 2003).

Table 5 (bottom row) shows that over  $\frac{3}{4}$  of interaction descriptions had one *query term occurrence* as the subject and the other as the object of an interaction-indicating verb phrase. However here again a sizeable fraction of over 20% did not, rising to nearly half (0.13-0.14 out of 0.28-0.29) of interaction descriptions that crossed phrase boundaries (middle row). An example of how a textual structure can describe an interaction between biomolecules A and B without using a subject-object term relationship is, “A activates X, which inhibits B.”

|                         |  | Terms in co-occurrence have an S-O relationship ( $p=1$ ) |             | Terms in co-occurrence have a non-S-O relationship |              |             | Either S-O or not S-O |             |             |
|-------------------------|--|---|-------------|--|--------------|-------------|-----------------------|-------------|-------------|
|                         |  | <i>n</i>  | <i>r</i>    | <i>n</i>   | <i>p</i>     | <i>r</i>    | <i>n</i>              | <i>p</i>    | <i>r</i>    |
| Within a phrase         | Triples                                | 260   | 0.63        | 37   | 0.064        | 0.090       | <b>297</b>            | <b>0.35</b> | <b>0.72</b> |
|                         | co-occurrences                         | 207   | 0.62        | 31   | 0.21         | 0.093       | <b>237</b>            | <b>0.68</b> | <b>0.71</b> |
|                         | phrases                                | 198   | 0.65        | 28   | 0.24         | 0.092       | <b>224</b>            | <b>0.72</b> | <b>0.74</b> |
| Across phrases          | Triples                                | 61  | 0.15        | 52   | 0.056        | 0.13        | <b>113</b>            | <b>0.11</b> | <b>0.28</b> |
|                         | co-occurrences                         | 51  | 0.15        | 46   | 0.19         | 0.14        | <b>97</b>             | <b>0.33</b> | <b>0.29</b> |
|                         | sentences                              | 44  | 0.15        | 40   | 0.23         | 0.13        | <b>84</b>             | <b>0.38</b> | <b>0.28</b> |
| <b>Within or across</b> | <b>Triples co-occurrences passages</b> | <b>321</b>  | <b>0.78</b> | <b>89</b>  | <b>0.059</b> | <b>0.22</b> | <b>410</b>            | <b>0.22</b> | <b>1</b>    |
|                         |  | <b>258</b>  | <b>0.77</b> | <b>77</b>  | <b>0.20</b>  | <b>0.23</b> | <b>334</b>            | <b>0.52</b> | <b>1</b>    |
|                         |  | <b>238</b>  | <b>0.79</b> | <b>68</b>  | <b>0.27</b>  | <b>0.22</b> | <b>303</b>            | <b>0.62</b> | <b>1</b>    |

**Table 5. Triples, co-occurrences, and passages associated with interaction descriptions. Recalls (*r*) are relative to the set of all interactions between *query term occurrences* described within the bounds of a single sentence.**

### 3.3 Comparison of triples, co-occurrences, and passages

System designers must choose whether the structure to be tagged as associated with an interaction is the triple, the co-occurrence, the passage, or something else. Table 5 allows a comparison of these. Some of the bolded numerical totals (*n*) are smaller than the sum of their unbolded constituents. This is because the same co-occurrence or passage can be counted in multiple unbolded cells of the table, as a result of multiple triples associated with one co-occurrence or multiple co-occurrences associated with one passage.

All entries in the column for S-O relationships have precisions of 1 because classification of a case into that column means the two query term occurrences were the subject and object of an interaction-indicating verb phrase. This, along with recall values of 0.77, 0.78 or 0.79, suggests the S-O relationship category as a target for system design. Addressing non-S-O cases as well has the potential for substantive further improvement in recall. However their more complex syntactic structure should make them more difficult to mine. Non-syntax-based methods such as template matching or term vector based analysis are simpler, but finding the type of interaction with such methods may be infeasible in general because only 0.22 of interactors (Table 5, last cell) were in *interaction descriptions*. (We did not analyze the extent to which the remaining 0.78 may have been involved in describing interactions involving non-query-related terms.) The presence but not the type of an interaction may be identified with considerable higher precision, 0.52 overall (0.72 for phrases, see Table 5 last column). Predicting the presence but not the type of interactions, which can be useful for such important applications as screening and ranking passages for examination by human curators, is thus a system design decision worth considering.

