2010

Text Mining for Systems Biology and MetNet

Lifeng Zhang
Iowa State University

Text mining for systems biology and MetNet

by

Lifeng Zhang

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Computer Engineering

Program of Study Committee:
Daniel Berleant, Co-Major Professor
Julie Dickerson, Co-Major Professor
Eve Wurtele
Diane Rover
Heike Hofmann

Iowa State University

Ames, Iowa

2010

Table of Contents

Abstract

The rapidly expanding volume of biological and biomedical literature motivates demand for more friendly access. Better automated mining of this literature can help find useful and desired citations and can extract new knowledge from the massive biological "literaturome." The research objectives presented here, when met, will provide comprehensive text mining utilities within the MetNet (Metabolic Network Exchange) (Wurtele et al., 2007), platform to help biologists visualize, explore, and analyze the biological literaturome. The overarching research question to be addressed is how to automatically extract biomolecular interactions from numerous biomedical texts. Here are the specific aims of this work.

1. Research on the text empirics of interaction-indicating terms to find more clues to improve the current algorithm applied in PathBinder to more precisely judge whether biomolecular interaction descriptions are present in sentences from the biological literature.
2. Based on these research results, extract interacting biomolecule pairs from literature and use those pairs to construct a biomolecule interaction database and network.
3. Integrate biomolecular interaction-indicating term extraction into MetNet's existing metabolomic network database.
4. Apply all of the above results in PathBinder software.
5. Quantitatively evaluate the success of algorithms developed based on the text empirics results.

This work is expected to advance systems biology by answering scientific questions about biological text empirics, by contributing to the engineering task of building MetNet and key constituent subsystems of MetNet, and by supporting the MetNet project through selected maintenance tasks.

Chapter 1. Background: text mining of biological literature for interaction extraction

In recent years more and more collections of gene sequence and expression data have appeared. Each gene's function and role in a pathway needs to be clarified and organized systematically. A lot of works involve creation of biomolecular interaction databases that are useful for understanding biological processes (Ono et al., 2001). A database could be populated through laboratory research, like MIPS (Pagel et al., 2005) and KEGG (Minoru et al., 2006). But a lot of information exists throughout the scientific literature about interactions. Considerable work has been done to extract interactions from literature, e.g. the Database of Interacting Proteins (Salwinski et al., 2004) and BIND (Bader et al., 2001). However many of these databases are manually populated. The interactions existing in the natural language of the literature are not easily extracted by humans, so the task to create such a database is time and labor intensive. Therefore, an increasing amount of work has focused on automatic interaction extraction from scientific literature based on text-mining technology in order to help researchers find knowledge from information already in the literature and even to construct interaction databases automatically.

Current topics of research in text mining of biomedical literature include named entity recognition, text classification, synonym and abbreviation extraction, and relationship extraction (Cohen et al., 2005). For example, Cohen et al. (2005), and Yu et al. (2002, 2003) used different methods including statistical methods, Supporte Vector machines, and pattern matching to extract synonyms. Similar methods also were used to extract genes and proteins from the literature (Nenadic et al., 2003; Tanabe et al., 2002). Most of this research is based on the MEDLINE literature collection.

MEDLINE, provided by the National Library of Medicine (http://www.nlm.hih.gov/pubs/factsheets/MEDLINE.html), and developed by the National Center for Biotechnology Information (NCBI), is one of the most well-known biomedical text information resources. It contains approximately 18 million references to articles in the life sciences, especially biomedicine. This database is growing at a high rate, challenging people to keep up with new scientific information. NCBI provides a query interface, PubMed (http://www.pubmed.com), to let users search stored citations and do many other tasks. It also provides the Entrez Programming Utilities (http://www.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html) to let developers write software that accesses this data.

A considerable amount of research has been published on improving retrieval from MEDLINE and other biomedical literature. SLIM (Muin et al., 2005) was developed by NCBI itself. It uses user-friendly slide bars to help users filter search results. GENIA (Kim et al., 2003) annotated MEDLINE citations with more detailed tags to make text mining on the corpus more convenient. Textpresso (Muller et al., 2004) maintains a database where full texts about *Caenorhabditis elegans* tagged with ontology terms are stored, so it can provide ontology-aware search on its corpus.

Some researchers have investigated new methods to retrieve citations besides occurrences of keywords. eTBLAST (Pertsemlidis et al., 2004; Lewis et al., 2006) breaks user queries and citations into vectors representing keyword occurrence rates and computes similarities among the vectors. It also investigates BLAST-like methods of aligning queries and documents. Similar investigations improving similarity rankings in retrieval include PubFinder (Goetz et al., 2005), SGO (Homayouni et al., 2005), and the work of Rubin et al. (2005).

Some works focus on analyzing search results from Entrez Programming Utilities provided by PubMed. XplorMed (Perez-Iratxeta et al., 2001, 2002, 2003) classifies results into MeSH categories and then displays keywords in all results where the keywords appear. The keywords are determined by their co-occurrence frequencies with other words. In addition, XplorMed finds words related to each keyword based on co-occurrences and finds documents where keywords co-occur. MedMiner (Tanabe et al., 1999) mainly aims to find good citations in MEDLINE for users. But it also can search for individual sentences with a desired keyword co-occurrence. The co-occurrence can be of biomolecules, and the keywords can include interaction-indicating terms. Thus the results of a search can be sentences containing a tri-occurrence of biomolecules and an interaction-indicating term.

Analysis of passages containing biological term co-occurrences or tri-occurrences enables the extraction of relations among biological entities. Starting mostly from the late 1990s, considerable research on interaction extraction from literature has been reported. There are different kinds of information that can be extracted from literature such as identifying synonyms (Cohen et al., 2005) and contrastive relationships, which are usually stated as contrastive negation patterns like "A but not B has some properties" (Kim et al. 2005). Most studies focus on extracting interaction relationships among genes or proteins. The resulting interaction databases can support annotation of gene function, which will be increasingly useful as more and more gene expression data become available. Interaction extraction also may identify information that is only implicitly present in the literature (Wren et al., 2004). Moreover, the extracted interactions can be further processed into interaction networks and other advanced resources (Yuryev et al., 2006).

1.1 Review of interaction extraction methods

To extract the right information, many systems that mine a literature database do part-of-speech (POS) tagging and then named entity recognition. Recognizing biological entities facilitates automatic biological relationship extraction. Usually this process detects an appearance of a pair of biological entities, sometimes plus a relationship-indicating word, and then extracts the relationship between the pair of entities. The type of extracted relationship may be decided by the entities and the relationship-indicating word. Most of them focus on the relationships among genes or proteins. The scale of such systems is potentially greater than manually inputting interactions because it can save a lot time and labor.

There are different methods to automatically extract the interactions between a pair of biomolecules from literature, including simple co-occurrence extraction methods, parsing and pattern matching methods, including natural language processing (NLP) based methods and statistical analysis methods (Hirschman et al., 2002; Yeh et al., 2003). Boundaries between the categories are not always clear. Some works in this area are reviewed next.

Extraction based on co-occurrences of biomolecules

The most direct method is to find an appearance of a co-occurrence of two objective biomolecules in literature. Dragon Plant Biology Explorer (DPBE) (Vladimir et al., 2005) parses documents provided by users using basic co-occurrence criteria. It shows the results to users in, among other forms, a network graph of interactions. Albert et al. (2003) applied this method to create a protein interaction database for nuclear receptors, and post-processed this database by manual curation to delete false interactions. PDQ Wizard (Grimes et al., 2005) and the work of Hofmann et al. (2005) also use co-occurrence to extract relations between biomolecules plus later filtering of extraction results. Whatizit (Rebholz-Schuhmann et al., 2008) provide a Web service to find co-occurrences of two biomolecules in submitted text. Kabiljo et al. (2009) used perl regular expressions to extract two gene/protein names occuring together within a sentence that has an interaction word between them. The interaction words list is from other projects.

The co-occurrence of biomolecule names in the literature also can be used in other ways. ACS (Jelier et al., 2005; Van der Eijk et al., 2004) placed biomolecules in a Euclidean space by mapping the co-occurrence relations between biomolecules to the space. It started putting nodes in the space randomly and used the co-occurrence of two nodes in one document to move their position in the space document by document. Then, it can find or evaluate interactions based on the distance between two biomolecules in the space, thus improving recall by finding those pairs not found to be co-occurring in literature. (*Recall* is the proportion of relevant items that are retrieved out of all relevant items available. This is compared to *Precision*, which is the proportion of relevant items in a retrieved set to all the items retrieved.) LMMA (Li et al., 2006) combines co-occurrence mining results and microarray analysis to construct a biological network. Bibliometrics (Stapley et al., 2000) use gene pairs co-occurring at least three times to construct a gene network.

Extraction based on parsing text

Co-occurrence methods are relatively simple and cannot, in theory, match the performance of methods that incorporate more information such as template matching and sentence parsing. Usually in the final part of these kinds of approaches, a sentence or an abstract or the parsed result is matched against predefined patterns to check whether an interaction is described. The parsing method can be shallow parsing, which usually outputs a sequence of units, or full parsing, which outputs a full parse tree.

Yakushiji et al. (2001) used a term recognizer to identify multi-word terms and a shallow parser to reduce lexical ambiguity. Then they did full parses of the sentences. Domain-specific knowledge, including a set of target verbs and mapping rules provided by domain specialists, was used to construct frame representations, which show interaction structures.

GENIES (Friedman et al., 2001) extracted semantic patterns by observing typical semantic and syntactic co-occurrence patterns in a sample corpus using semantic relationship categories and biological objects. It did full parsing over documents and output a frame structure if parsing was successful based on the patterns.

MedScan (Daraselia et al., 2003; 2004; Novichkova et al., 2003) broke sentences into tokens and recognized them based on the corresponding lexemes. Its syntactic parser then built a number of alternative syntactic structures for a given sentence using the rules defined in its grammar. A semantic interpreter transformed the syntactic tree into a normalized semantic tree, which represents logical relationships between objects. Finally, an ontology interpreter utilized ontology and a knowledge base to evaluate each semantic tree and convert the valid ones into ontological representations. An ontology is defined as a collection of concepts representing domain-specific entities, a set of relationships between the concepts, and the range of admissible values for each concept. It was constructed after analyzing about 2000 MEDLINE abstracts.

Temkin et al. (2003) created a context free grammar by manually analyzing a corpus of 500 non-topic-specific scientific abstracts from MEDLINE and used it to parse documents to extract interactions.

PathwayFinder (Yao et al., 2004) combined adjacent features, grammar features and pattern features to extract interactions. The patterns were manually defined and can be enriched through interaction with users. Three features are combined to compute the pattern score for documents that have been subjected to full parsing.

RelEx (Fundel et al., 2007) used public-domain NLP tools to preprocess and fully parse sentences. This results in Chunks Dependency Parse Trees, in which the nodes can be chunks composed of several words. Then, it extracts paths in the trees that match three predefined rules about descriptions of relations. If there is a match, an interaction has been extracted. Santos et al. (2005), Natarajan et al. (2006), Fundel et al. (2007) and Rinaldi et al. (2007) used full parsing to verify matches to predefined rules about descriptions of relations. Miyao et al. (2009) used different natural language parsing tools to extract interactions and compared the results.

Some researchers have used third party parsing tools to analyze documents. Santos et al. (2005) used a third-party parser to fully parse sentences to find subject-verb-object relationships. They maintained a list of verbs derived manually. Natarajan et al. (2006) used a commercial parsing tool to do full parsing over biomedical literature to find relationships among genes about which they were interested.

Although full parsing analyzes syntax in depth, the final results are still not reliable because full parsing usually returns multiple, ambiguous results and is domain specific.

In addition, full parsing is slow (Dershowitz, 2006). Shallow parsing that is faster and outputs a series of tokens is therefore used in other research instead. Ono et al. (2001) manually defined a set of patterns containing any of four interaction-indicating keywords: “interact,” “associate,” “bind” and “complex,” and used them to match tagged and shallow parsed sentences. These patterns, augmented by more verbs, were also utilized in PPLook (Zhang et al., 2010) to extract interaction descriptions. Arizona Relation Parser (McDonald et al., 2004) broke sentences into words and then transformed them to recognized tokens. Shallow parsing was used to create phrase chunks from those tokens based on hundreds of parsing rules and used predefined knowledge patterns to match a chunks sequence. Koike et al. (2005) used an NLP model to extract gene/protein interactions. It first recognizes the gene/protein name and Gene Ontology (GO) function name in the literature and then extracts ACTOR-OBJECT (gene-function) relationships through shallow parsing, noun phrase bracketing and sentence structure analysis. Hunter et al. (2008) also utilized manually extracted rules in a shallow parsing procedure.

There are also methods of matching predefined patterns without parsing procedures in published works. GIS (Chiang et al., 2004) generated sentence expression patterns, which are distributions of words and terms in relation descriptions, from sample sentences. Then it judged sentences according to the sentence expression patterns to determine whether relations between biomolecules are described.

GIFT (Domedel-Puig et al., 2005) tried to match substrings of sentences to predefined simple patterns including a predefined verb list.

Huang et al. (2004) and Yao et al. (2005) used dynamic programming to automatically create patterns from a training document set and to align the citations to the patterns to extract interactions.

Extraction based on statistical analysis (data mining)

Although pattern-matching methods can yield relatively high precision, recall may be low because the patterns defined do not include all situations in which interactions appear. To improve recall, some researchers have used statistical methods based on the simple method of counting tri-occurrences mentioned earlier. Marcotte et al. (2001) found discriminating words using a training set of 260 MEDLINE abstracts describing interactions, based on differences in frequencies of occurrence of those words. They used these words' probabilities of appearing in documents describing interactions to train Naïve Bayesian classifiers to score a document and then judge whether a document describes an interaction. Ramani et al., (2005) also used this classifier approach to improve the precision of its interaction extraction in its interaction database ID-Serve. It used protein co-citation analysis and then computed the probability of each co-citation based on the total number of abstracts, the number of abstracts in which both proteins appear, and the number where one of the pair appears, and used a threshold to select a set of final interaction candidates.

Some statistical applications apply statistical machine learning methods on parsed results. Huang et al. (2008) trained a SVM (support vector machine) with a vector representing different features of documents by 0 or 1 and then uses this SVM to classify interaction description. Zhou et al. (2008) also used a machine learning method to estimate probabilities that help parse the document. Niu et al. (2010) used a linear kernel based SVM to extract interactions. The features used were contextual, lexical, syntactic, keyword-based, pattern-based, and so on. Fayruzov et al. (2009) also used the SVM technique to extract protein interactions from natural text based on deep syntactic features (i.e., grammatical relations), shallow syntactic features (part-of-speech information), and lexical features, then evaluated the different features' impacts on extraction results. Li et al. (2010) used a coupling technology to couple a number of lexical and syntactical features as kernel input, and then the kernel was used by an SVM to classify sentences regarding their descriptions of interactions. Sætre et al. (2008) used syntactic shallow dependency parsing to parse a sentence and then SVM-extracted rules from a training corpus. Kim et al. (2010) used a method to parameterize non-contiguous syntactic structures as well as semantic roles and lexical features to make learning structural aspects from a small amount of training data effective.

Besides SVM, other machine learning methods have also been used. Bundschus et al. (2008) treat each word x and relation label y as a pair, and from a sequence of these pairs use the CRF (conditional random field) technique to extract interaction descriptions. They trained on a corpus to find a probability function for inferred interactions, using a CRF based on features including orthography, word shape, n-gram properties, dictionary membership, a context dictionary window, start window, and negation. Chowdhary et al. (2010) chose a Bayesian network (BN) to extract protein-protein-interaction word triples in sentences. They manually selected 12 features to let Bayesian network learn the rules behind a manually created corpus.

Wren et al. (2003, 2004) assigned a weight to the potential relationship between co-occurring biomolecule names. This value was $1 - r^n$, where r is 0.83 when the co-occurrence is in a sentence and 0.58 when the co-occurrence is in an abstract but not the same sentence. These figures are based on the research of Ding et al. (2002), who estimated the difference between the probabilities that an interaction is described given a co-occurrence of biomolecules in a phrase vs. in a sentence. Based on this weight, they judge the relatedness between genes. After that, they separated genes into sets and ranked the cohesion of a gene to a gene set by comparing the observed frequencies of the gene's co-occurrences with the nodes in the set to the expected frequencies. The expected frequencies are calculated by summing of the probabilities that this gene is connected to each gene of the gene set. This probability is derived by dividing the number of genes connected to a gene by the total number of genes in the set.

1.2 Existing problems

Despite the advances noted above, a major issue that needs to be solved in order to properly understand the reliability of extracted interactions is specifying a good ranking policy. Such a policy would enable better assessment of putative interactions based on both assessments of individual sentences containing the relevant biomolecule names, and on combination of the evidence provided by these sentences.

In addition, for extraction of interactions from the biological literature, the two categories of approaches typically used have limitations.

- For protein-protein co-occurrence counting or protein-protein-interaction tri-occurrence counting methods, high recall but low precision tends to occur (Albert et al., 2003). A lot of systems involve manual post-processing to improve the precision. This way does not eliminate the time and labor costs of the manual annotation method. However, it is computationally simpler and faster. Furthermore, it can be improved by better understanding and use of evidence provided by an understanding of empirical properties of relevant texts.
- The methods that parse and match interactions patterns can get a higher precision. These patterns can be predefined or determined with a machine learning procedure. Sentences matching the patterns are identified and the pattern determines the specifics of the interaction. The patterns are induced from existing sentences describing interactions so if a new sentence matches a pattern, a relatively high precision is likely. However, it is not practical to manually find all patterns of biomolecular interaction (Huang et al., 2004) and NLP is not very precise so far (Yao et al., 2004). Not all interaction descriptions satisfy any pattern in a given set of patterns, so some interactions will not be extracted by this method, (i.e., there will be a lower recall). For example, MedScan (Daraselia et al., 2003) obtained a recall of 21% with relatively restrictive templates, while Koike et al. (2005) achieved 54% with more unconstraining, inclusive templates that assume some syntactic analysis.
- The statistical methods reviewed focus on the overall evaluation of biomolecule interactions. They do not try to find the specific text unit where the interactions of a specific pair of biomolecules are described, which could be the interest or requirement of users. For this reason, we cannot easily compare them to other methods.

Chapter 2. Research method and result

2.1 Text empirics research overview

As discussed earlier, either simply counting tri-occurrences or using pattern matching has limitations we wish to overcome. We seek to advance understanding about the properties of biomedical texts and to apply this knowledge to automatic identification of biomolecular interactions. Properties of texts were identified empirically (i.e., by examining actual sentences) and used to evaluate the probability that a given sentence describes an interaction between a specific biomolecule pair. A major issue in evaluating such extracted interactions is how to specify a good ranking policy. Such a policy would facilitate assessment of putative interactions.

By *empirical* we refer to knowledge about text properties derived from "experience or observation" (e.g., dictionary.reference.com/browse/empirical). We derive our observations by manually examining corpora, and tabulating and analyzing the passages therein. This differs from other common approaches for extracting knowledge from text, such as natural language processing (NLP), which deduces knowledge from passages based on syntactic and semantic rules, and machine learning (ML). ML offers a corpus-based, statistical approach like the text empirics approach, but differs in that, with ML, text properties are found automatically by a computer. This has the following shortcomings compared to using text empirics.

1) Classification rule sets (typically arranged in decision trees) derived by ML usually include uninteresting junk mixed in. As a result,
2) the rules derived by ML are typically omitted from publications in favor of conclusions about the parameters of the ML process itself. As a result,
3) the outcome of ML can be harder to apply than the results of an empirical text analysis because ML-derived knowledge tends to be less readily available in a directly usable form, while text-empirics-derived results must necessarily be disseminated in an explicit form readily used by software designers.

2.2 Advancing text empirics and its application for extracting biomolecular interactions from the literature

2.2.1 Previous work on sentence structure properties

The first step is identifying sentences containing a particular pair of biomolecules. The structure of these sentences can give clues that help enable extraction of the biomolecular interactions they describe (Ding et al., 2002). For example, consider these two sentences:

> *3'-O-(4-Benzoyl)benzoyl-ATP (Bz2ATP), an analog of ATP (Bz2ATP), containing a photoreactive benzophenone moiety, was used as a probe of the ATP (Bz2ATP), binding site of myosin subfragment 1 (SF1). (*Mahmood *et al., 1984)*

> *We also found that the rate constants of elementary steps become progressively slower starting from ATP binding to the myosin head and ending by ADP isomerization, and this stepwise slowing may be the essential and integral part of the energy transduction mechanism by muscle. (Kawai et al., 2003)*

These sentences differ in the number of appearances of the biomolecules, their positions in the same or different phrases, the position of the interaction indicator word (usually a verb — here, “bind”), and so on. The second sentence describes the interaction between ATP and myosin more directly. Clues to automatic recognition of interactions in such sentences are provided by various such properties of sentences.

We chose MEDLINE as the repository to analyze. While these records may not completely reflect the idea that an article tries to convey, they usually contain the abstract and thus the most important information the authors wish to convey. Using MEDLINE, Ding et al. (2002) showed that sentences are useful text units for automatically extracting interactions. Therefore, we collected sentences containing biomolecule co-occurrences to analyze as the basis of this work.

To analyze passage properties for interaction extraction, some operational definitions need to be given:

Term	Definition
sentence	Either an article title or a word sequence beginning with a capital letter and ending with a period.
phrase	A word sequence occurring inside a *sentence*, and begins and ends with: , | ; | : | . | <the beginning of the sentence> | <whitespace>-<whitespace> | (|).
IIT	*Interaction-indicating term.* A word, often a verb, that can describe an interaction between two biomolecules.

We have manually created a list of IITs based on reading several hundred MEDLINE abstracts. For example, *activate*, *activation*, etc. can describe an interaction between two biomolecules, as in “the activation of A by B.”

To extract an interaction, we require a sentence to contain two biomolecules of interest. However, such a sentence does not necessarily describe an interaction. For example, the sentence

“Both A and B can bind to C.”

does not describe an interaction between A and B, even though it describes interactions between A and C and between B and C. Our hypothesis is that we can find properties of sentences from the MEDLINE collection that can support automatic interaction extraction. The first goal, therefore, is to advance understanding of relevant sentence properties. The second, related goal is to better understand properties of IITs. The third goal is to use the results of the first and second goals to predict whether a sentence describes an interaction. The fourth goal is to scale up by generating and evaluating a database of biomolecular interactions.

By analyzing typical passages from MEDLINE, it is possible to advance toward those goals by empirically investigating certain questions, such as the following:

1) How can the presence of IITs be used to infer the type of interaction between two specific biomolecules?
2) If p_{phrase} is the likelihood that biomolecules co-occurring in the same phrase are described by the phrase as interacting, how does p_{phrase} differ from p_{sentence}, the analogous situation where they are in different phrases of the same sentence?

3) How does the order of appearance of three important words — two biomolecules and an IIT — in a phrase or sentence affect the probability that the biomolecules are described as interacting?
4) How do properties of IITs occurring near two biomolecule names — such as their identities, inflections, roots, and semantic categories — affect the probability that they help describe an interaction between the biomolecules?

For questions 1–4, Ding et al. (2002) collected 303 MEDLINE abstracts and extracted 664 sentences, based on ten queries to PubMed. Each query consisted of two biomolecule names known to interact. They were elicited from biologists who were requested to suggest queries typical of those likely to be made. Some further details about this corpus appear in Ding et al. (2002), and a list of the abstracts in the corpus may be downloaded from http://ifsc.ualr.edu/jdberleant/IEPA/IEPA.htm. Each sentence was manually analyzed with respect to the properties related to questions 1–4 above and tagged as to whether or not it described an interaction between the two query biomolecules.

The results for questions 1 and 2 (Table 1) indicate that the probability that an interaction is described when two biomolecules co-occur in a phrase is higher than when they are in different phrases in a sentence (67 percent versus 33 percent). Secondly, if an IIT appears with the two biomolecules, the probability that an interaction is described is higher than without an IIT present (55 versus 7.99 percent and 71 versus 0 percent). These two comparisons are statistically significant ($p<0.001$, $\chi2$ test).

Table 1. Biomolecule co-occurrences in sentences and phrases, with and without IITs.

	# (%) that describe the interaction	**Total number**
Sentences where two biomolecules tri-occur with at least one IIT	331 (55%)	606
Sentences where two biomolecules co-occur without any IIT	3 (7.9%)	38

Table 1 continued

All sentences where two biomolecules co-occur	334 (52%)	644
Phrases where two biomolecules tri-occur with at least one IIT	236 (71%)	334
Phrases where two biomolecules co-occur without any IIT	0 (0%)	17
All phrases where two biomolecules co-occur	236 (67%)	351
Sentence co-occurrences that are not also phrase co-occurrences	98 (33%)	293

For question (3), we investigated how the presence of an IIT between the two biomolecules differs from the presence of an IIT but not between the biomolecules. The results are shown in Table 2.

Table 2. Percentages (i.e., precisions) of sentences and phrases describing interactions, by IIT location.

	IIT intervening	**IIT elsewhere in sentence**	**IIT in either place**
Phrases in which two biomolecules co-occur	63%	24%	45%
Sentence co-occurrences that are not also phrase co-occurrences	30%	9.1%	21%
Both phrase and sentence co-occurrences	48%	17%	34%
Percent of interaction descriptions	77%	23%	100%

Table 2 shows that the presence of an IIT intervening between the two biomolecule names is associated with a relatively high likelihood that an interaction is described. Consequently, for descriptions in which one or more IIT was present, most (77 percent) had an IIT between the biomolecule names.

2.2.2 Interaction terms and probability that an interaction is described

2.2.2.1 Interaction term properties for interaction extraction

In previous research (Berleant et al., unpublished work), we found that the presence of an IIT with a co-occurrence of biomolecule terms suggests an interaction with higher precision than a co-occurrence of biological terms without an IIT. IITs thus can play a role in predicting interactions. However, interaction terms have their own properties, such as verb tense and part of speech (e.g. noun, adjective, or verb). In addition, verbs can indicate several different categories of biomolecular interactions, such as association, regulation, and so on. These different properties of interaction terms may affect the description of the biomolecular interactions.

Our objectives in identifying the effect of interaction term properties can be divided into two parts: understanding IIT categories and understanding IIT forms. For example, in the following three sentences, the verb “bind” appears in three different forms in describing the interaction between “ATP” and “myosin.”

> *(1)* ***ATP*** ***binds*** *single-headed* ***myosin*** *VI following a two-step reaction mechanism with formation of a low affinity collision complex (1/K(1)' = 5.6 mm) followed by isomerization (k(+2)' = 176 s-1) to a state with weak actin affinity. (*Robblee *et al., 2004)*

> *(2)* ***ATP*** ***binding*** *to* ***myosin*** *subfragment 1 (S1) induces an increase in tryptophan fluorescence. (*Reshetnyak *et al., 2000)*

> *(3) The data demonstrate that unlike any previously characterized* ***myosin*** *a single-headed myosin V spends most of its kinetic cycle (>70%) strongly* ***bound*** *to actin in the presence of* ***ATP****. (*De La Cruz *et al., 1999)*

In these cases, the forms of “bind” are present, present continuous, and past tense in these three sentences, respectively. Sentences (1) and (2) describe the interaction between myosin and ATP with the verb root “bind,” but sentence (3) does not describe the interaction between myosin and ATP, even though it has the same verb root.

It would be useful to be able to determine the probabilities of interactions that might be described in a sentence based on the properties and identity of a given IIT in it.

Here is a five-category list suggested by Ding (personal communication), based on the previous analysis of 644 sentences.

i Association

ii Modification

iii Negative regulation

iv Positive regulation

v Transportation

An alternative set of categories used by colleagues at Procter & Gamble (P & G) is:

i. Activate directly

ii. Activate indirectly

iii. Activate ligand mediated

iv. Inhibit directly

v. Inhibit indirectly

vi. Inhibit ligand mediated

vii. Bind to

From a biological perspective, the categories of interaction in MetNetDB (http://metnet.vrac.iastate.edu/MetNet_db.htm, Wurtele et al., 2007), are:

i. Enzymatic reaction

ii. Translation

iii. Transcription

iv. Composition-AND

v. Composition-OR

vi. Transport

vii. Transport/channel-type facilitators

viii. Bind

ix. Transport/ATP-driven transporters

x. Transport/PEP-dependent transporters

xi. Transport/decarboxylation-driven transporters

xii. Transport/electron-flow-driven transporters

xiii. Transport/light-driven transporters

xiv. Transport/mechanically driven transporters

xv. Positive regulation

xvi. Positive regulation/allosteric activation

xvii. Positive regulation/competitive activation

xviii. Positive regulation/covalent modification

xix. Positive regulation/complex formation.

xx. Positive regulation/transcriptional activation

xxi. Positive regulation/direct

xxii. Positive regulation/coactivation

xxiii. Positive regulation/translational activation

xxiv. Positive regulation/indirect

xxv. Negative regulation

xxvi. Negative regulation/allosteric inhibition

xxvii. Negative regulation/competitive inhibition

xxviii. Negative regulation/covalent modification

xxix. Negative regulation/complex formation

xxx. Negative regulation/transcriptional inhibition

After reviewing all these categories, the P & G categories seem too detailed in some ways, as well as not comprehensive. MetNetDB is too biologically oriented and detailed, and not all of its categories are appropriate for text mining. In addition, through experience, we have found that to extract interactions from sentences, some terms, like “interact,” “influence,” “form,” etc., do not reflect specific interactions but only state that an interaction exists. Two new categories, *create* and *vague*, were added. The category *create* holds instances of “form,” “produce,” “construct,” and other similar words describing that one biomolecule can be created from another biomolecule. The category *vague* holds instances of “interact,” “influence,” “affect,” and similar words describing that an interaction occurs between two biomolecules without a detailed description of it, like “oxidize.”

