Learning Anchor Verbs for Biological Interaction Patterns from Published Text Articles

Vasileios Hatzivassiloglou and *Wubin Weng*
Department of Computer Science
Columbia University, New York

Abstract: Much of knowledge modeling in the molecular biology domain involves interactions between proteins, genes, various forms of RNA, small molecules, etc. Interactions between these substances are typically extracted and codified manually, increasing the cost and time for modeling and substantially limiting the coverage. In this paper, we describe an automatic system for learning from text interaction verbs; these verbs can then form the core of automatically retrieved patterns that model classes of biological interactions. We investigate text features relating verbs with genes and proteins, and apply statistical tests and a logistic regression statistical model to determine whether a given verb belongs to the class of interaction verbs. Our system, AVAD, achieves over 87% precision and 82% recall when tested on an 11 million word corpus of journal articles.

INTRODUCTION

Almost every day, new biological substances such as genes, proteins, and other molecules are discovered, and interactions between them are studied. The results are reported in numerous publications of papers. Even a biologist who works in this fast-developing field cannot keep track of all these newly identified interactions without the help of an effective knowledge extraction computer system. Researchers have developed systems to extract automatically interaction relationships among proteins, genes, and other biological molecules. These systems apply patterns that are manually pre-constructed, in terms of pre-defined interaction verbs and/or pre-specified protein and gene names (Blaschke et al., 1999; Proux et al., 2000), or even are fully instantiated in a knowledge database or by a semantic grammar (Park et al., 2001; Yakushiji et al., 2001).

Thus, current approaches perform automatic interaction extraction based on patterns that are already known. Their power is greatly limited by the small set of pre-defined interaction verbs used in the patterns. For instance, Blaschke and colleagues (Blaschke et al., 1999) used a set of 14 pre-specified verbs that denoted actions related to protein interactions; Proux and colleagues (Proux et al., 2000) limited interaction verbs by presenting them explicitly in "request scenarios".

One way to ease this limitation is to enlarge the size of the interaction verb set automatically. Discovering interaction verbs automatically would allow substantial improvements in the performance and power of current systems. It would also balance current manually built verb lists, which tend to contain the most common interaction verbs, with other rarer members of this class (e.g., *co-localize* and *synergize*, both of which were automatically discovered by the system presented in this paper).

Finding the interaction verbs is also an important step in the automatic discovery of relationship patterns from large biological text corpora. Interaction verbs naturally link their subject and object, which are the participants in the interaction. Sekimizu and colleagues (Sekimizu, et al., 1998) built a system to find the subjects and objects for the frequently seen verbs in the genome domain, as the basis for a genome-related thesaurus. The verbs they used, however, were still pre-defined. To discover interaction patterns automatically, we can start from a set of automatically discovered interaction verbs and use text mining techniques to extract the initial patterns and corresponding tuples of genes or proteins that participate in the relationships indicated by the interaction verbs. We can then generalize the evidence obtained for individual proteins and genes by using clustering techniques on the proteins and genes in these tuples to recover automatically subclasses that have a similar functional behavior. As a result, we can propose appropriately restricted versions of the patterns for inclusion in a database of relations between finely grained subclasses of biological substances.

In this paper, we present AVAD, a system that uses a novel automatic method to discover interaction verbs that code for gene and protein interactions in molecular biology articles. We treat the discovery of such verbs as a two-category classification problem: among all verbs appearing in the text, automatically determine those that code for biological interactions and those that serve a normal discourse purpose (e.g., *say*, *report*, *be*). The features that AVAD uses include the frequency of a verb *before* gene or protein names (for convenience, we denote "gene or protein name" as

GPN), the frequency of that verb *after* GPNs, and the frequencies of the verb in different domains (biological, medical, and financial). First, we apply statistical tests to the features. Then we use either a rule-based combination or a fitted linear model to decide whether the verb is an interaction verb.

In Section 2, we outline the structure of AVAD and describe the methods we use for preprocessing text and recognizing verbs, GPNs, and associations between them. In Section 3, we discuss the statistical methods used over the word pair counts obtained earlier. Section 4 presents our analysis of the results generated from a large collection of biological journal articles by different versions of AVAD.

1. EXTRACTING INFORMATION FROM TEXT

The basic premise of our approach for determining if a verb is an interaction verb is to extract from the text the subjects and objects in its various occurrences over a large biological corpus. We reason that for an interaction verb these are likely to be entities from the biological domain (most commonly, genes and proteins), while for discourse verbs the subjects and objects are often not biological substances (e.g., authors *report* and *believe*, a study or another paper is *cited*, etc.).

