Pacific Symposium on Biocomputing 7:326-337 (2002).

MINING MEDLINE: ABSTRACTS, SENTENCES, OR PHRASES?

J. DING[a], D. BERLEANT[a,d], D. NETTLETON[b], AND E. WURTELE[c]
[a]*Department of Electrical and Computer Engineering,*
[b]*Department of Statistics,*
[c]*Department of Botany,*
[d]*berleant@iastate.edu*
Iowa State University, Ames, Iowa 50011, USA

A growing body of works address automated mining of biochemical knowledge from digital repositories of scientific literature, such as MEDLINE. Some of these works use abstracts as the unit of text from which to extract facts. Others use sentences for this purpose, while still others use 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.

1 Introduction

The rapid growth of digitally stored scientific literature provides increasingly attractive opportunities for text mining. Concurrently, text mining is becoming an increasingly well-understood alternative to manual information extraction. Most reports on text mining of scientific literature for biochemical interactions have used the MEDLINE repository. Such mining activities have great potential for tasks such as extracting networks of protein interactions as well as for benefiting researchers who need to efficiently sift through the literature to find work relating to small sets of biochemicals of interest. 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, constituent sentences, or phrases. The most basic way to "mine" MEDLINE is simply to use the PUBMED Web interface.[8] The user can submit a query to the database consisting of the AND of two biochemical terms, and abstracts in MEDLINE containing both terms 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.[16] A related approach by Shatkay *et al.*[14] infers functional relationships among genes based on similarities among abstracts. Neither of those works identifies the type of interaction (e.g. inhibit, activate, etc.), which is desirable for applications such as automatic construction of networks of interactions. Because an abstract is a relatively large processing unit

which contains a great deal of material besides the query terms, it is relatively difficult to automatically determine the type of interaction between the terms without methods that are sensitive to smaller text units such as sentences or phrases.

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

> `...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...`[21]

The interactions in this example (gibberellin regulates alpha-Amy2/A and alpha-Amy2/D) are described over two sentences, so 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 many interactions, they will miss others interactions that are expressed over longer passages. Consequently, information retrieval recall must decrease with decreasing text unit size. However a clean qualitative relationship between text unit size and information retrieval precision 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. Four text units are investigated here: abstracts, adjacent sentence pairs, sentences, and phrases, from the perspective of three standard information retrieval (IR) performance measures: recall, precision, and effectiveness. Recall is the fraction of the relevant items in a test set that are retrieved. Precision is the fraction of retrieved items that are also relevant. Effectiveness is a composite measure combining the recall and the precision. The benefit of the present investigation of the relationships between text unit type and information retrieval performance measures is better understanding of the ability of the different text units to support mining of scientific abstract repositories for interactions among biochemicals.

2 Experimental Procedure: The Data

To compare the merits of different text processing units, a corpus of slightly over three hundred abstracts, termed the Interaction Extraction Performance Assessment (IEPA) corpus, was manually analyzed. The corpus consists of abstracts retrieved

from MEDLINE using ten queries (Table 1) to its PUBMED interface.[8] Each query was the AND of two biochemical nouns. The 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 studied 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 at the time the corpus was defined were studied, enough so that the studied set included approximately forty abstracts describing interaction(s) between the biochemicals in the query, plus those that contained the biochemicals but did not describe interactions between them that were also encountered. Thus the ten queries yielded ten sets of abstracts, with each abstract in a set containing both terms in the query corresponding to that set.

Although each studied abstract contained both biochemical terms in a query, only some of them described interaction(s) between them. 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 necessarily also present in any larger text unit 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 Ca2+ 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. All but the first and last sentence in an abstract therefore appeared in two sentence pairs, once as the first of the pair and once as the second. The text between two successive periods was defined to be a sentence. In addition, the title was defined to be a sentence, as was the body up to the first period. The text between any two successive punctuation marks {. : , ;} was defined as a phrase. The title up to its first punctuation mark was also defined as a phrase, as was a complete title containing no punctuation mark, and also the body of the abstract up to the first punctuation mark.

While both members of the query occurred in each abstract, in only some of the abstracts did both terms or their synonyms 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 the same phrase.

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,[15] while a generalized form of effectiveness includes the relative weights of recall and precision as a parameter in the calculation.[19]

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 other text unit under consideration:

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

where A and B are query terms or their synonyms.

Intuitively, recall here measures the capacity of a given text unit to contain the interactions present in MEDLINE abstracts. Any interaction described within a particular text unit is also described within all larger text units. Therefore, since the largest unit considered here is the abstract the recall for abstracts is exactly 1.