We finally settled on the following categories:

vi Association

vii Modification

viii Negative regulation

ix Positive regulation

x Transportation

xi Create

xii Translation & transcription

xiii Vague

The forms of an interaction term that we account for include:

i Present tense

ii Present continuous (i.e., present progressive) tense

iii Past tense

iv Perfect tense

v Noun

vi Adjective

vii Adverb

3.2.2.2 Training corpus creation and analysis

Corpus creation

The research to judge the effects of various properties on the probability of an interaction being described was conducted through analysis of sample data. A hand-annotated corpus of sentences was used for exploration of this problem. The corpus is from the PUBMED query results for 10 pairs of biomolecules (*nitrite & xanthine*, *pyruvate dehydrogenase & phosphofructokinase*, *indole acetic acid & starch*, *glucose & starch*, *glucose-6-p & starch*, *carotenoid & IPP*, *cre & cytokinin*, *acetyl-CoA & leucine*, *glucose & pyruvate,* and *ATP & myosin*). Each pair was sent to PUBMED (http://www.pubmed.com) as a query. and the citation results were saved. The sentences in the result citations in which both members of a pair of biomolecules co-occur with an interaction-indicating term (total of 320 sentences) were collected and analyzed. The last publication date of those citations was June 2007.

Each objective property, such as the syntactic or semantic category of an interaction-indicating term in each sentence, was recorded numerically in one table for each biomolecular pair (see Table 3). All IITs investigated in this corpus and also used in the later real application for interaction extraction with their category information are in Appendix II.

For each sentence, whether an interaction is described between the pair of biomolecules (that is, the query which retrieved the sentence) was determined. I did the determination, except the difficult sentences, for which my advisor and biologists were asked to make a final decision. Whether there is an interaction description in a sentence will be a column in the final table (i.e. Table 3) of the corpus and will be expressed by 0 (if the sentence does not describe the interaction) or 1 (if the sentence does describe an interaction). The distance measurements are described by the number of words (5 in the example of Table 2). The detailed description and the full example table are in Appendix I.

Table 3. An example of an analysis of a sentence from the query: Nitrite & Xanthine.
(1 means one property is positive and 0 means negative in one sentence, and a number greater than 1 may mean the times a property appears or the position property).

PMIDro	Verb root	Noun	Adjective	Adverb	Past tense	…..	Number of words from the nearest biomolecule	between biomolecules?	All in a phrase?
17202421	reduce	0	0	0	1	…..	5	1	0

In detail, for each sentence, its PMID and whether it describes an interaction are recorded in addition to each interaction-indicating term appearing in it. For each interaction-indicating term in the sentence, its verb root, its exact spelling in the sentence, whether it is the correct interaction description term for the specific biomolecule pair, whether it describes the interaction in the sentence, its form (noun, adjective,…), its semantic category (association, modification,…), its distance from the nearest biomolecule in the specific pair, its distance from the other biomolecule of the specific pair (choosing the nearest one if more than one), whether it appears between the two biomolecules, whether it appears in a phrase with the two biomolecules, and whether it appears in a phrase with the two biomolecules and between them in this phrase are recorded.

Some sentences have very vague structures. The results of their analysis are also indefinite.

- Some biomolecules are parts of other biomolecules. For example here are two examples:

In contrast, the combination of lytB and a cDNA encoding ***IPP*** *isomerase (ipi) was no more effective in enhancing* **carotenoid** *accumulation than ipi alone, indicating that the ratio of* ***IPP*** *and DMAPP produced via the DOXP pathway is influenced by LytB.* (Cunningham et al., 2000)

The enzyme isopentenyl pyrophosphate ***(IPP)*** *isomerase catalyzes the reversible isomerization of* ***IPP*** *to produce dimethylallyl pyrophosphate, the initial substrate leading to the biosynthesis of* **carotenoids** *and many other long-chain isoprenoids*. (Sun et al., 1998)

Here we are interested in the biomolecule pair *IPP* and *carotenoid*. But *IPP* appears in these examples as a part of another biomolecule, *IPP isomerase*, which can be confusing when analyzing the sentence because *isomerase* is a very common suffix word and the combination of it with another bimolecule may not be a special biomolecule thus not collected in the dictionary. In these two examples, *IPP isomerase* is *IPI*, which is an important biomolecule. However, when faced with some combinations, it would be wrong to extract a part of the combination and use it as part of an interaction. An improved biomolecule dictionary would be a solution for keeping track of which common suffix or prefix combinations are to be considered breakable. The correct analyses of these two sentences are that the first one does not describe an interaction between IPP and carotenoid, but the second one does.

A good dictionary also is useful in a situation involving several different words that actually indicate varieties of the same biomolecule or are synonyms. A good dictionary should be able to provide the information about this kind of relationship.

Different uptake of pools of **14C-acetyl CoA***, synthesized from injected 14C-acetate, and* **3H-acetyl CoA***, synthesized through metabolic pathways of* **3H-leucine***, indicates the compartmentalization of* **acetyl CoA** *in the synthesis of saturated and unsaturated fatty acids.*

We want to judge whether this sentence describes an interaction between *acetyl-coA* and *leucine*. But this sentence only gives *acetyl-coA* and *3H-leucine*, which is a subtype of leucine. If we do not know or the algorithm does not know the relation between *3H-leucine* and *leucine*, the analysis result for this sentence will be that there is no interaction described between *leucine* and *acetyl-coA*. Actually, *14C-acetyl CoA* and *3H-acetyl CoA* are also subtypes of *acetyl-coA*, so the correct analysis of this sentence is that it does describe an interaction between *leucine* and a*cetyl-coA* and the interaction-indicating term is synthesize.

- Sometimes, the combination of a biomolecule and a common word is not a special biomolecule.

The Arabidopsis thaliana AHK4 histidine kinase (also known as **CRE1** *or WOL) acts as a* **cytokinin** *signal transducer, presumably, in concert with downstream components, such as histidine-containing phosphotransfer factors (AHPs) and response regulators (ARRs), through the histidine-to-aspartate (His-->Asp) phosphorelay.*

Here we are interested in *CRE1* and *cytokinin* and we can see that *CRE1* is a *cytokinin* signal transducer. Then the direct interaction described in this sentence is that *CRE1* transduces the signal of *cytokinin*. But consider the definition of signal transduction, which is the movement of signals from outside the cell to inside (http://www.med.unibs.it/~marchesi/signal.html). *CRE1* is described as the first component in the whole pathway of *cytokinin* signal transduction in this sentence, along with other downstream components. So we can conclude that *CRE1* must interact with *cytokinin* to propagate the *cytokinin* signal. Therefore, the proper analysis of this sentence is that there is an interaction between *CRE1* and *cytokinin*.

- Sentences may have verbs that imply an interaction but do not directly describe an interaction.

Conversely, **cytokinin** *regulation of the early nodulin Nodule Inception1 (Mt NIN) depends on Mt* **CRE**1.

Xanthine *nitration by myeloperoxidase required hydrogen peroxide and* **nitrite.**

Suicide inactivation of **xanthine oxidoreductase** *during reduction of inorganic* **nitrite** *to nitric oxide*

In the first sentence, we are interested in interactions between *cytokinin* and *CRE1*. There is a relation between them described in this sentence: *cytokinin* regulation of something depends on *CRE1*; in other words, without *CRE1*, *cytokinin* can't regulate. Based on biological knowledge, we know *CRE1* is a *cytokinin* receptor, so we can infer *cytokinin* needs to be bound by *CRE1* to regulate something. But the sentence doesn't say that. And we can't know that from the sentence. The sentence says nothing about binding or receptors, and "depends" does not imply binding or receptors. What we can get from this sentence is that there is a relation between *cytokinin* and *CRE1*, but this sentence does not describe what the interaction is.

The same situation occurs in the second sentence. *Xanthine* nitration requires *nitrite*, and we can tell *xanthine* and *nitrite* interact but the sentence does not explain what “nitration” consists of. So we cannot tell what the interaction is from the sentence unless “nitration” itself is a suitable answer.

The key words in these sentences are *depend* and *require*, and there are a lot of similar words like *relate* and *respect*.

In the third sentence, it looks like probably the reduction process inactivates *xanthine oxidoreductase*, but the sentence doesn't say that. So we can infer that *nitrite* might have something to do with inactivating *xanthine oxidoreductase*, but it is only might. For “might” situation, see the following sections.

To analyze sentences of this kind and be sure of a correct conclusion, we must conclude we cannot find a correct interaction-indicating term for the interaction between the two specified biomolecules in these sentences. Whether the sentence is considered to describe an interaction between the specified two biomolecules depends on the biologist. Some biologists think this kind of sentence does provide information about interaction between the biomolecules.

- Some interactions are described using several interaction-indicating terms, and then it is hard to identify the right interaction term.

Glucose *was found to repress alpha-amylase and, more severely, maltase activity, thus repressing* **starch** *degradation by L. gongylophorus, so that we hypothesize that:*

The climacteric respiration burst was reduced by the action of ***IAA****, and* **starch** *degradation and sucrose formation were delayed.*

Here is the analysis of the right interaction-indicating term which will be discussed further in section 2.2.4. We can see that glucose represses and *IAA* delays starch degradation. *Glucose* and *IAA* downregulate *starch* degradation. But we cannot say *glucose* represses *starch* or *glucose* degrades *starch*, which means the actual interaction term is not stated in the sentence. We can, however, infer that *glucose* and *IAA* preserve starch. It seems that, if we are interested in interactions among biomolecules A (*glucose* or *IAA*) and C (*starch*), these sentences imply one (preserve) but do not describe one. It would be hard to design an algorithm to determine that. For our purposes, if we want to know whether A and C interact or not, this sentence is evidence that they do. But if we want to determine through software what the interaction is, this sentence is likely to mislead the algorithm, for neither repress nor degrade is right. The relationship between A and C is not one of either repressing or degradation.

We could say the relation between A and C is “repress degradation of,” but in the corpus analysis we cannot record these kinds of interaction terms because they are combinations of terms. Actually biologists also are interested in this information when retrieving interactions. To make the system comprehensive and precise, in the analysis of the corpus, both “repress” and “degrade” will not be counted as correct interaction terms but the sentence is counted as describing the interaction between A and C.

In fact we can summarize these kinds of situations in four categories:

1. A downregulates change of B
2. A upregulates change of B
3. A downregulates production of B
4. A upregulates production of B

The first one is like the example sentences. A downregulates change in B: because A does not downregulate B, downregulate is not the interaction term. Because A does not change B, change is not the interaction term. We might conclude that A preserves B. For our analysis, we record that there is an interaction between A and B and but do not record either downregulate or change as their interaction term.

For A upregulates change of B: here, A changes B so the interaction term is change.

For A downregulates production of B: here A decreases B. It depends on whether downregulates means decreases. If it does, then the interaction term is downregulates. If it does not, then there is no interaction term stated.

The last one is "A upregulates production of B." Here, the interaction term is upregulates. Therefore:
For A downregulates change of B, the interaction is preserve
For A upregulates change of B, the interaction is change
For A downregulates production of B, the interaction is downregulate
For A upregulates production of B, the interaction is upregulate

This relatively complex situation occurs because the interaction is actually between a biomolecule and a process. If in the future, we can treat such interactions (processes) as we currently do biomolecules, then we may look at processes as actors in interactions and the problem would be solved.

- Some interactions include several interaction-indicating terms, but terms are too vague to be sure an interaction is in fact described.

Hence, **cytokinin** *signaling mediated by a single receptor, Mt* **CRE1***, leads to an opposite control of symbiotic nodule and lateral root organogenesis.*

The **CTK***-mediated repression of LR initiation is transmitted through the two-component signal system and mediated by the receptor* **CRE***1.*

Studies suggested that organic **nitrite** *(R-O-NO) is produced from* **XO***-mediated organic nitrate reduction.*

In these sentences, *mediate* is describing the relation between *cytokinin* (*CTK*) and *CRE1* or between *nitrite* and *XO*. However, not only two biomolecules are involved in the interactions, but also other biochemical actions like signaling in the first sentence. “Mediate” is too vague to conclude an interaction is described, and we cannot just say that *CRE1* mediates *CTK*. This situation does not describe any specific interaction between *CRE1* and *cytokinin*. Yet it does give some information about interaction between these two biomolecules. We could say it *might* describe the interaction between

CRE1 and *CTK*. To differentiate it from clearly stated interaction descriptions, we could count this sentence as 1/2 when we record its interaction score. However, it is very hard to define clear criteria to decide which situations to count 1/2. To make the system more objective and easily specified, for these situations, we count that there is an interaction described and score 1 for them.

Some vague interaction-indicating terms, however, can describe interactions between two biomolecules.

Our results shed light on a novel role of the recovery stroke: fine-tuning of this reversible equilibrium influences the functional properties of **myosin** *through controlling the effective rates of* **ATP** *hydrolysis and phosphate release.*

Here we can find that *ATP* hydrolysis influences myosin. Or we also can say *ATP* influences *myosin*. Therefore, for this sentence, we consider that it describes an interaction between *ATP* and *myosin* and the interaction-indicating term is *influence*.

- Some sentences contain interaction-indicating terms, but the terms do not indicate the right interaction but have the relation with the interaction between the two biomolecules. Those sentences usually describe the interactions but do not have the right interaction-indicating terms.

A comparison of the **glucose-6-phosphate** *isotope patterns in different pathways of the synthesis with the experimental data on the distribution of carbon isotopes in* **starch** *glucose of storing plant organs led to the conclusion that the* **starch** *resources are predominantly formed at the expense of* **glucose-6-phosphate** *of photorespiration.*

We can see the direct description between starch and glucose-6-phosphate is “starch is formed at the expense of glucose-6-phosphate.” There is no interaction term “formed at the expense of” in our lexicon of interaction terms. We can analyze it as describing that *starch* decreases *glucose-6-phosphate*.

- Some interactions are described in very vague ways in some sentences.

Mutations in **CRE***1 reduced but did not eliminate the effect of* **cytokinin** *on gene expression for a subset of* **cytokinin***-responsive genes and had little or no effect on others, suggesting functional redundancy among the* **cytokinin** *receptors.*

Because mutations of biomolecules usually lessen the effect of the original biomolecule, we can infer that *CRE1* mediates *cytokinin*'s effect on *cytokinin*-responsive genes. But we still don't know the interaction between *CRE1* and *cytokinin*. However, from the last part of the sentence, far away from where it says that "mutations in *CRE1* reduced but did not eliminate" *cytokinin*'s effect, it says this suggests "functional redundancy among the *cytokinin* receptors." Then we can infer that *CRE1* is a *cytokinin* receptor and therefore that the interaction is *CRE1* binds *cytokinin*. Some publications pointed this out: "*Identification of CRE1 as a cytokinin receptor from Arabidopsis*" (Inoue et al., 2000).

Here, we demonstrate that **myosin** *VI gating is achieved instead by blocking* **ATP** *binding to the lead head once it has released its ADP.*

From the first glance, we did not find a direct interaction between *myosin* and *ATP*. There may be implicit interaction between *ATP* and *myosin*, which is that "blocking *ATP* binding to the lead head" can result in "*myosin* VI gating." This implicit interaction is very vague. However, if we check the whole paragraph, what is really happening becomes clarified.

A processive molecular motor must coordinate the enzymatic state of its two catalytic domains in order to prevent premature detachment from its track. For myosin V, internal strain produced when both heads of are [sic] attached to an actin track prevents completion of the lever arm swing of the lead head and blocks ADP release. However, this mechanism cannot work for myosin VI, since its lever arm positions are reversed.

Here, we demonstrate that myosin VI gating is achieved instead by blocking ATP binding to the lead head once it has released its ADP.

We can see the head that binds to *ATP* is the *myosin* head in fact, so this sentence does describe a direct interaction between *ATP* and *myosin*: *ATP* binds to *myosin*. This interaction is extracted only after reviewing the whole paragraph, which is not expected for the automatic extraction algorithms this research is intended to support.

The interactions in the sentences about *CRE1* & *cytokinin* and *ATP* & *myosin* are direct interactions even though its description is very vague. However, some sentences contain interactions that are not direct, but implicit.

Astrocyte-selective expression of **pyruvate** *carboxylase (PC) enables synthesis of glutamate from* **glucose***, accounting for two-thirds of astrocytic* **glucose** *degradation via combined* **pyruvate** *carboxylation and dehydrogenation*

In this sentence, the directly described interaction related to glucose is that *PC* enables conversion of *glucose* to glutamate. However, it mentions *glucose* degradation is via combined *pyruvate* carboxylation and dehydrogenation. In other words, *pyruvate* is carboxylated, which may mediate degradation of *glucose*. Therefore, pyruvate carboxylation has something to do with glucose degradation. But we don't know what the connection is. Maybe some other process or chemical degrades glucose and carboxylates pyruvate, but this sentence does not give enough information so we do not count there is an interaction between glucose and pyruvate described in this sentence.

To be consistent, for implicit interactions or uncertain interactions described in sentences, we count them and record those interaction terms mentioned as the correct interaction description, if such an interaction term is present in the sentence. For those sentences indicating interactions but there is no appropriate interaction term inside the sentences, we count the sentences describing the interactions but do not record the interaction terms.

- Some interactions are not described directly in sentences.

The greater suppressive effect of lactate as compared to **pyruvate** *suggests that alteration of the NAD(+)/NADH ratio underlies the suppression of* **glucose** *oxidation by lactate.*

This sentence directly described the interaction between *glucose* and *lactate* (lactate suppresses glucose oxidation, so, lactate preserves glucose). However, it gives the information that the *pyruvate* has a comparable effect on *glucose*. Thus we can infer that *pyruvate* also suppresses *glucose* oxidation, and so *pyruvate* preserves *glucose*.

- Some interaction descriptions involve negative words.

It was found that **acetyl-CoA** *produced from L-acetylcarnitine or by oxidation from either pyruvate, octanoate or palmitylcarnitine but not from* **leucine** *led to a stimulation of pyruvate carboxylation.*

The addition of the long neck domain of **myosin** *Va to the Chara motor domain largely increased the velocity of the motility without increasing the* **ATP** *hydrolysis cycle rate, consistent with the swinging lever model.*

These two sentences contain negative words (not, without) for the interaction between the specific biomolecules. However, the results of their influence on interaction description are different. In the first sentence, from the information provided in the sentence, *leucine* is similar to *pyruvate*, *octanoate* and *palmitylcarnitine* in producing *acetyl-coA* by oxidation. So this sentence describes an interaction between *acetyl-coA* and *leucine* and the interaction-indicating term is *produce*. “Not” does not give the negative effect on the interaction description between *acetyl-CoA* and *leucine*. But the second sentence describes the reality that *myosin* increased the velocity of the motility **without** increasing the *ATP* hydrolysis cycle rate, so no interaction between *ATP* and *myosin* is described in this sentence. If we delete “without,” then the reality in the sentence becomes *myosin* can increase *ATP* hydrolysis that is an interaction description. However, the appearance of “without” denies it, so there is no interaction between *myosin* and *ATP* described here.

We can see negative words (e.g. not, without) can negate the interaction description (*increasing the **ATP** hydrolysis* in the second sentence) or alternatively might not influence the interaction description (as in the first sentence), which means the influence of negative words on interaction description is uncertain. Therefore, in our analysis, we do not count the effect of negative words on interaction descriptions.

Result, data analysis and validation

To see the influence that different verb forms and categories have, we need to analyze the data from this corpus.

Summarization: The data were summarized based on different properties of interaction terms involved in each sentence. First, each pair's data were summarized based on each column or each property by calculating the sum of the numbers in each column of each pair's table (e.g., Table 3). However, the sum calculation was conducted conditionally. For each pair, there are different sums for different purposes. For sentence interaction description purposes, the sum of sentences where an interaction-indicating term with each form or category appears and the sum of sentences describing interactions where a verb with each form or category appears are calculated, plus the fraction of the second sum over the first sum was calculated.

Table 4. Summary of the sentence data in the corpus

pairs	**# (%) of sentences describing interactions**	**sentence**
ATY & myosin	30/65%	46
cre & cykotintin	35/81%	43
nitrite & xanthine	29/64%	45
glucose-6-p & starch	20/71%	28
glucose & starch	19/46%	41

Table 4 continued

glucose & pyruvate	16/41%	39
acetyl-coa & leucine	21/62%	34
indole acetic acid & starch	3/38%	8
carotenoid & ipp	6/75%	8
pyruvate dehydrogenase & phosphofructokinase	1/4%	28
Total	180/56%	320

Table 4 gives an overview of the corpus results. Some pairs don't have enough sentences in all of MEDLINE, and some have more than 30, in which case we used the 30 most recent ones. This resulted in 320 sentences of which 180 describe interactions between the query biomolecule pair. This is 56.25%. In this table, the first seven pairs have about 30 qualified sentences, and the proportion that describes an interaction between the specific pair seems to cluster around 50%. The pair "indole acetic acid & starch" and the pair "carotenoid & ipp" don't appear in many citations in MEDLINE, so all sentences (8 for each) from all publications containing them were collected. The pair "pyruvate dehydrogenase & phosphofructokinase" has enough sentences mentioning them, but most of them do not describe interactions between them even though they are technically in a same pathway.

To analyze the sentences for different pairs, we need to do a summary for each factor influencing whether an interaction is described. As mentioned earlier, previous research exists, but here we mainly focus on interaction-indicating terms' properties for the first time.

Table 5. The appearance of interaction-indicating terms' properties in sentences of the corpus

form or category	ATP & myosin	cre & cytokinin	nitrite & xanthine	g-6-p & starch	glucose & starch	glucose & pyruvate	acetyl-coa & leucine	indole acetic acid & starch	carotenoid & ipp	pyruvate dehydrogenase & phosphofru-ctokinase	Total
Noun	29	35	36	23	25	31	28	7	4	19	237
Adj	0	2	1	3	0	3	6	2	0	3	20
Adv	0	0	0	0	0	0	0	0	0	0	0
Present	15	7	12	2	11	15	6	1	3	4	76
-ing	25	3	10	4	5	5	5	2	2	8	69
past/perfect	16	16	25	16	14	9	24	5	4	12	141
Association	29	33	2	7	1	3	8	0	4	2	89
Modification	25	1	36	4	20	12	15	3	0	5	121
negative regulation	3	9	5	7	16	15	6	5	1	17	84
Positive regulation	10	16	19	11	15	15	9	2	3	12	112
transportation	7	5	0	3	0	3	1	1	0	1	21
transcription	0	2	0	0	1	1	0	0	3	0	7
Create	3	1	18	19	13	16	19	4	3	0	96
Vague	13	6	7	3	5	14	11	3	1	13	76

Table 6. The appearance of interaction-indicating terms' properties in sentences describing interactions

form or category	ATP & myosin	cre & cytokinin	nitrite & xanthine	g-6-p & starch	glucose & starch	glucose & pyruvate	acetyl-coa & leucine	indole acetic acid & starch	carotenoid & ipp	pyruvate dehydrogenase & phosphofru-ctokinase	Total
Noun	20	31	25	17	12	13	17	3	2	1	141
Adj	0	1	0	3	0	1	3	1	0	0	9
Adv	0	0	0	0	0	0	0	0	0	0	0
Present	10	5	9	2	7	9	4	1	3	0	50
-ing	13	1	7	3	4	2	2	1	1	1	35
past/perfect	10	13	15	11	6	3	15	2	2	0	77
association	18	30	2	4	0	0	3	0	3	0	60
modification	17	0	26	4	16	4	11	2	0	0	80
Negative regulation	1	7	2	5	6	5	2	3	1	1	33
Positive regulation	4	12	13	5	4	3	4	0	1	1	47
transportation	5	5	0	3	0	1	0	0	0	0	14
transcription	0	2	0		0	1	0	0	2	0	5
create	3	1	9	16	8	8	14	2	2	0	63
vague	10	6	5	3	1	8	7	1	0	0	41

Table 5 gives the summary of some of each interaction-indicating term properties in the corpus. We can see that of all interaction-indicating term forms, the noun appears in more sentences than any other form, and two forms (adverb and perfect tense) do not appear in any sentences. In fact, the form "perfect" could appear in some sentences, but a lot of interaction-indicating terms have the same perfect form as past form, and so far the design of our algorithm does not include a method to recognize the difference between perfect and past tense when the spellings are the same. For example, “bound” is both the past form and perfect form of the term “bind.” Moreover, after investigating the corpus and following evaluation part, there are much fewer appearances of perfect tense than past tense in corpus and other sentence sets we used. Therefore, in the corpus construction and analysis, we integrated the perfect form’s appearance data into the past form. Note that some IITs have same spelling for nouns and present tense. Here we can manually differentiate them but not in algorithm. POS tagging is assumed to be available to use the results.

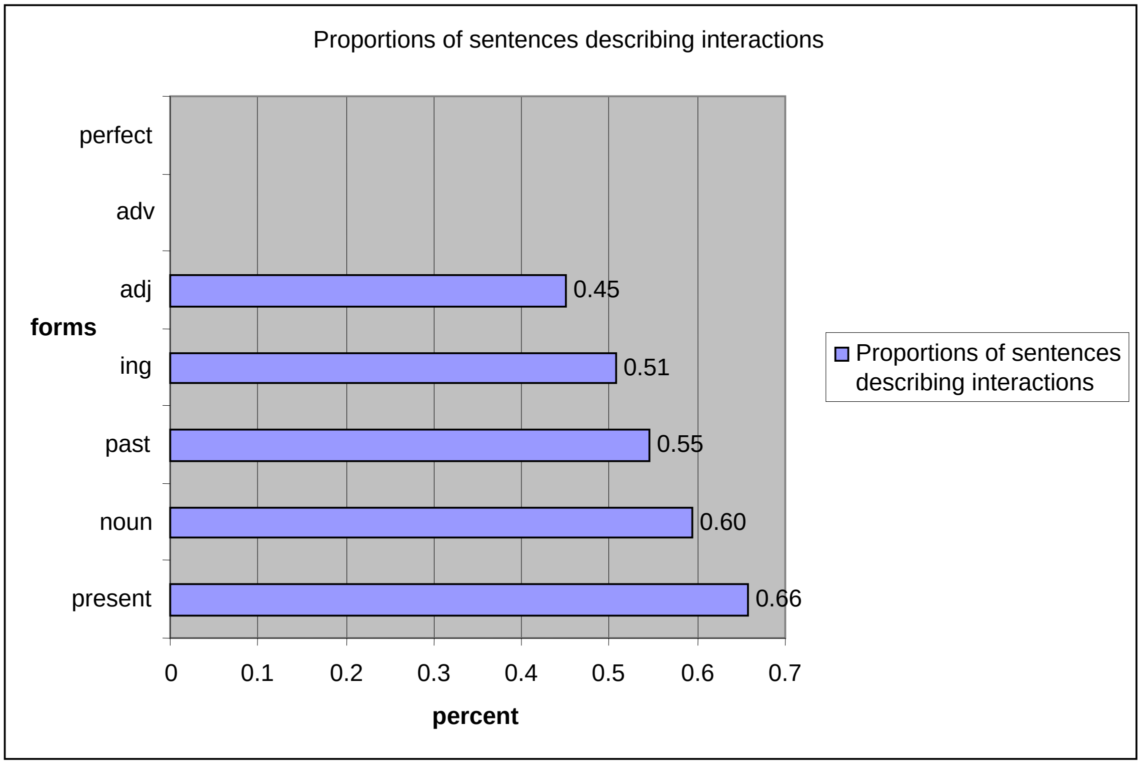
For each sentence, we determined whether an interaction between the query biomolecule pair was described. Then we checked what forms or categories of interaction-indicating terms the sentence contains, to investigate the relationship between the interaction-indicating terms’ forms/categories and interaction descriptions. For those sentences describing the interactions, let’s see the situation for different forms/categories. Table 6 gives the data on different interaction-indicating term properties in sentences describing the specified interactions. We found that the form “noun” and the category “modification” appear more times than other forms and categories in sentences describing interactions. But they also appear in more sentences overall than others. To see the true situation, let’s see their ratio of appearance in the sentences describing interactions over appearance in all sentences for different interaction-indicating term properties. In Table 7, the total number of sentences where different interaction-indicating terms’ forms/categories appear, and the number of sentences describing interactions where those interaction-indicating terms’ forms/categories appear, is given with the proportions of sentences describing interactions. Of all forms, the sentences containing an interaction-indicating term in the present form have the highest likelihood of describing interactions between the query biomolecule pair. The same situation occurs for the interaction-indicating term category “transcription” out of all categories. A more direct comparison can be found in Figure 1 and Figure 2.

Table 7. Data on likelihoods that sentences describe interactions when they contain biomolecule co-occurrences that are not in the same phrase.

Forms	# (%) of sentences describing interactions	Total sentences
Noun	141 (59%)	237
Adjective	9 (45%)	20
Present	50 (66%)	76
-ing	35 (51%)	69
Past/Perfect	77 (55%)	141
Categories		
Association	60 (67%)	89
Modification	80 (66%)	121
Negative regulation	33 (39%)	84
Positive regulation	47 (42%)	112
Transportation	14 (67%)	21
Transcription	5 (71%)	7
Create	63 (66%)	96
Vague	41 (54%)	76

As in the earlier research (section 2.2.1), these probabilities are integrated into the score algorithm to help evaluate a given sentence's probability of describing an interaction. This integration will be introduced in the following sections. To match the previous research results, which give different results for co-occurrence in phrases compared to sentences, here we also analyzed the interaction-indicating terms' influence when co-occurrence is in a phrase. Table 8 shows information about the number of phrases where interaction-indicating terms appear with different forms/categories. Similar to the situation for sentences, form "noun" and category "modification" appear in more phrases that other forms/categories. But in those phrases where those forms or categories appear, how many describe interactions? Table 9 gives the numbers of phrases in which different interaction-indicating terms that describe interactions appear. The number of phrases overall and phrases describing interactions are less than for sentences because a phrase has a lower probability of containing two biomolecules together. However, a phrase has a higher probability in our corpus of describing an interaction between the specific biomolecule pair when the pair co-occurs in it. We can see the fraction data in Table 10, which is similar to the sentence situation in Table 7. The proportion of phrases describing interactions is higher for almost each interaction-indicating term property than for sentences, but the ranks of the interaction-indicating term forms/categories are similar. The "present" form and "transcription" category are still the form and category having the highest probability in the corpus that the phrases containing them describe interactions. A more direct comparison can be found in Figure 3 and Figure 4.

