ELSEVIER

International Journal of Medical Informatics 67 (2002) 19–32

International Journal of Medical Informatics

www.elsevier.com/locate/ijmedinf

Learning anchor verbs for biological interaction patterns from published text articles

Vasileios Hatzivassiloglou *, Wubin Weng

Department of Computer Science, Columbia University, 1214 Amsterdam Avenue, New York, NY 10027, USA

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 of the resulting knowledge base. In this paper, we describe an automatic system that learns from text interaction verbs; these verbs can then form the core of automatically retrieved patterns which 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. In addition, we compare the automatically obtained results with a manually constructed database of interaction verbs and show that the automatic approach can significantly enrich the manual list by detecting rarer interaction verbs that were omitted from the database.

Keywords: Protein–protein interactions; Protein–gene interactions; Interaction verbs; Computer analysis of biological text; Text mining; Machine learning

1. 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. Even a molecular biologist working 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 [1,2], or even are fully instantiated in a knowledge database or by a semantic grammar [3,4].

Thus, current approaches perform automatic interaction extraction based on patterns that are already known. Their power is greatly

* Corresponding author
E-mail address: vh@cs.columbia.edu (V. Hatzivassiloglou).

1386-5056/02/$ - see front matter
PII: S1386-5056(02)00054-0

limited by the small set of pre-defined interaction verbs used in the patterns. For instance, Blaschke et al. [1] used a set of 14 prespecified verbs that denoted actions related to protein interactions; Proux et al. [2] 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 et al. [5] 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 automatically recover 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.

Although AVAD populates a knowledge base with interaction verbs and such a knowledge base can be used for information extraction with either pre-defined or automatically learned patterns, AVAD is not itself an information extraction system or even a pattern learning system. AVAD operates earlier in the knowledge modeling pipeline. It learns without human intervention the anchor words (interaction verbs) with which patterns can be extracted or automatically learned. Those patterns can then be used in a follow-up information extraction system in similar ways as the more limited manually constructed patterns are currently used.

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. Section 5 contains a discussion of experimental factors that affect the evaluation and AVAD's performance; we note limitations in the data and the language tools used by AVAD that reduce the observed precision and recall, but do not reflect fundamental limitations of the approach. Finally, Section 6 presents a comparison between a manually constructed database of interaction verbs and several sets of verbs produced by AVAD; we establish that the output of the automated system can complement the manual list by detecting interaction verbs missed during the knowledge engineering process.

2. 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 (e.g. 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 online 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 2783 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 [6], 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 MXTERMINATOR [7] (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 part-of-speech tagger [8] assigns a part of speech label, such as noun, adjective or verb, to each word in the text. We use this information to detect verbs and verb groups, 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 Fig. 1.

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 separately 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 automatic 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. Fig. 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 valid. We have experimented with values in the range of 0–4 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, such as Refs. [9] or [10], would require.

```
<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>
```

Fig. 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. In this example, VBZ stands for 'verb in present tense, third person singular', VBN for 'verb, past participle' and IN for 'preposition'.

3. 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-test and its one-sided variant, commonly known as the proportions test [11]. Under the latter, we assume:

> *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.*

(1)

