

## MINING MEDLINE: ABSTRACTS, SENTENCES, OR PHRASES?

J. DING<sup>a</sup>, D. BERLEANT<sup>a,d</sup>, D. NETTLETON<sup>b</sup>, AND E. WURTELE<sup>c</sup>

<sup>a</sup>*Department of Electrical and Computer Engineering,*

<sup>b</sup>*Department of Statistics,*

<sup>c</sup>*Department of Botany,*

<sup>d</sup>*berleant@iastate.edu*

*Iowa State University, Ames, IA 50011, USA*

A growing body of work addresses automated mining for biochemical information from digital repositories of scientific literature such as MEDLINE. Some of this work uses abstracts as the unit of text from which to extract facts. Other work uses sentences for this purpose, while still other work uses phrases. Here, we compare abstracts, sentences, and phrases in MEDLINE using the standard information retrieval performance measures of recall, precision, and effectiveness for the task of mining interactions among biochemical terms based on term co-occurrence. Results show statistically significant differences that can impact the choice of text unit, although no one of these three text units is unambiguously superior to the others.

### 1 Introduction

The rapid growth of electronic scientific literature is providing increasingly attractive opportunities for text mining. Concurrently, text mining is becoming an increasingly well understood alternative to manual information extraction. Most reports on automatic extraction of biochemical interactions from scientific literature have used the MEDLINE repository. Mining such literature is of great potential benefit to researchers who need to efficiently sift through the literature to find work relating to specific small sets of biochemicals. Mining the literature can also address tasks of larger scope, such as inferring networks of protein interactions. While deep, fully automated literature analysis via natural language understanding (NLU) is an intriguing long-term objective, shallower and human-assisted analysis is both achievable and valuable.

The text processing units from which facts are extracted in MEDLINE mining systems may be the full abstracts, sentences, or phrases. The most basic way to “mine” MEDLINE is simply to use the PUBMED Web interface.<sup>8</sup> The user can submit a query to the database consisting of the AND of two biochemical names. Abstracts in MEDLINE containing both names are returned. Such abstracts can be used as monolithic data items in systems that automatically search for interactions among genes based on term co-occurrence within an abstract, as in Stapley and Benoit 2000.<sup>16</sup> A related approach by Shatkay *et al.*<sup>14</sup> infers functional relationships among genes based on similarities among abstracts. Neither of those works identified the type of interaction (e.g. inhibit, activate, etc.), which is desirable for

applications such as automatic construction of networks of interactions. However Thomas *et al.*<sup>18</sup> used sophisticated text processing to extract protein interactions from abstracts. Because an abstract is a relatively large processing unit which consequently contains a great deal of material besides any query terms, it is relatively difficult to automatically determine the type of interaction between the terms without methods that are sensitive to smaller structures such as sentences or phrases.

Easier inference of interaction type might be expected if retrieval is limited to cases in which the terms of interest co-occur in the same sentence (Craven and Kumlien 1999,<sup>2</sup> Dickerson *et al.* 2001,<sup>4</sup> Ng and Wong 1999,<sup>6</sup> Rindfleisch *et al.* 1999 & 2000,<sup>9,10</sup> Sekimizu *et al.* 1998,<sup>12</sup> Tanabe *et al.* 1999<sup>17</sup>), or even in the same phrase (Blaschke *et al.*,<sup>1</sup> Humphreys *et al.*,<sup>5</sup> Ono *et al.*<sup>7</sup>). But such systems will miss interactions that are described over a longer passage. For example, consider the following passage:

...in wild oat aleurone, two genes, alpha-Amy2/A and alpha-Amy2/D, were isolated. Both were shown to be positively regulated by gibberellin (GA) during germination...<sup>21</sup>

The interactions in this example (gibberellin regulates alpha-Amy2/A and alpha-Amy2/D) are described over two sentences. To extract the interactions in this example, a system needs to process text units longer than a sentence. Thus, while smaller text units might make it easier to infer interactions, they will tend to miss more interactions because some of them will be expressed over passages longer than the unit being used. Consequently recall must decrease with decreasing text unit size. Unlike for recall, a clean qualitative relationship between precision (and consequently effectiveness) and text unit size cannot be inferred from first principles.