AVAD includes a collection of modules that preprocess HTML input to produce annotated XML files with information about word and sentence breaks and part of speech labels. Further analysis of the text (for example, to detect co-occurring verbs and GPNs) is performed on the annotated text. We assume that the input to our system comes in HTML form, as most journal articles available already are already in this format. Additional preprocessing modules can be activated to handle ASCII text or PDF files.

In the preprocessing phase we start with the HTML::TreeBuilder perl module from CPAN (http://www.cpan.org) to parse the HTML files. Then, we discard the HTML tags that are used for graphic display purposes but carry no useful information for text analysis. We output the contents of the HTML files as raw text, and transform that to XML files via a pipeline containing five additional phases:

1. **GPN tagger**. We need to detect names of proteins and genes, since we base our verb statistics on the verb's associations with these words and phrases. We use a small dictionary of 2,783 GPNs, which provides us with a manually built, high-quality, but relatively small set of GPNs. Since we use these GPNs as seed points for the detection of interaction verbs, high precision in the labeling of GPNs is more important than high recall—if desirable, another source of GPNs such as GenBank (Benson et al. 1999) can be used. We maximally match phrases from the text against the dictionary, and perform this step first because of some gene names that contain punctuation marks (e.g., "Inositol (1,4,5) P3 receptor 1"), which would otherwise confuse our sentence boundary detector and tokenizer.
2. **Sentence boundary detector**. We use MX-TERMINATOR (Reynar and Ratnaparkhi, 1997; http://www.cis.upenn.edu/~adwait/statnlp.html) to detect sentence boundaries.
3. **Tokenizer**. We use a tokenizer for arbitrary raw text, a sed script developed for the Penn Treebank (http://www.cis.upenn.edu/~treebank/tokenizer.sed)
4. **Part-of-speech (POS) tagger**. The statistical POS tagger (Brill 1995) assigns a part of speech label to each word in the text. We use this information to detect verbs as explained later.
5. **XML generator**. The XML generator transforms the output of the part-of-speech tagger to XML. We use only four tags: (1) PAPER, which is the root tag for each file; (2) S, for "Sentence"; (3) W, for "Word," which has a POS attribute; and (4) GPN, for "gene or protein name". A very simple example XML file is shown in Figure 1:

```
<PAPER>
 <S>
   <GPN>A</GPN>
   <W POS="VBZ">is</W>
   <W POS="VBN">activated</W>
   <W POS="IN">by</W>
   <GPN>B</GPN>
   <W POS=".">.</W>
 </S>
</PAPER>
```

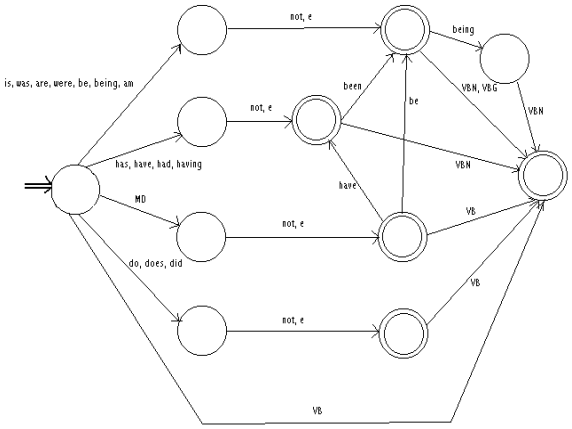

Figure 1: An XML File for an artificially simple article. The article has only one sentence, "A is activated by B." A and B are GPNs; PAPER is the root tag; S stands for "sentence"; and W stands for "word", which has a POS (part-of-speech) attribute.

Once all files in a corpus of biological texts have been annotated and transformed to XML as described above, our system detects verb groups and subsequently finds GPNs that are close to these verb groups, either before or after the verb. AVAD collects the "before" and "after" counts for each verb in the corpus. Similar counts can also be obtained from corpora in other domains, to compare with the frequencies of verbs in the biology domain.

