BIOINFORMATICS

Vol. 17 no. 4 2001
Pages 359–363

Mining literature for protein–protein interactions

Edward M. Marcotte[1, 2, 3,*], Ioannis Xenarios[1,*] and David Eisenberg[1]

[1]Molecular Biology Institute, UCLA-DOE Laboratory of Structural Biology & Molecular Medicine, University of California at Los Angeles, PO Box 951570, Los Angeles, CA 90095-1570, USA, [2]Protein Pathways Inc., 1145 Gayley Avenue, Ste. 304, Los Angeles, CA 90024, USA and [3]Institute of Cellular and Molecular Biology, Department of Chemistry and Biochemistry, University of Texas at Austin, 2500 Speedway, Austin, TX 78712, USA

Received on August 3, 2000; revised on November 16, 2000; accepted on November 22, 2000

ABSTRACT

Motivation: A central problem in bioinformatics is how to capture information from the vast current scientific literature in a form suitable for analysis by computer. We address the special case of information on protein–protein interactions, and show that the frequencies of words in Medline abstracts can be used to determine whether or not a given paper discusses protein–protein interactions. For those papers determined to discuss this topic, the relevant information can be captured for the Database of Interacting Proteins. Furthermore, suitable gene annotations can also be captured.

Results: Our Bayesian approach scores Medline abstracts for probability of discussing the topic of interest according to the frequencies of *discriminating words* found in the abstract. More than 80 discriminating words (e.g. complex, interaction, two-hybrid) were determined from a training set of 260 Medline abstracts corresponding to previously validated entries in the Database of Interacting Proteins. Using these words and a log likelihood scoring function, ∼2000 Medline abstracts were identified as describing interactions between yeast proteins. This approach now forms the basis for the rapid expansion of the Database of Interacting Proteins.

Contact: marcotte@icmb.utexas.edu; ixenario@mbi.ucla.edu; david@mbi.ucla.edu

INTRODUCTION

Mining biological literature for information is essential for transforming discoveries reported in the literature into a form useful for computation. Already, databases of protein–protein interactions (Xenarios *et al.*, 2001; Bader *et al.*, 2001) and signaling pathways (Kanehisa and Goto, 2000; the Signal Transduction Knowledge Environment: http://www.171.66.122.61/) are generating new insights into how cells are organized, such as demonstrating that protein–protein interactions link many proteins in the cell into just a few, large, connected interaction networks (Marcotte, 2000; Xenarios *et al.*, 2001).

Efforts to expand such databases have been hindered in part by the sheer volume of biological literature: over 10 million entries exist in Medline, the archive of abstracts of biological articles (http://www.ncbi.nlm.nih.gov/PubMed/). Since many databases, such as the Database of Interacting Proteins (DIP, Xenarios *et al.*, 2001, http://www.dip.doe-mbi.ucla.edu/) and SwissProt (Bairoch and Apweiler, 2000), are hand-curated to assure valid entries, evaluating literature normally becomes the rate limiting step in the growth of the database. Thus, automatic methods are needed to speed up this step of database construction.

Protein interactions have been discovered automatically in the literature by methods involving natural language processing to parse sentences in abstracts into grammatical units (Thomas *et al.*, 2000; Humphreys *et al.*, 2000) and by methods analyzing sentences discussing interactions using frequency analysis of individual words (Blaschke *et al.*, 1999). Due to the complexity and variety of the English language, such approaches are inherently difficult. Even the language used to describe specific proteins is ambiguous, such that efforts are required to collect synonymous protein names into databases (Yoshida *et al.*, 2000). Also, simple regular expression searches for abstracts containing relevant words, such as ‘interact*’, poorly discriminate true hits from abstracts using the words in alternate senses and miss abstracts using different language to describe the interactions. For example, searches of a small test set of interaction and non-interaction abstracts for the word root ‘interact*’ are 65% accurate and recover 65% of the interaction articles. However, given that we desire a curated database, each article must be manually evaluated, and the high

*These authors contributed equally to this work.

false positive rate (35%) greatly reduces the efficiency of database entry. Looking for specific Medline MESH terms fares no better: searches for the most effective MESH term ('protein binding') are 73% accurate but have only 33% coverage. Most importantly, such simple searches provide no criterion for prioritizing hits. For a database in which a curator examines each article prior to entry of an interaction into the database, we find the following approach most useful.