Considerations like these revolve around the issue of what the advantages and disadvantages are of different text units from the standpoint of systems that automatically extract interactions among biochemical terms. This is important when a choice of text processing unit must be made for a text mining system design. To help support and justify text unit choices for systems that mine biochemical literature we have carried out the present investigation of the information retrieval (IR) properties of text processing units in MEDLINE. Four text units are examined: abstracts, adjacent sentence pairs, sentences, and phrases, from the perspective of three standard information retrieval (IR) measures: recall, precision, and effectiveness. Recall is the fraction of the useful items in a test set that are retrieved. Precision, in contrast, is the fraction of retrieved items that are also relevant. Effectiveness is a composite measure that weights recall and precision equally. The benefit of this exploration is better understanding of the ability of the different text units to support mining of scientific literature for interactions among biochemicals.

## 2 Experimental Procedure: The Data

To compare the merits of different text processing units, slightly over three hundred abstracts were manually analyzed. The abstracts were retrieved from MEDLINE using ten queries (Table 1) to its PUBMED interface.<sup>8</sup> Each query was the AND of two biochemical nouns. Queries were suggested by colleagues who are actively performing research in diverse biological areas, to help make them representative of the kinds of queries users of text mining systems would be interested in. A suggested query was used only if the number of abstracts retrieved by PUBMED was ten or more to facilitate statistical analysis of results. If more than 100 abstracts conforming to a given query were retrieved, only the most recent abstracts were studied, enough so that the set included approximately forty abstracts describing interaction(s) between the biochemicals in the query, plus any others that contained the biochemicals but did not describe interactions between them that were also encountered during the process of analyzing retrieved abstracts from most to less recent. Thus the ten queries yielded ten sets of abstracts, with each abstract in a set containing both terms in the corresponding query.

Although each abstract we studied contained both biochemical terms in the query, only some of them described interaction(s) between them (deemed “relevant”). An interaction between two terms was defined as a direct or indirect influence of one on the quantity or activity of the other. Examples of interactions between terms A and B include the following.

- A increased B.
- A activated C, and C activated B.
- A-induced increase in B is mediated through C.
- Inhibition of C by A can be blocked by an inhibitor of B.

The following examples do not indicate an interaction between A and B.

- A increases C, and B also increases C.
- C decreases A and B.

Below are some examples taken from MEDLINE abstracts. Only the smallest text unit containing an interaction is noted, but the interaction is also present in all larger text units as well.

**...whereas a combination of gibberellin plus cycloheximide treatment was required to increase alpha-amylase mRNA levels to the same extent.** (PMID is 10198105, query is gibberellin AND amylase, interaction is described within a phrase.)

...the regulation of hypothalamic NPY mRNA by leptin may be impaired with age. (PMID is 10868965, query is leptin AND NPY, interaction is described within a phrase.)

We investigated mechanisms underlying the control of this movement by acetylcholine using an insulinoma cell line, MIN6, in which acetylcholine increases both insulin secretion and granule movement. The peak activation of movement was observed 3 min after an acetylcholine challenge. The effects were nullified by the muscarinic inhibitor atropine, phospholipase C (PLC) inhibitors (D 609 and compound 48/80), and pretreatment with the Ca<sup>2+</sup> pump inhibitor, thapsigargin. (PMID is 9792538, query is insulin AND PLC, interaction is described within the abstract.)

An abstract was defined to consist of both title and body. A sentence pair was defined as two adjacent sentences. Each sentence therefore appeared in two sentence pairs, once as the first in the pair and once as the second. The text between two adjacent periods was defined to be a sentence. A title was also defined as a sentence, as was the body up to the first period. The text between any two punctuation marks { . : , ; }, with no intervening punctuation mark, was counted as a phrase. A title without punctuation was also counted as a phrase, as was the body of the abstract up to the first punctuation mark.

While in each abstract both members of the query occurred, in only some of the abstracts did both terms or synonyms of those terms occur within adjacent sentences. In only some of these sentence pairs did both occur within just one sentence of the pair. Finally, in only some of those sentences did both occur in a single phrase within (or comprising) the sentence.