To apply the test, we need to estimate the

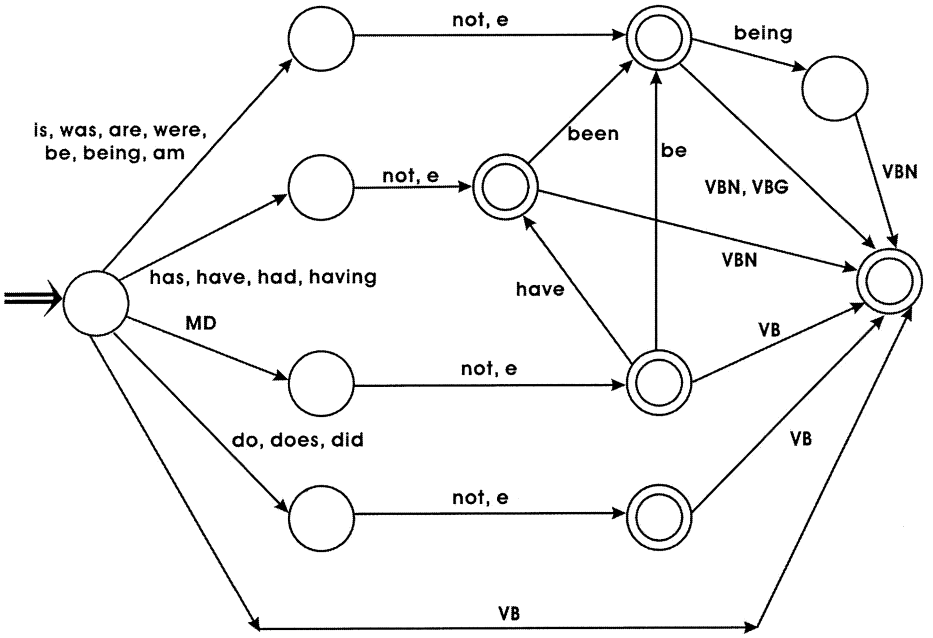

Fig. 2. The finite state machine for finding the head verb after a GPN. When the FSM stops at one of its end (double-circled) states, it returns the last-met verb as the head verb. 'e' Stands for the empty transition.

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. For each verb, we apply the proportions test twice, for the before and after counts. The test hypotheses are given below

$$H_{0,\text{position}}\colon\ r_{\text{verb,position}} = r_{\text{common,position}}$$

$$H_{1,\text{position}}\colon\ r_{\text{verb,position}} > r_{\text{common,position}} \qquad (2)$$

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 $H_{1,\text{before}}$ and $H_{1,\text{after}}$ 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 [12], 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 [13] that constitutes another way to combine information from the different indicators and predict whether a verb

belongs to the interaction verb class. In classical linear modeling, the response variable y is modeled as $y = \mathbf{b}^{\mathrm{T}}\mathbf{x} + e$, where $\mathbf{b}$ is a vector of weights, $\mathbf{x}$ is the vector of the values of the predictor variables and e is an error term which is assumed to be normally distributed with zero mean and constant variance, independent of the mean of y. The log-linear regression model generalizes this setting to binomial sampling where the response variable follows a Bernoulli distribution (corresponding to a two-category outcome); note that the variance of the error term is not independent of the mean of y anymore. In Section 4, we present the results of fitting regression models of different interaction orders to these features and compare the performance of the regression models to the conjunction and disjunction rules.

4. Results and evaluation

For the experiments reported in this paper, we used 1381 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 1 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 1346 verbs in the EMBO corpus. Only verbs occurring more than 15 times in the corpus were supplied to the experts. This was carried out to alleviate the data sparseness issue, as verbs with very low counts would likely have unreliable statistics. 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 [14] combines the usually competing measures of precision and recall in a single number with equal weights.

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 two-

Table 1
The results of the χ^2 test for Experiment I

	Precision (%)	Recall (%)	F-measure (%)
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 for Experiment I

	Precision (%)	Recall (%)	F-measure (%)
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

thirds of the verbs as the training set to fit the model on and then use the fitted model on the test set, the remaining one-third of the verbs. We repeat the procedure ten 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 regression-based model utilizes information from both within and outside the corpus of biological articles (by comparing verb frequencies with the corresponding frequencies in the Wall Street Journal and medical corpus) and optimizes the weights for the various counts and test probabilities according to the training data. The combined model offers the best performance, outperforming any single test or feature or the conjunction or disjunction rules alone.

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.

Table 3
Average results of the log-linear model with interaction terms of order 2 on all the features for Experiment I

	Precision (%)	Recall (%)	F-measure (%)
Training	71.7	68.9	70.3
Test	61.1	58.0	59.5

We randomly selected 150 verbs, and supplied to experts ten 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, namely 'aggregate', 'assemble', 'associate', 'attach', 'bundle', 'cleave', 'co-express', 'co-immunoprecipitate', 'co-migrate', 'co-precipitate', 'co-localize', 'disrupt', 'inactivate', 'relocate', 'repress', 'synergize' and 'translocate'.

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. As the significance level α increases, the precision rates drop, but the recall rates increase. As in Experiment I, the precision of conjunction is higher than that of disjunction and recall exhibits the opposite behavior. Table 5 shows the corresponding results for the χ^2-test, which achieves consistently high recall (as in Experiment I), but significantly lower precision and overall F-measure than the proportions test.