Here, we introduce a method to scan the biological literature and select and rank just those abstracts discussing a given topic. Our approach is based on that used to identify James Madison as the author of 12 of the Federalist papers of disputed authorship, written in 1787–1788 (Mosteller and Wallace, 1984). Discriminating words are identified that appear at unexpectedly high or low frequencies in abstracts discussing the topic of interest. Using a Bayesian approach, each of the many Medline abstracts can then be scored for its probability of discussing the topic of interest according to the frequencies of the discriminating words observed in the abstract. We apply the method to Medline abstracts to identify and rank-order several thousand abstracts discussing protein–protein interactions; interactions in these abstracts are now being incorporated into the Database of Interacting Proteins.

METHODS

For this analysis, we chose to focus on yeast, a well-studied system with many known protein interactions and for which the Medline entries were available to us. 88 921 Medline entries were acquired from PubMed that contain the term '*Saccharomyces cerevisiae*' in the title, abstract, or MESH terms. Of these 88 921 Medline entries, only 65 807 contained abstracts. All analyses in this paper use the abstracts from this subset of Medline entries, termed 'Yeast Medline'. As of June 2000, the Database of Interacting Proteins cited 260 papers reporting protein–protein interactions involving yeast proteins; these 260 Medline abstracts were set aside as a training set of positive examples of interaction abstracts. All analyses are case insensitive, and all punctuation marks other than hyphens, apostrophes, and plus and minus signs were substituted by spaces.

First, a dictionary was constructed containing the frequencies of the 60 000 most common words in Yeast Medline abstracts, which includes every word used more than three times in Yeast Medline. Next, words from the training set of 'interaction' abstracts were tested for frequencies unexpectedly higher or lower than calculated dictionary frequencies, indicating words that would be useful for discriminating the training abstracts from other abstracts. For each word in the training abstracts, the number of occurrences n was counted, and the probability $p(n|N, f)$ of finding the word the observed number of times given the known dictionary frequency f and the total number of words N in the training abstracts, was calculated from the Poisson distribution as

$$p(n|N, f) \approx \mathrm{e}^{-Nf}\frac{(Nf)^n}{n!}. \qquad (1)$$

This approximation is valid when the total number of words used to generate the dictionary is much greater than N and when f is small. In practice, to avoid floating point errors, the log of the probability was calculated as $\ln p(n|N, f) \approx -Nf + n\ln(Nf) - \ln(n!)$, where $n!$ was estimated using Stirling's approximation for large n.

The 500 words in the training abstracts with the most negative log probability scores were selected as *discriminating words*. A property of any discriminating algorithm is that it will be biased by its training set. However, we attempted to minimize the most obvious source of bias by removing gene and protein names (e.g. actin, tup1, sup35p, etc.), as well as names of specific cellular systems or pathways (e.g. cytoskeleton, anaphase-promoting, bud-site), thereby making the approach general to any abstracts describing protein interactions. This curation step left 83 general words that discriminate abstracts discussing protein–protein interactions from other abstracts. These discriminating words were all statistically significant (all had $\ln p < -13$) and included both under- and over-represented words. In practice, most discriminating words were over-represented in interaction abstracts.

Armed with these discriminating words, each abstract in Yeast Medline could then be scored for its likelihood of discussing protein–protein interactions in the following manner: in an abstract with N total words, the number of occurrences n_i of each discriminating word i is counted. Given n_i, we would like to know if the abstract is likely to discuss interactions. Casting this in Bayesian form gives the following two probability expressions:

$$p(\text{InteractionAbstract}|n_i) = \frac{p(n_i|\text{InteractionAbstract}) * p(\text{InteractionAbstract})}{\text{NormalizationFactor}}$$

$$p(\text{NonInteractionAbstract}|n_i) = \frac{p(n_i|\text{NonInteractionAbstract}) * p(\text{NonInteractionAbstract})}{\text{NormalizationFactor}}$$

The normalization factor, equal in each equation, is the sum of the numerators of the two equations.