Figure 1. Proportions of sentences with different interaction-indicating terms' forms describing interactions



As mentioned for the sentence data discussion, these proportions were integrated into our algorithm, which will be introduced later, to evaluate the likelihood that a given sentence describes an interaction. More detailed proportion data for each pair in each form and category can be found in Appendix III.

Figure 2. Proportions of sentences with different interaction-indicating term categories describing interactions

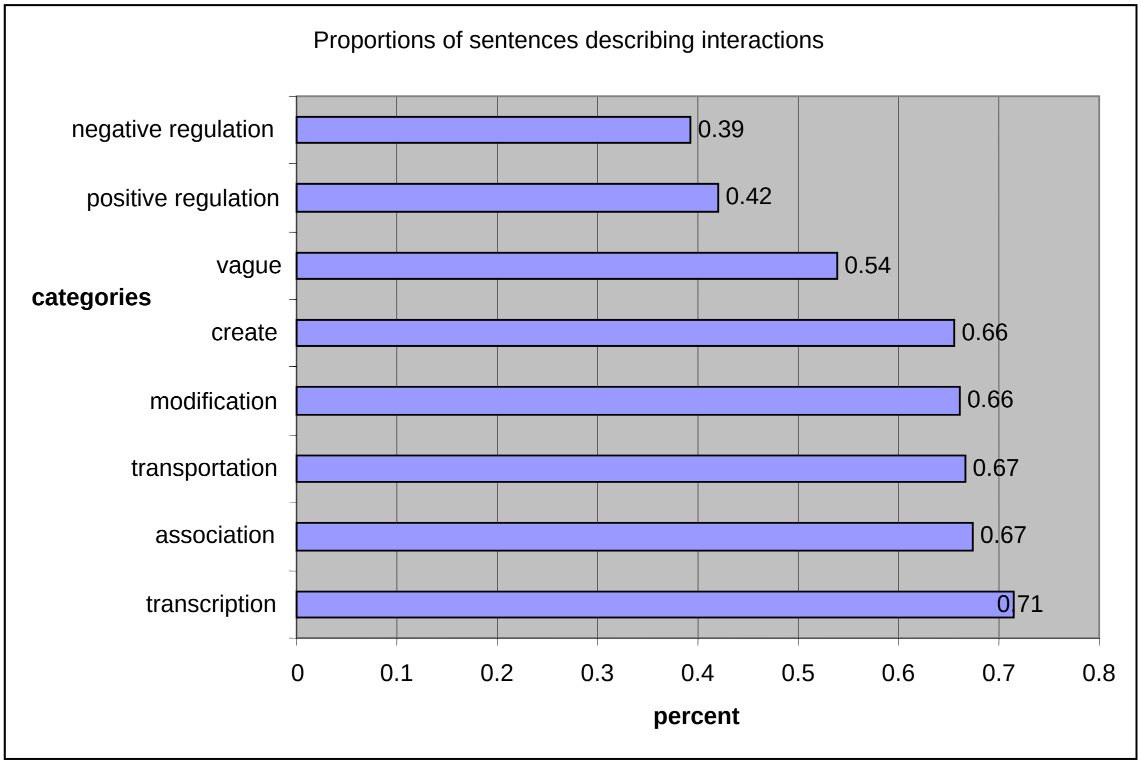


Table 8. The appearance of interaction-indicating term properties in phrases of the corpus

form or category	ATP & myosin	cre & cytokinin	nitrite & xanthine	g-6-p & starch	glucose & starch	glucose & pyruvate	acetyl-coa & leucine	indole acetic acid & starch	carotenoid & ipp	pyruvate dehydrogenase & phosphofru-ctokinase	Total
noun	26	24	23	18	12	23	12	1	1	8	148
adj	0	1	0	0	0	4	2	0	0	0	7
adv	0	0	0	0	0	0	0	0	0	0	0
present	12	2	7	0	7	9	2	1	2	0	42
-ing	16	1	4	0	2	5	0	0	0	1	29
past/perfect	15	10	13	15	7	12	12	0	0	2	86
association	22	21	0	5	0	1	4	0	2	0	55
modification	22	1	24	1	13	7	7	1	0	1	77
negative regulation	2	4	3	6	11	12	4	1	0	6	49
positive regulation	9	9	10	8	6	12	2	0	1	1	58
transportation	7	2	0	2	0	2	0	0	0	0	13
transcription	0	0	0	0	0	1	0	0	1	0	2
create	2		9	17	7	9	7	0	0	0	51
vague	10	4	6	2	3	13	6	0	0	4	48

Table 9. The appearance of interaction-indicating term properties in phrases describing interactions

form or category	ATP & myosin	cre & cytokinin	nitrite & xanthine	g-6-p & starch	glucose & starch	glucose & pyruvate	acetyl-coa & leucine	indole acetic acid & starch	carotenoid & ipp	pyruvate dehydrogenase & phosphofru-ctokinase	Total
noun	18	20	18	14	6	12	7	1	0	1	96
adj	0	0	0	0	0	2	1	0	0	0	3
adv	0	0	0	0	0	0	0	0	0	0	0
present	10	1	5	0	5	5	2	1	2	0	31
ing	8	0	2		2	3	0	0	0	1	15
past	10	8	9	12	5	4	8	0	0	0	56
perfect	0	0	0	0	0	0	0	0	0	0	0
association	14	19	0	3	0	0	3	0	2	0	41
modification	15	0	20	1	13	4	6	1	0	0	60
negative regulation	1	3	1	5	6	4	2	1	0	1	23
positive regulation	4	5	8	4	3	3	1	0	1	1	29
transportation	4	1	0	2	0	0	0	0	0	0	7
transcription	0	0	0	0	0	1	0	0	1	0	2
create	2	0	5	14	6	6	4	0	0	0	37
vague	8	4	4	2	1	8	4	0	0	0	31

Table 10. The percentages of phrases describing interactions by interaction-indicating term form and category.

forms	phrases describing interactions	all phrases	percentage
noun	97	148	66%
adj	3	7	43%
adv	0	0	0%
present	31	42	74%
-ing	16	29	55%
Past/perfect	56	86	65%
categories			
association	41	55	75%
modification	60	77	78%
negative regulation	24	49	49%
positive regulation	30	58	52%
transportation	7	13	54%
transcription	2	2	100%
create	37	51	73％
vague	31	48	65%

Figure 3. Proportions of phrases with different interaction-indicating term forms describing interactions

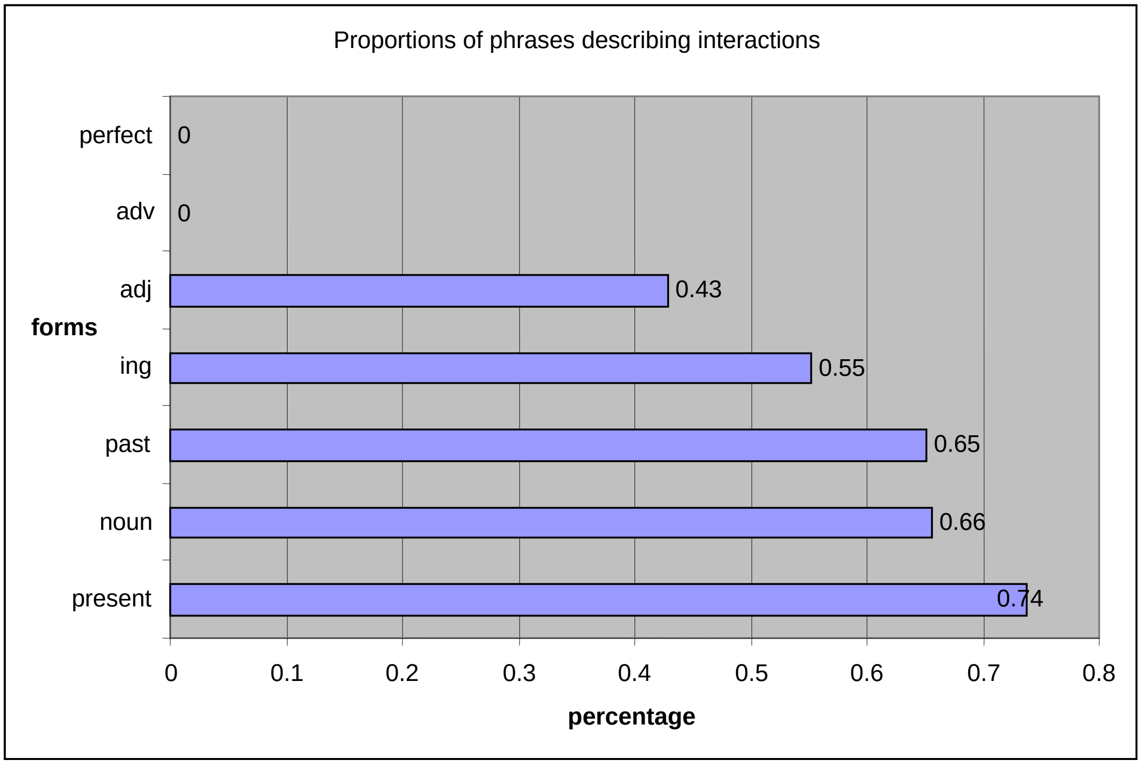
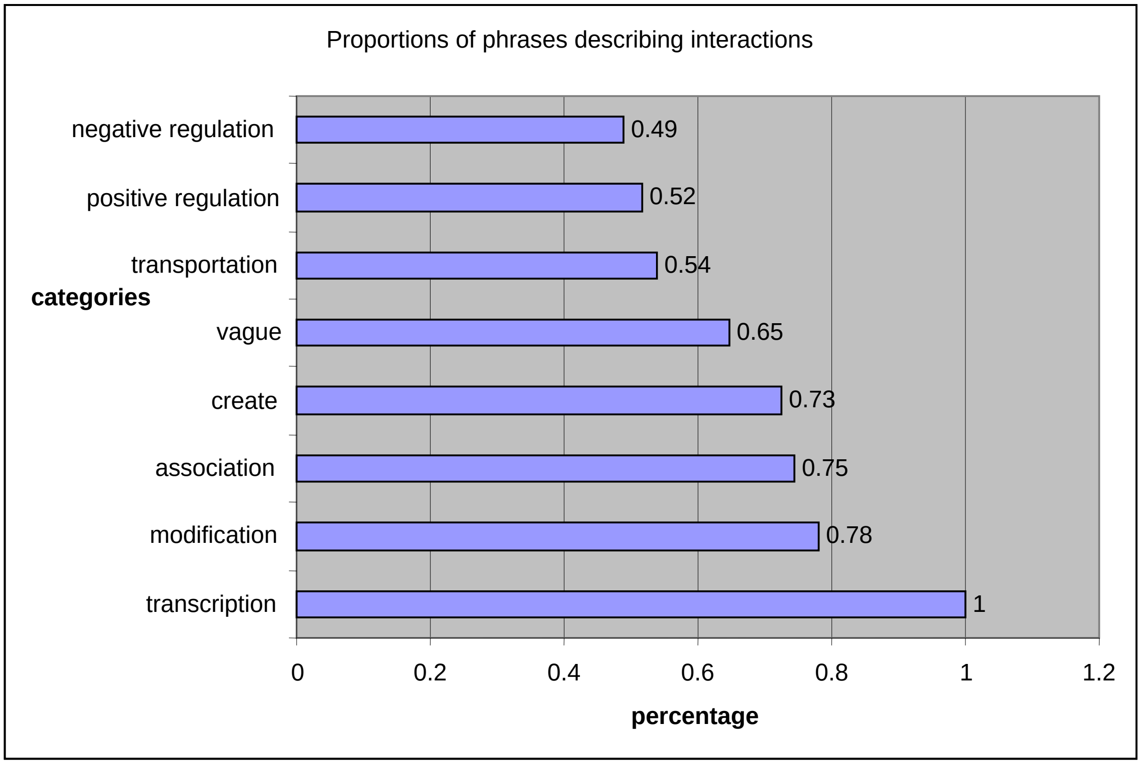


Figure 4. Proportions of phrases with different interaction-indicating term categories describing interactions



RCBD analysis

To check whether the different interaction-indicating terms' properties have different influences on the probability that a sentence describes an interaction between a specific biomolecule pair, we can use statistical methods to analyze our data. Randomized Complete Block Design (RCBD) is one of the most useful experimental designs. It groups similar experimental units into blocks (replicates) and the criterion of grouping is to make the experimental units in one group be maximally uniform. Each treatment must appear at least once in each block, and in each block, the appearance of each treatment or each experiment unit is randomized (http://www.ndsu.nodak.edu/ndsu/horsley/RCBD.pdf). These characteristics of RCBD match our corpus and following data analysis results very well so our data set can be treated as an RCBD design. The biomolecule pairs can be looked at as the blocks or the replicates. Each sentence containing them in our corpus can be looked at as an experimental unit. Each interaction-indicating term property can be looked at as a treatment in an RCBD design. The experimental result in each unit is whether the sentence describes the interaction between the biomolecule pair, but the data for each treatment to analyze are the summary of all experimental units in each block (each biomolecule pair). In fact, there are two RCBD designs, one for the interaction-indicating term forms and another for the categories.

First, let's look at the influence of interaction-indicating forms on likelihood that a sentence describes an interaction.

Table 11 gives our RCBD experimental results. The interaction-indicating forms in the table are the treatments, and the 10 biomolecule pairs are the blocks. The adjective, adverb and perfect forms don't have enough data available, so here we only analyze the difference significance among noun, present, present continuous (-ing), and past forms. In addition, the sentence set for the pair "pyruvate dehydrogenase" and "phospho-fructokinase" does not have many members describing interactions, so in the analysis part, we neglect it. First, we calculate the sum of the data in each block and each treatment, as shown in Table 12.

Table 11. The proportions of sentences in each pair's set describing specific interactions, broken out by different interaction-indicating term forms

	noun	adj	adv	present	-ing	past	perfect
ATY MYOSIN	0.69	0	0	0.67	0.52	0.63	0
cre cytokinin	0.89	0.50	0	0.71	0.33	0.81	0
nitrite xanthine	0.69	0	0	0.75	0.70	0.60	0
glucose-6-p starch	0.74	1	0	1	0.75	0.69	0
glucose-starch	0.48	0	0	0.64	0.80	0.43	0
glucose pyruvate	0.42	0.33	0	0.60	0.40	0.33	0
acetyl-coa leucine	0.61	0.50	0	0.67	0.40	0.63	0
indole acetic acid starch	0.43	0.50	0	1	0.50	0.40	0
carotenoid ipp	0.50	0	0	1	0.50	0.50	0
pyruvate dehydrogenase phosphofructokinase	0.05	0	0	0	0.13	0	0

Table 12. Influence of interaction-indicating term form on likelihood that a sentence describes an interaction: RCBD analysis step 1

Rep	noun	present	-ing	past	Total (pair)
1	0.69	0.67	0.52	0.63	2.50
2	0.89	0.71	0.33	0.81	2.75
3	0.69	0.75	0.70	0.60	2.74
4	0.74	1.00	0.75	0.69	3.18
5	0.48	0.64	0.80	0.43	2.34
6	0.42	0.60	0.40	0.33	1.75
7	0.61	0.67	0.40	0.63	2.30
8	0.43	1.00	0.50	0.40	2.33
9	0.50	1.00	0.50	0.50	2.50
Total (form)	5.44	7.03	4.90	5.01	22.39

The RCBD analysis result is in Table 13 for the forms noun, present, -ing and past.

Table 13. RCBD analysis results for interaction-indicating term form influence on the likelihood that a sentence describes an interaction. SOV: source of variance; DF: degree of freedom; SS: sum of squares; MS: mean squared; F: F test statistic

SOV	DF	SS	MS	F
pairs	8	0.307284	0.038411	1.718383
forms	3	0.323536	0.107845	4.82471
error	24	0.536465	0.022353	
total	35	1.167286		

From the F test table, we found the relevant F value: $F_{.05;8,24}$ = 2.34, $\mathrm{F}_{.05;3,24}$ =3.01. Therefore, for pairs, since F_{pairs} = 1.718383 < $F_{.05;8,24}$ = 2.34 at the 95% level of confidence, we fail to reject H_0: all pairs are equal or there is no clear difference among pairs' influences on the likelihood that sentences describe the interaction. For different interaction-indicating term forms, since F_{forms} = 4.82471 > $F_{.05;3,24}$ =3.01 at the 95% level of confidence, we reject H_0: interaction-indicating term forms are the same. In other words, we are 95% confident that different interaction-indicating term forms have different influences on the likelihood that a sentence describes an interaction.

For interaction-indicating term categories, the same analysis was used.

Table 14. The proportions of sentences in each pair's set describing specific interactions where different interaction-indicating term categories appear

	association	modification	negative regulation	positive regulation	transportation	transcription	create	vague
ATY MYOSIN	0.62	0.68	0.33	0.40	0.71	0	1	0.77
cre cykotinin	0.91	0	0.78	0.75	1	1	1	1
nitrite xanthine	1	0.72	0.40	0.68	0	0	0.50	0.71
glucose-6-p starch	0.57	1	0.71	0.45	1	0	0.84	1
glucose-starch	0	0.80	0.38	0.27	0	0	0.62	0.20
glucose pyruvate	0	0.33	0.33	0.20	0.33	1	0.50	0.57
acetyl-coa leucine	0.38	0.73	0.33	0.44	0	0	0.74	0.64
indole acetic acid starch	0	0.67	0.60	0	0	0	0.50	0.33
carotenoid ipp	0.75	0	1	0.33	0	0.67	0.67	0
pyruvate dehydrogenase phosphofructokinase	0	0	0.06	0.08	0	0	0	0

Table 14 gives proportion data for each interaction-indicating form category (treatment) in each biomolecule pair (block). The categories "transportation" and "transcription" and the pairs "indole acetic acid & starch", "carotenoid & ipp" and "pyruvate dehydrogenase & phosphofructokinase" have too many data of value 0, so they are not used in the analysis procedure, as too many 0s make the data useless. The data about term categories are listed in Table 15.

Table 15. Interaction-indicating term category influence on likelihood that a sentence describes an interaction: RCBD analysis step 1

Rep	association	modification	negative regulation	positive regulation	create	vague	Total (pairs)
1	0.62	0.68	0.33	0.40	1	0.77	3.80
2	0.91	0	0.78	0.75	1	1	4.44
3	1	0.72	0.40	0.68	0.50	0.71	4.02
4	0.57	1	0.71	0.45	0.84	1	4.58
5	0	0.80	0.38	0.27	0.62	0.20	2.26
6	0	0.33	0.33	0.20	0.50	0.57	1.94
7	0.38	0.73	0.33	0.44	0.74	0.64	3.26
Total (categories)	3.48	4.27	3.27	3.20	5.19	4.89	24.30

The RCBD analysis results in Table 16 are for the categories association, modification, negative regulation, positive regulation, create and vague.

Table 16. RCBD analysis result for interaction-indicating term category influence on sentence interaction description likelihood

SOV	DF	SS	MS	F
Pairs	6	1.074816	0.179136	3.076793
Categories	5	0.53288	0.106576	1.830521
Error	30	1.74665	0.058222	
Total	41			

From the F test table, we found the relevant F values: $F_{.05;6,30} = 2.42$, $F_{.05;5,30} =2.53$. Therefore, for pairs, because $F_{\text{pairs}} = 3.0767793 > F_{.05;5,30} = 2.42$ at the 95% level of confidence, we reject H_0: all pairs are equal. That is, from the test results, we can say there is a difference among pairs' influences on likelihoods that sentences describe interactions with 95% confidence. For different interaction-indicating term categories, because $F_{\text{categories}} = 1.830521 < F_{.05;5,30} =2.53$ at the 95% level of confidence, we fail to reject H_0: all interaction-indicating terms' categories have the same influence on the likelihood of sentences' describing interactions. So we cannot say that different IIT categories have an influence. It appears the interactions between some pairs are unique or exceptional.

We also show the phrase-based experimental data about the forms and categories in Table 17. In this table we see that because the numbers of phrase examples where two biomolecules and at least one interaction-indicating term tri-occur are less than the number of sentence examples, the result data have a lot of 0s and 1s. Too many 0s and 1s may not really reflect the true situation, so we did not do RCBD analysis for phrase experimental data.

Table 17. The proportions of phrases in each pair's set describing the specific interaction for different interaction-indicating term forms/categories

	noun	adj	adv	Present	-ing	past	Perfect	Association	Modification	Negative regulation	Positive regulation	transportation	transcription	create	vague
ATY MYOSIN	0.69	0	0	0.83	0.5	0.67	0	0.64	0.68	0.5	0.44	0.57	0	1	0.8
Cre Cykotinin	0.83	0	0	0.5	0	0.8	0	0.9	0	0.75	0.56	0.5	0	0	1
nitrite xanthine	0.78	0	0	0.71	0.5	0.69	0	0	0.83	0.33	0.8	0	0	0.56	0.67
Glucose-6-p Starch	0.78	0	0	0	0	0.8	0	0.6	1	0.83	0.5	1	0	0.82	1
Glucose Starch	0.5	0	0	0.71	1	0.71	0	0	1	0.55	0.5	0	0	0.86	0.33
Glucose Pyruvate	0.52	0.5	0	0.56	0.6	0.33	0	0	0.57	0.33	0.25	0	1	0.67	0.62
acetyl-coa leucine	0.58	0.5	0	1	0	0.67	0	0.75	0.86	0.5	0.5	0	0	0.57	0.67
indole acetic acid starch	1	0	0	1	0	0	0	0	1	1	0	0	0	0	0
Carotenoid Ipp	0	0		1	0	0	0	1	0	0	1	0	1	0	0
Pyruvate dehydrogenase phosphofructo- Kinase	0.13	0		0	1	0	0	0	0	0.17	1	0	0	0	0

Therefore, for sentence interaction evaluation, interaction-indicating terms' categories give little help so it is not appropriate to use it to evaluate interaction description. Our algorithm will not use interaction-indicating terms' categories.

Sentence Evaluation Algorithm

At this point, we wish to create an algorithm for evaluating the probability that a sentence describes an interaction between two given biomolecules. We will use odds from information about location of biomolecule names in sentences (from previous research, see Appendix IV) and about interaction-indicating term forms (from my research). We combined the evidence contributed by different sentence attributes by multiplying the odds of the attributes, and normalizing. Done appropriately this returns the composite probability that the sentence describes the hypothesized interaction (Manning et al. 2008, sections 11.1, 11.3; Davis 1990, : 128-130). The precise formula we used is $O(h|f_1,...,f_n)=O(h|f_1)O(h|f_2)...O(h|f_n)\ /\ O(h)^{n-1}$, which expresses the odds of hypothesis h (here that a given passage describes an interaction between a given pair of biomolecules), given n sources of evidence and a default odds $O(h)$ modeling the entire corpus. The $O(h|f_k)$, $k=1,\ldots,n$, are the odds of the hypothesis, given sentence feature or attribute k as evidence. See Dickerson et al. (2005 section 2.3.3) and Berleant (2004). Specifically, in our method, given a sentence and two biomolecule names:

1) Identify the co-occurrences of these two biomolecules in the sentence. Each co-occurrence is treated separately. To rate a sentence containing more than one co-occurrence:
    a. For each co-occurrence, determine the odds that it is part of an interaction description based on the locations in the sentence of the co-occurring biomolecules, and similarly, determine the odds based on the morphological form of the interaction-indicating term that the co-occurrence is part of a description of an interaction between the given biomolecule pair. (See steps 2 and 3 respectively below.)
    b. Combine the odds of the co-occurrences based on location to get a probability that the sentence describes an interaction based on the location information. Also combine the odds of the co-occurrences based on interaction-indicating term form information to get a probability that the sentence describes an interaction based on the interaction-indicating term form information. There are two methods of combination, each of which can provide the pair of probabilities we seek (this pair will be combined together later, in step c).

Method 1:

Convert each of the two odds for each co-occurrence into a probability, using p=odds/(1+odds). Separate them into two groups based on the sources of evidence (which are the location and interaction-indicating term form.) In each evidence group (location and interaction-indicating term form), combine the probabilities of the co-occurrences to get the probability that at least one of the co-occurrences in the sentence is in an interaction description based on the corresponding source of evidence (location or interaction-indicating term form) using the following equation:

$$\begin{aligned}
&p(\text{at least one co-occurrence is in an interaction description})\\
&\quad =1\text{-}p(\text{no co-occurrence is in an interaction description})\\
&\quad\quad =1\text{-}p(\text{co-occurrence \#1 is not in an interaction description})*p(\text{co-occurrence \#2 is not in an interaction description})*p(\text{co-occurrence \#3 is not in an interaction description})*\ldots\\
&\quad\quad =1-\prod_{i} p(\text{ co-occurrence } i \text{ is not in an interaction description})\\
&\quad\quad =1-\prod_{i} (1-p(\text{ co-occurrence } i \text{ is in an interaction description}))
\end{aligned}$$

Note that the degenerate case where a sentence has only one co-occurrence works with this formula too.

Method 2:

As in method 1, convert each co-occurrence odds (some based on location and some on interaction-indicating term forms) into a probability using p=odds/(1+odds). Separate them into two groups, one based on the location evidence and the other on interaction-indicating term form evidence. In each evidence group, take the maximum probability as an estimate of the sentence's probability that it is in an interaction description based on the corresponding source of evidence (location or interaction-indicating term form):

p (at least one co-occurrence is in an interaction description)

=Max(p(co-occurrence i is in an interaction description)) for location evidence, and similarity for interaction-indicating term form evidence,

c. For the two probabilities from two groups of sources of evidences combine them according to:

p (the sentence describes the interaction)

$=1 - (1- p_{\text{position}})(1-p_{\text{form}})$.

In the two probability combination methods, method 1 assumes each co-occurrence is independent, and then each co-occurrence's probability can be multiplied directly to give the final probability that there is a co-occurrence that is the interaction description. However, it is risky to assume they are independent, as there is typically no particular reason to believe they are independent. Moreover, from previous research (Berleant et al., see Appendix IV), multiple co-occurrences do not necessarily improve the probability that a sentence describes an interaction. Method 2 avoids assuming independence by using the best co-occurrence in a given sentence. Therefore, in my implementation, method 2 is used to combine evidence from the co-occurrences in a sentence. Note that, in method 2, choosing the best interaction-indicating term with highest odds is also used to assign the odds that a sentence describes an interaction when there are multiple interaction-indicating terms appearing with the co-occurrence in the sentence. Thus if we meet an interaction-indicating term and we cannot determine its form (noun or present when their spellings are the same,) we will use the form that brings larger odds.

2) To calculate the odds O(co-occurrence i is in an interaction description) from location evidences, do the following.

   a. If the co-occurrence is within a phrase:

      i. If no interaction-indication term is in the phrase, estimate p=0.1

      ii. If an interaction-indicating term is in the phrase,

         1. if there is 1 co-occurrence in the phrase, estimate O_{2aii1}=0.7/0.3=2.33

2. if there is >1 co-occurrences in the phrase, estimate $O_{2aii2}=0.86/0.14=6.1$
3. if there is not an interaction-indicating term between the co-occurring terms, estimate $O_{2aii3}=0.24/0.76=0.316$, except if separation=0. In that case, $O_{2aii3}=1/9$ (as an estimate of 0+epsilon).
4. if there is an interaction-indicating term between the co-occurring terms, and separation>0, estimate $O_{2aii4}=(-0.03k+0.9)/(1-(-0.03k+0.9)=(-0.03k+0.9)/(0.1+0.03k))$, where k is the number of words between the co-occurring terms. However, if this is below 0, set $O_{2aii4}=0$. If separation=0 then $O_{2aii4}=17$.
5. Let prior odds $O_{phrase}=0.68$.
6. Compute the product of all the $O_{2aii\_}$ that apply.
7. Divide by O_{phrase}^{n-1} where n is the number of $O_{2aii\_}$ that apply.

b. If the co-occurrence is within a sentence but not a phrase:

1. if the sentence has 1 co-occurrence, estimate $O_{2b1}=0.4/0.6=0.67$
2. if the sentence has >1 co-occurrences, estimate $O_{2b2}=0.32/0.68=0.47$
3. if there is an interaction-indicating term between the co-occurring terms, and separation>0, estimate $O_{2b3}=(-0.01k+0.6)/(1+0.01k-0.6)=(-0.01k+0.6)/(0.4+0.01k)$ where k is the number of words between the co-occurring terms. However, if this is below 0, set $O_{2b3}=0$). If separation=0, then $O_{2b3}=1/9$ (this is an estimate of 0+epsilon).
4. if there is not an interaction-indicating term between the co-occurring terms, and separation>0, estimate $O_{2b4}=(-$

$0.0033k+0.2)/(0.8+0.0033k)$ $=(-0.0033k+0.2)/(0.8+0.0033k)$ where k is the number of words between the co-occurring terms. However, if this is below 0, set $O_{2b4}=0$). If separation=0, then $O_{2b4}=4/7$.

5. Let prior odds $O_{sentencee}=0.33$.
6. Compute the product of all the $O_{2b\_}$ that apply.
7. Divide by $O_{sentence}^{n-1}$ where n is the number of $O_{2b\_}$ that apply.

3) To calculate the odds O(co-occurrence i is in an interaction description) from interaction-indicating term evidence, do the following.
   a. If the co-occurrence is within a phrase:
      i. If no interaction-indicating term is in the phrase, return null (i.e., do not return interaction-indicating term form odds)
      ii. If there are interaction-indicating terms in the phrase, for each term, find its odds O_{3aii} based on the following Table 18.

Table 18. Interaction-indicating term form odds in phrase

Form	Odds
Noun	1.902
Adj	0.75
Present	2.818
-ing	1.231
past/perfect	1.867

      iii. Find the highest odds of all of the interaction-indicating terms and return that as the final odds.
   b. If the co-occurrence is within a sentence (but not a phrase):
      i. If no interaction-indicating term is in the sentence, return null (i.e., do not return interaction-indicating term form odds).

ii. If there are interaction-indicating terms in the sentence, for each term, find its odds O_{3bii} based on the following Table 19.

Table 19. Interaction-indicating term form odds in sentence

Form	Odds
Noun	1.469
Adj	0.818
Present	1.923
-ing	1.029
past/perfect	1.203

iii. Find the highest odds of all of the interaction-indicating terms and return that as the final odds.