### 3 Experimental Procedure: Measuring Information Retrieval Quality

Recall and precision measure the completeness and correctness of information retrieval, respectively. Effectiveness assesses overall performance by combining both recall and precision.<sup>15</sup> Generalized effectiveness allows the relative weights of recall and precision to be varied as a parameter in the calculation.<sup>19</sup>

In the present case, recall is the fraction of all those interactions between two biochemical terms in the corresponding set of abstracts that are stated within a sentence, phrase, or whatever text unit is under analysis:

$$\text{recall} = \frac{\text{\# of interactions between A and B occurring within a specific text unit}}{\text{\# of interactions between A and B within abstracts}}$$

where A and B are query terms or their synonyms.

Any interaction described within a particular text unit is also described within all larger text units. Since the largest unit considered here is the abstract, the recall for abstracts is exactly 1.0. Intuitively, recall here measures the capacity of a given text unit to contain the interactions present in MEDLINE abstracts.

Precision refers to the fraction of abstracts, sentences, phrases, etc. containing both biochemical terms that also describe an interaction between them. In the present case, precision is determined from the abstracts, sentences, or whatever text unit is under consideration containing both query terms. The fraction that also describe an interaction is the precision:

$$\text{precision} = \frac{\text{\#of interactions between A and B in a specific text unit}}{\text{\#of times A and B co - occur in the same text unit}}.$$

where A and B are query terms or their synonyms. Intuitively, precision here measures the quality of a given text unit as “ore” from which to mine interactions from biochemical term co-occurrences.

Effectiveness combines recall and precision with the harmonic mean (the reciprocal of the arithmetic mean of the reciprocals, suitable e.g. for calculating average travel speed of a trip):

$$\text{effectiveness} = \frac{1}{\frac{1}{2} \cdot \frac{1}{\text{recall}} + \frac{1}{2} \cdot \frac{1}{\text{precision}}} = \frac{\text{recall} \times \text{precision}}{(\text{recall} + \text{precision})/2}$$

which has recall and precision each contributing ½ of the mean. Generalized effectiveness (G) parameterizes effectiveness with a weight coefficient  $w$  specifying the relative importances of recall and precision:

$$G = \frac{1}{w \cdot \frac{1}{\text{recall}} + (1-w) \cdot \frac{1}{\text{precision}}} = \frac{\text{recall} \times \text{precision}}{w \times \text{precision} + (1-w) \times \text{recall}}, \quad 0 \leq w \leq 1.$$

Generalized effectiveness can account for differences among applications and users in their needs for recall compared to precision.

#### 4 Data Analysis

Information retrieval performances for abstracts, sentence pairs, sentences, and phrases were assessed by tabulating term co-occurrences and the subset of co-occurrences describing interactions, for each query and each text unit, then calculating the recall, precision, and effectiveness of each (Tables 1 and 2).

Table 1. Queries and the recall, precision, and effectiveness for each, given abstracts (Ab), sentences (Se), and phrases (Ph) as text units from which to extract interactions between the query terms or their synonyms in MEDLINE abstracts containing both query terms. (The last query is discussed further in Appendix A.)

Query terms	Recall			Precision			Effectiveness		
	Ab	Se	Ph	Ab	Se	Ph	Ab	Se	Ph
insulin & PLC	1	.80	.54	.38	.58	.69	.55	.68	.61
leptin & NPY	1	.88	.53	.52	.46	.53	.69	.60	.53
AVP & PKC	1	.85	.60	.83	.65	.78	.91	.74	.68
Beta-amyloid & PLC	1	.86	.71	.67	.83	.89	.80	.85	.79
prion & kinase	1	.79	.71	.70	.79	.77	.82	.79	.74
UCP & leptin	1	.96	.69	.53	.57	.73	.69	.71	.71
insulin & oxytoxin	1	.89	.65	.45	.63	.73	.62	.74	.69
gibberellin & amylase	1	.89	.71	.95	.94	.96	.97	.92	.82
oxytoxin & IP	1	.98	.80	.68	.73	.77	.81	.83	.79
flavonoid & cholesterol	1	.25	.10	.55	.50	.50	.71	.33	.17

Table 2. Information retrieval measures for different text units. Averages are over the 10 queries.