The system achieves the best overall result of F-measure 84.9% when using the proportions test by conjunction at a significance level of 10%. This is significantly higher than the result using the χ^2-test whose best F-measure is 55.6%. It is also better than the best result of Experiment I, F-measure 59.5% on the test set

Table 4
Performance of AVAD using the proportions test and conjunction/disjunction rules at different significance levels on the gold standard of Experiment II

		Precision (%)	Recall (%)	F-measure (%)
$\alpha = 1\%$	Conjunction	100	58.8	74.1
	Disjunction	45.5	88.2	60.0
$\alpha = 5\%$	Conjunction	86.7	76.5	81.3
	Disjunction	39.5	88.2	54.5
$\alpha = 10\%$	Conjunction	87.5	82.4	84.9
	Disjunction	37.2	94.1	53.3

using the linear model (Table 3). This demonstrates that when the labels of the verbs become more accurate, the performance of AVAD improves in both precision and recall when using the proportions test. The log-linear model performed slightly worse than the proportions test on this data, which we attribute to the small number of positively labeled samples. Table 6 lists the average scores obtained by the log-linear model on the training and test data; note that the model achieves a much better performance on the training data, indicating that overfitting is taking place. Because of the low number of verbs labeled as interaction verbs in Experiment II (11%), only a first-order log-linear model can be fitted (models of higher order are singular).

Table 6
Performance of AVAD using the log-linear regression model with no interaction terms on the gold standard of Experiment II

	Precision (%)	Recall (%)	F-measure (%)
Training	92.7	78.9	85.3
Test	73.4	76.3	74.8

Table 4 shows that AVAD achieves precision, recall and F-measure above 80% using the conjunction rule and a relatively high confidence level for the statistical test. These numbers represent a significant improvement over our results for Experiment I, thus indicating the importance of providing the experts with enough information to properly assess whether a relatively not so frequent verb is an interaction verb. Note that it would have been worthwhile to measure the consistency of expert decisions during both experiments and in particular for the first experiment. One such measure of consistency is the rate of agreement between experts on

Table 5
Performance of AVAD using the χ^2 test and conjunction/disjunction rules at different significance levels on the gold standard of Experiment II

		Precision (%)	Recall (%)	F-measure (%)
$\alpha = 1\%$	Conjunction	52.6	58.8	55.6
	Disjunction	29.4	88.2	44.1
$\alpha = 5\%$	Conjunction	40.6	76.5	53.1
	Disjunction	22.4	88.2	35.7
$\alpha = 10\%$	Conjunction	37.1	76.5	50.0
	Disjunction	21.1	88.2	34.1

the same data. However, for Experiments I and II reported above, the two participating experts marked separate subsets of verbs and each verb was marked by only one expert, due to limited time and availability of experts. We analyze a shared sample of verbs marked by multiple experts to measure expert agreement in Section 6.

5. An analysis of AVAD's errors

In this section, we examine AVAD's errors according to the data from Experiment II ($\approx$ 15% of the system's decision are in error according to the gold standard of that experiment). We are looking for systematic patterns of errors and attempt to classify them into those that reflect a true limitation of AVAD's statistical approach and those that can be attributed to the limited amount of data we have for training or represent limitations of the tools that AVAD uses, such as the part-of-speech tagger.

To find the reasons for misclassification, we check in context the misclassified verbs. We ranked the verbs in ascending order of their proportions test P-values. Then we check those verbs with top ranks and bottom ranks. If there are common verbs among those verbs with top ranks, the precision will be damaged and if there are interaction verbs among those verbs with bottom ranks, the recall will be reduced. We find several factors that may affect the precision:

1) Some common, non-interaction verbs have high frequency in biological texts, either in the 'before' case or in the 'after' case, such as 'detect', 'identify', 'play', 'characterize', etc. They are frequently used to describe an experiment, to analyze results, to state a fact, etc. They take a GPN as a subject or object, but they do not indicate any interaction. An example is shown in Fig. 3(a), where *detect* is not an interaction verb but it is a 'before' head verb for the GPN. Among those verbs, some appear almost equally significantly in the 'before' and 'after' cases. We call them 'balanced' verbs. An example is 'detect', which appears 535 times in the 'after' case and 198 times in the 'before' case out of a total frequency 6464. The P-values of the proportions test are $< 10^{-16}$ and 0.016 for the 'before' and 'after' cases, respectively. Both are significant at the 5% significance level. These verbs cannot be eliminated by the conjunction rule at the significance level of 5%. If they are unbalanced, however, like 'play', which is frequently used in the active form—'GPN plays a role in...'—and is seldom used in the passive form, they will be rejected by the conjunction rule. But the disjunction rule does not work for either of the 'balanced' and 'unbalanced' cases. This also explains why the precision of the disjunction rule is significantly lower than that of the conjunction rule.
2) The errors caused by the part-of-speech tagger (with 95% reported accuracy, but trained on a completely different domain) can also affect the precision because part-of-speech labels are used to find the 'before' verb and the 'after' verb (Section 2). For example, in the sentence in Fig. 3(b), 'was' is wrongly found as the 'after' head verb because the part of speech of 'co-expressed' is mistakenly reported as an adjective, 'JJ', instead of a past participle, 'VBN'. When the FSM meets 'co-expressed', it stops and returns the last-met verb, 'was', as the head verb.
3) The 'before' verb or the 'after' verb found by the algorithm in Section 2 is not always the actual head verb for the GPN in the sentences because our head verb detection