4) Possible Odds normalization. Our corpora were made of sentences containing pairs of biomolecules that do interact in reality. Thus the odds obtained from analysis of these corpora may deviate from the real odds in MEDLINE because many pairs of biomolecules do not interact in reality. Therefore we created another corpus made of 300 sentences randomly chosen from MEDLINE (details are in Appendix V). These sentences contain at least two biomolecules different from those used in the corpora above.

We get another odds from this corpus, called O_{300}, and want to use this odds to normalize the above algorithm. The relative probability p_{300} is the probability that a sentence describes an interaction between two biomolecules in it, based on these 300 sentences. Assuming it represents the whole of MEDLINE, then this corpus, which is independent from the location and form corpora, can be used to normalize odds and probabilities from steps 2) and 3) above.

For example, the probability p that a sentence describes an interaction if it contains a noun form of an interaction-indicating term can be normalized as:

$p = p_{\text{noun}} * p_{300}$

Then we can reason that $O = p / (1\text{-}p)$ implies that

$O = (p_{\text{noun}} * p_{300}) /(1\text{-}(p_{\text{noun}} * p_{300}))$.

Since p=O/(1+O), the last equation becomes

$$O = \frac{\frac{O_{\text{noun}}}{1+O_{\text{noun}}} \times \frac{O_{300}}{1+O_{300}}}{1-(\frac{O_{\text{noun}}}{1+O_{\text{noun}}} \times \frac{O_{300}}{1+O_{300}})}$$

$$= \frac{O_{\text{noun}} \times O_{300}}{1+O_{\text{noun}}+O_{300}}.$$

We found the values $O_{300} = 0.723$ and $p_{300} = 0.419$.

In this way, we can normalize the results from the algorithm. There are two ways to normalize the odds. One is to normalize the prior odds only, used in step 2), because they represent the default odds that come from a background sentence set. Another way is to apply the normalization to each odds, in the hope that this makes each odds more likely to reflect the truth. However, there is a problem for the normalization. In section 2).a.i, the algorithm does not calculate the odds but instead gives the probability 0.1 directly when no interaction-indication term is in a phrase, therefore there is no place to apply O_{300} because here there are neither prior odds nor individual numerator odds (e.g. $O_{2\text{aii}1}$.) For this case, we will use equation (1) to normalize final probability directly with p_{300}.

Overall, then, we have three candidate methods for calculating the final probability that a sentence describes an interaction between two given biomolecules. The first one is the original one of steps 1), 2), and 3). The second one is to normalize the prior odds in step 2) and the third one is to normalize all odds in steps 2) and 3) by O_{300}. We call them ORI, PRI, and ALL. The results of each method will be used to choose the one that best matches the reality. Since the normalization of step 2).a.i is the same for the first and the second methods, we will not compare the results from this step of the first and second methods.

Note 1: if the numbers should change in the future, the software will need to be changed accordingly.

Note 2: ***Semi-Naïve Evidence Combination.*** This method is scalable in the number of features, like Naïve Bayes, but has the advantage of making fewer independence assumptions. The most parsimonious formula for semi-naïve evidence combination is

$$O(i|f_1,...,f_n)=O_1...O_n/O_p^{\mathrm{n-1}} \qquad (1)$$

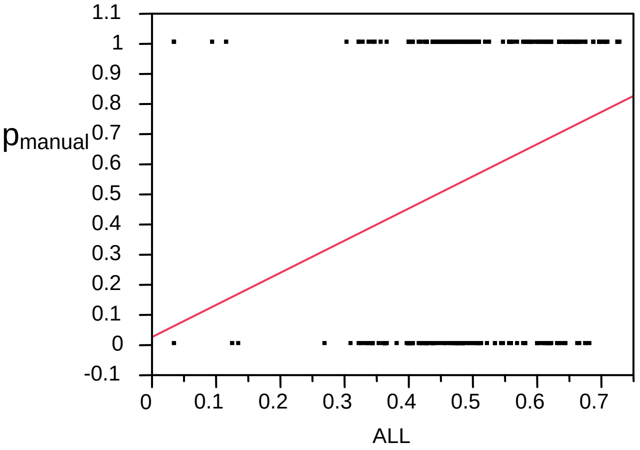
where $O(i|f_1,...,f_n)$ is the odds that a passage describes an interaction if it has features $f_1,...,f_n$, O_k is the odds that a passage with feature k contains an interaction and O_p is the prior odds (i.e. over all passages in the test set irrespective of their features) that a passage contains an interaction. Eq. (1) is in terms of odds, but these are easily converted to the more familiar probabilities by substituting $O=p/(1\text{-}p)$; thus the odds of flipping a head are 1/1=1 (1 expected success per failure), while the odds of rolling a six are 1/5 (one success expected per five failures).

Example 1. As a simple case, consider a set of sentences, 4 with interactions and 4 without. Feature f_1 is associated with 4 sentences that describe interactions and 2 that don’t. The same holds for feature f_2. Then O_1=4/2=2=2/1, or an expectation of drawing two sentences with interactions (a hit) for each one without (a miss). Likewise for O_2. O_0=4/4=1, so by Eq. 1, $O(i|f_1,f_2)$=2·2/1=4=4/1, or odds of 4 to 1 that a sentence with both features describes an interaction. This corresponds to probability $p(i|f_1,f_2)$=4/5.

Evaluating the Evidence Combination Algorithm

We have devised three different evidence combination methods and wish to choose the best one. For each sentence in our corpus of sentences containing one of the 10 pairs of biomolecules, the three evidence combination methods were used to calculate the probability that the sentence describes an interaction between the biomolecule pair. We have 320 sentences so we have 320 probability estimates for each method. In creating the corpus, we manually judged whether the sentence describes the interaction, recording 1 if so or 0 if not. Combining these 4 (3 automatic + 1 manual) scores, we get a table of 4 scores for the corpus of 320 sentences (Appendix VI). For each automatic method's 320 probability scores, we did a linear regression fit to the manual data. Ideally, the number of sentences describing interactions divided by the total number of sentences should exactly equal the probability. Therefore, the ideal linear regression result should be the line *Y*=*X*. Let's see the actual regression results using JMP software:

Figure 5. The linear regression results using the "ALL" method. Note that the 320 manually determined data points, all with values of 0 or 1, often overlap.



Fit of interaction By ALL

—— Linear Fit

Linear Fit

p_{manual} **= 0.0288512 + 1.0660049*ALL**

Summary of Fit

RSquare	0.071729
RSquare Adj	0.0688
Root Mean Square Error	0.479626
Mean of Response	0.561129
Observations (or Sum Wgts)	319

Analysis of Variance

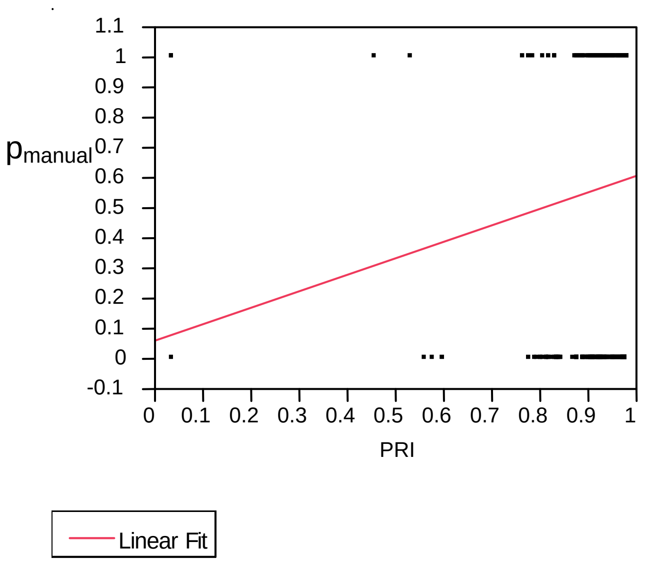
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	5.634864	5.63486	24.4950
Error	317	72.923129	0.23004	**Prob > F**
C. Total	318	78.557994		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.0288512	0.110849	0.26	0.7948
ALL	1.0660049	0.215387	4.95	<.0001

Figure 6. The linear regression results using the "PRI" method. Note that the 320 manually determined data points, all with values of 0 or 1, often overlap.

Fit of interaction By APPLYING ODDS ON PRI ONLY



Linear Fit

$p_{manual} = 0.0587154 + 0.5489084*PRI$

Summary of Fit

RSquare	0.014581
RSquare Adj	0.011473
Root Mean Square Error	0.49417
Mean of Response	0.561129
Observations (or Sum Wgts)	319

Analysis of Variance

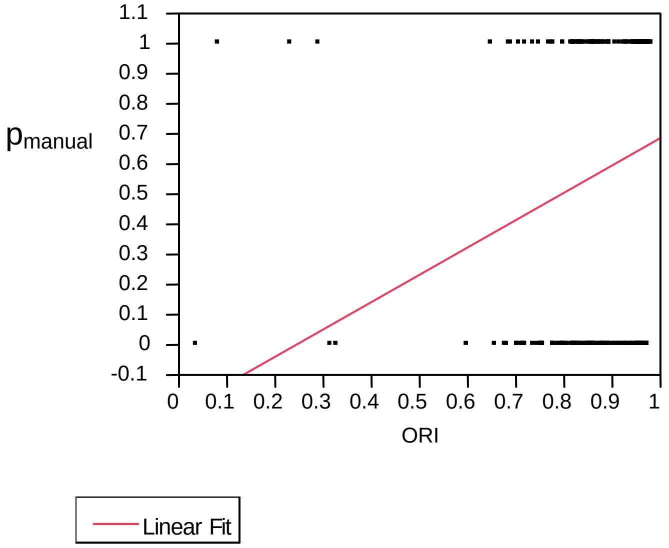
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Ratio**
Model	1	1.145470	1.14547	4.6906
Error	317	77.412524	0.24420	**Prob > F**
C. Total	318	78.557994		0.0311

Parameter Estimates

Term	**Estimate**	**Std Error**	**t Ratio**	**Prob>|t|**
Intercept	0.0587154	0.233621	0.25	0.8017
PRI	0.5489084	0.253445	2.17	0.0311

Figure 7. The linear regression results using the "ORI" method. Note that the 320 manually determined data points, all with values of 0 or 1, often overlap.

Fit of interaction By ORI without applying O300



Linear Fit

p_{manual} = -0.220473 + 0.9057469*ORI

Summary of Fit

RSquare	0.048583
RSquare Adj	0.045581
Root Mean Square Error	0.485569
Mean of Response	0.561129
Observations (or Sum Wgts)	319

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	3.816549	3.81655	16.1871
Error	317	74.741445	0.23578	**Prob > F**
C. Total	318	78.557994		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-0.220473	0.196161	-1.12	0.2619
ORI	0.9057469	0.225124	4.02	<.0001

From these regression results, we can see that the ALL method achieved the best regression result:

$$p_{\text{manual}} = 0.0288512 + 1.0660049^{*}\ p_{\text{ALL}} \qquad (2)$$

which is nearest to the line *Y*=*X*. In other words, the ALL method best reflects the observed probability that a sentence describes an interaction. Therefore, the ALL method was used as the input to the final version of our sentence evaluation algorithm.

To make our final score more accurate, however, we need to adjust the ALL method to make the final score regression line become exactly Y=X. Because we wish to adjust p_{ALL} to equal p_{manual}, we defined

$$p_{\mathrm{adjusted}}=0.0288512 + 1.0660049^{*}\ p_{\mathrm{ALL}} \qquad (3)$$

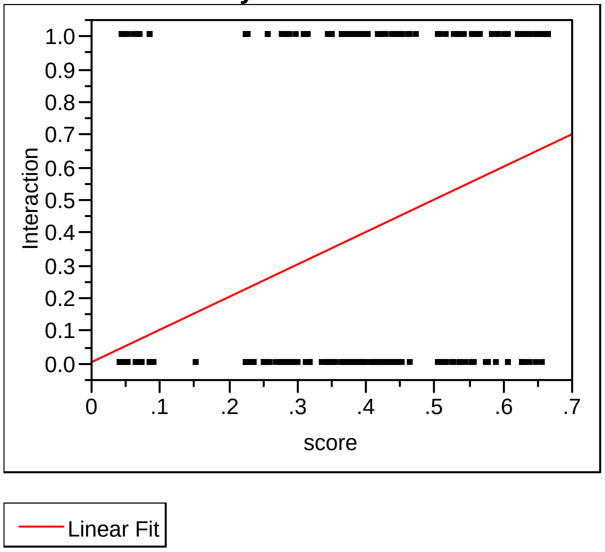
This modification to the ALL method, called Adjusted All, for the corpus we used, gives results that regress to exactly the line Y=X.

To test the validity of the $p_{adjusted}$ calculation, we collected a test set of 123 sentences in which the 10 biomolecule pairs we used to create the corpus appear, but which were not already in the 320-sentence experimental set. For each test set sentence, PathBinder calculated $p_{adjusted}$. Whether it described an interaction between the two biomolecules of interest in it was also judged manually. In addition, to make the test more comprehensive, we collected another test set of sentences with values of $p_{adjusted}$ of 0, 0.1±0.01, 0.2±0.02, 0.3±0.03, 0.4±0.04, 0.5±0.05, 0.6±0.06, 0.7±0.07 and 0.739±0.07 (the $p_{adjusted}$ computation gives results up to about 0.739). About 50 sentences for each value were collected. Some of them came from the 123 sentences noted above, and others were from search results using the pairs: *ethanol & acetaldehyde, acetyl-CoA & NADH, dynamin & GTP, adenylate cyclase & ATP,* and *ATP & creatine*. Overall, there were 600 sentences in this set, test set B. PathBinder computed $p_{adjusted}$ for each of these sentences, and whether they really described the interaction also was judged manually and recorded as 0 (no) or 1 (yes). Then we did a line regression as we did earlier. The regression results are In Figure 8.

The regression line we get is *Interaction* = 0.0069822 + 0.9943749**score*, or Y = 0.0069822 + 0.9943749X, which is very close to the ideal line Y=X. We can compare this line (based on the Adjusted ALL method) with the line Y=X and the regression line of the ALL method (Figure 9).

Figure 8. The linear regression results for the test set of 600 sentences. Note that the manually determined data points, all with values of 0 or 1, often overlap.

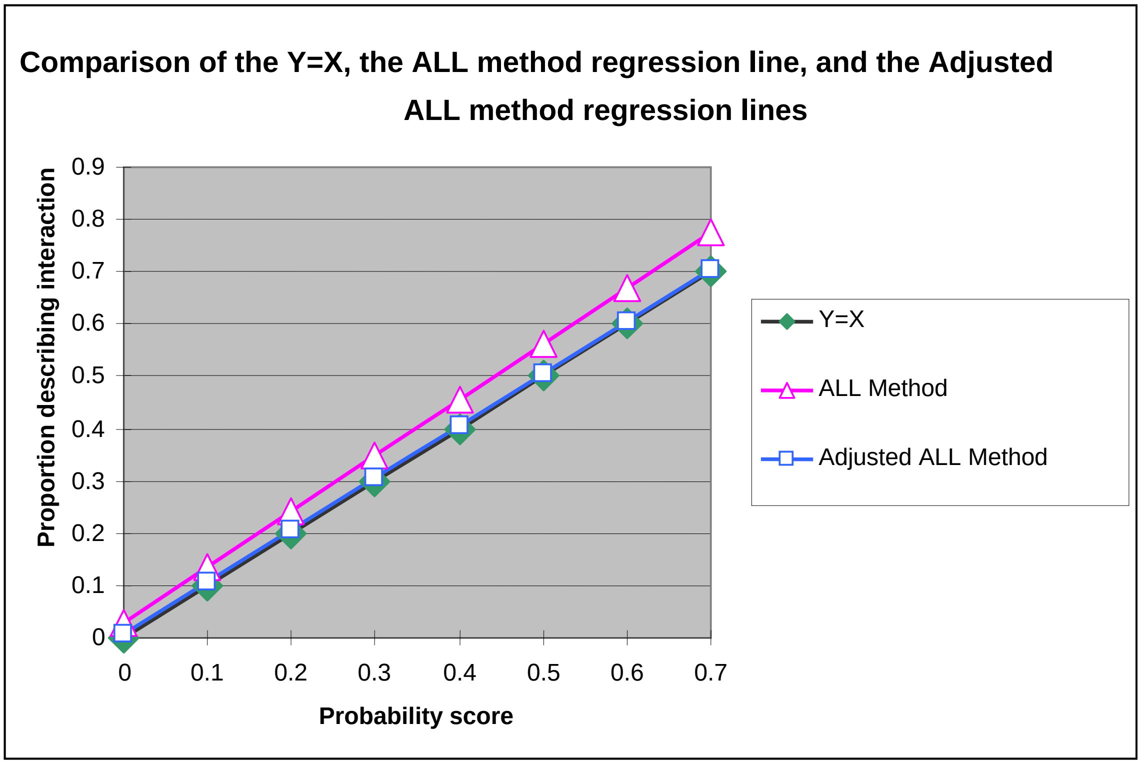
Fit of Interaction By score



```
Linear Fit
Interaction = 0.0069822 + 0.9943749 score

Summary of Fit

RSquare                             0.138873
RSquare Adj                         0.137429
Root Mean Square Error              0.439626
Mean of Response                    0.337793
Observations (or Sum Wgts)               598

Analysis of Variance
Source          DF     Sum of Squares     Mean Square      F Ratio
Model            1           18.57653         18.5765      96.1166
Error          596          115.18936          0.1933     Prob > F
C. Total       597          133.76589                       <.0001

Parameter Estimates
Term              Estimate      Std Error     t Ratio     Prob>|t|
Intercept        0.0069822       0.038233        0.18       0.8552
score            0.9943749       0.101426        9.80       <.0001
```

From Figure 9, we can see the Adjusted All algorithm generates a regression line from the test set very close to the ideal result *Y*=*X*. In fact, they are almost identical. On the other hand, before adjusting to *Y*=*X*, the ALL method is quite distinct from the *Y*=*X* line. After adjustment, therefore, the algorithm gives very good results. Thus the PathBinder system can give nearly ideal probabilities for sentences describing interactions between biomolecule pairs.

Figure 9. The Comparison among *Y=X*, ALL method regression line, and Adjusted ALL method regression line.



2.2.3 Applying the results of section 2.2.2.2 to identify interactions between biomolecules and create interaction network

2.2.3.1 Methods for identifying interactions between two biomolecules

For each candidate sentence, we can assess its probability of describing an interaction. This enables assessing the probability of interaction between two biomolecules by combining the evidence provided by multiple sentences containing the same co-occurrence of biomolecules. To do this, a method for combining the quality and quantity of sentences containing a given pair of biomolecules is needed. The basic idea is that the more sentences describe the interaction between the two biomolecules well, the higher the probability the two biomolecules really interact. I will use different methods and then evaluate which one provides the best results.

Here is the *comprehensive method* of combining evidence. Let p be the probability that a sentence describes an interaction. Then $q = 1 - p$ is the probability that the sentence does not describe an interaction. Given n such independent sentences and assume the probability p that each sentence describes an interaction is same, q^n is the probability that none of them describe an interaction. So the probability that there is an interaction is described is $1 - q^{n.}$ Therefore,

p(*there is an interaction described between these two entities*)

$$= 1 - q^n$$

$$= 1 - (1 - p)^n \qquad (4)$$

Based on this formula, we can compute the probability that there is an interaction for a pair of biological entities from n relevant sentences. In the more typical case of n sentences each with its own value p_i for the probability that it describes an interaction, the formula generalizes to: given n independent sentences where the two entities appear,

p(*there is an interaction described between these two entities*)

$$= 1 - (1 - p_1)(1 - p_2)(1 - p_3)\ldots(1 - p_n) \qquad (5)$$

Scientific knowledge is developing continuously. New theories and discoveries appear daily. Some new interactions may be mentioned only in the most recent publications, which means the number of publications describing these interactions is limited compared to more typical interactions. Thus, particularly for a recent discovery, few sentences may not be evidence against interaction. Therefore, we also assessed the ability of two variant methods for identifying the probability of an interaction between two biomolecules. These are as follows:

- Use the average of the scores of the top 5 sentences (those having the highest score for the probability of describing an interaction between the two biomolecules).
- Use the average of the scores of the top 2 sentences.

These two methods are called “Best 2” and “Best 5” respectively. The method of formula (1) is called the “All” method here. If a biomolecule pair has less than 2 or 5 sentences where they co-occur, probability 0 is used for missing sentences to reach 2 or 5 when calculating the average probability scores. The number of sentences does not influence the computation in these cases.

2.2.3.2 Interaction network creation

As just described, for any pair of biomolecules, we can estimate a probability of interaction. But how do we find or enumerate those biomolecule candidates? We will use an existing database about genome-wide plant mRNA, protein, and metabolite profiling data, MetNetDB, in which there is a biomolecule entity table. The table currently contains gene, protein, and biochemical entities focusing especially on Arabidopsis and soy. Each entity in the table is a biomolecule candidate. A direct method for getting biomolecular pairs is to select any two biomolecules in the table. Then for each such pair, the sentences where they co-occur can be collected and analyzed to estimate the probability that they are described as interacting.

Using this method, we could get all pairs of biomolecules having a high probability of interaction. All these interacting pairs would form a biomolecular interaction graph or network. The biomolecules would be the vertices in the network, and if two biomolecules are found to interact, there will be an edge between these two biomolecules (vertices). Therefore choosing biomolecules and then biomolecule pairs from a database, we can create a corresponding interaction network. The goal here is to create the interaction network from MetNetDB. It can be one for Arabidopsis, soybean, both, or future new species.

If the number of candidates were small, enumerating pairs for them would be a reasonable job. However, the entity table in MetNetDB has more than $2*10^6$ records so that the enumeration of pairs would be more than 100 billion. If each pair needs to be evaluated through the whole literature, it would not be affordable. However, not all pairs co-occur in the literature, and currently they are only in the tens of millions of sentences in MEDLINE. So, instead of enumerating biomolecule pairs from the entity table directly, the pairs will come from parsing sentences. Sentences in MEDLINE will be scanned one by one, and all biomolecular pairs in both a sentence and the entity table will be extracted to enumerate the biomolecular pairs. In other words, each biomolecular pair is extracted from sentences that contain co-occurring biomolecules, so those biomolecules in the entity table that do not occur in any sentences or co-occur with other biomolecules not in the table will not be counted. This will save a lot of time.

For the comprehensive method, $1 - p_i$ is stored in the PathBinder database for each sentence *i* and for each pair of biomolecules co-occurring in the sentence *i*. Finally, eq. (5) is applied to the set of sentences containing a given biomolecular pair. The result is the probability of interaction between the biomolecule pair as evaluated from the literature. This comprehensive method is currently employed in PathBinder as a demo. The whole procedure is shown in Figure 10.

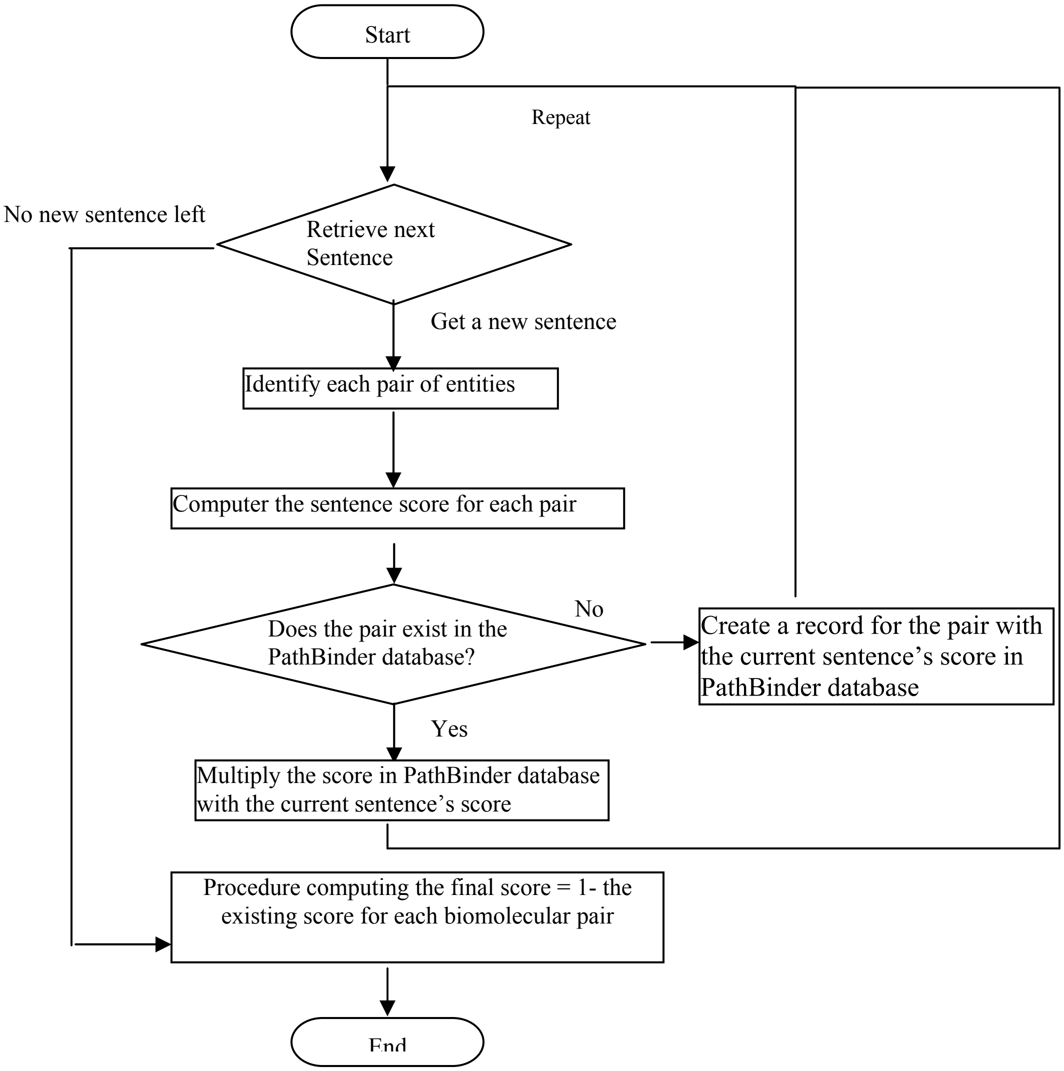
For the top part methods, which only use the top-scoring sentences, we will still scan the sentences one by one. But a cache will be maintained for each extracted biomolecule pair when scanning all sentences. Each biomolecule pair's cache stores up to five top scoring sentences for the second method and up to two sentences for the third method. After scanning the whole literature, the average score of the final top five or two sentences will be computed and will become the final score for the interaction between the biomolecule pair.

The overall system structure is shown in Figure 11. There are two main parts.

1. Extracting interactions.
   a. The system examines each sentence in MEDLINE for keywords (biomolecules, IITs, & cellular locations) stored in MetNetDB, tags them and stores tagged sentences into PathBinderDB.
   b. When scanning each sentence, the system combines the interaction evidence for each biomolecule pair inside the sentence using the formula (1), and accumulates those scores with different methods when meeting same pair in the future. We have two tables about biomolecules, one for each appearance in citations and one for entities recognized by biologists but which may not appear in citations. The relationship between them is n to m. Some terms

appearing in different citations may be a same biomolecule. When accumulating the scores, we first calculate the score for the appeared term pari, find the appeared term's relative entity record, and then accumulate the score for the actual biomolecule pair.

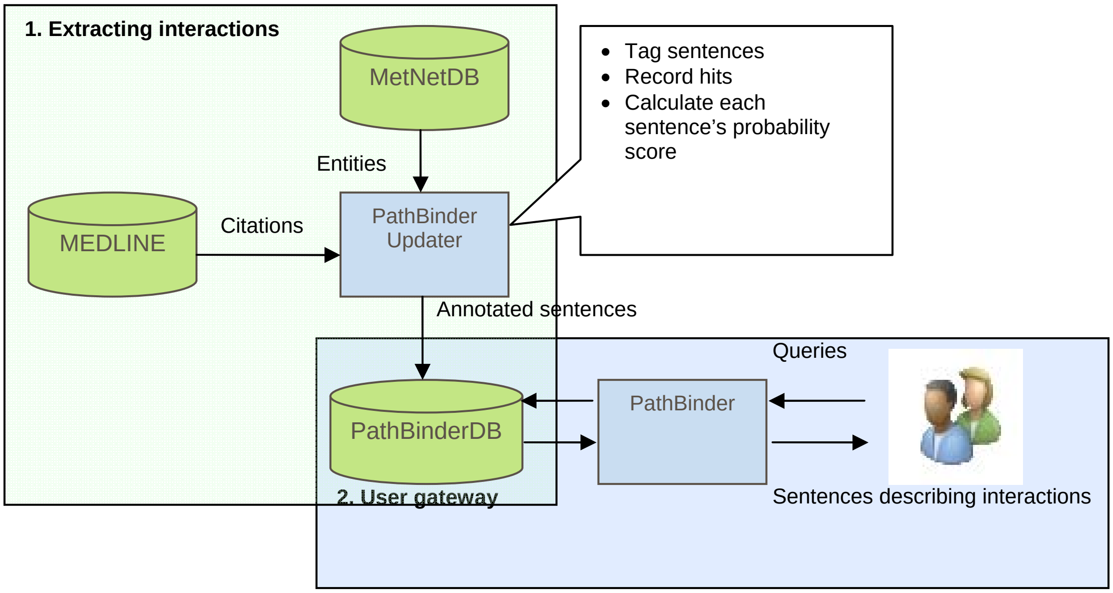
Figure 10. The interaction score computation procedure for the comprehensive method. (A sentence score for a biomolecular pair is defined as the probability that the sentence describes an interaction between these two biomolecules.)