TEXT UNIT → ↓ IR MEASURE	Abstracts	Sentence pairs	Sentences	Phrases
<b>Recall</b>	1	0.90	0.81	0.61
<b>Precision</b>	0.63	0.39	0.67	0.74
<b>Effectiveness</b>	0.76	0.53	0.72	0.65

Table 2 suggests a trend of increasing precision for smaller text units, except for sentence pairs which rated poorly overall. Phrases, the smallest unit, had the highest precision. Precision differences were significant at the 0.05 level in all cases except abstracts vs. sentences (see Appendix B).

For effectiveness, sentences were significantly more effective than phrases at the 0.05 level, indicating that the advantage of phrases over sentences in precision is outweighed by the disadvantage in recall. Although abstracts were measured as more effective than sentences, this may easily be due to chance as the difference was far from significant ( $p=0.84$  two-tailed). The measured effectiveness advantage

of abstracts over phrases also did not reach significance ( $p=0.17$  two-tailed). Abstracts, sentences, and phrases all rated significantly higher than sentence pairs.

Application of the generalized effectiveness formula to the figures in Table 2 rates abstracts as the most effective when recall is of overriding concern, phrases as the most effective when precision is of overriding concern, and sentences as the most effective for some values of  $w$  (Table 3).

Table 3. Ranges of weight parameter  $w$  for which each text unit measured as best in generalized effectiveness. Sentences had higher generalized effectiveness than phrases for all  $w>0.26$ . But note the discussions of statistical significance.

TEXT UNIT →	Abstract	Sentence pair	Sentence	Phrase
$w$ →	$w>0.29$	–	$0.26<w<0.29$	$w<0.26$

## 5 Discussion and Conclusion

In view of the results reported here it is not surprising that researchers have reported interesting results for text mining in MEDLINE based on abstracts, sentences, and phrases. Tables 2 and 3 and the statistical significance summary in the preceding section indicate that each of these units has advantages and disadvantages compared to the others. Sentence pairs did not fare well, suggesting that anaphora resolution or other sophisticated approaches, applied just to pairs of adjacent sentences, would not form a particularly productive research path toward the objective of mining for interactions among biochemicals in scientific literature.

Increasing the sophistication of text processing can raise precision without degrading recall, raising effectiveness as well, as implied by Craven and Kumlein's<sup>2</sup> Figure 2 and Thomas *et al.*'s<sup>18</sup> Table 5. Sophisticated text processing seems likely to benefit smaller text units more than larger ones because of their generally shorter lengths, simpler structures, and higher proximity of relevant verbs and biochemical nouns, making their processing more tractable. For example, appropriate verbs in close proximity to biochemical terms are likely to be better indicators of an interaction than more distant verbs. Ease of analysis would not be an issue if complete automatic natural language understanding were available, which would in principle provide precisions of 1.0 for all text units. This would swing the advantage back to longer text units because the principle of decreasing recall for smaller text units, in conjunction with the theoretical possibility of equal precisions for all text units, in principle implies superiority of longer text units in effectiveness. However, complete automatic natural language understanding is not possible currently or in the foreseeable future. Effectiveness figures for the current state of the art for biochemical interaction extraction using sophisticated text processing is typified by Thomas *et al.*<sup>18</sup> whose best recall and precision results (their "TOP1" condition) imply an effectiveness of 0.66 for abstracts, Rindfleisch *et al.*<sup>9</sup> whose results section

reported recall and precision values implying a higher effectiveness of 0.75 for sentences, and Ono *et al.*'s<sup>7</sup> Table 3 for which the average total recall and precision figures imply a still higher effectiveness of 0.89 for phrases. In other words, the admittedly limited applicable body of literature from which effectiveness figures can be derived suggests a trend of higher effectivenesses for smaller text units, given sophisticated text analysis algorithms.