(a) *The 9E10 monoclonal antibody was used to* ***detect*** *myc-tagged* GPN.
(b) *FLAG-tagged BAG-1 and Siah-1A* GPN ***was co-expressed*** *in GM701 cells and their distribution patterns were determined by indirect immunofluorescence.*
(c) *The homology of BCAR3 with other SH2 domain-containing* GPN *is* ***limited*** *to this domain.*

Fig. 3. Some example sentences indicating where AVAD can be misled.

algorithm does not analyze the structure of the sentences (e.g. prepositional phrases embedded in complex noun phrases). An example can be found in the sentence in Fig. 3(c) where *limited* is not the 'after' head verb for the GPN but is the head verb for *homology*. On the other hand, the reason why the recall drops is that there is not enough evidence, that is, either the 'before' or the 'after' frequency of a verb is not large enough to get a significant test result. The following factors may affect the recall:

4) The GPN dictionary is not large enough to cover all the GPNs in the corpus. For instance, the interaction verb 'cleave' appears 1112 times in the corpus, but for only 27 times it is caught as an 'after' verb and for 24 times as a 'before' verb. In many cases, the GPNs near it are not tagged as such by the GPN tagger. For those interaction verbs with a low total frequency, for instance, lower than 50, the situation will be worse when the GPNs near it are not tagged because it will greatly affect the result of the proportions test since the denominator of the ratio is small and a little fluctuation of the numerator may cause a big fluctuation of the ratio.
5) The errors caused by the POS tagger may also affect the recall rate. Take the sentence in Fig. 3(b) as an example again. Because 'co-expressed' is wrongly tagged as 'JJ' instead of 'VBN', the searching algorithm returns 'was' as the head verb. This reduces the 'after' frequency of 'co-express' by one. If there were many cases in the corpus like this or too few occurrences of 'co-express' to start with, the frequency of 'co-express' appearing near a GPN would not be large enough to obtain a significant test result.

This analysis indicates that a large part of the current errors made by AVAD are due to either errors made by external tools (factors 2 and 5), errors made by our verb detection finite state grammar (factor 3) or the limited size of the GPN dictionary we currently use (factor 4). Only the first of the items above represents an inherent limitation of the method. We expect that as the field of text analysis with biological documents matures, specialized tools, such as part-of-speech taggers trained for this domain and even efficient finite-state parsers that partially recover sentence structure more accurately will become available. The existence of such tools ([15]) has helped improve the accuracy of text analysis systems in other information extraction applications. Finally, obtaining a larger high-quality collection of GPNs, whether manually built or filtered from entries in databases, such as GenBank, is certainly a possibility.

6. Comparison with a manually built list of activation verbs

AVAD automatically produces a list of interaction verbs by examining a large number of biology articles. As noted in the introduction, such lists are useful for modeling

protein–protein and gene–protein interactions and consequently have been built by hand before. In this section, we compare the sets of interaction verbs produced by AVAD under different configurations to a reference list of interaction verbs, the one used in the GeneWays system.

GeneWays, currently under construction at Columbia University, is a system for the automatic analysis of large amounts of textual data in the biology domain. GeneWays annotates journal articles, locating genes and proteins, extracting relationships between them, validating those relationships across articles and representing them graphically. In addition to the validation, database management and visualization components, its text analysis component includes algorithms for the detection and disambiguation of genes and proteins in text and for the detection of relationships that match patterns known to the system. Currently, the patterns that represent the system's understanding of biological relationships are encoded as part of GeneWays detailed knowledge model, which was built by hand by consulting experts in the domain of pathway analysis [16]. These patterns are then represented as rules in a semantic grammar and a parser adapted to the biological domain [17], extracts new instances of gene and protein relationships as new texts arrive.