Start
Repeat
No new sentence left
Retrieve next Sentence
Get a new sentence
Identify each pair of entities
Computer the sentence score for each pair
Does the pair exist in the PathBinder database?
No
Create a record for the pair with the current sentence's score in PathBinder database
Yes
Multiply the score in PathBinder database with the current sentence's score
Procedure computing the final score = 1- the existing score for each biomolecular pair
End

2. PathBinder is the user portal to PathBinderDB. PathBinder serves as a query gateway for interaction descriptions stored in PathBinderDB. Users can provide two biomolecules to PathBinder, which will access PathBinderDB and return all

sentences in which the two biomolecules appear. It calculates a probability score for each returned sentence, ranks them based on the score, and then shows them to the user. On the other hand, if a user provides just one biomolecule, PathBinder returns all biomolecules potentially interacting with it.

Figure 11. PathBinder system structure



2.2.3.3 Assessment of the interaction network

As mentioned in the method section, we can combine evidence from multiple sentences to evaluate the probability score that a pair of biomolecules interacts (see equation (1) and the "Best 2" and "Best 5" methods). The result is an interaction graph (network) composed of thousands of biomolecules and the interaction relationships among them. After scanning each sentence of MEDLINE and synonym processing, we collected 1,981,796 biomolecule co-occurrences from MEDLINE and then 7,706,968 pairs of biomolecules that may co-occur multiple times (for MEDLINE collection till October 2008). The key information retrieval measures of precision and recall were used to compare the three methods for combining evidence provided by multiple sentences. To access the recall and precision, we randomly sampled a set of 400 pairs of biomolecules co-occurring in MEDLINE from the previously generated interaction network. The set of sentences for each pair was evaluated by the three methods (All, Best 5 and Best 2), and the resulting computationally estimated probabilities of interaction were recorded for each pair. Meanwhile, these 400 pairs were manually analyzed to see whether they do in fact interact. One hundred eight of them did interact. The overall precision was thus 108/400=0.27 for this random set. More importantly, we calculated the precisions for 7 subsets of the 400 pairs meeting 7 different thresholds for interaction probability. This was done separately for each of the three methods (making 7*3=21 subsets). Thus each subset was associated with a threshold and a calculation method, and with a recall that was the fraction of the 108 interacting pairs meeting the threshold using the calculation method. For each set, the precision and recall are defined as following:

Precision = Number of pairs interacting having the score over the threshold in the set/ Number of pairs in the set

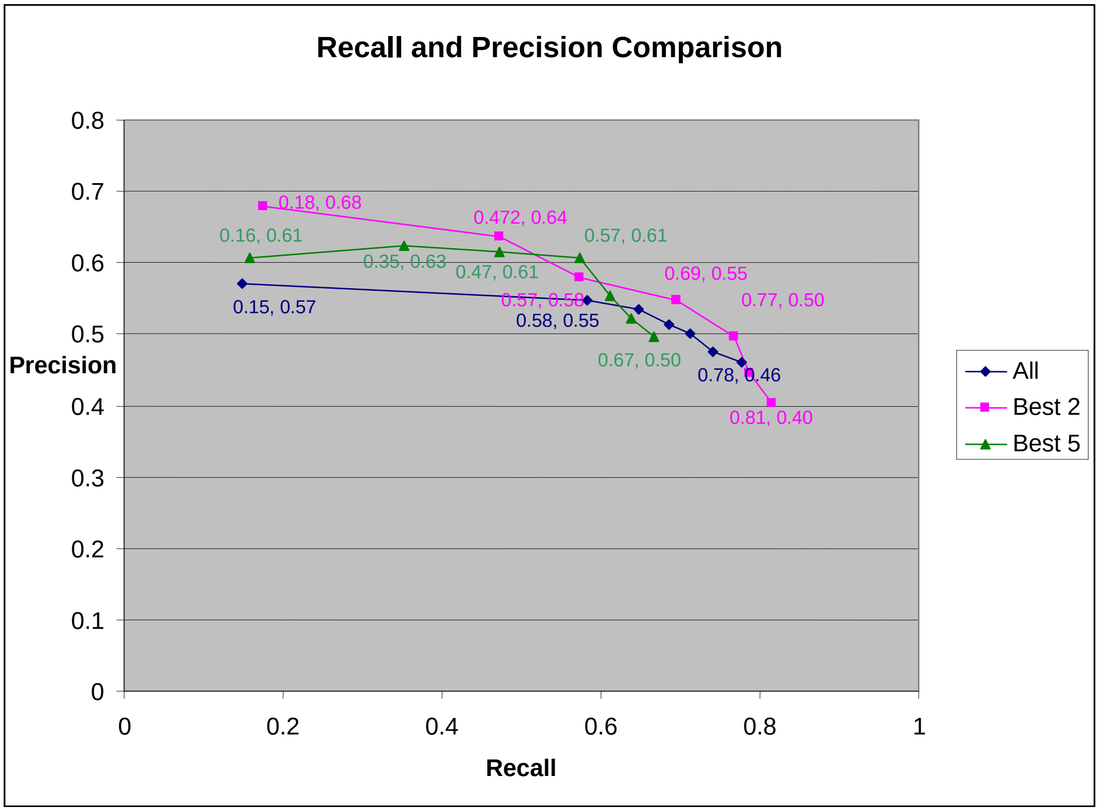
Recall = Number of pairs interacting having the score over the threshold in the set/ Number of pairs having the score over the threshold in the set

The overall recall for the whole set is necessarily 1. The detail results are shown in Table 20 and Figure 12.

Table 20. The recalls and precisions of different interaction network creation methods in different score threshold

All			**Best 2**			**Best 5**		
Score	**Recall**	**Precision**	**Score**	**Recall**	**Precision**	**Score**	**Recall**	**Precision**
1.00	0.15	0.57	**0.60**	0.18	0.68	**0.58**	0.16	0.61
0.95	0.58	0.55	**0.55**	0.47	0.64	**0.53**	0.35	0.62
0.90	0.65	0.53	**0.50**	0.57	0.58	**0.48**	0.47	0.61
0.85	0.69	0.51	**0.45**	0.69	0.55	**0.43**	0.57	0.61
0.80	0.71	0.50	**0.40**	0.77	0.50	**0.38**	0.61	0.55
0.75	0.74	0.48	**0.35**	0.79	0.45	**0.33**	0.64	0.52
0.70	0.78	0.46	**0.30**	0.81	0.40	**0.28**	0.67	0.50

Figure 12. Recall and precision comparison among three methods



Some other aspects of Figure 12 are worth considering further. For the “All” method, the leftmost data point occurs for a computed probability of 1 that a co-occurrence interacts. Such a high value can happen when there are a lot of sentences providing evidence. Combining that evidence using equation (5) leads to score values that are effectively 1 (for example, the co-occurrence of “bilirubin” and “cytochrome P450” including their synonyms was computed to have a score of $1\text{-}10^{-11}$). We considered any score over $1\text{-}10^{-6}$ to be effective 1. This occurred for 342,492 biomolecule pairs (for MEDLINE collection till October 2008). However, because the “Best 5” and “Best 2” methods only look at average scores of sentences, scores tend to be lower for these methods than for the “All” method. Thus, for these methods, it is possible to set score thresholds that are more selective such that the number of qualifying pairs is smaller than 342,492.

Another aspect of Figure 12 is that the curves are not always monotonic. For example, the first part of the “Best 5” curve is not monotonic. The leftmost point on that curve, (0.16, 0.61) is based on the 28 pairs meeting or exceeding a threshold score value of 0.58, computed by the “Best 5” method. This was the most selective threshold used to generate the curve. Yet the 62 pairs qualified by a lower threshold of 0.53 actually had a higher precision, giving point (0.35, 0.63) in Figure 12. One possible reason is noise from the limited data. A less likely possibility is that “Best 5” actually does produce this effect for some unknown reason.

Recall and precision are often combined to get a single, composite measure of information retrieval quality using the *F*-measure, or *effectiveness* of an information retrieval method:

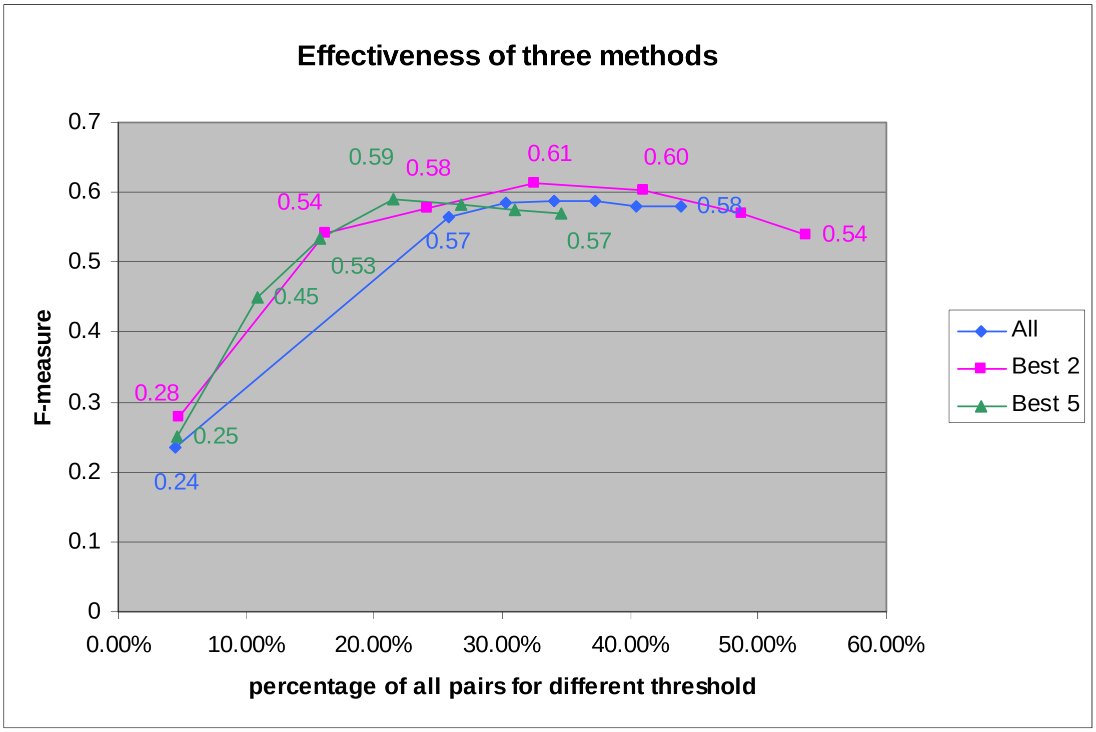
F = 2*(recall*precision)/(recall+precision).

Figure 13 and Table 21 shows the effectiveness for the three methods as a function of the size of set meeting a given threshold with the percentage of pairs of different thresholds in our final pairs set

Table 21. The effectiveness of different interaction network creation methods in different score threshold with data of the percentage of pairs of different thresholds in all pairs set.

All			Best 2			Best 5		
Score	**Effectiveness**	**Percentage**	**Score**	**Effectiveness**	**Percentage**	**Score**	**Effectiveness**	**Percentage**
1.00	0.24	4.44%	**0.60**	0.28	4.68%	**0.58**	0.25	4.60%
0.95	0.57	25.80%	**0.55**	0.54	16.18%	**0.53**	0.45	10.91%
0.90	0.59	30.32%	**0.50**	0.58	24.17%	**0.48**	0.53	15.78%
0.85	0.59	34.00%	**0.45**	0.61	32.47%	**0.43**	0.59	21.55%
0.80	0.59	37.29%	**0.40**	0.60	41.08%	**0.38**	0.58	26.77%
0.75	0.58	40.52%	**0.35**	0.57	48.65%	**0.33**	0.58	31.04%
0.70	0.58	43.93%	**0.30**	0.54	53.69%	**0.28**	0.57	34.65%

Figure 13. Effectiveness comparison among three methods



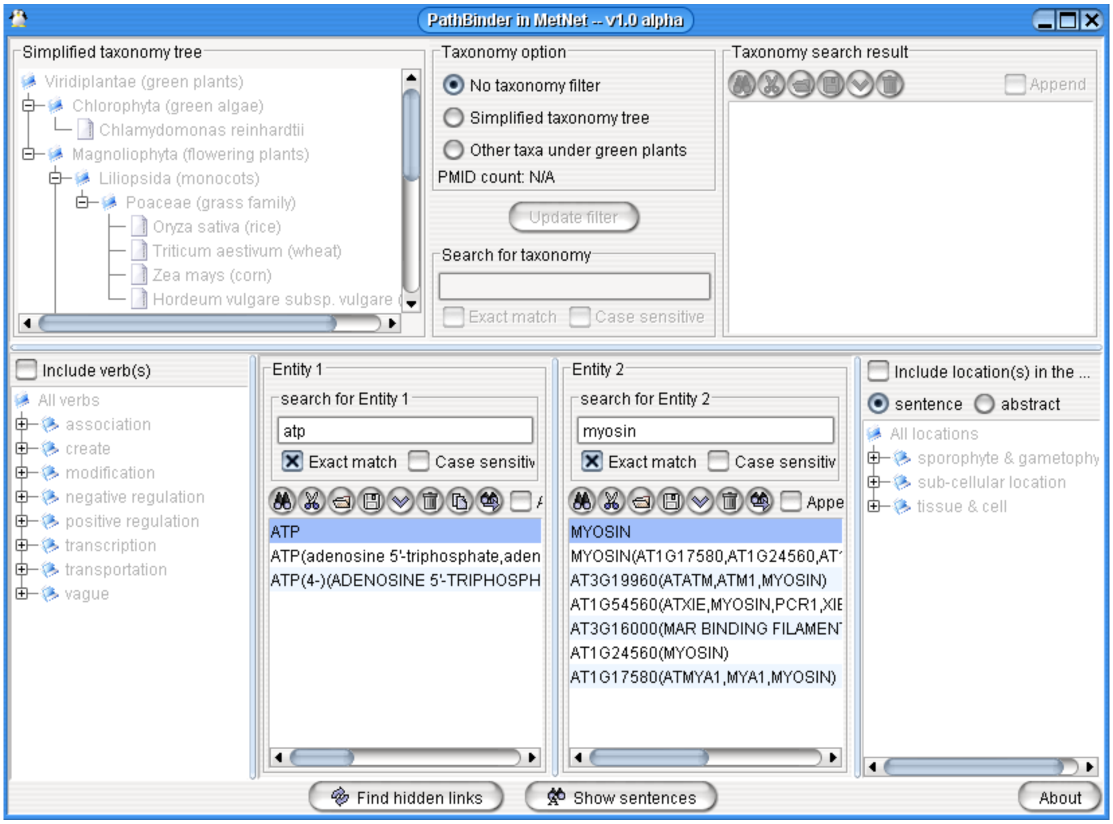
For *F* value, the "Best 2"method gave the highest peak value, for a threshold that qualifies 137 pairs. If we apply this threshold score to the full result interaction pair graph, there are 1,646,337 pairs that it qualifies. Therefore, if effectiveness is the key point, the "Best 2" method should be used. Otherwise, Figure 4 would be appropriate to use.

Our technique has been applied in the PathBinder System, which also provides a query gateway to users. If a user provides a biomolecule, PathBinder can find other biomolecules potentially interacting with it. Users can choose a biomolecule pair as a query for sentences describing interactions, as illustrated in Figure 14. Users also can attach more query conditions, such as cellular locations (e.g., nucleus, mitochondrion, etc.), categories of IITs appearing with the co-occurring biomolecule names (e.g. association, modification, etc.), specific IITs appearing with a co-occurrence (e.g. bind, increase, etc.) and Linnaean taxonomic categories. All these data are obtained when processing MEDLINE and are already recorded in the database. When PathBinder gets a query, it will search for all sentences satisfying the query and display them in a new window as in Figure 15. It can rank the result sentences by PMID or by their estimated probability of describing an interaction between the query biomolecule pair. Users can click the PMID to read the citation containing the sentence directly on the PUBMED Website.

2.2.3.4 Discussion of the interaction network result

As explained earlier, we calculate a rather precise probability estimate that a sentence describes an interaction between a given biomolecule pair. However, this precision can be misleading. There are some sentences that PathBinder mistakenly gives high scores to and some that get too low a score. For overly high scores, a typical problem is that an IIT describes the interaction of one biomolecule in the given pair with another biomolecule not in the pair, but the non-syntactic approach of PathBinder mistakenly concludes the interaction is between the co-occurrence of interest. For example, consider the sentence

Figure 14. PathBinder Main Interface



Simplified taxonomy tree

- Viridiplantae (green plants)
  - Chlorophyta (green algae)
    - Chlamydomonas reinhardtii
  - Magnoliophyta (flowering plants)
    - Liliopsida (monocots)
      - Poaceae (grass family)
        - Oryza sativa (rice)
        - Triticum aestivum (wheat)
        - Zea mays (corn)
        - Hordeum vulgare subsp. vulgare (

Taxonomy option

● No taxonomy filter
○ Simplified taxonomy tree
○ Other taxa under green plants

PMID count: N/A

Search for taxonomy

□ Exact match □ Case sensitive

Taxonomy search result

□ Include verb(s)

- All verbs
  - association
  - create
  - modification
  - negative regulation
  - positive regulation
  - transcription
  - transportation
  - vague

Entity 1

search for Entity 1: atp

☒ Exact match □ Case sensitiv

ATP
ATP(adenosine 5'-triphosphate,aden
ATP(4-)(ADENOSINE 5'-TRIPHOSPH

Entity 2

search for Entity 2: myosin

☒ Exact match □ Case sensitiv

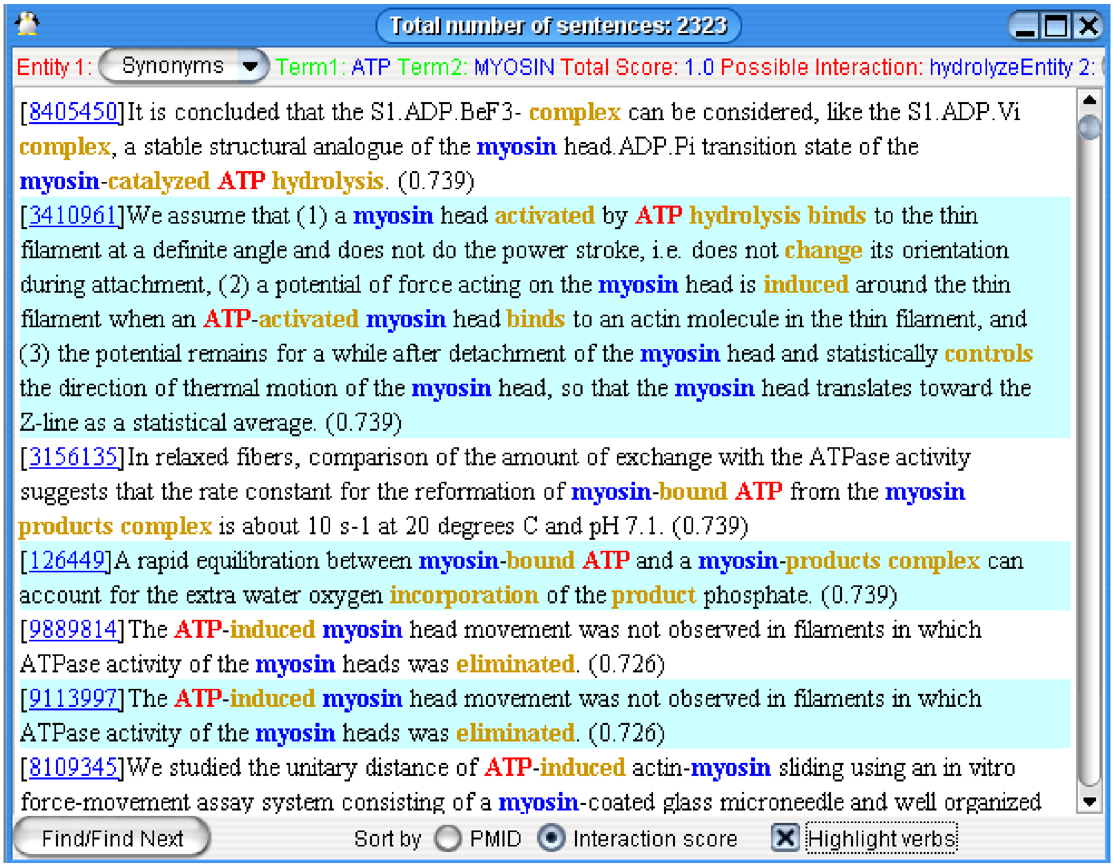
MYOSIN
MYOSIN(AT1G17580,AT1G24560,AT
AT3G19960(ATATM,ATM1,MYOSIN)
AT1G54560(ATXIE,MYOSIN,PCR1,XIE
AT3G16000(MAR BINDING FILAMEN
AT1G24560(MYOSIN)
AT1G17580(ATMYA1,MYA1,MYOSIN)

□ Include location(s) in the ...

● sentence ○ abstract

- All locations
  - sporophyte & gametophy
  - sub-cellular location
  - tissue & cell

Figure 15. PathBinder Search Results Windowbiomolecules pair with other



Entity 1: Synonyms ▾ Term1: ATP Term2: MYOSIN Total Score: 1.0 Possible Interaction: hydrolyzeEntity 2:

[8405450]It is concluded that the S1.ADP.BeF3- complex can be considered, like the S1.ADP.Vi complex, a stable structural analogue of the myosin head.ADP.Pi transition state of the myosin-catalyzed ATP hydrolysis. (0.739)

[3410961]We assume that (1) a myosin head activated by ATP hydrolysis binds to the thin filament at a definite angle and does not do the power stroke, i.e. does not change its orientation during attachment, (2) a potential of force acting on the myosin head is induced around the thin filament when an ATP-activated myosin head binds to an actin molecule in the thin filament, and (3) the potential remains for a while after detachment of the myosin head and statistically controls the direction of thermal motion of the myosin head, so that the myosin head translates toward the Z-line as a statistical average. (0.739)

[3156135]In relaxed fibers, comparison of the amount of exchange with the ATPase activity suggests that the rate constant for the reformation of myosin-bound ATP from the myosin products complex is about 10 s-1 at 20 degrees C and pH 7.1. (0.739)

[126449]A rapid equilibration between myosin-bound ATP and a myosin-products complex can account for the extra water oxygen incorporation of the product phosphate. (0.739)

[9889814]The ATP-induced myosin head movement was not observed in filaments in which ATPase activity of the myosin heads was eliminated. (0.726)

[9113997]The ATP-induced myosin head movement was not observed in filaments in which ATPase activity of the myosin heads was eliminated. (0.726)

[8109345]We studied the unitary distance of ATP-induced actin-myosin sliding using an in vitro force-movement assay system consisting of a myosin-coated glass microneedle and well organized

Sodium dichloroacetate increased ***glucose*** *oxidation and* ***pyruvate*** *oxidation in hearts from fed normal or alloxan-diabetic rats perfused with glucose and insulin.* (McAllister et al., 1973)

The term "oxidation" is between the biomolecules "glucose" and "pyruvate" but it does not describe an interaction between them. PathBinder, however, will give a high score to this sentence anyway. Analyzing the syntactic structure of the sentence, as with full parsing, would help solve this problem, but is computationally more expensive.

For scores that are too low, a typical problem is that some IITs are not recognized. An unusual IIT might not be stored in our database and so would not be recognized. For example, consider the following sentence.

GTP*-dependent twisting of* ***dynamin*** *implicates constriction and tension in membrane fission.* (Roux et al., 2006)

If we try to find an interaction between GTP and dynamin, there is no obvious IIT describing their interaction. But the word “dependent” describes a relation between “GTP” and “twisting of dynamin,” so there is indeed an interaction described. However, neither “dependent” nor “twist” are known to the system as IITs and so this sentence gets a low score.

Another problem is due to biomolecules appearing very often in MEDLINE. The chance that two of them co-occur in one sentence can be correspondingly elevated even if they do not interact. The sentence they co-occur in may not get a high estimated probability of describing an interaction. But when we construct the interaction network from MEDLINE based on thousands of sentences, the chance some pairs get listed with a high probability of interaction is high. An example is “ATP” and “starch.” A different problem in network construction is posed by biomolecules that look like common words in English. For example, “no” and the abbreviation of nitrous oxide have the same spelling and “no” appears very often in MEDLINE. A naïve analysis will find that nitrous oxide has interactions with thousands of biomolecules. In addition, some entities tend to creep into dictionaries of biomolecules that are not really biomolecules, like “resistance.” Such terms tend to then become members of invalid “interactions.” In fact, if we eliminate the effects of words like “no” and “resistance,” the precision of our results increases significantly, as shown in Table 22. The effectiveness also was improved with improved precision, as shown in Table 23 and Figure 16. Note that, because no new interacting pairs appear, the recall keeps same.

Our precision results are higher than for some other interaction extraction applications. For example, for a direct co-occurrence counting method, Albert et al. (2003) obtained a precision of about 35%. Our highest precision 95% is among the best results for extracting interactions so far. NLP methods in principle should be capable of obtaining close to 100% precision and recall. Avoiding NLP, however, our system saves considerable time. Our approach could be improved more by investigating and using empirics for more text features. Finally, we note that even when full NLP becomes available at some future time, easily computed text empirics will still have potential value as an ancillary evidence source that could speed up NLP-based analyses.

Table 22. The updated precision of different interaction network creation methods

All		Best 2		Best 5	
Score Threshold	**Precision**	**Score Threshold**	**Precision**	**Score Threshold**	**Precision**
1	0.84	0.6	0.95	0.58	0.89
0.95	0.74	0.55	0.84	0.53	0.90
0.9	0.71	0.5	0.8	0.48	0.86
0.85	0.67	0.45	0.73	0.43	0.83
0.8	0.65	0.4	0.64	0.38	0.74
0.75	0.63	0.35	0.57	0.33	0.69
0.7	0.6	0.3	0.51	0.28	0.65

Table 23. The updated effectiveness of different interaction network creation methods in different score threshold with data of the percentage of pairs of different thresholds in all pairs set.

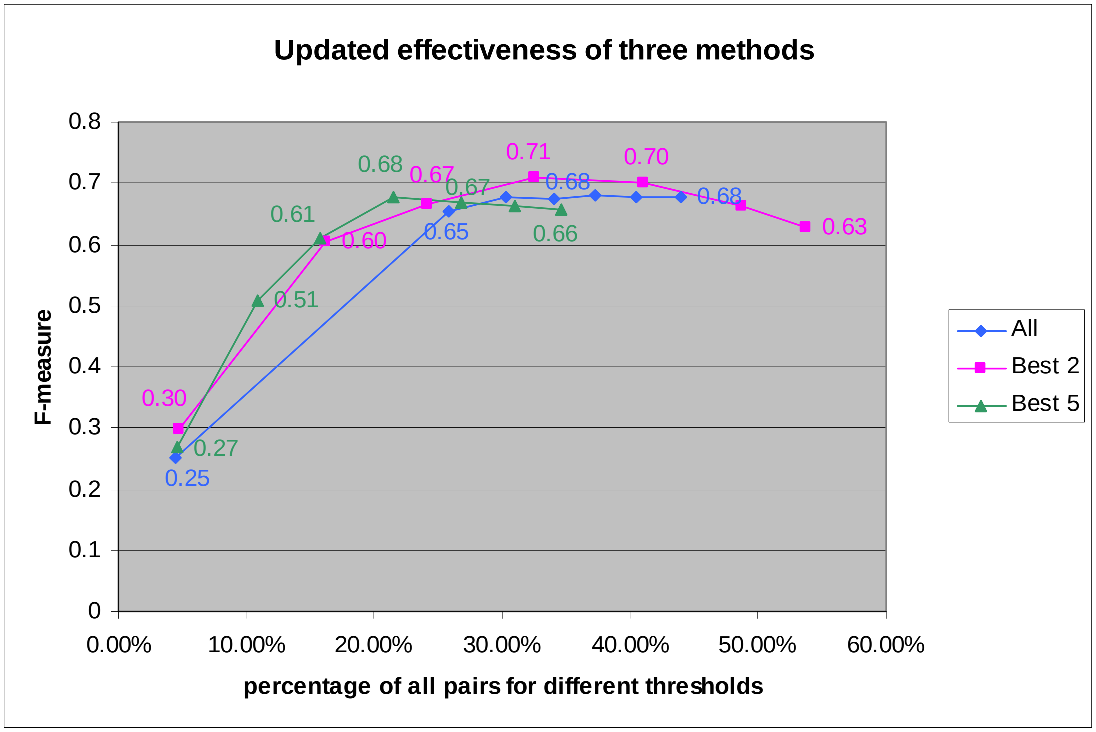
All			Best 2			Best 5		
Score	**Effectiveness**	**Percentage**	**Score**	**Effectiveness**	**Percentage**	**Score**	**Effectiveness**	**Percentage**
1.00	0.25	4.44%	**0.60**	0.30	4.68%	**0.58**	0.27	4.60%
0.95	0.65	25.80%	**0.55**	0.60	16.18%	**0.53**	0.51	10.91%
0.90	0.68	30.32%	**0.50**	0.67	24.17%	**0.48**	0.61	15.78%
0.85	0.68	34.00%	**0.45**	0.71	32.47%	**0.43**	0.68	21.55%
0.80	0.68	37.29%	**0.40**	0.70	41.08%	**0.38**	0.67	26.77%
0.75	0.68	40.52%	**0.35**	0.66	48.65%	**0.33**	0.66	31.04%
0.70	0.68	43.93%	**0.30**	0.63	53.69%	**0.28**	0.66	34.65%

2.2.4 Interaction-indicating terms extraction based on text empirics

As mentioned before, there is a lot of research about automatic interaction extraction. However, these works tend to focus on whether an interaction exists between two biomolecules than on the type of interaction. The type of interaction is important and also should be extracted. The co-occurrence and template matching methods may yield the type of interaction if an interaction triple like "A activates B" is extracted (Yakushiji et al., 2001; Ono et al. 2001; Domedel-Puig et al., 2005; Fundel et al., 2007). There is some research on classifying interactions within several predefined groups. Rosario et at. (2005) use an existing database of biomolecular interactions to train several probabilistic graphical models to classify sentences into several interaction groups. More research is about extract gene-disease relationship types. Theodosiou et al. (2005) used linear discriminant analysis to assign a gene a function for a disease from text. Bundschus et al.