Using the part of speech labels, we have built finite state machines (FSMs) to detect combinations of verbs and auxiliaries that comprise a single verb group. We automatically detect the head (main verb) in a verb group, and associate it with any GPNs to the left and right of the verb group. Detected verbs are normalized to a canonical form, using the SCOL stemmer available from http://www.sfs.nphil.uni-tuebingen.de/~abney, so that statistics for all morphological variants of the same verb will be collected together. Figure 2 shows the finite state machine used to detect verb groups starting from an observed GPN. The detection algorithm uses a parameter that controls how close the GPN and the verb group must be to consider their association a valid one. We have experimented with values in the range of 0 to 4

Figure 2: The Finite State Machine for finding the head verb after a GPN. When the FSM stops at one of its end states, it returns the last-met verb as the head verb.

intervening tokens, observing little difference in the final results of AVAD. Note that our algorithms for detecting an association between verbs and GPNs simulate locally a dependency parser to find the head verb for a GPN subject (*after)* or a GPN object (*before)*. We have found that these finite-state methods offer reasonable accuracy for this specialized task, thus avoiding the intensive computation that a full parser would require.

2. CLASSIFYING VERBS

After association counts have been collected for all verbs in the corpus, we have a big table in which each verb has a row with "GPN before" and "GPN after" frequencies, as well as the total frequency of the verb. Next, an appropriate statistical test is needed to rank the verbs in descending order of their likelihood of being an interaction verb. We have applied Pearson's χ^2 (chi-square) test and its variant commonly known as the proportions test. Under the latter, we assume:

(1) The ratio of the "before" *(or* "after"*)* frequency to the total frequency of an interaction verb is higher than the corresponding ratio for a common (non-interaction) verb.

To apply the test, we need to estimate the ratio for a common verb. We estimate the "before", "after", and total frequency of a common verb by summing all the frequencies of the verbs in the table, except those of the verb in question. We can use this estimation method because we assume that the interaction verbs form a small subset of all the verbs, and that the sum of the frequencies actually reflects the true distribution of the frequencies for a common verb. For each verb, we apply the proportions test twice, for the before and after counts. The test hypotheses are given below

$$H_{0,position}: \quad r_{verb,position} = r_{common,position}$$

$$H_{1,position}: \quad r_{verb,position} > r_{common,position}$$

(2)

where r *means ratio and* position *is either "before" or "after". Using a contingency table with four cells corresponding to the before/after and total frequencies of the verb in question and all other verbs, we can calculate the* χ^2 *statistic for both the original* χ^2 *test and the proportions test.*

We combine the results of the "before" and "after" tests in two ways: either by requiring that both $H_{1,before}$ *and* $H_{1,after}$ *are true (conjunction) or that either of them is true (disjunction). We would normally expect conjunction to perform better, as an interaction verb normally has biological substances as both subject and object. However, due to the limited GPN dictionary and possible verb-GPN link detection errors, we tested the disjunction rule as an alternative.*

In addition to the two tests involving the before or after frequencies of each verb, we also consider the difference between the rate of occurrence of a verb between a corpus of biological articles and other collections of text in other domains. We measure differences in these rates of occurrence with the log-likelihood test (Rayson and Garside, 2000), calculating that value for each verb and each other domain that we examine. We use the log-likelihood values together with our previously computed results of the before/after tests as features in a logistic regression model that constitutes another way to combine information from the different indicators and predict whether a verb belongs to the interaction verb class.

3. RESULTS AND EVALUATION

For the experiments reported in this paper, we used 1,381 HTML articles extracted from the European Molecular Biology Organization (EMBO) Journal Online (http://www.emboj.org/) to form our corpus of biological articles. This corpus contains 10,931,907 words. For the purpose of comparing verb frequencies with those in other domains, we used two additional corpora: a collection of one year of articles from the Wall Street Journal, including general news articles but focusing primarily on financial news (22,503,667 words), and a set of 29,784 articles from 20 cardiology journals (88,944,123 words).

4.1 Experiment I

In this experiment, without looking at context, experts with M.S. or Ph.D. degrees in biology and related disciplines such as mathematical genetics labeled 647 (48% of the total) verbs as positive (interaction verbs) out of 1,346 verbs in the EMBO corpus. Only verbs occurring more than 15 times in the corpus were supplied to the experts. Using the "after" test, the "before" test, and the conjunction and disjunction of the "after" and "before" tests at the significance level of 5%, we give the precision, recall, and F-measure of the χ^2 test and the proportions test in Table 1 and Table 2 respectively. Precision is the percentage of correctly classified interaction verbs among those that the system reports as interaction verbs; recall is the percentage of correctly classified interaction verbs among all verbs labeled as interaction verbs by the experts. The F-measure (vanRijsbergen 1979) combines the usually competing measures of precision and recall in a single number with equal weights.