The authors of the present paper are part of the multidisciplinary team that is building GeneWays and in fact, our work on AVAD was motivated by the need to augment and adapt the manually built knowledge base of interaction verbs and relationship patterns that is currently used in the GeneWays system. Selecting the interaction verbs for that knowledge base and specifying restrictions on their subjects and objects took a considerable amount of time during the design of the knowledge base; updates to the knowledge base either as new biological subdomains are considered or simply as the field evolves are difficult to perform. Therefore, there is a significant benefit from having access to automatically produced lists of interaction verbs, even if the results of an automated system, such as AVAD, need curation by experts before being incorporated in the knowledge base.

In Table 7 we examine six different configurations of AVAD (using the proportions test and either the conjunction or disjunction rule and three different significance levels: 1, 5 and 10%). These configurations produce different numbers of interaction verbs, with those being most productive expected to have lesser accuracy. We compare AVAD's output with the manually built list of interaction verbs in GeneWays, which contains 96 single-word verbs and 15 multiword verb expressions (phrasal verbs and other verb groups, such as 'suppress activity of'); we base our comparisons only on the single-word verbs, since AVAD only detects interaction verbs of this type. We asked two biologists, both with Ph.D. degrees in molecular biology or a related discipline, to judge each of the verbs proposed by AVAD and mark it as either an interaction verb or a common, discourse verb; to keep the number of judgments manageable, these decisions were made without looking at text examples of the use of the verbs, so the caveats mentioned earlier during our discussion of Experiments I and II apply. As a control, we included the manually constructed entries in the GeneWays knowledge base, for a total of 462 verbs or verb groups (447 single-word verbs). For each AVAD configuration, we list how many interaction verbs it reports, how many of those are judged as correct by at least one or both evaluators, how many are found among the 96 single-word verbs in the

Table 7
Comparison of verbs in AVAD's output, the manually built knowledge base for GeneWays and two expert biologists' assignments of status as interaction verbs or not

	and, 1%	and, 5%	and, 10%	or, 1%	or, 5%	or, 10%
Verbs proposed by AVAD	95	110	131	338	384	423
Verbs found in knowledge base	34	37	39	55	57	57
Coverage of knowledge base (%)	35.4	38.5	40.6	57.3	59.4	59.4
Correct interaction verbs proposed[a]	87	99	120	268	306	333
Correct interaction verbs proposed[b]	70	76	90	184	204	219
Overall precision[a] (%)	91.6	90.0	91.6	79.3	79.7	78.7
Overall precision[b] (%)	73.7	69.1	68.7	54.4	53.1	51.8
Verbs not in the knowledge base	61	73	92	283	327	366
Correct interaction verbs not in KB[a]	53	62	81	214	250	277
Correct interaction verbs not in KB[b]	40	43	55	136	154	169
Precision on verbs not in KB[a] (%)	86.9	84.9	88.0	75.6	76.5	75.7
Precision on verbs not in KB[b] (%)	65.6	58.9	59.8	48.0	47.1	46.2

The six system configurations are denoted by 'and' or 'or' for conjunction and disjunction, respectively, and the significance level applied on the output of the proportions test.
[a] For measures involving a gold standard a verb is considered an interaction verb if at least one of the two experts rate it as such.
[b] For measures involving a gold standard a verb is considered an interaction verb only if both experts agree.

manually built GeneWays knowledge base and how many of the verbs *not* in GeneWays' knowledge base at least one or both of the experts consider correct.

We observe from Table 7 that AVAD covers 35–40% of the knowledge base with the conjunction rule and more than half with the more permissive disjunction rule. Its precision is 80–90% if we apply the broader interpretation considering a verb an interaction verb if one of the experts says so and 70% on the stricter gold standard that requires agreement between the experts; precision remains $>50\%$ in all cases even with the disjunction rule. Most significantly, AVAD proposes a substantial number of verbs not in the knowledge base and its precision on those verbs is only about 5% less than its overall precision. This means that AVAD's output can be used to effectively quadruple the size of the knowledge base after curation, while only proposing a verb that will be rejected by the experts about one-third of the time. By using a corpus larger than our current collection of 11 million words, it is likely that many more interaction verbs can be reliably extracted.

We also measured the agreement between the two experts and found that they assign the same label to verbs in 355 of the 462 cases, or 77% of the time. This shows that the problem of determining the status of a verb as an interaction verb or not is a non-trivial one. It is not a simple task for the biologists to unambiguously choose the interaction verbs from a list, much less construct the list from scratch in the first place. Interestingly, there were several cases where the experts considered the entries in the manually built knowledge base to be invalid: in four of the 96 single-word entries, both experts labeled the verb as a common verb, and in an additional 11 cases, one of them did so (there were two and one additional such cases, respectively, among the 15 multiword verb phrases in the knowledge base). This results in a precision for the manually built list of 94.6% under the

first standard and 83.8% under the second, stricter standard.