Precision refers to the fraction of abstracts, sentences, phrases, etc. containing both biochemical terms that also describe an interaction between them:

$$\text{precision} = \frac{\#\text{of interactions between A and B occurring within a type of text unit}}{\#\text{of times A and B co - occur in that type of text unit}}$$

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

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

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

Generalized effectiveness (G) parameterizes effectiveness with a weight coefficient w specifying the relative weights given to recall and precision:

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

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, for each query and each text unit, term co-occurrences and the subset of co-occurrences describing interactions. The recall, precision, and effectiveness of each were then tabulated (Tables 1 and 2). Because preliminary study showed that often an interaction is described using a synonym of a query term rather than the query term itself, occurrences of synonyms were treated as occurrences of query terms.

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, an outlier, 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 types of text units. Recall and precision figures are means over the relevant figures for each query (shown in Table 1 for all text unit types except sentence pairs). Each figure was appropriately weighted, by the number of abstracts in the set associated with that query (in the case of precision of abstracts), the number of co-occurrences for that query within the text unit under consideration (in the case of precision of sentence pairs, sentences, and phrases), or by the number of interactions described for that query within the associated set of abstracts (for recall).

TEXT UNIT → ↓ IR MEASURE	Abstracts	Sentence pairs	Sentences	Phrases
Recall	1	0.916	0.849	0.621
Precision	0.571	0.345	0.638	0.743
Effectiveness	0.727	0.501	0.729	0.677

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 except in the case of abstracts vs. sentences (Appendix B). (*Comment added 11/12/04*: Wren and Garner[23] (2004, p. 193 col. 2) corroborated the value we found for precision of abstracts with an independently derived figure of 0.58, but found a higher figure of 0.83 for precision of sentences. The co-occurring words they examined included phenotypes and diseases (p. 193 col. 1) in addition to biomolecules of the type we studied.)

With respect to effectiveness, sentences were significantly better than phrases at the 0.05 level, indicating that the advantage of phrases over sentences in precision is outweighed by the disadvantage in recall. Abstracts measured about equal to

sentences in effectiveness. The measured effectiveness advantage of abstracts over phrases 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 most effective when recall is of overriding concern, phrases as most effective when precision is of overriding concern, and sentences as most effective over an intermediate range of weightings (Table 3).

Table 3. Ranges of weight parameter w for which each text unit measured as best in generalized effectiveness (w can range between 0 and 1).

TEXT UNIT →	Abstract	Sentence pair	Sentence	Phrase
$\boldsymbol{w} \rightarrow$	$w>0.511$	–	$0.339<w<0.510$	$w<0.338$

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 fared so poorly in precision that an analysis was undertaken to understand why. Although considering pairs of sentences nearly doubled (99%) the number of distinct co-occurrences found compared to limiting consideration to sentences, the number of distinct interactions went up by only 8%. In fact, the dominant contributor to the already low precision of sentence pairs is interactions that are actually described in a single sentence within the pair. For the remaining interactions, those for which each term was in a different sentence, the precision was a mere 0.05. This in turn suggests that compared to the effort it would take to build a system to extract biochemical interactions from sentences, it might not be worth much additional effort to deal with sentence pairs as well. Even large expenditures of computation time or system development effort to achieve quality anaphora resolution across adjacent sentences would result in only modest benefit.

Regardless of the text unit chosen for a system that extracts biochemical interactions from MEDLINE, interactions contained in an abstract were often described using a synonym of the query term. Thus we counted synonyms as query term instances in deriving the retrieval performance measures reported here.

Increasing the sophistication of text processing can raise precision without degrading recall, thereby raising effectiveness, as suggested by Craven and Kumlein's[2] Figure 2 and accompanying discussion, and by Thomas *et al.*'s[18] Table 5. Sophisticated text processing techniques seem likely to benefit smaller text units

more than larger ones because their generally shorter lengths, simpler structures, and higher proximity of relevant verbs and biochemical nouns make their processing more tractable. For example, appropriate verbs ("bind," "activate," etc.) in close proximity to biochemical terms are likely to be better indicators of an interaction than more distant verbs. However, ease of analysis would not be an issue if complete automatic natural language understanding were available, which would in principle enable precisions of 1 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 the same precision, 1, for all text units implies potential superiority of longer text units. However, complete automatic natural language understanding is currently not possible nor is it likely to be for some time. Effectiveness figures for the current state of the art for biochemical interaction extraction using sophisticated text processing were derivable from two reports, summarized in table 4.