Table 1: The Results of the χ^2 Test.

	Precision	Recall	F
Before	51.4%	32.9%	40.1%
After	54.3%	36.8%	43.9%
Conjunction	53.9%	21.5%	30.7%
Disjunction	52.5%	48.2%	50.3%

Table 2: The Results of the Proportions Test.

	Precision	Recall	F
Before	64.4%	23.2%	34.1%
After	70.5%	28.4%	40.5%
Conjunction	78.2%	13.3%	22.7%
Disjunction	64.6%	38.3%	48.1%

Generally, the precision of the proportions test is higher than that of the χ^2 *test but the recall is lower. Also, as expected, the conjunction rule between the before and after tests leads to higher precision (and lower recall) than either test alone, while the opposite is true for the disjunction rule.*

We subsequently fit a log-linear (logistic regression) model on the features of a verb, including the total frequency, the before and after

frequency, the proportions and χ^2 test statistics, the ranks in the two sorted lists, and the log-likelihood tests between the biology and other domains. We randomly select 2/3 of the verbs as the training set to fit the model on, and then use the fitted model on the test set, the remaining 1/3 verbs. We repeat the procedure for 10 times with different random splits and compute the averages. We analyzed models of various orders of feature interaction; Table 3 shows the results for an order 2 model on all features. The combined model offers the best performance, outperforming any single test or feature or the conjunction or disjunction rules alone.

Table 3: Average Results of the Log-Linear Model with Interaction Term Order 2 on All the Features.

	Precision	Recall	F
Training	71.7%	68.9%	70.3%
Test	61.1%	58.0%	59.5%

4.2 Experiment II

Our best results from Experiment I (Table 3) indicate around 60% precision and recall on unseen data. We analyzed the cases where the system disagreed with the labels assigned by the experts, and followed this analysis with discussions with them. We found, to our surprise, that the experts would often revise their decisions when presented with examples where verbs were used as interaction verbs (or the opposite). Thus, we designed a second experiment, aiming to create another gold standard where the experts would be more confident in their labels.

We randomly selected 150 verbs, and supplied to experts 10 example sentences where each occurred. By viewing the verbs in context, the experts were more certain of their status as interaction or non-interaction verbs. Using a strict criterion that interaction verbs act as such in almost all the supplied example sentences, only 17 of the 150 verbs were labeled as interaction verbs.

We then repeated the calculations of the statistical tests and the training and testing of the log-linear models. We show in Table 4 results from the proportions test (which performed better than the χ^2 test) at different levels of confidence. The log-linear model performed slightly worse than the proportions test on this data, possibly because of the small number of labeled samples.

Table 4: Performance of AVAD Using the Proportions Test and Conjunction/Disjunction Rules at Different Significance Levels.

		Precision	Recall	F
α=1%	Conjunction	100%	58.8%	74.1%
	Disjunction	85.7%	76.5%	80.9%
α=5%	Conjunction	86.2%	78.4%	82.1%
	Disjunction	39.5%	88.2%	54.5%
α=10%	Conjunction	87.5%	82.4%	84.9%
	Disjunction	37.2%	94.1%	53.3%

4. CONCLUSION

We have described AVAD, a system that automatically discovers interaction verbs between genes and proteins. The system achieves respectable precision (61.1%) and recall (58.0%) when it categorizes interaction verbs marked by experts out of context. But when the evaluation is focused on the cases where the experts can safely label the verbs by checking their contexts, performance rises to 87.5% precision and 82.4% recall.

The system is in addition able to recover interaction verbs that are relatively infrequent or specialized, and are thus unlikely to be captured during manual knowledge engineering. For example, AVAD automatically classified *co-localize* and *synergize* as interaction verbs, both of which do not appear in the detailed knowledge model for interaction verbs constructed for the GeneWays system (Rzhetsky et al. 2000). In fact, AVAD grew out of our desire to increase GeneWays' coverage for interaction verbs.[1]

Our approach may be used by current interaction extraction systems as an extension or refinement by automatically enlarging the size of the interaction verb sets they use. It is also an important step in our automatic discovery of interaction patterns from large biological corpora. We plan to extend its coverage to interactions among other biological substances in addition to genes and proteins, such as tRNA, mRNA, and other molecules, by including the names of these substances in the dictionary. Extending our current coverage of verb forms to deverbal

[1] The authors are part of the interdisciplinary team that is building GeneWays at Columbia University.

nominal forms (e.g., *activation*) is another goal of future work.