More sophisticated text processing techniques can be important for reasons other than increasing IR performance. For example, automatic construction of signal transduction pathways is an application that requires accounting for verbs. An application that clearly favors smaller text units is the simultaneous display of as rich a collection of targeted passages as possible from the often unwieldy body of scientific literature. It is better for this purpose to display sets of relevant sentences or phrases taken from numerous abstracts on a screen than it is to display one or two entire abstracts with occasional embedded relevant passages, particularly if the user interface makes it convenient move from a short relevant passage to its containing abstract, as by clicking.

In summary, abstracts, sentences, and phrases are all competitive for automatic extraction of interactions among biochemicals from MEDLINE, depending on the objectives of the system and the user. Not surprisingly, sophisticated text processing tends to increase IR performance relative to more basic text processing as explained earlier in this Section. However, a very large range of choices is possible in designing systems with advanced text processing capabilities. For example, just defining a set of verbs that indicate interactions will be difficult to characterize definitively. Thus to provide relatively clean tabulations we avoided verb analysis, using co-occurrence of biochemical terms to retrieve abstracts and within retrieved abstracts also classified use of synonyms of these terms as occurrences as well. As noted earlier, accounting for verbs would be expected to generally increase precisions and hence effectivenesses particularly for smaller text units.

### **Appendix A: An Outlier Query**

It is interesting to consider an outlier from among our ten queries. For the query "cholesterol AND flavonoid," smaller text units fared more poorly than for other queries (Table 1). Closer inspection of these abstracts showed that flavonoid is a large family of chemicals, and the name of a specific flavonoid is usually stated in the first sentence of an abstract. In the rest of the abstract, the name of the specific flavonoid is used instead of the general term "flavonoid." Therefore the term "flavonoid" tends to be distant from the term "cholesterol" in the abstracts, leading to relatively low recall, precision, and hence effectiveness for sentence pairs,

sentences, and phrases. This factor should be considered in the context of very general chemical terms.

## Appendix B: Statistical Procedure

We conducted separate analyses for precision and effectiveness. The structure of the data suggests an analysis based on the usual linear model for a block design, where each query serves as a block. The model often used for such data is

$$Y_{ij} = \mu + \alpha_i + \beta_j + e_{ij} \quad (1)$$

Here  $Y_{ij}$  denotes a measure of information retrieval quality (recall, precision, or effectiveness) for the method using the  $i^{\text{th}}$  processing unit ( $i = 1, 2, 3$ , or 4 corresponding to abstract, sentence pair, single sentence, or phrase, respectively) on the  $j^{\text{th}}$  set of abstracts. The  $e_{ij}$  represent independent random errors with mean zero and variance  $\sigma^2 / w_j$ , where  $w_j$  is a weight equal to the number of abstracts used in the determination of  $Y_{ij}$ . The parameters  $\alpha_1, \dots, \alpha_4$  represent the statistical effects associated with the processing units. These are the quantities of interest. The  $\alpha_i$  are typically constrained to sum to zero for easier interpretation, and the  $\mu$  parameter is introduced as an intercept. Thus  $\alpha_i$  greater (less) than zero implies above (below) average performance for the  $i^{\text{th}}$  method relative to the others for any particular chemical pair. The  $\beta_1, \dots, \beta_{10}$  quantities are the statistical effects associated with each of the 10 sets of abstracts corresponding to the 10 queries.

For the IR performance measures of precision and effectiveness, we are interested in testing for differences among pairs of text units. For two different text units indexed by  $i$  and  $i'$ , we may formally write our null and alternative hypotheses as

$$H_{ii'} : \alpha_i = \alpha_{i'} \text{ and } K_{ii'} : \alpha_i \neq \alpha_{i'},$$

respectively. To test  $H_{ii'}$  against  $K_{ii'}$  we will compute the usual weighted t-statistic using the differences  $D_{ii'j} = Y_{ij} - Y_{i'j}$  with weights  $w_j$  ( $j = 1, \dots, 10$ ). The formula for the weighted t-statistic is

$$t_{ii'} = \bar{D}_{ii'} / s(\bar{D}_{ii'}), \text{ where}$$

$$\bar{D}_{ii'} = \sum_{j=1}^{10} w_j D_{ii'j} / \sum_{j=1}^{10} w_j \text{ and}$$

$$s(\bar{D}_{ii'}) = \sqrt{\sum_{j=1}^{10} w_j (D_{ii'j} - \bar{D}_{ii'})^2 / \{(10-1) \sum_{j=1}^{10} w_j\}}.$$