A careful examination of Table 5 raises numerical questions. One is the difference in precisions between co-occurrences and passages (0.20 vs. 0.27) which, while modest, is higher in the bottom row than in other rows. Furthermore, the precision values in the bottom row for the **passage** category (0.27 and 0.62) are higher than for the **phrase** and **sentence** categories in the top and middle rows respectively, rather than being close to their mean as might be expected. The reason is that a substantive number of passages contain two co-occurrences, one in a constituent phrase and the other across phrases, of which one is in an interaction description and the other is not. These reduce the pool of passages without any interaction description at all, thereby raising the passage precision in the bottom row. The difference between 0.20 and 0.27 (last row, middle column) may appear smaller than the difference between 0.52 and 0.62 (last row, last column) but as might be expected is actually proportionately larger:  $(0.27 - 0.20) / 0.20 > (0.62 - 0.52) / 0.52$ .

#### 4. Discussion and Conclusion

Properties of phrases and sentences containing co-occurring protein names were presented. These properties may be useful in designing systems that mine protein interactions from text. Systems that perform sophisticated parsing could use statistical properties as a complementary source of information; template and term occurrence based systems could use the results more directly, and the complex decisions involved in system design could be impacted by results such as are given here in addition to the intended application.

An important issue in any study of this type is the representativeness of the analyzed corpus, because the general applicability of the findings depend on its being representative of the overall body of literature. This is difficult to resolve definitively, however the corpus we used (described in more detail in Ding et al. 2002) was designed to be representative of the interests of biological researchers. Nevertheless, no generic corpus can be assumed to be representative of any specific body of texts of interest to a given researcher or for a given purpose, as that set itself might not be representative of the overall literature.

Serious shortcomings in capabilities of NLU help motivate statistical results as a resource for practical system development. Even if complete NLU should some day be achieved, statistical results such as those presented here may still contribute to system accuracy, efficiency, or both.

#### References

- Agresti A (1990), *Categorical Data Analysis*, Wiley, New York.
- Bader GD, D Betel, and CW Hogue (2003), BIND: the Biomolecular Interaction Network Database, *Nucleic Acids Research* **31** (1):248-250.
- Blaschke C, M Andrade, C Ouzounis, and A Valencia (1999), Automatic extraction of biological information from scientific text: protein-protein interactions, *AAAI Conference on Intelligent Systems in Molecular Biology*, 60-67.
- Cowie J and W Lehnert (1996), Information extraction, *Communications of the ACM* **39** (1):80-91.
- Craven M and J Kumlien (1999), Constructing biological knowledge bases by extracting information from text sources, *Proc. 7<sup>th</sup> Int. Conf. on Intelligent Systems for Molecular Biology (ISMB)*, 77-86.
- Ding J (2003), PathBinder: a sentence repository of biochemical interactions extracted from MEDLINE. Master's thesis, Dept. of Electrical and Computer Engineering, Iowa State University. The PathBinder repository is at <http://www.vrac.iastate.edu/~berleant/MedRep/>.
- Ding J, D Berleant, J Xu, and AW Fulmer (2003), Extracting biochemical interactions from MEDLINE using a Link Grammar parser, *Proc. Fifteenth IEEE Conference on Tools with Artificial Intelligence (ICTAI)*, Sacramento, 467-471.
- Ding J, D Berleant, D Nettleton, and E Wurtele (2002), Mining MEDLINE: abstracts, sentences, or phrases? *Pacific Symposium on Biocomputing* **7** (PSB), Kaua'i, Hawaii, 326-337.
- Donaldson I, J Martin, B de Bruijn, C Wolting, V Lay, B Tuekam, S Zhang, B Baskin, GD Bader, K Michalickova, T Pawson, and CW Hogue (2003), PreBIND and Textomy – mining the biomedical literature for protein-protein interactions using a support vector machine, *BMC Bioinformatics* **4** (11), [www.biomedcentral.com/1471-2105/4/11](http://www.biomedcentral.com/1471-2105/4/11).
- Fedorowicz J (1982), A Zipfian model of an automatic bibliographic system: an application to MEDLINE, *J Am Soc Inf Sci.*, **33** (4):223-232.
- Humphreys K, G Demetriou, and R Gaizauskas (2000), Two applications of information extraction to biological science journal articles: enzyme interactions and protein structures, *Pacific Symposium on Biocomputing* **5**, 502-513.
- Li W, Zipf's Law, [linkage.rockefeller.edu/wli/zipf/](http://linkage.rockefeller.edu/wli/zipf/), as of 11/6/03.
- Marcotte EM, I Xenarios, and D Eisenberg (2001), Mining literature for protein-protein interactions, *Bioinformatics* **17** (4):359-363.
- Ng S-K and M Wong (1999), Toward routine automatic pathway discovery from on-line scientific text abstracts, *Proc. 10<sup>th</sup> Workshop on Genome Informatics*, 104-112.
- Ono T, H Hishigaki, A Tanigami, and T Takagi (2001), Automated extraction of information on protein-protein interactions from the biological literature, *Bioinformatics* **17** (2):155-161.
- Park JC, HS Kim, and JJ Kim (2001), Bidirectional incremental parsing for automatic pathway identification with combinatory categorial grammar, *Pacific Symposium on Biocomputing* **6**, 396-407.
- Rau LF, PS Jacobs, and U Zernik (1989), Information extraction and text summarization using linguistic knowledge acquisition, *Information Processing and Management* **25** (4):419-428.