To evaluate which is more likely, that the abstract discusses interactions or does not discuss interactions, the ratio of the two probability expressions is evaluated, allowing the normalization factors to be cancelled. The observed number of occurrences n_i is then tested for its probability of being drawn from the distribution characterized by the frequency $f_{I,i}$ of the discriminating word i in the training abstracts, or from the distribution characterized by frequency $f_{N,i}$, the dictionary frequency of discriminating word i. Assuming flat prior probabilities of having interaction or non-interaction abstracts means that these terms cancel as well, and modeling $p(n_i|\text{AbstractSet})$ with a Poisson distribution gives the expression:

$$\frac{p(\text{InteractionAbstract}|n_i)}{p(\text{NonInteractionAbstract}|n_i)} = \frac{\mathrm{e}^{-Nf_{I,i}}(f_{I,i})^{n_i}}{\mathrm{e}^{-Nf_{N,i}}(f_{N,i})^{n_i}} = \mathrm{e}^{-N(f_{I,i}-f_{N,i})}\left(\frac{f_{I,i}}{f_{N,i}}\right)^{n_i}.$$

To convert this into an additive score, the log of the probability ratio is taken, giving:

$$\ln\left(\frac{p(\text{InteractionAbstract}|n_i)}{p(\text{NonInteractionAbstract}|n_i)}\right) = n_i \ln\frac{f_{I,i}}{f_{N,i}} - N * (f_{I,i} - f_{N,i}).$$

Finally, the expression is summed for all of the 83 discriminating words to give a log likelihood score S that the abstract discusses protein–protein interactions:

$$S = \sum_i \left(n_i \ln\frac{f_{I,i}}{f_{N,i}} - N * (f_{I,i} - f_{N,i})\right) \qquad (2)$$

in which the sum is over discriminating words. The first term is positive when the frequency of the discriminating word i is larger in abstracts discussing interactions than in those abstracts not discussing interactions. The second term is negative under the same circumstances. The sharpness of the distribution of log likelihood scores as a function of $f_{I,i}$ increases as the observed frequency n_i/N of the discriminating word i increases.

RESULTS

As described in the Methods, words that discriminated interaction abstracts from other abstracts were identified. The discriminating words with the 20 most negative (most discriminating) ln p-scores are listed in Table 1. Among the 83 discriminating words are those clearly related to interactions, such as ‘binds’, ‘interacts’, ‘complexes’ and ‘associates’, and those words related to experimental methods that measure interactions, such as ‘2-hybrid’ and ‘co-immunoprecipitation’. Other words are related to the language used to describe interactions, such as ‘with’ and ‘together’. Some of the words, such as ‘protein’, ‘domain’, and ‘[in] vitro’, simply insured that the abstract was concerned with molecular studies, rather than topics such as clinical or microbiological studies. Most of the discriminating words were over-represented in the interaction abstracts, although a few, such as ‘enzyme’, ‘sequences’ and ‘cDNA’ were significantly under-represented.

Table 1. The 20 words that most discriminate abstracts discussing protein interactions from other abstracts include words describing interactions, names of experimental methods, and general molecular terms. The ln p-scores are calculated from equation (1)

Discriminating word	Word frequency in interaction abstracts (n/N)	Frequency in yeast Medline (f)	ln p-score
Complex	6.1e−03	1.1e−03	−245
Interaction	3.6e−03	7.0e−04	−141
Two-hybrid	2.1e−03	2.2e−04	−133
Interact	2.1e−03	2.6e−04	−124
Proteins	7.0e−03	2.6e−03	−121
Protein	1.1e−02	5.3e−03	−103
Domain	3.6e−03	1.2e−03	−75
Interactions	1.9e−03	3.9e−04	−73
Required	3.0e−03	9.2e−04	−68
Kinase	2.8e−03	8.9e−04	−60
Interacts	1.1e−03	1.7e−04	−57
Complexes	1.4e−03	3.2e−04	−49
Function	3.0e−03	1.2e−03	−49
Essential	2.1e−03	6.6e−04	−49
With	1.5e−02	1.0e−02	−45
Binding	3.6e−03	1.7e−03	−41
Component	1.1e−03	2.7e−04	−35
Suggesting	1.5e−03	5.0e−04	−35
From	3.2e−03	5.8e−03	−35
Demonstrate	1.3e−03	3.7e−04	−35