(2007) also used conditional random fields, another probabilistic graphical model, to classify a sentence into several broad gene-disease interaction groups. Rindflesch et al. (2003) constructed a gene-disease relationship vocabulary and matched sentences to predefined rules to find relationship words and to classify them into groups. Yen et al. (2006) extracted gene-disease relationship directly by applying a bigram probabilities method and choosing those words with higher probabilities.

Figure 16. Updated effectiveness comparison among three methods



Here we introduce a method to automatically extract from literature interaction-indicating terms (IITs) that connect biomolecules. Similar to the method introduced above to extract interactions, we first sought empirical rules implicit in the biomedical texts and then applied this knowledge to design an algorithm to evaluate which IIT is most likely to correctly describe the interaction between a given pair of biomolecules. However, besides empirical rules based Bayesian method, our IIT extraction method also involves information retrieval theory to find the best IITs for two biomolecules based on all

MEDLINE collections. In contrast to the classification task, we instead extract actual interaction-indicating words rather than interaction categories. We also find the interaction description between two biomolecules based on multiple sentences identified within an entire corpus, rather than on a single sentence or other text unit. Our software, PathBinder, applies this method and provides query functions. Users not only can search for sentences describing interactions from MEDLINE by giving a pair of biomolecules based on the algorithm introduced in 3.2.2, but also a list of IITs associated with the given biomolecular pair and ranked based on their probabilities of properly describing the interaction.

2.2.4.1 Extraction method

To investigate automatic extraction of the correct IITs (inter-action-indicating terms) for biomolecule pairs of interest, we analyzed text units consisting of individual sentences (Ding et al. 2002). Titles were counted as sentences. For example, consider the following sentences.

Measurement of the reversibility of ***ATP binding*** *to* ***myosin*** *in calcium-****activated*** *skinned fibers from rabbit skeletal muscle.* (Bowater et al., 1989)

A parallel pathway model of ***regulation*** *simulated the effects of Ca(2+) and* ***ATP****-free* ***myosin binding*** *on both equilibrium* ***binding*** *of myosin-nucleotide* ***complexes*** *to actin and the general features of ATPase activity.* (Gafurov et al., 2004)

In rigor (in the absence of ***ATP****, when all the* ***myosin*** *heads are rigidly* ***bound*** *to the thin filament), a slight decay was observed in the first few microseconds, followed by no* ***change*** *in the anisotropy.* (Ramachandran et al., 1999)

These sentences contain possible interactions between ATP and myosin. These sentences also contain several verbs or other interaction-indicating terms that might describe the interaction between ATP and myosin. Their canonical forms are: “bind,” “activate,” "regulate," "complex," and "change." "Bind" appears more frequently than the others. Thus we might hypothesize "bind" as the interaction between ATP and myosin (careful reading shows this is indeed the case).

Note the distinction between an IIT that describes an interaction, and an IIT that describes an interaction in a given sentence. The latter refers to what a particular sentence says, while the former refers to a general fact about two biomolecules. Sentence 3 for example does *not* describe the interaction between ATP and myosin as "bind," even though the interaction is in fact one of binding.

Different text properties may play different roles in determining the probability that a particular IIT describes a biomolecular pair. Thus, we aimed to analyze sentences from the literature to manually identify useful properties that could help automatically extract correct IITs for a given biomolecular pair.

We used the same corpus used for interaction extraction introduced above. But other than focus on whether an interaction is described, we investigated manually the following features about the text and their relationship to the probability that an IIT is the correct IIT between two given biomolecules.

a) Whether a given IIT correctly describes the interaction between two given biomolecules.

b) The position of that IIT relative to the two biomolecules, including:

   i) whether the tri-occurrence of two biomolecules and the IIT is within a phrase or not,

   ii) how many words are between the biomolecules and the IIT,

   iii) whether the verb appears between the two molecules or somewhere else, and

   iv) the frequency of the verb's appearance in the sentence.

c) The differences among IITs properties like syntactic forms and semantic categories.

Each feature about an IIT in each sentence was recorded along with whether the IIT is the correct IIT between the pair of biomolecules specified by the query judged by biological experts. Their overall IIT distribution over sentences of ten biomolecules pairs is in Table 24.

Table 24. IIT distribution over biomolecules pairs

Biomolecules pair	a	b	c	d	e	f	g	h	i	k	Total
Total IITs appearing	113	87	112	65	79	110	93	24	20	67	770
Correct IITs for each pair	61	42	64	31	35	39	60	0	5	1	338
IITs describing interaction in sentences	33	31	30	20	23	15	25	0	4	1	182

In these 320 sentences, there are 770 times of IIT appearance and the correct IITs appear 338 times. The probability is less then 50%. We also differentiated all IIT appearances in two groups based on their position relative to the biomolecules as mentioned in a.i and a.iii above. The result is in Table 25. If an IIT appears between the two biomolecules, it has a higher probability to be the correct IIT describing interaction between the two biomolecules than not appearing between (50% vs. 39%) and if an IIT appears with the biomolecules pair together (tri-occur) in a phrase, it has a higher probability to be the correct IIT describing interaction between the two biomolecules than not appearing in a same phrase with the biomolecules than otherwise (50% vs. 37%).

Table 25. IIT position properties results

	Total	Between biomolecules	Not between biomolecules	Tri-occurring in a phrase	Tri-occurring not in a phrase
Total IIT	770	327	443	417	353
Correct IIT	338	164	174	209	129
percentage	44%	50%	39%	50%	37%

For the IIT properties themselves, we investigated the possibility that IIT form (noun, adjective, adverb, present, present continuous and past/perfect) and semantic category (association, modification, negative regulation, positive regulation, transportation, transcription, create, and vacuous) can be used as evidence to differentiate correct IITs from other IITs. “Vacuous” was used as the category when an IIT could not be clearly placed in another category (affected, influenced, etc.). The past and perfect forms of IITs are sometimes the same. The frequency of the perfect form is low, so we did not distinguish between them. Note that some IITs have the same spelling for both the noun and present tense forms. We can manually differentiate them but to use the results in automatic methods would require POS tagging. The results are shown in Table 26. We found that the correct IITs in the sentence distributes evenly among different pairs and among different forms, but not among different categories. The different IIT forms are significant to differentiate IITs for their probability to be correct IITs ($p<0.05$, F test). The IIT semantic categories in differentiating IITs with respect to probability of being correct IIT are closely related to the query biomolecules pair ($p_{pair}<0.05$, $p_{category}>0.05$, two combined F tests), which means IIT category may not be appropriate to help extract correct IIT.

Table 26. Data on likelihoods that IITs are correct, by interaction-indicating term form and category.

Forms	**# (%) correct IITs**	**All IITs**
noun	190/54%	353
adj	10/43%	23
adv	0/0%	0
present	23/25%	92
-ing	42/52%	81
past/perfect	67/32%	210
Categories		
association	82/74%	111
modification	110/71%	154
negative regulation	1/1%	104
positive regulation	5/4%	132
transportation	1/4%	23
transcription	0/0%	7
create	101/78%	130
vacuous	39/37%	105

We also investigated the influence of the distance between an IIT and the biomolecules pair. There are two distance varieties: one is the number of words between the IIT and the nearest biomolecule in the pair and another is the number of words between the IIT and another biomolecule in the pair (other than the biomolecule closer to the IIT). The sample data is in Table 27 and Table 28. The detailed information for each pair and each distance can be found in Appendix VII. To better understand the properties of the nearest biomolecule, we analyzed sentences with distances from the IIT to the nearest biomolecule of 38 intervening words or less. Similarly, to better understand the other biomolecule in the pair, we analyzed sentences where this distance was also 38 or less. Tables 27 and 28 appear to corroborate the hypothesis that the likelihood that an IIT is correct decreases as the distances become larger. Thus we hypothesized that the relationship between the probability and the distance can be represented by this equation:

$$p(\text{an IIT is the correct IIT}) = b_0 * e^{-b_1 * dis\tan ce} \qquad (6)$$

where b_0 and b_1 are determined from regression analyses on the data synopsized in Tables 25 and 26.

Table 27. Data on likelihoods that IITs are correct, by the different distances to the nearest biomolecule in the pair.

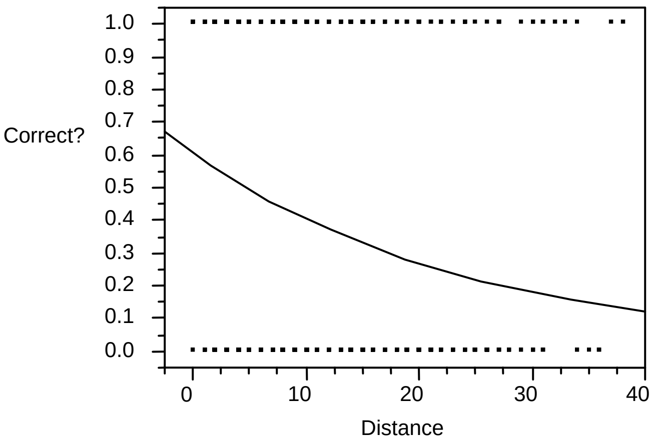
Distance to nearest biomolecule	#(%) correct IITs	All IITs
0	17/89%	19
1	23/70%	33
2	42/67%	63
3	42/65%	65
4	29/37%	78
5	26/33%	80
6	22/41%	54
…	…	…
38	0/0%	1

Table 28. Data on likelihoods that IITs are correct, by the different nearest distances to the biomolecule other than the one in the pair closer to the IIT.

Distance to nearest biomolecule	#(%) correct IITs	All IITs
0	191/63%	302
1	76/45%	168
2	38/38%	99
3	42/40%	106
4	21/29%	73
5	17/30%	57
6	4/11%	38
…	…	…
38	0/0%	1

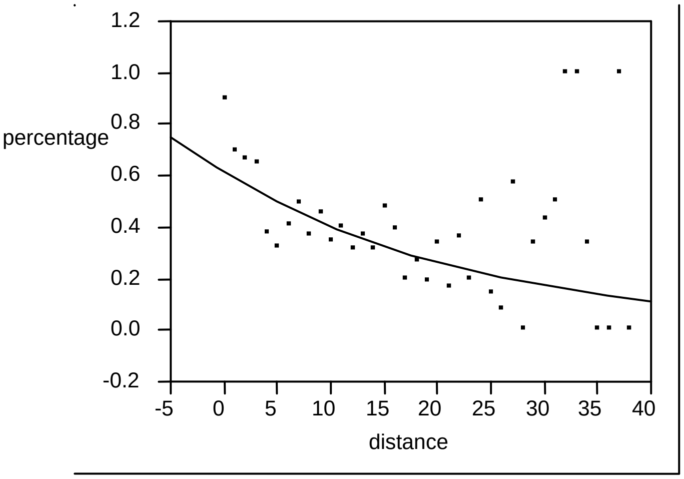
We did the regression analysis based on the data whether an IIT is the correct IIT related the two distances. The regression analysis is implemented in JMP software based on data in Tables 27 and 28 and also based on pure 1/0 (whether an IIT is the correct IIT) for different distance. The results are shown in Figures 17 and 18 for farther distances and Figure 19 and 20 for closer distances. Note that the y axis in the graph is the correct IIT probability with value only 0 or 1 because it is for each appearance of an IIT which is correct IIT or not in Figures 17 and 19. In fact they match the regression results based on the calculated percentage data above in Tables 27 and 28 very well in Figures 18 and 20. We chose the result regression curves in Figures 17 and 19 with smaller errors and relative relationship equations are equation (7) for the farther distance and (8) for the closer distance.

Figure 17. Regression curve for the relationship between the likelihood that an IIT is correct and the distance of the IIT to the *other* biomolecule in the pair (not the nearest). Each sentence is a separate data point in the graph, but many data points are superposed.



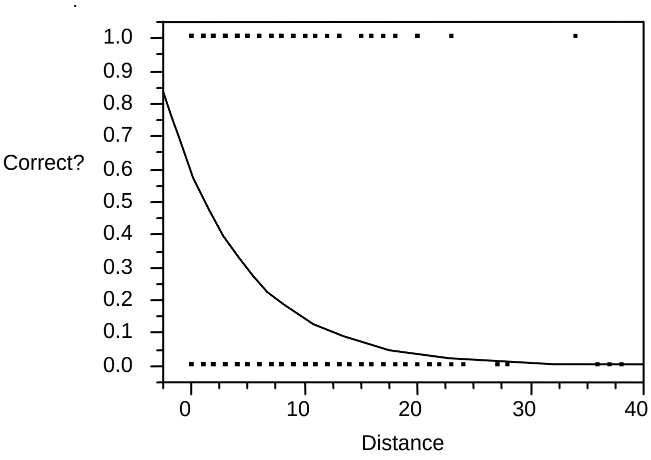
Parameter	Estimate	ApproxStdErr	Lower CL	Upper CL
b0	0.605186767	0.03461204	0.5339632	0.68251498
b1	-0.04073855	0.00655469	-0.0563802	-0.0268338

Figure 18. Regression curve for the relationship between the likelihood that an IIT is correct and the distance of the IIT to the *other* biomolecule in the pair (not the nearest). Each data point represents the set of sentences for a given distance.



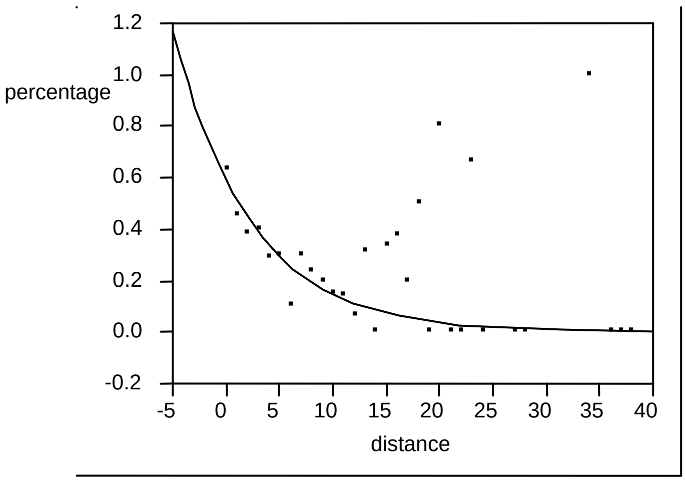
Parameter	Estimate	ApproxStdErr	Lower CL	Upper CL
b0	0.608727371	0.04712929	0.51078631	0.71874031
b1	-0.041680757	0.00893949	-0.0641114	-0.0226901

Figure 19. Regression curve for the relationship between the likelihood that an IIT is correct and the distance of the IIT to the nearest biomolecule. Each sentence is a separate data point in the graph, but many data points are superposed.



Parameter	Estimate	ApproxStdErr	Lower CL	Upper CL
b0	0.5882890188	0.02323824	0.54002426	0.63737176
b1	-0.14058278	0.01832478	-0.1885659	-0.1005924

Figure 20. Regression curve about the relationship between the likelihood that IITs are correct and the distance of the IIT to the nearest biomolecule. Each data point represents the set of sentences for a given distance.



Parameter	Estimate	ApproxStdErr	Lower CL	Upper CL
b0	0.5869431534	0.03088942	0.40161	0.77228
b1	-0.138346172	0.02414969	-0.2832	0.00655

$$p(\text{an IIT is the correct IIT}) = 0.605 * e^{-0.04*\mathit{dis\tan ce}} \qquad (7)$$

$$p(\text{an IIT is the correct IIT}) = 0.588 * e^{-0.14*\mathit{dis\tan ce}} \qquad (8)$$

Combining evidence and identifying the kind of interaction between two biomolecules

We combined the evidence contributed by the properties above to give a weight that each sentence contributes about the probability that an IIT is the correct IIT as the interaction extraction method introduced above by the same method used for interaction extraction (3.2.2.2). This way, for each IIT in a given sentence, we can calculate the chance that it correctly describes the interaction of the biomolecule pair of interest. To best determine the interaction between two biomolecules, however, requires looking at sets of sentences, rather than one sentence in isolation. With sets, in attempting to extract the type of interaction between two biomolecules, we can combine evidence provided by multiple sentences in the literature containing a given IIT in association with a biomolecular pair of interest.

If an IIT co-appears with the two biomolecules more frequently, this IIT might be expected to have a correspondingly higher probability of being a correct IIT for the biomolecular pair compared to other IITs. Confounding this, however, is the fact that the background frequencies with which different IITs appear in the literature differ. Some commonly appearing IITs may appear more frequently in association with a given biomolecule pair than the correct interaction-indicating term for that pair merely because they are so common overall. Thus misleading results could occur if we are not careful.

To correct for varied background frequencies of IITs, we applied the *tf-idf* weighting system, which is often used in information retrieval. The hypothesis for this phase of the study was at follows.

HYPOTHESIS: Using the properties of tri-occurrences discussed in Section 1.1, interaction-indicating terms characterizing the interaction between a given pair of interacting biomolecules may be extracted using a method that:

1. is an adaption of the tf-idf method;

2. makes intuitive sense for the same reasons as the tf-idf method; and

3. works.

The remainder of this subsection addresses this hypothesis. The tf-idf method is explained in most information retrieval textbooks. The term frequency (*tf*) of a word *i* in a document is:

$$tf_i = \frac{n_i}{\sum_k n_k} \qquad (9)$$

where n_i is the number of occurrences of the term *i* in the given text, and the denominator is the number of occurrences of all terms in the text.

The inverse document frequency (*idf*) measures the ability of a term to separate relevant from irrelevant texts:

$$idf_i = \log \frac{|D|}{|\{d : i \in d\}|} \qquad (10)$$

where |*D*| is the total number of documents or other texts in a corpus, and the denominator is the number of texts in which term *i* appears.

Thus a high value of *tf*idf* occurs if the term has a high frequency in a given document but a low frequency in the whole corpus, suggesting that if the term is a search term, it is correspondingly powerful for discriminating between relevant and non-relevant texts.

The *tf-idf* method can be applied to the tasks of finding the IIT(s) that are most relevant to a given biomolecule pair, and hypothesizing that these IIT(s) correctly describe the interaction.

Adaptation 1

Recall that *tf-idf* weighting often is used in information retrieval to score and rank a document's relevance to a given user query, when the query is composed of several key words. Here we want to retrieve the right interaction-indicating words for a query consisting of a biomolecule pair. We can look at the specific biomolecule pair as a query, and all interaction-indicating term candidates as documents. To do this, we can look at the set of all sentences in our corpus for which a specific interaction-indicating term appears, as a “document,” and then an occurrence of a word in a document becomes an occurrence of a biomolecule pair in one of the sentences in which a specific interaction-indicating term appears.

Then the calculation of *tf* becomes:

$$tf_i = \frac{s_i}{\sum_k s_k} \qquad (11)$$

Where s_i is the number of sentences where the specific pair of biomolecules appears with an interaction-indicating term, and s_k is the number of sentences where any pair of biomolecules appears with the interaction-indicating term.

The calculation of *idf* becomes:

$$idf_i = \log \frac{V}{\{v : b_i \in v\}} \qquad (12)$$

Where v is the number of different “documents” (each consisting of all sentences containing a particular interaction-indicating term) that a specific biomolecule pair b_i appears in, and V is the total number of different interaction-indicating terms in our corpus. Then combining eq. (11) and eq. (12) we can use this *tf-idf* strategy to rank interaction-indicating terms for a given biomolecule pair.

However, eq. (11) is not easy to calculate. $\sum_k s_k$ includes all pairs occurring with interaction-indicating term *i*, and the pairs could include millions of possible combinations of biomolecules. Calculating each interaction term's set of sentences equates to scanning the whole corpus.

To be simpler, we can assume that $\sum_k s_k = \rho s$, where *S* is the number of sentences in which the interaction-indicating term of interest appears, and ρ is the proportion of those containing some biomolecule pair. Because we need to compare *tf-idf* among different interaction-indicating terms, if we assume ρ is the same for all interaction-indicating terms, we can modify *tf* to *mtf*:

$$mtf_i = \frac{s_i}{S} . \qquad (13)$$

Using this we can simplify the calculation and still be able to compare interaction-indicating terms. However the assumption that ρ is the same for all interaction-indicating terms is only an assumption. We need to establish its validity to be confident.

Adaptation II

Recall that *tf-idf* is used to evaluate how important a word is to a document in a collection or corpus. In the same way, we can evaluate how important an interaction-indicating term is to a biomolecule pair. Because an interaction-indicating term cannot appear literally in a biomolecule pair, let's instead look at the sentences containing a given biomolecule pair as the "document" and the interaction-indicating terms as the terms of the *tf-idf* definition.

From the text empirics discussion earlier, we know that each occurrence of an IIT can be given a weight reflecting its likelihood of correctly describing the interaction of a given biomolecular pair in a sentence. This weight is determined from the sources of evidence described above, combined using the odds method (3.2.2.2). Thus, we can weight each IIT occurrence in the set of sentences and sum the weights, instead of merely counting the number of occurrences of an IIT in the set. This is even more reasonable because the *tf-idf* method presumes that instances of a keyword suggest relevance, so by calculating a weight for each occurrence we are building on that presumption. Thus we can restate the *tf* term of the *tf-idf* method as

$$tf_i = \frac{\sum_j \; p_{i,j}}{\sum_{k,j} p_{k,j}} \qquad (14)$$

where i is a specific IIT, k is some IIT, j is some sentence in the set containing the biomolecular pair of interest, and $p_{i,j}$ and $p_{k,j}$ are the likelihoods that IIT i and j respectively reflect the interaction between the biomolecular pair of interest in sentence j. The more often IIT i appears with the two biomolecules and the higher the weight (likelihood) associated with each appearance, the higher the value tf_i. This satisfies clauses 1 & 2 of the hypothesis for the *tf* term.

For inverse document frequency, based on the model underlying the *tf* term just described, the classical *idf* expression may be restated as

$$idf_i = \log \frac{B}{\left|\{b : i\ with\ b\}\right|} \qquad (15)$$

where B is the total number of biomolecular pairs in our corpus, and i *with* b is true if biomolecular pair b tri-occurs in at least one sentence with IIT i but false otherwise. We adjust this expression to account for biomolecular pairs that co-occur, but without ever tri-occurring in the same sentence with any IIT. No straightforward, sentence-based computation could extract a proper IIT to such a pair, even it they did interact. Therefore such pairs are outside the scope of the work presented here, and thus were removed from consideration. The expression for *idf* then becomes

$$idf_i = \log \frac{\{b : \exists t\ \ with\ b\}}{\{b : t_i\ \ with\ b\}} \qquad (16)$$

where b is a biomolecular pair, t can be any IIT and i is a specified IIT. Eq. (16) requires finding the number of biomolecule pairs that an IIT appears with, which is tedious. Therefore in order to facilitate computation, equation (16) was approximated as

$$idf_i = \log \frac{|\{b : \exists t\ \ with\ b\}|}{|\{b : t_i\ \ with\ b\}|} \approx \log \frac{|\{s : \exists t \in s, \exists b \in s\}|}{|\{s : t_i \in s, \exists b \in s\}|} \qquad (17)$$

where *s* is a sentence.

Based on eq. (14), and similar in meaning to the classical *idf* expression (eq. 10), eq. (17) satisfies clauses 1 & 2 of the hypothesis for the *idf* term. If we have identified all sentences with tri-occurrences, and all sentences with tri-occurrences of which a given IIT is a member, we can calculate *tf***idf* for each IIT occurrence with a given biomolecule pair. The IIT with the highest *tf-idf* value among all candidates would then be ranked as the term most associated with the biomolecule pair and thus, if the hypothesis holds, would be the correct IIT.

With this approach and its variations, if we are given a biomolecule pair, we can extract all possible IITs describing interaction between the pair based on a literature. As mentioned above, we use MEDLINE, an 18 million citation collection, as our literature base. Through the text empirical analysis above and biologists' suggestions, we constructed a list of 125 IITs in different semantic categories and their 558 variations. Then we can scan each sentence in MEDLINE against the biomolecules and IITs in our database to evaluate each sentence and then use the *tf-idf* based method to combine evidences from all sentences in MEDLINE to extract the correct IITs.

The whole procedure is implemented in our software, PathBinder. In the query part, users can provide two biomolecules to PathBinder, which will access PathBinderDB and return all sentences in which the two biomolecules appear. Our system will automatically find all synonyms of the input biomolecules based on information in MetNetDB, and the sentences containing those synonyms are also returned together. IIT extraction is based on sentences returned by PathBinder. When scanning each sentence, the software counts each tri-occurrence in the sentence and then records or updates the information about the number of sentences containing tri-occurrence used in our method. PathBinder then calculates a weight score for each IIT appearing in those sentences (i.e. in MEDLINE), ranks them based on the calculated weight score, and then shows them to the user. For the present study, PathBinder was used to gather data for an evaluation of clause 3 of the hypothesis that the information collected can be used to extract the IITs describing the relationship between bio-molecule pairs of interest.

2.2.4.2 Results and analysis

We have stored into a database more than 30 million sentences, extracted by PathBinder from MEDLINE, in which at least one biomolecule in our lexicon appears. There are more than 8 million sentences containing at least one tri-occurrence of a biomolecule pair and an IIT.

To evaluate our IIT extraction approach, we randomly chose 200 pairs of biomolecules that co-occurred in those sentences. Of these, biologists judged that 106 pairs of biomolecules occurring with IIT(s) interact. We processed these 106 pairs with PathBinder, which returned a ranked list of potential IITs for each pair. For example, for the biomolecule pair "chlordecone" and "cytochromes P-450," PathBinder returned the ranked IIT list shown in Table 29.

Table 29. List of IITs tri-occurring with biomolecule pair chlordecone and cytochromes P-450, ranked by likelihood of correctly describing their interaction as calculated using text empirics.

induce	affect
change	control
potentiate	produce
reduce	decrease
regulate	bind
increase	lower
alter	metabolize
amplify	derive

While this list contains 16 IITs, of which "induce," "regulate" and "increase" are correct, up to 87 IITs were returned for each pair. On the other hand, some pairs had only one IIT in their lists. We manually investigated each returned IIT list and noted the correct IITs for its corresponding biomolecule pair. Figure 21 shows that 90% of the pairs appeared in at least one sentence with a correct IIT in our IIT lexicon, and how good the rankings of the lists were when they did. On the other hand, 10% did not tri-occur with a correct IIT from the lexicon. The following sentence is an example.

We identified the terpene synthase ***TPS10*** *that* ***forms*** *(E)-beta-farnesene, (E)-alpha-bergamotene, and other herbivory-****induced*** *sesquiterpene hydrocarbons from the substrate* ***farnesyl diphosphate****.* (Schnee et al., 2006)

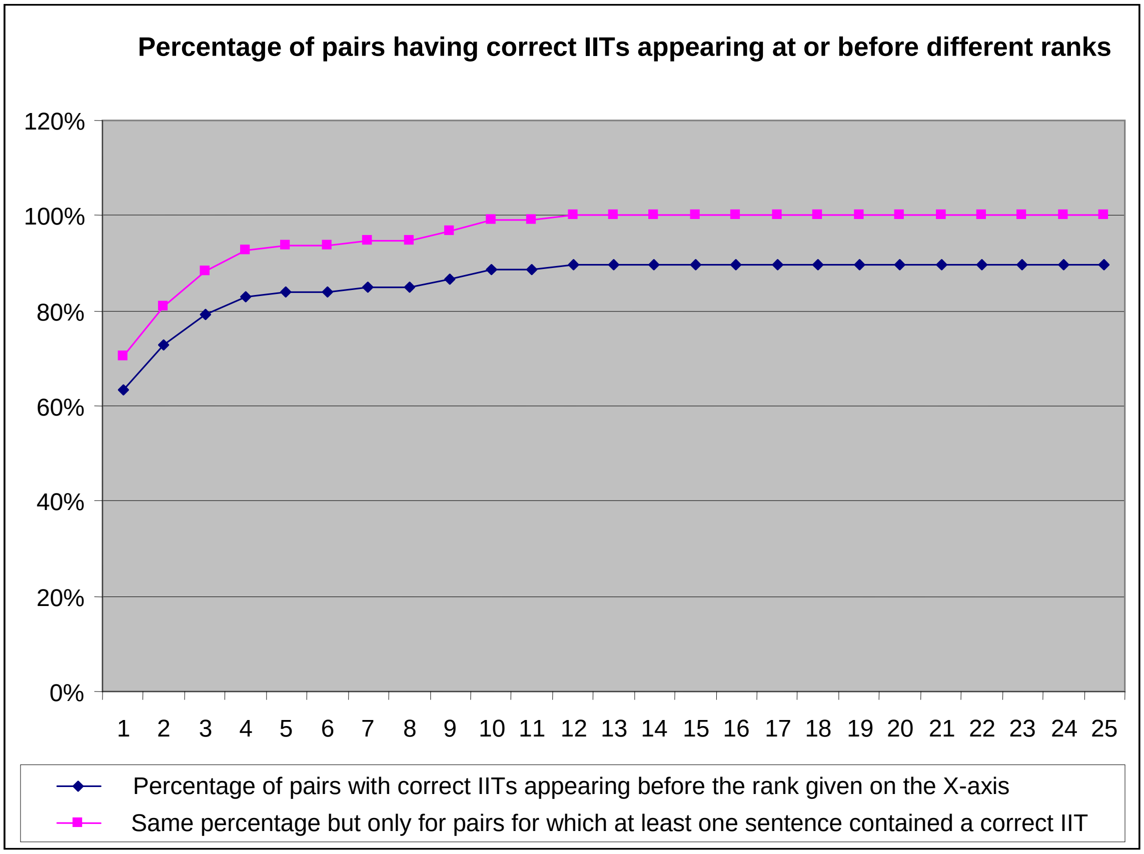
In this sentence, TPS10 forms sesquiterpene molecules from farnesyl diphosphate. Thus clearly there is an interaction between them. However, there is no IIT in this sentence indicating this interaction that our system can find because "form" is an ambiguous word and for that reason not in our IIT lexicon.

Of the 106 pairs, 67 (63%) ranked a correct IIT first in the associated retuned IIT list. Eleven pairs did not have any correct IIT appearing in the sentences in which they co-occur. Thus for the remaining 95 pairs, 71% of them had a correct IIT returned first in the IT list. Figure 21 shows that when the correct IIT appears, it always ranked 12 or better. In more detail, a correct IIT was returned first 71% of the time, first or second 80% of the time, in the top three 88% of the time, and in the top four 93% of the time. In other words, more than 90% of the time a correct IIT was in the first four IITs returned when it was present at all. Thus the results show that our system tends to rank correct IITs above incorrect ones.