- Salton G and CS Yang (1973), On the specification of term values in automatic indexing, *Journal of Documentation* **29** (4):351-372.
- Sekimizu T, HS Park, and J Tsujii (1998), Identifying the interaction between genes and gene products based on frequently seen verbs in MEDLINE abstracts, *Proc. 9<sup>th</sup> Workshop on Genome Informatics*, 62-71.
- Shatkay H, S Edwards, WJ Wilbur, and M Boguski (2000), Genes, themes, and microarrays: using information retrieval for large-scale gene analysis, *8<sup>th</sup> International Conference on Intelligent Systems in Molecular Biology (ISMB)*.
- Thomas J, D Milward, C Ouzounis, S Pulman, and M Carroll (2000), Automatic extraction of protein interactions from scientific abstracts, *Pacific Symposium on Biocomputing* **5**, 538-549.
- Usuzaka S, KL Sim, M Tanaka, H Matsuno, and S Miyano (1998), A machine learning approach to reducing the work of experts in article selection from database: a case study for regulatory relations of *S. cerevisiae* genes in MEDLINE, *Proc. 9<sup>th</sup> Workshop on Genome Informatics*, 91-101.
- Wong L, (2001), A protein interaction extraction system, *Pacific Symposium on Biocomputing (PSB)* **6**.
- Xenarios I, L Salwinski, XJ Duan, P Higney, S-M Kim, and D Eisenberg (2001), DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions, *Nucleic Acids Research* **30** (1):303-305.
- Yeh A, L Hirschman, and A Morgan (2003), Evaluation of text data mining for database curation: lessons learned from the KDD Challenge Cup, *Proc. 11<sup>th</sup> Int. Conf. on Intelligent Systems for Molecular Biology (ISMB)*.
- Zanzoni A, L Montecchi-Palazzi, M Quondam, G Ausiello, M Helmer-Citterich, and G Cesareni (2002), MINT: a Molecular INTeraction database, *FEBS Letters* **513**, 135-140.
- Zipf GK (1935), *The Psychobiology of Language*, Houghton Mifflin.
- Zipf GK (1945), The meaning-frequency relationship of words, *The Journal of General Psychology* **33**: 251-256.
- Zipf GK (1945), The repetition of words, time-perspective, and semantic balance, *The Journal of General Psychology* **32**: 127-148.