The prediction of abstracts discussing interactions was tested first on the set of abstracts from the Database of Interacting Proteins used to choose the discriminating words. The scores of these training abstracts are plotted in Figure 1, along with the scores of 10 000 randomly chosen yeast Medline abstracts. Although the set of 10 000 abstracts certainly contain abstracts discussing protein interactions, most abstracts do not discuss interactions, providing a fairly representative negative set. As seen in Figure 1, the log likelihood score predicts interaction abstracts very effectively. More than 88% of the interaction abstracts have positive scores. The mean log likelihood score for interaction abstracts is 11.1 ± 9.9, compared with −5.9 ± 9.9 for the 10 000 randomly chosen abstracts.

To perform an independent test of the method, 325 Yeast Medline abstracts were manually evaluated for their description of protein interactions. The abstracts were chosen as a sequential run of abstracts from an intermediate year of Medline abstracts. Seventy discussed interactions; 255 did not. The log likelihood scores for these two sets of abstracts are plotted in Figure 2a. Again, the scores effectively discriminate abstracts discussing interactions from those that do not. More than 77% of the interaction abstracts receive positive scores. The mean log likelihood score for the 70 interaction abstracts is 6.8±9.3, compared with −8.5 ± 7.6 for the 255 abstracts not discussing interactions. When the scores for this data set are re-plotted as

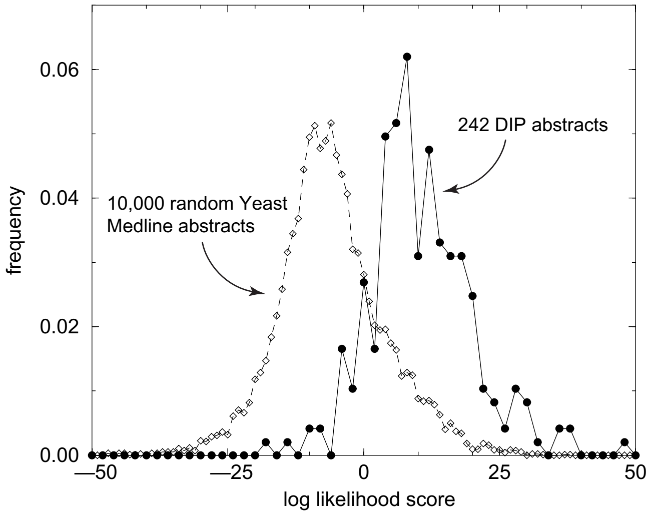

Fig. 1. Medline abstracts that discuss protein interactions (filled circles) receive considerably higher log likelihood scores (from equation (2)) than 10 000 randomly chosen Medline abstracts (open diamonds), which largely do not discuss abstracts. Here, the interaction abstracts are for the 242 articles available in June 2000 describing yeast protein interactions in the Database of Interacting Proteins (Xenarios *et al.*, 2001). For both Figures 1 and 2, data sets with <300 entries are analyzed in bins of size 2.

in Figure 2b, these independent test data show the coverage and accuracy of the method as evaluated on this test set of abstracts. More than 77% of abstracts receiving a log likelihood score of 5 discuss protein interactions, and 100% of the abstracts scoring 10 or higher discuss interactions.