To assess the significance of an observed value of  $t_{ii'}$ , we condition on the magnitudes of the observed differences and note that under the null hypothesis the

probability of a positive difference is equal to the probability of a negative difference. This follows from the fact that

$$D_{i'j} = Y_{ij} - Y_{i'j} = \alpha_i - \alpha_{i'} + d_{i'j} = d_{i'j}$$

when the null hypothesis is true. Now, under the null hypothesis, all  $2^{10}$  possible assignments of signs to  $|D_{i'1}|, \dots, |D_{i'10}|$  are equally likely assuming  $d_{i'j} = e_{ij} - e_{i'j}$  are independent for  $i \neq i'$ . Thus the conditional null distribution of  $t_{i'}$  places probability mass  $1/2^{10}$  on each of the  $2^{10}$  values obtained by computing  $t_{i'}$  for the  $2^{10}$  possible assignments of signs to  $|D_{i'1}|, \dots, |D_{i'10}|$ . The relevant two-tailed p-value is obtained by counting the proportion of those  $2^{10}$  values whose magnitudes match or exceed the observed value of  $|t_{i'}|$ . This is essentially the randomization test for matched pairs described, for example, in Section 5.11 of Conover.<sup>3</sup> We have augmented this slightly by using the number of abstracts as weights in our test statistic to account for variation in the number of abstracts used to compute the measures of performance.

To illustrate the testing procedure that we have used, consider testing for a difference between the effectiveness of sentence pairs and single sentences. The relevant differences (one for each query) are

$$-0.19, -0.23, -0.28, -0.18, -0.17, -0.28, -0.24, -0.22, -0.25, \text{ and } +0.14.$$

The preponderance of negative signs immediately suggests greater effectiveness for the single sentence method. The weighted t-statistic is  $t_{23} = -5.97$ . If we were to randomly assign signs to the observed differences, the chance of getting a weighted t-statistic as far from zero as -5.97 is only  $6/1024 \approx 0.0059$ . This is the p-value of the test, and it can be computed by calculating that there are only 6 sign configurations (among the 1024 possible configurations) that yield a t-statistic, weighted to reflect the number of examined abstracts associated with each query, as far from zero as -5.97.

Because it is so unlikely (probability 0.0059) to see a value of the test statistic as extreme as -5.97 when the null hypothesis is true, we reject the null hypothesis and conclude that single sentences are significantly more effective than sentence pairs. Other results for effectiveness, and results for precision, are shown in Table 4.

Two columns of Table 4 contain p-values that have been adjusted for multiple testing using the restricted step-down method.<sup>13</sup> A clear description of restricted step-down method for p-value adjustment is provided in Section 2.7 of Westfall and Young.<sup>21</sup> The use of adjusted p-values is conservative and reduces the chance of errantly rejecting a true null hypothesis simply because many hypotheses are being tested. Motivation for the use of adjusted p-values may be found in several statistical texts on the subject of simultaneous inference.

Table 4. Probabilities of rejection of null hypotheses of no difference between text units. Sentences and phrases are significantly different. Precision of phrases is significantly different from that of sentences. Other cells do not reach significance.

Comparison	Precision			Effectiveness		
	Weighted t-statistic	P-value	Adjusted p-value	Weighted t-statistic	P-value	Adjusted p-value
Abstract vs. sentence	-1.34	0.3516	0.3516	0.25	0.8398	0.8398
Abstract vs. phrase	-3.00	0.0488	0.0488	1.36	0.1719	0.1719
Sentence vs. phrase	-5.14	0.0078	0.0234	5.26	0.0039	0.0117