7. 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 [16]. Our analysis showed that AVAD proposes a large number of verbs that humans did not include in the knowledge base and that most of these verbs should in fact be included in the knowledge base.

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

Acknowledgements

The authors thank Andrey Rzhetsky for providing valuable input during the design of the system and serving as one of the expert judges. Michael Krauthammer and Pavel Morozov served as the other two judges and we are grateful for the hours they spent marking verb lists for us. We also thank Carol Friedman and Pauline Kra for providing us with the manually constructed list of interaction verbs currently used in the GeneWays system, which we compared to the automatically identified interaction verbs. Finally, we thank Adam P. Arkin who supplied us with the gene and protein dictionary used in the experiments reported here. This work was supported in part by National Science Foundation Information Technology Research Award No. 0121687, as well as by grants from the National Institutes of Health and the New York State Science and Technology Foundation. All opinions, findings and recommendations reported in this article are those of the authors and do not necessarily represent the views of the NSF, NIH or NYSSTF.

References

[1] C. Blaschke, M.A. Andrade, C. Ouzounis, A. Valencia, Automatic extraction of biological information from scientific text: protein–protein interactions, Proceedings of the Seventh Conference on Intelligent Systems in Molecular Biology, 1999, pp. 60–67.

[2] D. Proux, F. Rechenmann, L. Julliard, A pragmatic information extraction strategy for gathering data on genetic interaction, Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology, La Jolla, CA, 2000, pp. 279–285.

[3] J.C. Park, H.S. Kim, J.J. Kim, Bidirectional incremental parsing for automatic pathway identification with combi-

natory categorical grammar, Pacific Symp. Biocomput. 6 (2001) 396–407.
[4] A. Yakushiji, Y. Tateisi, Y. Miyao, J. Tsujii, Event extraction from biomedical papers using a full parser, Proceedings of Pacific Symposium on Biocomputing, 2000, pp. 408–419.
[5] T. Sekimizu, H.S. Park, J. Tsujii, Identifying the interaction between genes and gene products based on frequently seen verbs in MedLine abstracts, in: Genome Informatics, Universal Academy Press, Tokyo, 1998.
[6] D.A. Benson, M.S. Boguski, D.J. Lipman, J. Ostell, B.F. Ouellette, B.A. Rapp, D.L. Wheeler, GenBank, Nucl. Acids Res. 27 (1) (1999) 12–17.
[7] J.C. Reynar, A. Ratnaparkhi, A maximum entropy approach to identifying sentence boundaries. Proceedings of the Fifth Conference on Applied Natural Language Processing, Washington, DC, 1997.
[8] E. Brill, Transformation-based error-driven learning and natural language processing: a case study in part of speech tagging, Comput. Linguist. 21 (4) (1995) 543–565.
[9] M.J. Collins, Three generative, lexicalized models for statistical parsing, Proceedings of the Thirty-fifth Annual Meeting of the Association for Computational Linguistics, Madrid, Spain, 1997, pp. 16–23.
[10] E. Charniak, Immediate head parsing for language models, Proceedings of the Thirty-ninth Annual Meeting of the Association for Computational Linguistics, Toulouse, France, 2001.
[11] J.L. Fleiss, Statistical Methods for Rates and Proportions, second ed, Wiley, New York, 1981.
[12] P. Rayson, R. Garside, R, Comparing corpora using frequency profiling, Proceedings of the Workshop on Comparing Corpora, Thirty-eighth ACL, Hong Kong, 2000, pp. 1–6.
[13] J. Santner, D.E. Duffy, The Statistical Analysis of Discrete Data, Springer Verlag, New York, 1989.
[14] C.J. van Rijsbergen, Information Retrieval, second ed, Butterworths, London, 1979.
[15] D. Appelt, J. Hobbs, J. Bear, D. Israel, M. Tyson, Fastus: a finite state processor for information extraction from real world text, Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence, Chambery, France, 1993.
[16] A. Rzhetsky, T. Koike, S. Kalachikov, S.M. Gomez, M. Krauthammer, S.H. Kaplan, P. Kra, J.J. Russo, C. Friedman, A knowledge model for analysis and simulation of regulatory networks, Bioinformatics 16 (2000) 1120–1128.
[17] C. Friedman, P. Kra, H. Yu, M. Krauthammer, A. Rzhetsky, GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles, Bioinformatics 17 (2001) S74–82.