It should be noted that some biomolecule pairs have only one correct IIT in their IIT result list, and of these, some pairs only have one in their lists, which is the correct IIT. On the other hand, some pairs have multiple correct IITs in their IIT results lists. For each pair of biomolecules, we determined the information retrieval metrics of recall and precision as follows. Suppose for a given biomolecule pair, N IITs are returned, of which C are correct. Let c_n, n = 1…N, be the number of correct IITs among the n top-ranked IITs, and let the IIT recall rates $r(n)$ of a given IIT list be:

$$r(n) = c_n/C. \tag{18}$$

Figure 21. Percentage of correct IITs present in the top *n* ranks.



Similarly, let the IIT precisions $p(n)$ for a give IIT list be:

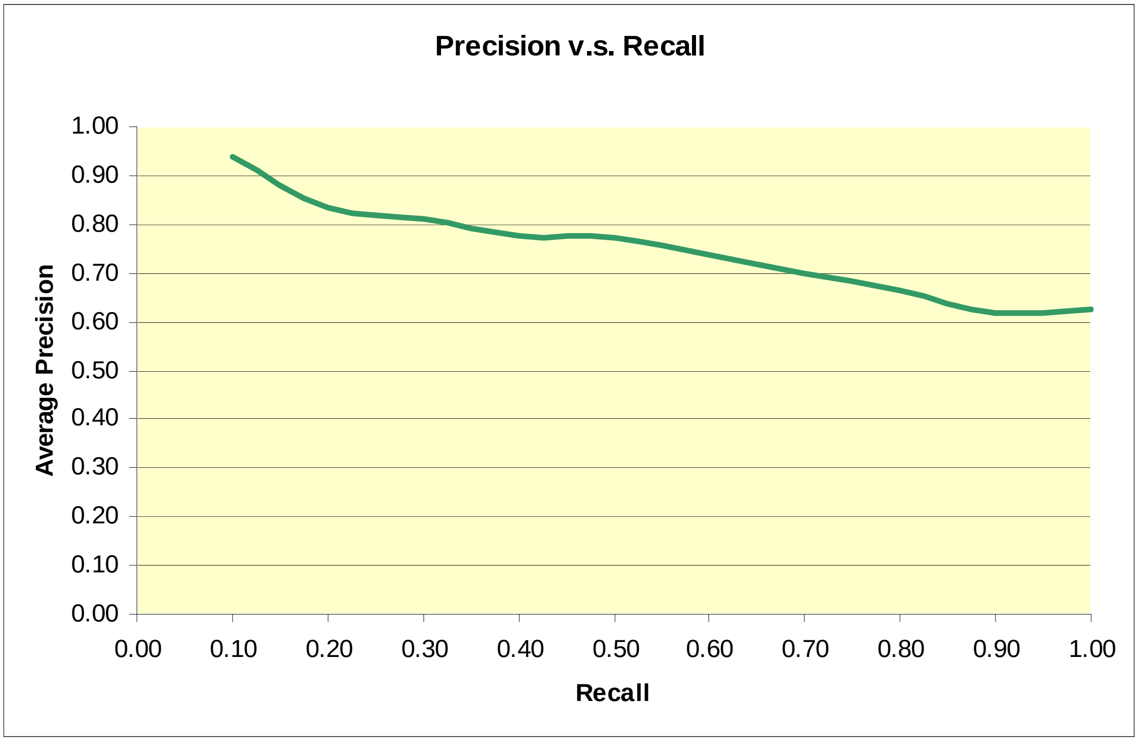
$$p(n) = c_n/N. \qquad (19)$$

Every value of n has associated IIT recall and precision values $r(n)$ and $p(n)$. We can therefore plot recall against precision for each IIT list. The mean precision, plotted vs. recall, is shown in Figure 22. We can see the precision is generally higher than 0.6 no matter how much the recall is.

The discussion thus far has not considered cases in which the interaction between two biomolecules is *not* described by a single IIT in the sentence. However, such cases occur. For example consider the following sentence.

S5. ***Glutathione peroxidase*** *(Se-GPx) is a selenoenzyme which* ***catalyzes*** *the* ***reduction*** *of hydroperoxides by* ***glutathione*** *(GSH), in most mammalian cells.* (Chaudière, 1986)

Figure22. Average precision vs. Recall of results.



The biomolecule pairs of interest in these two sentences are “glutathione peroxidase” and “glutathione.” The interaction between these biomolecules, as described in these sentences, is not named by a single IIT. Instead, the sentences follow the patterns “A catalyzes the reduction of B by C” and “A catalyzes the conjugation of C to B.” Regarding interactions between biomolecules A and C, these sentences imply that such relationships exist, but do not describe them explicitly and directly using an IIT.

For example, in this sentence, A catalyses a reduction process and C is involved in this process. We can infer that A causes oxidation of C, but because the sentence does not say this directly, it would be hard to design an algorithm to identify this interaction. For our purposes, if we want to know whether A and C interact or not, this sentence is evidence they do. But if we want to determine through software what the interaction is, then this sentence is likely to mislead the algorithm, for neither “catalyze” nor “reduce” truly describes the interaction: the relationship between A and C is not one of either catalysis or reduction. Therefore, in this sentence we do not count "catalyze" and "reduce" as correct IITs.

To be precise, we could say the relation between A and C in this sentence is "catalyze reduction to." However, here we have aimed to extract a single accurate IIT describing the interaction between two biomolecules, not a phrase or combination of several IITs. This relatively complex situation occurs because the interaction in the sentence is actually between a biomolecule and a process.

Alternatively, one might consider sentences such as that sentence as indicating a relation among three biomolecules. Then when searching for the interaction between A and C, the third term B would be extracted for the user, in addition to the two IITs. In fact, there are a series of IITs that can (but do not necessarily) act like "catalyze" in these examples: "inhibit," "stimulate," and so on. Like more typical IITs, they can appear early in a ranked result list. Indeed, they are helpful to biologists for identifying interactions between biomolecules. If we also count those words as correct in our 106-pair test corpus results, the percentage of pairs whose IIT lists have a correct IIT ranked first increased from 60% to 63% for all 106 pairs and from 68% to 71% for pairs with at least one correct IIT in the results list.

Some IITs indicate a relationship between two biomolecules, but not what kind. For example, in this sentence,

*Geminivirus **AL2** and L2 proteins **interact** with and inactivate SNF1 **kinase**.* (Hao et al., 2003)

the IIT "interact" does describe that there is an interaction between AL2 and kinase, but unfortunately, it does not say what kind of interaction. Other such IITs include "affect," "influence," and so on. We classify these IITs as vacuous.

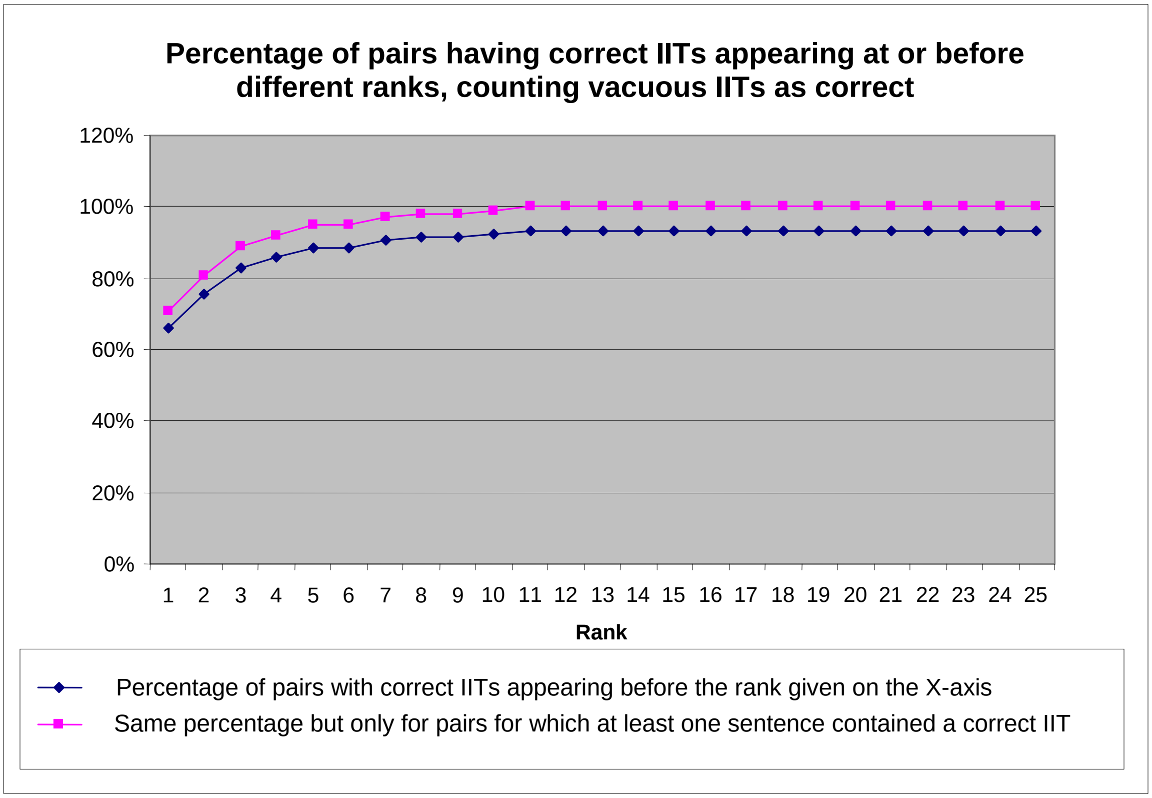
For the purpose of extracting correct IITs, we cannot deny the correctness of vacuous IITs. Therefore every vacuous IIT appearing in the results list should be counted as correct if the purpose is to find all correct IITs without regard for usefulness. However, vacuous IITs are not very useful because they are not specific about the type of interaction between the given pair of biomolecules. Nevertheless, because a vacuous IIT is correct, we may wish to count it as correct even it does not describe the interaction between the biomolecule pair of interest. For example, consider the sentence

While Zn2+ was capable of ***inhibiting*** *all the enzymes except the H+-ATPase, AlCl3 and Al-****citrate*** *had minimal* ***effects*** *except for with* **phospholipase A2** *where an* ***interaction*** *with AlCl3 occurred.* (Jones et al., 1997)

There are two vacuous IITs: effects and interaction. They do not describe an interaction between “citrate” and “phospholipase A2” in the sentence. But from other sentences, we know that “citrate” can inhibit (hence have an effect on and interact with) “phospholipase A2”, so “effect” and “interact” are counted as correct IITs. For this reason, if we count all vacuous IITs appearing in the sentences with given biomolecular pairs, the analysis result is slightly different from Figure 21, as shown in Figure 23. Compared to Figure 21, Figure 23 shows that when vacuous IITs are counted as correct, more correct IITs are found at earlier ranks in the IIT result lists (66% vs. 63% for rank 1, 75% vs. 73% for rank 2, 83% vs. 79% for rank 3, etc.). But for the subset of pairs that excludes pairs without at least one sentence containing a correct IIT, there is little difference. In this new analysis, the additional correct IITs add more data points so that Figure 25 has higher density than Figure 24, though they otherwise appear somewhat similar.

However, counting vacuous IITs as correct contradicts our original intent of investigating methods for extracting correct IITs, because these vacuous IITs often are not useful for biologists. Yet, so far we have treated them like other IITs, resulting in Figures 21 and 22. Alternatively, we can delete vacuous IITs from the results lists, in which case we get Figures 26 and 27. Removal of vacuous IITs makes fewer correct IITs appear in earlier ranks. However, for the subset that excludes pairs for which no sentences contain a correct IIT, the percentages of pairs having correct IITs appearing in the top three ranks are a little higher (72%, 82%, and 90%).

Figure 23. Percentage of correct IITs present in the top *n* ranks, counting vacuous IITs.



For incorrect IITs in the result lists, a typical situation is that the IIT describes the interaction of one biomolecule in the given pair with another biomolecule not in the pair, or it is involved in 3-way interaction as in the "*Xanthine oxidoreductase*" sentence examples above. Moreover, for 3-way interaction descriptions, IITs involved can get high rankings so our non-syntactic approach can mistakenly conclude that such an IIT is the correct IIT for the pair, like "reduce" in the example sentence about "gutathione peroxidase" and "glutathione" above. Analyzing the syntactic structure of the sentence, as with full parsing, may help solve this problem as it would help solve the direction of interaction ("reduced by" and "conjugation of"), but the extraction of an interaction between a biomolecule and another interaction is ultimately needed.

Figure 24. Precision vs. recall of results.

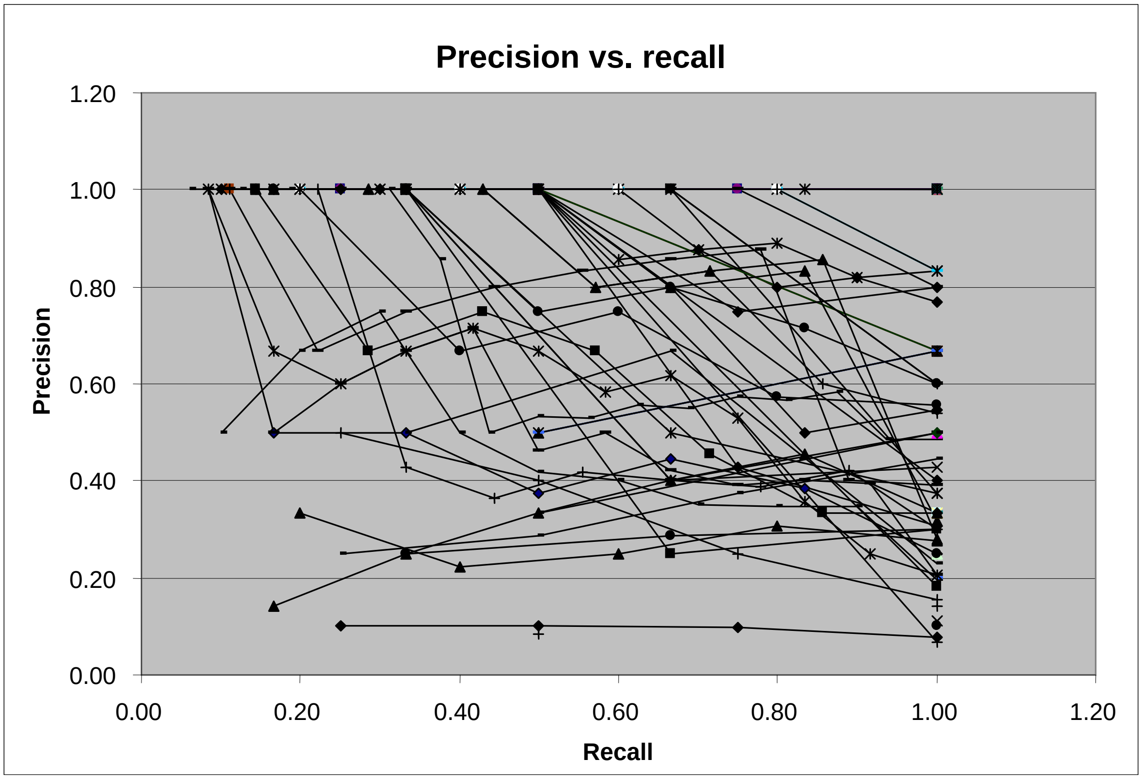


Figure 25. Precision vs. recall of results counting all vacuous IITs

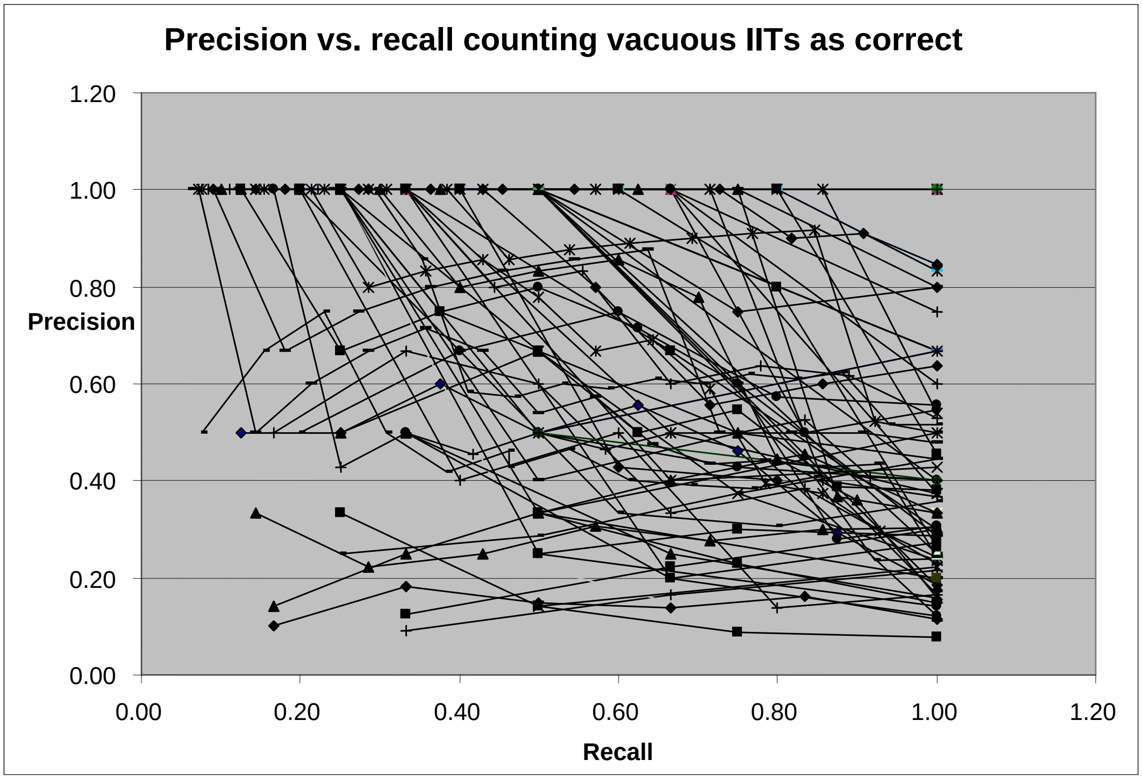


Figure 26. Percentage of correct IITs present in the top *n* ranks with vacuous IITs deleted from results lists.

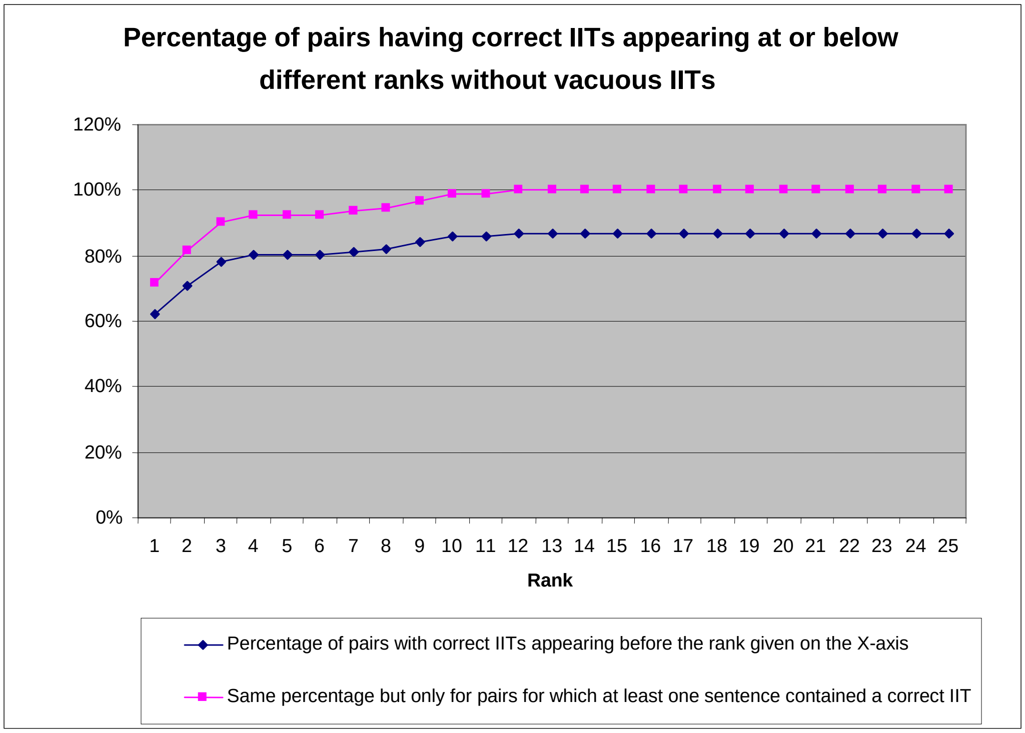
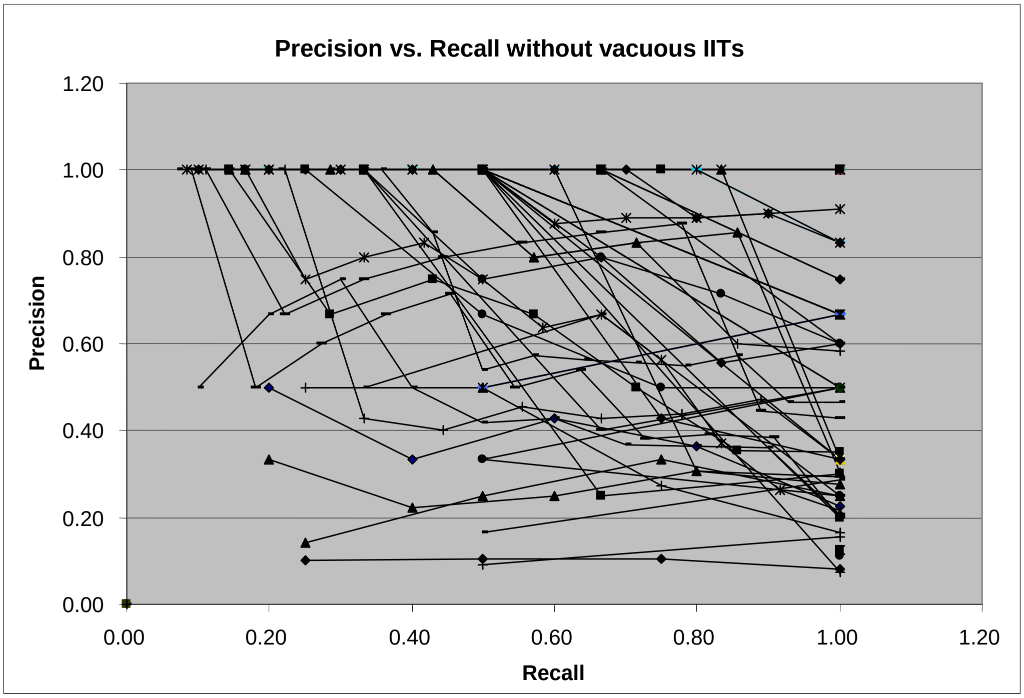


Figure 27. Precision vs. recall of results with vacuous IITs deleted from results lists.



There are some pairs in our results for which no correct IITs are in the returned list. A typical problem is that the correct IITs were not recognized as IITs at all. An unusual IIT might not be stored in our database and so would not be recognized. For example, consider the following sentence.

In the wild type, ***PGR5-dependent PSI*** *cyclic electron transport competed with NADP(+) photoreduction.* (Okegawa et al., 2008)

If we try to find an interaction between PGR5 and PSI, there is no obvious IIT describing their interaction. But the word “dependent” describes a relation between“PGR5” and “PSI cyclic electron transport,” so there is indeed an interaction described. However, “dependent” and its root word “depend” are not yet recognized by our system, which results in no correct IIT returned for PGR5 and PSI. Besides unrecognized IITs, there are also some unrecognized variants of otherwise recognized IITs. For example,

Geminivirus ***AL2*** *and L2 proteins interact with and* ***inactivate SNF1 kinase***. (Hao et al., 2003)

In this sentence, “inactivate” was not extracted because it is not in the IIT lexicon in our system, although “activate” is. A more comprehensive IIT lexicon could solve this problem.

2.3 Conclusion and future research

2.3.1 Conclusion and discussion

I developed a new method for extracting biomolecular interactions from natural text, in particular from scientific article abstracts in MEDLINE. To do this, I first investigated manually how different syntactic and semantic features influence biomolecular interactions and summarized the resulting empirical data about it. This data is public and anyone can use it. Based on this set of empirical data, we designed a naïve Bayes based algorithm to extract sentences that are likely to describe interactions between given biomolecules from the MEDLINE collection of scientific articles. Combining evidences from sentences across all of MEDLINE, I created a biomolecular interaction network for MetnetDB. Besides pure interaction pair extraction, I also performed research on how to extract the interaction indicating terms associated with a specific interacting biomolecular pair in order to describe the interaction between the pair. I also manually collected text empirics data for the IIT extraction task. We used the naïve Bayes method to utilize this empirical data to evaluate the weight contributed by each sentence in MEDLINE and then combined the weights of sets of sentences, using the tf-idf metric to find the best IITs for any biomolecular pair.

The research is implemented in the PathBinder software, which provides a search interface for users to input a biomolecules pair and other feature constraints to search interaction descriptions from MEDLINE. PathBinder is part of MetNet software suite. The MetNet bioinformatics platform is a suite of software applications designed for analysis of genomic, proteomic, transcriptomic, and metabolomic experimental data. MetNet applications use innovative visualization, statistical, and graphing techniques to help users analyze metabolic and regulatory networks.

For interaction description extraction, we tried a naïve Bayes based method to give an interaction likelihood score for each sentence and rank them based on this score. We randomly chose several hundred sentences from MEDLINE for different randomly chosen biomolecular pairs and PathBinder calculated interaction scores for them. Meanwhile we manually judged each sentence. Then we did linear regression between the PathBinder scores and the manual judges, and found that the likelihood score reflects the proportion of sentences describing an interaction for different score ranges correctly. Our linear regression result is very close to the line $y=x$.

There are different ways to extract interactions from natural text. Simple counting of co-occurrences of two biomolecules, or of tri-occurrences of two biomolecules and one interaction indicating term, is the first method we tested. It can get a high recall since it covers most interaction descriptions. However, many of the cases it finds are not real interaction descriptions so that the method usually has low precision. Thus most systems that use this type of approach have involved manual post-processing to improve precision. Our method used rules behind samples of real natural texts and applied these rules on the co-occurrences and tri-occurrences to extract interactions. In addition, our likelihood score ranking policy returns ranked co-occurrences and tri-occurrences with higher precision and without decreasing recall.

The method I used compares manual judgments of whether sentences describe interactions or not. The judgments are used to rate the results provided by rules based text parsing method. Depending on the construction of these rules, the recalls and precisions of are different. Looser rules have more comprehensive results (high recall) but with more mistakes (low precision). On the contrary, constricting rules extract fewer results but usually have higher precision. This helps explain that MedScan (Daraselia et al., 2003) obtained a recall of 21% and a precision of 90% with relatively restrictive templates, while Koike et al. (2005) achieved 54% recall with relative simple rules. In addition, to improve the precision of extracted interactions, some methods have used full parsing procedures to implement detailed sentence syntactic structure analysis. In principle, full parsing can bring more precision. However current NLP technology cannot solve ambiguous parsing results well, and these often occur. Therefore the precision of full parsing cannot be assured presently. In addition, parsing text, especially full parsing, has a higher computation resource requirement. In comparison, our method has lower resource requirements, due to the light text parsing. We use a ranking policy to return all possible sentences describing interactions but in order of interaction description likelihood score. In this way, we can assure a higher recall but keep high precision for those sentences with high scores. This is the same ranked results philosophy used by standard Web search engines.

Besides parsing and matching methods, applying statistical analysis on parsed result is another way to extract interactions. Statistics based machine learning has become popular for many applications. Instead of manually predefined rules, computer can use different methods, e.g. SVM, CRF, and BN, to learn the rules behind training corpora and then use these rules to extract interactions. However, the biggest problem of machine learning is that the rules are related to a specific algorithm and usually are not very readable for users. Also the rules are not very applicable to other applications. We manually analyzed samples from MEDLINE, and the data used for the rules (interaction description probabilities for different features) are direct, simple, and easy to be reused by other method and researchers.

For interaction extraction, we not only developed a method for finding sentences describing interactions between biomolecule pairs, but also a method for finding a specific term describing the interactions between a biomolecule pair. In other words, we can both extract interacting pairs and also the interactions between them. There are relatively few reports on extracting interaction terms or interaction types for interacting pairs in the literature. For most of them, instead of extracting interacting pairs in texts they try to extract triples consisting of an interacting pair of biomolecules plus a verb, where the verb describes the interaction between the pair of biomolecules. Here we not only extract sentences where a triple appears but combine evidence provided by multiple sentences together based on statistics and information retrieval to extract the right interaction indicating terms for pair of biomolecules. Our evaluation results showed that our average precision is always over 60% no matter how high the recall is in our ranked IIT lists.

We have investigated a number of syntactic and semantic features about sentences and interaction indicating terms, and then uses a Naïve Bayes related method to combine them. However only a limited number of features were investigated. In general, the more features we analyze, the more precise the extracted interactions may be expected to be. Besides more features to analyze, there are more directions we can continue research on. The following sections provide more details on these directions for future research.

2.3.2 Interactions involving several interaction-indicating terms

As mentioned for IIT extraction, all of the methods above assume the interaction can be expressed by an interaction-indicating term. However, some interactions are described by several IITs and even several biomolecules. Currently we count them as interaction descriptions and some IITs involved as correct IITs. For example:

Xanthine oxidoreductase **catalyse***s the anaerobic* **reduction** *of glyceryl trinitrate (GTN), isosorbide dinitrate and isosorbide mononitrate to inorganic* **Nitrite** *using xanthine or NADH as* **reducing** *substrates.* (Doel et al., 2001)

Starch*, d-glucose, salinomycin and monensin* **inhibited** *the* **production** *of skatole and indole from Trp, and skatole from* **indoleacetic acid** *by rumen bacteria.* (Mohammed et al., 2001)

In the first sentence, the interaction can be described as “A catalyses reduction of B to C.” For the biomolecule pair A and C, there is no single IIT to describe their interaction. We could say the relation between A and C is “catalyze reduction to” or “stimulate production of,” but in the corpus analysis we cannot record these kinds of interaction terms because they are combinations of terms.