## References

1. C. Blaschke, M. Andrade, C. Ouzounis, and A. Valencia, "Automatic extraction of biological information from scientific text: protein-protein interactions" *AAAI Conf. on Intelligent Systems in Molecular Biology*, 60-67 (1999).
2. M. Craven and J. Kumlien, "Constructing biological knowledge based by extracting information from text sources" *7<sup>th</sup> International Conference on Intelligent Systems for Molecular Biology Biology (ISMB-99)*.
3. W. Conover, *Practical Nonparametric Statistics, 2nd Edition*, (Wiley, New York, 1980).
4. J. Dickerson, D. Berleant, Z. Cox, W. Qi, D. Ashlock, and E. Wurtele, "Creating metabolic network models using text mining and expert knowledge" *Atlantic Symp. on Computational Biology and Genome Information Systems & Technology (CBGIST 2001)*, 26-30.
5. K. Humphreys, G. Demetriou, and R. Gaizauskas, "Two applications of information extraction to biological science journal articles: enzyme interactions and protein structures" *Pacific Symposium on Biocomputing 5*, 502-513 (2000).
6. S.-K. Ng and M. Wong, "Toward routine automatic pathway discovery from on-line scientific text abstracts" *Genome Informatics 10*, 104-112 (1999).
7. T. Ono, H. Hishigaki, A. Tanigami, and T. Takagi, "Automated extraction of information on protein-protein interactions from the biological literature" *Bioinformatics 17*, 155-161 (2001).
8. PUBMED interface to MEDLINE, U.S. National Library of Medicine, <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed>.

9. T. Rindflesch, L. Hunter, and A. Aronson, "Mining molecular binding terminology from biomedical text" *Proceedings of the AMIA '99 Annual Symposium*.
10. T. Rindflesch, L. Tanabe, J. Weinstein, L. Hunter, "EDGAR: extraction of drugs, genes and relations from the biomedical literature" *Pacific Symposium on Biocomputing 5*, 514-525 (2000).
11. W. Salamonsen, K. Mok, P. Kolatkar, and S. Subbiah, "BioJAKE: a tool for the creation, visualization and manipulation of metabolic pathways" *Pacific Symposium on Biocomputing 4*, 392-400 (1999).
12. T. Sekimizu, H. Park, and J. Tsujii, "Identifying the interaction between genes and gene products based on frequently seen verbs in MEDLINE abstracts" *Genome Informatics* (Universal Academy Press, Inc., 1998).
13. J. Shaffer, "Modified sequentially rejective multiple test procedures" *Journal of the American Statistical Association* **81**, 826-831 (1986).
14. H. Shatkay, S. Edwards, W. Wilbur, and M. Boguski, "Genes, themes, and microarrays: using information retrieval for large-scale gene analysis" *8<sup>th</sup> International Conference on Intelligent Systems for Molecular Biology (ISMB 2000)*, (La Jolla, CA, Aug. 19-23).
15. W. Shaw, R. Burgin, and P. Howell, "Performance standards and evaluations in IR test collections: cluster-based retrieval models" *Information Processing and Management* **33** (1), 1-14 (1997).
16. B. Stapley, and G. Benoit, "Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in medline abstracts" *Pacific Symposium on Biocomputing 5*, 529-540 (2000).
17. L. Tanabe, U. Scherf, L. Smith, J. Lee, L. Hunter, and J. Weinstein, "MedMiner: an internet text-mining tool for biomedical information, with application to gene expression profiling" *BioTechniques* **27**, 1210-1217 (1999).
18. J. Thomas, D. Milward, C. Ouzounis, S. Pulman, and M. Carroll, "Automatic extraction of protein interactions from scientific abstracts" *Pacific Symposium on Biocomputing 5*, 538-549 (2000).
19. C. Van Rijsbergen, *Information Retrieval*, Butterworths (1979).
20. P. Westfall, and S. Young, *Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment* (Wiley, New York, 1993).
21. R. Willmott, P. Rushton, R. Hooley, and C. Lazarus, "DNase1 footprints suggest the involvement of at least three types of transcription factors in the regulation of alpha-Amy2/A by gibberellin" *Plant Molecular Biology* **38** (5), 817-825 (1998).
22. L. Wong, "A protein interaction extraction system" *Pacific Symposium on Biocomputing 6*, (2001).