The algorithm was trained on abstracts reporting only yeast protein interactions. As a second independent check on the method we tested the algorithm's performance on abstracts reporting non-yeast protein interactions. Abstracts reporting interactions where neither protein partner was from *Schizosaccharomyces*, *Candida*, or *Saccharomyces* were taken from the DIP database; in the 353 such abstracts, the majority of interactions involved human proteins (58%), mouse proteins (19%), or fly proteins (11%). More than 71% of these abstracts receive positive log likelihood scores; the mean log likelihood scores for these abstracts is 6.2 ± 9.1. The distribution of scores, plotted as a dashed line in Figure 2a, closely resembles the distribution of yeast interaction abstract scores plotted as a bold line in the same figure.

To identify abstracts discussing protein interactions, to facilitate the addition of entries to the Database of Interacting Proteins, we calculated log likelihood scores for all 65 807 Yeast Medline abstracts and rank-ordered the abstracts by their scores. 7021 abstracts receive a score of one or more, potentially discussing protein–protein interactions. Many of these abstracts are high-scoring, with 1747 abstracts scoring higher than 10. These 7000 abstracts now provide curators a ready source of protein interactions for entry into the Database of Interacting Proteins.

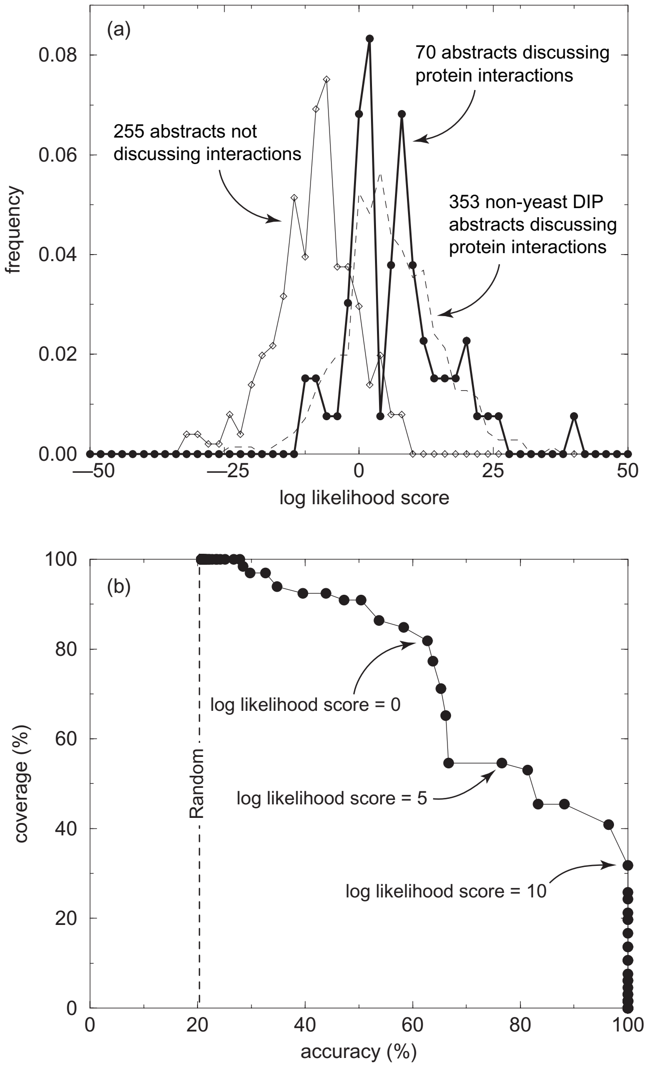

Fig. 2. Good discrimination between interaction and non-interaction abstracts is found for an independent test set of abstracts not used either to choose discriminating words or to generate the frequencies of words tallied in the frequency dictionary. A set of 325 abstracts was divided among the 70 which discuss protein interactions (filled circles) and the 255 which do not (open diamonds). In (a), the log likelihood scores of these abstracts are plotted. For comparison of the method's performance on abstracts from systems other than yeast, accompanying the scores of these abstracts are the scores of 353 non-yeast abstracts discussing protein interactions (dashed line). In (b), the scores of the 325 abstracts are re-plotted to show the coverage and prediction accuracy of the algorithm. Above a log likelihood score of ∼10, virtually all abstracts discuss protein interactions. The performance of an algorithm that randomly categorizes abstracts is plotted as a vertical dashed line.