REFERENCES

Benson, D. A., M. S. Boguski, D. J. Lipman, J. Ostell, B. F. Ouellette, B. A. Rapp, and D. L. Wheeler (1999) GenBank. *Nucl. Acids Res.* **27**(1):12–17.

Blaschke, C., M. A. Andrade, C. Ouzounis, and A. Valencia (1999) Automatic Extraction of Biological Information from Scientific Text: Protein-Protein Interactions. *In Proceedings of the 7th Conference on Intelligent Systems in Molecular Biology,* 60–67.

Brill, E. (1995) Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging. *Computational Linguistics* **21**(4):543–565.

Park, J. C., H. S. Kim, and J. J. Kim (2001) Bidirectional Incremental Parsing for Automatic Pathway Identification with Combinatory Categorical Grammar. *Pacific Symposium on Biocomputing,* **6**, 396–407.

Proux, D., F. Rechenmann, and L. Julliard (2000) A Pragmatic Information Extraction Strategy for Gathering Data on Genetic Interaction. *In Proc. of 8th International Conference on Intelligent Systems for Molecular Biology*, La Jolla, Calif., pp 279– 285.

Rayson, P. and Garside, R. (2000). Comparing Corpora Using Frequency Profiling. *In Proc. of the Workshop on Comparing Corpora, 38th ACL*, Hong Kong, pp. 1–6.

Reynar, J. C. and A. Ratnaparkhi (1997) A Maximum Entropy Approach to Identifying Sentence Boundaries. *In Proceedings of the Fifth Conference on Applied Natural Language Processing*, Washington, D.C.

van Rijsbergen, C. J. (1979) *Information Retrieval* (2nd edition). London:Butterworths.

Rzhetsky, A., T. Koike, S. Kalachikov, S. M. Gomez, M. Krauthammer, S. H. Kaplan, P. Kra,

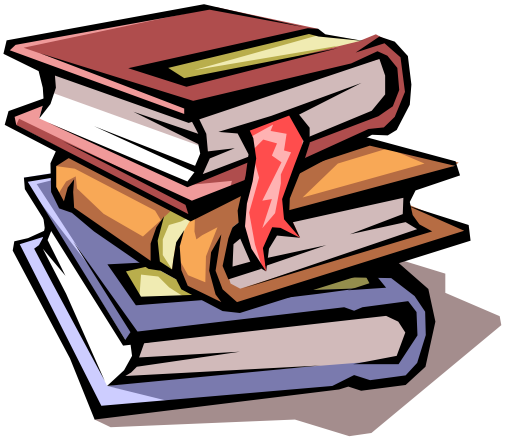

Figure 1. Caption, Times 12pt, Bold

For more complex tasks, where state variables must be maintained throughout a sequence of screens, we have developed a simple procedural language. This language can control the sequencing of HTML documents, execute validation rules, save and restore arbitrary data elements, access the environment variables, and trigger actions in the…

For example, for the process of discharging patients treated for acute myocardial infarctions, the Cardiology Service uses this technique (Figure 1): several HTML forms are used to capture information about key aspects of the hospitalization, risk factors, future appointments, discharge medications, and various recommendations for the patient. Physicians planning the discharge can be asked to justify why certain medications (such as aspirin or a beta-blocker) were not prescribed. As a result, structured data useful for quality assurance is captured. Incentives for resident-physician end-users include the automated generation of prescriptions, discharge instructions for nurses, a customized letter for the patient, and a discharge note which becomes immediately available, at a time before the complete discharge summary can be dictated.

ANOTHER CHAPTER NAME

Some decision-support tools require a high level of interactivity, which cannot be provided by the…

CONCLUSION

The maintenance of a clinical decision-support system's knowledge base can be effectively distributed to its various stakeholders. A formal mechanism…

References

1. Forsythe DE, Buchanan BG, Osheroff JA, Miller RA. Expanding the concept of medical information: an observational study of physicians' information needs. Comput Biomed Res 1992;25:181-200
2. McDonald CJ, Murray R, Jerus D,

Bhargava B, Seeger J, Blevins L. A Computer-based record and clinical monitoring system for ambulatory care. Am J Pub Health. 1977;67:240-245