Table 4. Effectiveness of sophisticated text processing techniques is higher than the baseline figures in Table 2 above for both the sentence and phrase text units. For phrases, sophisticated techniques led to an effectiveness higher than that of any entry in Table 1 above. (However comparisons across reports should be interpreted with caution.)

Report	Comment	Text unit	Best effectiveness	Baseline average	Baseline range
Rindflesch *et al.*[9]	"RESULTS" section	Sentence	0.75	0.72	0.33-0.92
Ono *et al.*[7]	Their Table 3	Phrase	0.89	0.65	0.17-0.82

Sophistication in text processing techniques can be important for reasons other than improving IR performance. For example, automatic construction of signal transduction pathways is an application that requires accounting for verbs.

Another application that clearly favors small text units is the simultaneous display of targeted passages 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 it is convenient to move from a short relevant passage to its containing abstract, such as by clicking.

In summary, abstracts, sentences, and phrases are all competitive for automatic extraction of interactions among biochemicals from MEDLINE. Not surprisingly, sophisticated text processing appears to increase IR performance relative to more basic text processing. 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. To provide a relatively clean baseline we avoided verb analysis, although a suitable accounting of verbs might be expected to increase precision 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 particularly 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

$$Y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}$$

where Y_{ij} denotes a measure of information retrieval quality (recall, precision, or effectiveness) for the method using the i^{th} processing unit (i=1, 2, 3, or 4 corresponding to abstract, sentence pair, single sentence, or phrase, respectively) on the j^{th} set of abstracts. The e_{ij} represent independent random errors with mean zero and variance σ^2/w_j, where w_j is a weight equal to the number of abstracts used in the determination of Y_{ij}. We assume that the distribution of $d_{ii'j}=e_{ij}-e_{i'j}$ is symmetric for all $i \neq i'$ and j=1,…,10. The parameters $\alpha_1,...,\alpha_4$ represent the statistical effects associated with the processing units. These are the quantities of interest. The α_i are typically constrained to sum to zero for easier interpretation, and the μ parameter is introduced as an intercept. Thus α_i greater (less) than zero implies above (below) average performance for the i^{th} method relative to the others for any particular chemical pair. The $\beta_1,...\beta_{10}$ 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 $H_{ii'}: \alpha_i = \alpha_{i'}$ 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 $w_j\,(j = 1,...,10)$. The formula for the weighted t-statistic is

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

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

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

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 $D_{ii'j}=Y_{ij} - Y_{i'j}=\alpha_i - a_{i'} + d_{ii'j}=d_{ii'j}$ when the null hypothesis is true. Now, under the null hypothesis, all 2^{10} possible assignments of signs to $|D_{ii'1}|, \ldots |D_{ii'10}|$ are equally likely assuming $d_{ii'j} = e_{ij} - e_{i'j}$ are independent for $i \neq i'$. Thus the conditional null distribution of $t_{ii'}$ places probability mass $1/2^{10}$ on each of the 2^{10} values obtained by computing $t_{ii'}$ for the 2^{10} possible assignments of signs to $|D_{ii'1}|, \ldots |D_{ii'10}|$. 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 $|t_{ii'}|$. This is essentially the randomization test for matched pairs described, for example, in Section 5.11 of Conover.[3] 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 we 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, 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 5.

Two columns of Table 5 contain p-values that have been adjusted for multiple testing using the restricted step-down method,[13] for which a clear description is provided in Section 2.7 of Westfall and Young.[20] 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 5. Tests of null hypotheses of no difference between text units. Sentences and phrases are significantly different. Precision of phrases is significantly different from that of abstracts, while other cells do not reach significance. (Ab=abstract, Se=sentence, and Ph=phrase.)

Comparison	Precision			Effectiveness		
	Weighted t-statistic	P-value	Adjusted p-value	Weighted t-statistic	P-value	Adjusted p-value
Ab vs. Se	-1.34	0.3516	0.3516	0.25	0.8398	0.8398
Ab vs. Ph	-3.00	0.0488	0.0488	1.36	0.1719	0.1719
Se vs. Ph	-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 Conference 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" *7th International Conference on Intelligent Systems for Molecular Biology (ISMB-99).*
3. W. Conover, *Practical Nonparametric Statistics, 2nd Ed.* (Wiley, NY, 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 Symposium 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" *8th Int. Conf. on Intelligent Systems for Mol. Bio. (ISMB 2000)*, La Jolla, 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).
23. J. Wren and H. Garner, "Shared relationship analysis: ranking set cohesion and commonalities within a literature-derived relationship network" *Bioinformatics* **20** (2), 191-198 (2004).