DISCUSSION

In summary, we describe a method to sort through the large number of scientific articles and to choose those that are relevant to protein interactions. By first statistically identifying words that discriminate relevant abstracts from other abstracts, each new abstract can then be assigned a log likelihood score for discussing protein interactions. Although we chose to hand curate the discriminating words, thus limiting ourselves to the top-scoring 500 words, this curation step could have been automated and all statistically-significant words included in the analysis.

The method can easily be generalized to other topics, and should prove useful to other groups recovering data from scientific literature on a large scale. We find that once the frequency dictionary is calculated, the method can be rapidly applied to scan for other topics, for example cell signaling or protein–DNA interactions. The method could potentially be used to flag new articles of interest as they appear in Medline.

We also find that the calculation of discriminating words is useful for generating annotation about a specific topic, and we have used this method to automatically annotate yeast genes in the following manner: each gene is annotated by the over-represented discriminating words that appear in abstracts citing that gene. This method of annotation has the benefit of: (1) being as current as the set of abstracts used; (2) being unbiased; and (3) automatically including the names of many genes which are co-cited with the gene of interest, thereby including functional linkages (Stapley and Benoit, 2000) as part of the annotation.

For example, in a trivial but more general application of this method of generating annotation, we identified words that discriminate abstracts of articles published in 1999 from older abstracts. Beyond the obvious changes in experimental techniques and systems, this analysis reveals that the use of the past tense (‘was’, ‘were’) is underrepresented relative to older abstracts ($\ln p = -64$ and -95) and that the personal pronouns ‘we’ and ‘our’ are over-represented, as are currently popular words like ‘novel’ and ‘database’ ($\ln p = -147, -35, -72$, and -42, respectively).

Finally, we have applied the method to choose several thousand Medline abstracts that discuss protein interactions. Following calculation of log likelihood scores, the abstracts are sorted by score and written directly to an HTML file that contains the abstract, the Medline ID code hyper-linked to Medline, and hyper-links from gene names appearing in the abstract to sequence databases. This output file can be opened by a web browser and then be rapidly scanned by curators of the Database of Interacting Proteins to speed entry of protein interactions into the database.

ACKNOWLEDGEMENTS

We thank Rob Grothe and Parag Mallick for suggestions, NIH GM31299 and DOE DE-FC03-87-60615 for support, and the Swiss National Fund for a fellowship for I.X. Computer code, discriminating words, and Medline abstract scores are available from the authors on request.

REFERENCES

Bader,G.D., Donaldson,I., Wolting,C., Ouellette,B.F., Pawson,T. and Hogue,C.W. (2001) BIND—The Biomolecular Interaction Network Database. *Nucleic Acids Res.*, **29**, 242–245.

Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.

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

Humphreys,K., Demetriou,G. and Gaizauskas,R. (2000) Two applications of information extraction to biological sequence journal articles: Enzyme interactions and protein structures. *Pacific Symposium on Biocomputing,* Oahu.

Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.

Marcotte,E.M. (2000) Computational genetics: finding protein function by nonhomology methods. *Curr. Opin. Struct. Biol.*, **10**, 359–365.

Mosteller,F. and Wallace,D.L. (1984) *Applied Bayesian and Classical Inference: the Case of the Federalist Papers*. Springer, New York.

Stapley,B.J. and Benoit,G. (2000) Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in Medline abstracts. *Pacific Symposium on Biocomputing* Oahu.

Thomas,J., Milward,D., Ouzounis,C., Pulman,S. and Carroll,M. (2000) Automatic extraction of protein interactions from scientific abstracts. *Pacific Symposium on Biocomputing* Oahu.

Xenarios,I., Fernandez,E., Salwinski,L., Duan,X.J., Thompson,M.J., Marcotte,E.M. and Eisenberg,D. (2001) DIP: the database of interacting proteins: 2001 update. *Nucleic Acids Res.*, **29**, 239–241.

Yoshida,M., Fukuda,K. and Takagi,T. (2000) PNAD-CSS: a workbench for constructing a protein name abbreviation dictionary. *Bioinformatics*, **16**, 169–175.