To be precise, this situation should not treat these several interaction-indicating terms like interaction terms in simple situations such as “A binds B.” In addition, when several interaction-indicating terms appear together in one sentence, the probability that any given one is the correct interaction term should be lower than if there is only one interaction-indicating term appearing in the sentence. Therefore, to adjust the probability, if there are *n* interaction-indicating terms in one sentence, each interaction term’s probability used to give an overall evaluation in equation (11) will be divided by *n*. Then equation (14) becomes:

$$tf_i = \frac{\sum_j p_{i,j} / n_j}{\sum_{k,j} p_{k,j} / n_j} \tag{18}$$

where $p_{m,j}$ is the probability that the interaction-indicating term *m* gets from sentence *j* in the “document” consisting of sentences containing the given biomolecule pair, and n_j is the times of occurrence of all interaction-indicating terms in the sentence *j*.

However, to divide the probability by the number of occurrences of interaction-indicating terms in the sentence is only a proposal. Its validity needs to be verified in the future. Analyzing the text by hand and finding the empirical results is the only precise way to find the influence of multiple IITs on the probability that a correct interaction-indicating term appears. For a corpus containing sentences where different numbers of interaction-indicating terms occur, we could divide them into groups based on the numbers of interaction-indicating terms in each sentence. Then the proportion of sentences containing the correct interaction-indicating terms in each group could be determined. Based on the results, appropriate statistical methods could be used to find the relation between the likelihood of correct IIT and the number of IITs. Finally the empirical results could be used in section 2.2.4.1 or in equation (18). The properties dealt with in this empirical analysis are different from those we are analyzing in this research. In the future, it will be useful to conduct experiments to help with handling of the situation in which several interaction-indicating terms appear together in conjunction with simple situations like “A binds B.”

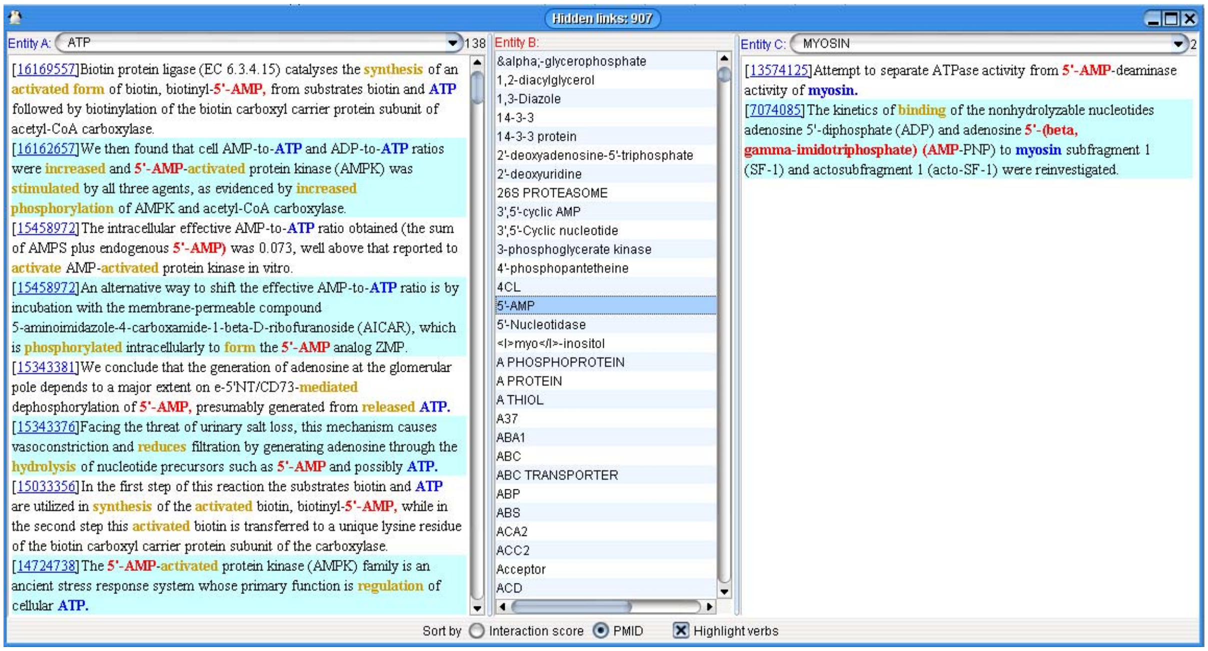
Actually, this relative complex situation occurs because the interaction is actually between a biomolecule and a process. If, in the future, we can identify such interactions, then we may look at processes as actors in interactions and the problem would be solved. Alternatively, we can look at the situation as a relation among three biomolecules. Then when searching for the interaction between A and C, the third one B also would be extracted for the user in addition to the two interaction-indicating terms.

2.3.2 Hidden interaction extraction and integration into interaction networks

Most current interaction extraction research focuses on direct interactions, which are based on sentences containing two biomolecules (such as names A and C) as evidence for the interaction of these two biomolecules. However, sometimes these two biomolecules may not appear in the same sentence or abstract, but they may be connected by another biomolecule (call it B) that appears in the same sentence or abstract with each of them even though A and C might never appear together. Then we can say A and C have an implied interaction described by the sentences describing A and B and the sentences describing B and C. The module to find the hidden links between A and C has been implemented in PathBinder (Figure 28) and is named the “ABC” module. Given two biological entities, the ABC module finds a list of B terms. When users choose any word in that list, PathBinder will show the sentences containing A and the selected B, and the sentences containing the selected B and C. The terms A, B and C come from PathBinder’s dictionary.

As mentioned in the previous section, the interaction evidence assessment can be computed from all sentences containing both terms. The ABC module can give a complementary way to assess the evidence for interaction using indirect co-occurrence. The integration of the ABC module into the interaction computation would give a more comprehensive assessment of the evidence for an interaction.

Figure 28. ABC module in PathBinder.



Entity A: ATP 138

[16169557]Biotin protein ligase (EC 6.3.4.15) catalyses the synthesis of an activated form of biotin, biotinyl-5'-AMP, from substrates biotin and ATP followed by biotinylation of the biotin carboxyl carrier protein subunit of acetyl-CoA carboxylase.

[16162657]We then found that cell AMP-to-ATP and ADP-to-ATP ratios were increased and 5'-AMP-activated protein kinase (AMPK) was stimulated by all three agents, as evidenced by increased phosphorylation of AMPK and acetyl-CoA carboxylase.

[15458972]The intracellular effective AMP-to-ATP ratio obtained (the sum of AMPS plus endogenous 5'-AMP) was 0.073, well above that reported to activate AMP-activated protein kinase in vitro.

[15458972]An alternative way to shift the effective AMP-to-ATP ratio is by incubation with the membrane-permeable compound 5-aminoimidazole-4-carboxamide-1-beta-D-ribofuranoside (AICAR), which is phosphorylated intracellularly to form the 5'-AMP analog ZMP.

[15343381]We conclude that the generation of adenosine at the glomerular pole depends to a major extent on e-5'NT/CD73-mediated dephosphorylation of 5'-AMP, presumably generated from released ATP.

[15343376]Facing the threat of urinary salt loss, this mechanism causes vasoconstriction and reduces filtration by generating adenosine through the hydrolysis of nucleotide precursors such as 5'-AMP and possibly ATP.

[15033356]In the first step of this reaction the substrates biotin and ATP are utilized in synthesis of the activated biotin, biotinyl-5'-AMP, while in the second step this activated biotin is transferred to a unique lysine residue of the biotin carboxyl carrier protein subunit of the carboxylase.

[14724738]The 5'-AMP-activated protein kinase (AMPK) family is an ancient stress response system whose primary function is regulation of cellular ATP.

Entity B:

&alpha;-glycerophosphate
1,2-diacylglycerol
1,3-Diazole
14-3-3
14-3-3 protein
2'-deoxyadenosine-5'-triphosphate
2'-deoxyuridine
26S PROTEASOME
3',5'-cyclic AMP
3',5'-Cyclic nucleotide
3-phosphoglycerate kinase
4'-phosphopantetheine
4CL
5'-AMP
5'-Nucleotidase
<I>myo</I>-inositol
A PHOSPHOPROTEIN
A PROTEIN
A THIOL
A37
ABA1
ABC
ABC TRANSPORTER
ABP
ABS
ACA2
ACC2
Acceptor
ACD

Entity C: MYOSIN 2

[13574125]Attempt to separate ATPase activity from 5'-AMP-deaminase activity of myosin.

[7074085]The kinetics of binding of the nonhydrolyzable nucleotides adenosine 5'-diphosphate (ADP) and adenosine 5'-(beta, gamma-imidotriphosphate) (AMP-PNP) to myosin subfragment 1 (SF-1) and actosubfragment 1 (acto-SF-1) were reinvestigated.

Sort by ○ Interaction score ● PMID ☒ Highlight verbs

The plan for this integration includes two parts. The first part is for co-occurrence search results. There will be a way to show not only actual co-occurrences but also indirect co-occurrences. The display of sentences should not simply assume indirect co-occurrences give less evidence of interaction than direct co-occurrences. We cannot tell whether a direct or an indirect co-occurrence has a higher probability to describe an interaction without computing the probabilities. A method to give the probability that an indirect co-occurrence describes an interaction needs to be specified. The factors should include those factors that compute the probability P(AB) that A and B interact, the probability P(BC) that B and C interact, and a way to combine them into a composite probability P(AC):

$$P(AC) = P(AB) * P(BC) \qquad (16)$$

Incorporating evidence for interactions provided by the ABC module, the interaction database will be changed.

- The interaction score of existing pairs will be increased by accounting for the additional indirect interaction evidence.
- New interaction pairs may be found through the indirect co-occurrences. These pairs need to be added into the interaction database.

However, several issues arise:

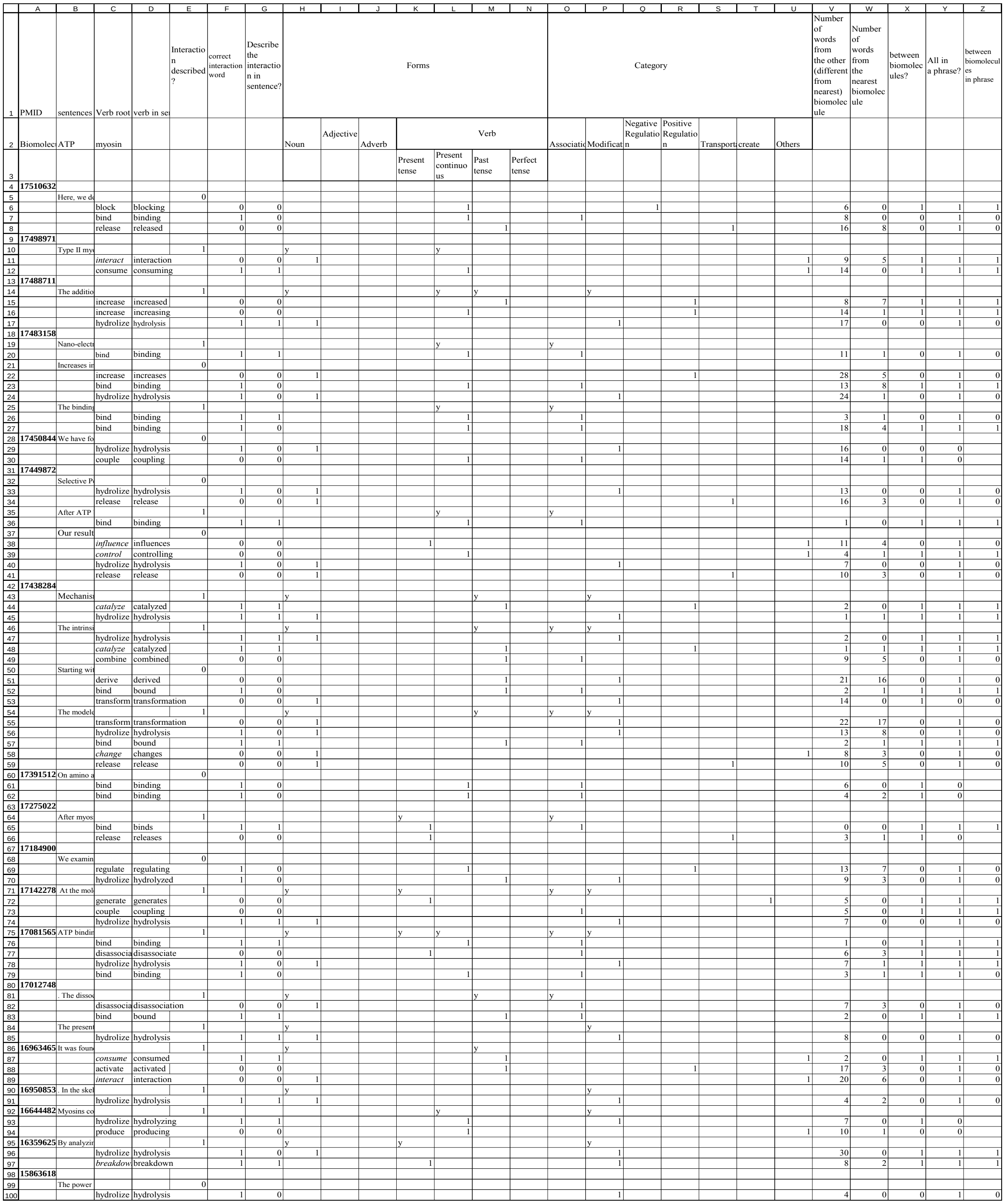
- When computing the interaction evidence score for biomolecules A and C, the scores for interaction evidence between A and B, and between B and C, are important. But the AB score could be influenced by not only sentences containing A and B, but also sentences containing A and D, and other sentences containing D and B. Which kind of score for AB should be used?
- To find indirect evidence of interaction, how many links should be used in the computation? For example, A-BCD-E indicates an interaction between A and E. In fact, limiting ourselves to 3-node paths like ABC cannot assure accuracy because evidence for interaction between A and C due to longer paths, like ABDC, is not accounted for.
- Should we differentiate the scores for direct interaction evidence and indirect interaction evidence when computing a final probability score for interaction of a pair?

To solve these problems, we can make some requirements for applying the ABC model. To compute the interaction score between A and C using an implicit relation through B, the score of AB and BC must come from the direct interaction relation in literature and only allow one intermediate biomolecule to compute the implicit relation score. To combine the direct and implicit score, we can give different weight to different scores and then test the results to see which weight is better.

When these problems are resolved, the interaction database will be more comprehensive because some new interactions may be found from the literature without ever being explicitly stated in the literature.

Appendix I Text empirics data sample

Text empirics data sample table “ATP” & “Myosin”



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	PMID	sentences	Verb root	verb in se	Interaction described?	correct interaction word	Describe the interaction in sentence?	Forms							Category							Number of words from the other (different from nearest) biomolecule	Number of words from the nearest biomolecule	between biomolecules?	All in a phrase?	between biomolecules in phrase
2	Biomolec	ATP	myosin					Noun	Adjective	Adverb	Verb				Associatio	Modificat	Negative Regulation	Positive Regulation	Transport	create	Others					
3											Present tense	Present continuous	Past tense	Perfect tense												
4	17510632																									
5		Here, we d			0																					
6			block	blocking		0	0					1					1					6	0	1	1	1
7			bind	binding		1	0					1			1							8	0	0	1	0
8			release	released		0	0						1						1			16	8	0	1	0
9	17498971																									
10		Type II my			1			y				y														
11			*interact*	interaction		0	0	1													1	9	5	1	1	1
12			consume	consuming		1	1					1									1	14	0	1	1	1
13	17488711																									
14		The additio			1			y				y	y			y										
15			increase	increased		0	0						1					1				8	7	1	1	1
16			increase	increasing		0	0					1						1				14	1	1	1	1
17			hydrolize	hydrolysis		1	1	1								1						17	0	0	1	0
18	17483158																									
19		Nano-electr			1							y			y											
20			bind	binding		1	1					1			1							11	1	0	1	0
21		Increases in			0																					
22			increase	increases		0	0	1										1				28	5	0	1	0
23			bind	binding		1	0					1			1							13	8	1	1	1
24			hydrolize	hydrolysis		1	0	1								1						24	1	0	1	0
25		The binding			1							y			y											
26			bind	binding		1	1					1			1							3	1	0	1	0
27			bind	binding		1	0					1			1							18	4	1	1	1
28	17450844	We have fo			0																					
29			hydrolize	hydrolysis		1	0	1								1						16	0	0	0	
30			couple	coupling		0	0					1			1							14	1	1	0	
31	17449872																									
32		Selective P			0																					
33			hydrolize	hydrolysis		1	0	1								1						13	0	0	1	0
34			release	release		0	0	1											1			16	3	0	1	0
35		After ATP			1							y			y											
36			bind	binding		1	1					1			1							1	0	1	1	1
37		Our result			0																					
38			*influence*	influences		0	0				1										1	11	4	0	1	0
39			*control*	controlling		0	0					1									1	4	1	1	1	1
40			hydrolize	hydrolysis		1	0	1								1						7	0	0	1	0
41			release	release		0	0	1											1			10	3	0	1	0
42	17438284																									
43		Mechanis			1			y					y			y										
44			*catalyze*	catalyzed		1	1						1					1				2	0	1	1	1
45			hydrolize	hydrolysis		1	1	1								1						1	1	1	1	1
46		The intrinsi			1			y					y		y	y										
47			hydrolize	hydrolysis		1	1	1								1						2	0	1	1	1
48			*catalyze*	catalyzed		1	1						1					1				1	1	1	1	1
49			combine	combined		0	0						1		1							9	5	0	1	0
50		Starting wi			0																					
51			derive	derived		0	0						1			1						21	16	0	1	0
52			bind	bound		1	0						1		1							2	1	1	1	1
53			transform	transformation		0	0	1								1						14	0	1	0	0
54		The model			1			y					y		y	y										
55			transform	transformation		0	0	1								1						22	17	0	1	0
56			hydrolize	hydrolysis		1	0	1								1						13	8	0	1	0
57			bind	bound		1	1						1		1							2	1	1	1	1
58			*change*	changes		0	0	1													1	8	3	0	1	0
59			release	release		0	0	1											1			10	5	0	1	0
60	17391512	On amino a			0																					
61			bind	binding		1	0					1			1							6	0	1	0	
62			bind	binding		1	0					1			1							4	2	1	0	
63	17275022																									
64		After myos			1						y				y											
65			bind	binds		1	1				1				1							0	0	1	1	1
66			release	releases		0	0				1								1			3	1	1	0	
67	17184900																									
68		We examin			0																					
69			regulate	regulating		1	0					1						1				13	7	0	1	0
70			hydrolize	hydrolyzed		1	0						1			1						9	3	0	1	0
71	17142278	At the mol			1			y			y				y	y										
72			generate	generates		0	0				1									1		5	0	1	1	1
73			couple	coupling		0	0								1							5	0	1	1	1
74			hydrolize	hydrolysis		1	1	1								1						7	0	0	1	0
75	17081565	ATP bindin			1			y			y	y			y	y										
76			bind	binding		1	1					1			1							1	0	1	1	1
77			disassocia	disassociate		0	0				1				1							6	3	1	1	1
78			hydrolize	hydrolysis		1	0	1								1						7	1	1	1	1
79			bind	binding		1	0					1			1							3	1	1	1	0
80	17012748																									
81		. The disso			1			y					y		y											
82			disassocia	disassociation		0	0	1							1							7	3	0	1	0
83			bind	bound		1	1						1		1							2	0	1	1	1
84		The present			1			y								y										
85			hydrolize	hydrolysis		1	1	1								1						8	0	0	1	0
86	16963465	It was foun			1			y					y													
87			*consume*	consumed		1	1						1								1	2	0	1	1	1
88			activate	activated		0	0						1					1				17	3	0	1	0
89			*interact*	interaction		0	0	1													1	20	6	0	1	0
90	16950853	. In the skel			1			y								y										
91			hydrolize	hydrolysis		1	1	1								1						4	2	0	1	0
92	16644482	Myosins co			1							y				y										
93			hydrolize	hydrolyzing		1	1					1				1						7	0	1	0	
94			produce	producing		0	0					1									1	10	1	0	0	
95	16359625	By analyzin			1			y			y					y										
96			hydrolize	hydrolysis		1	0	1								1						30	0	1	1	1
97			*breakdow*	breakdown		1	1				1					1						8	2	1	1	1
98	15863618																									
99		The power			0																					
100			hydrolize	hydrolysis		1	0									1						4	0	0	1	0

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
101		Upon bindi			1						y	y			y											
102			bind	binding		1	1					1			1							3	2	0	0	
103			disassocia	disassociates		0	0				1				1							1	0	0	0	
104	**15848287**	A crucial p			1			y					y			y										
105			*generate*	generation		0	0	1												1		13	9	0	1	0
106			release	released		0	0						1						1			5	1	0	1	0
107			hydrolize	hydrolysis		1	1	1								1						2	0	1	1	1
108		We assume			1			y				y	y		y	y										
109			hydrolize	hydrolysis		1	1	1								1						14	0	1	1	1
110			*transmit*	transmitted		0	0						1						1			11	3	1	1	1
111			bind	binding		1	1					1			1							11	3	1	1	1
112			*transmit*	transmission		0	0	1											1			28	12	0	0	
113	**15504033**	Myosin is a			1			y								y										
114			*catalyze*	catalysis		1	1	1										1				8	1	1	0	
115			hydrolize	hydrolysis		1	1	1								1						10	0	0	0	
116	**15489300**	Upon bindi			0																					
117			bind	binding		1	0					1			1							12	2	0	0	
118			*change*	change		0	0				1										1	18	8	0	0	
119	**15454448**	Myosin pro			1			y			y	y			y											
120			produce	produces		0	0				1									1		14	0	1	0	
121			interacte	interaction		0	0	1													1	9	5	1	0	
122			bind	binding		1	1					1			1							10	4	1	0	
123	**15345561**																									
124		Binding of			0																					
125			bind	binding		1	0					1			1							9	7	0	1	0
126			bind	binding		1	0					1			1							3	1	0	1	0
127			inhibit	inhibition		0	0	1									1					6	4	0	1	0
128		Ca(2+) and			0																					
129			stabilize	stabilize		0	0				1							1				2	0	0	1	0
130		The positio			1						y	y			y											
131			bind	binding		1	0					1			1							3	1	0	1	0
132			*affect*	affects		0	0				1										1	5	3	0	1	0
133			*affect*	affect		0	0				1										1	1	0	1	1	1
134			bind	binding		1	1					1			1							1	0	0	1	0
135		Ca(2+) and			0																					
136			*alter*	alter		0	0				1										1	1	0	0	1	1
137			bind	binding		1	0					1			1							9	7	0	1	0
138		A parallel p			0																					
139			regulate	regulation		0	0	1										1				8	6	0	1	0
140			*affect*	effect		0	0	1													1	5	3	0	1	0
141			bind	binding		1	0					1			1							2	0	1	1	1
142			bind	binding		1	0					1			1							6	4	1	1	1
143			complex	complexes			0	1							1							10	7	0	1	0
144	**15247304**																									
145		ATP binds			1			y			y				y	y										
146			bind	binds		1	1				1				1							2	0	1	1	1
147			form	formation		0	0	1								1						12	7	0	1	0
148			complex	complex		0	0	1							1							18	13	0	1	0
149		In the prese			0																					
150			bind	bound		1	0						1		1							14	8	0	0	
151			inhibit	inhibits		0	0				1						1					22	16	0	0	
152			bind	binding		1	0					1			1							24	18	0	0	
153	**15205456**																									
154		Blebbistati			0																					
155			bind	binding		1	0					1			1							6	1	0	1	0
156			induce	induced		0	0						1					1				5	0	0	1	0
157			disassocia	dissociation		0	0	1							1							7	3	0	1	0
158	**15196563**	Myosins ar			1			y					y			y										
159			hydrolize	hydrolysis		1	1	1								1						8	0	0	1	0
160			mediate	mediated		0	0						1					1				9	1	0	1	0
161	**12726731**	Most biomo			1							y			y											
162			bind	binding		1	1					1			1							9	0	0	0	
163	**12482135**	Actin and m			1			y			y		y		y	y										
164			*interact*	interact		0	0				1										1	11	0	1	1	1
165			link	linked		0	0						1		1							7	4	1	1	1
166			hydrolize	hydrolysis		1	1	1								1						10	1	1	1	1
167	**12471887**	Solution me			1			y			y				y	y										
168			bind	bind		1	0				1				1							4	1	1	1	1
169			dissociate	dissociate		0	0				1				1							3	2	1	1	1
170			hydrolize	hydrolysis		1	1	1								1						7	0	0	1	0
171		At high sl			1			y					y			y										
172			*interact*	interactions		0	0	1													1	4	4	1	1	1
173			hydrolize	hydrolized		1	1						1			1						10	0	0	1	0
174	**12381388**	The putativ			1							y			y											
175			bind	binding		1	1					1			1							16	2	0	0	
176			Total sentences		Total interaction	Total interactor apperance	Total Verb apperance	Total Noun																	Total verbs appearing in phrase	between biomolecules in phrase
177		=COUNT	46		30	62	113	41	0	0	18	33	19	0	43	28	3	12	8	3	16			50	90	40
178							Total Verb describing interaction in sentence	Total noun interactor	Total adjinteractor	Total adv interactor	Total present tense interactor	Total -ing interactor	Total PAST interactor	Total Perfect interactor	Total association interactor	Total modification interactor	Total negative interactor	Total positive interactor	Total transporation interactor	Total create interactor	Total OTHERS interactor				Total interactors	Total interactors
179							34	20	0	0	4	28	9	0	32	24	0	4	0	0	2			31	49	25
180								0.4878			0.22222	0.84848	0.47368		0.74419	0.85714	0	0.33333	0	0	0.125	Percent:			0.54444	0.625
181								Total noun describing interaction	Total adf describing interaction	Total adv describing interaction	Total present describing interaction	Total -ing describing interaction	Total past describing interaction	Total perfect describing interaction	Total association describing interaction	Total modification describing interaction	Total negative describing interaction	Total postivie describing interaction	Total transportation describing interaction	Total create describing interaction	Total others describing interaction			Total verbs between biomolecules out of phrase	Total verbs appearing out of phrase	not between biomolecules in phrase
182								13	0	0	3	12	6	0	14	15	0	3	0	0	2			50-40=10	113-90=2	90-40=50
183								0.31707			0.16667	0.36364	0.31579		0.32558	0.53571	0	0.25	0	0	0.125			Total inte	Total inte	Total inte
184																								31-25=6	62-49=13	49-25=24
185							Total sentence containing:	29			15	25	16		29	25	3	11	7	3	12			0.6	0.56522	0.48
186							Interactio	21			10	13	11		18	18	0	6	4	3	8			Total verbs out of biomolecules out of phrase		
187								0.72414	#DIV/0!	#DIV/0!	0.66667	0.52	0.6875	#DIV/0!	0.62069	0.72	0	0.54545	0.57143	1	0.66667			23-10=13		
188							Total phrase containing	25			11	15	15		21	21	2	10	6	2	9			Total interactors		
189							Interactio	20			8	8	11		14	16	0	6	3	2	5			13-6=7		
190								0.8	#DIV/0!	#DIV/0!	0.72727	0.53333	0.73333	#DIV/0!	0.66667	0.7619	0	0.6	0.5	1	0.55556			0.53846		
191							Total phrases where verbs are between biomolecules	9			9	10	10		15	7	1	3	1	1	7					
192							Interactio	9			8	6	9		11	7	0	3	1	1	6					

Analysis of the sample data table

Data:
For the sentence set retrieved for each biomolecule pair, the properties of the constituent of sentences and all interaction-indicating terms involved were tabulated.
In this table, note the following observations.

- Column C (interaction-indicating term root) shows the roots of the interaction-indicating terms appearing in the sentences, and column D (interaction-indicating term in sentence) shows the exact appearance of each interaction-indicating term in the sentence, whose form can be one of columns H through N.
- Column E (interaction described) records whether the sentence describes the interaction between the two biomolecules (1 means yes, 0 means no. The same also applies to those of the following columns which require values of 1 or 0.)

Columns F and G are about interaction-indicating terms.

- Column G (Describe the interaction in sentence?) shows whether the interaction-indicating term is used to describe an interaction between the two biomolecules of concern. For example, consider the sentence "The binding of ATP to myosin-S1 could be observed in the presence of up to 60 microM of excess MgATP without non-specific binding of MgATP to the myosin." In that sentence "bind" describes the interaction between ATP and myosin.
- Column F (correct interaction word) shows whether the interaction-indicating term is the right interaction description for the two biomolecules, regardless of what the sentence says. For example, in the sentence "Increases in the molecular mass of myosin-S1 of 425 +/- 10 were obtained with the binding of ADP to the active site and by 530 +/- 10 with either ATP or hydrolysis products ADP and phosphate", "bind" is not used to describe an interaction between ATP and myosin. However, from the last example we know "bind "is indeed the interaction between ATP and myosin, so column F will show for "bind" in this sentence.
- Columns O through U are for the categories of the interaction-indicating terms involved in the sentences. They are "association," "modification," "positive regulation," "negative regulation," "transportation," "transcription," "creation," and "vague." If an interaction-indicating term belongs to one of them, the corresponding column contains 1.
- Columns V through Z contain position data for the interaction-indicating terms. V and W are the distances from a term to the nearest biomolecule and to the other biomolecule in the pair of biomolecules. X, Y and Z give information about whether the interaction-indicating term is in the same phrase as the pair of biomolecules, whether it is between the two biomolecules in a phrase, and whether it is between the two in a sentence or not.

Analysis part:

Each table supports computing the following numbers.

(1) Total number of sentences examined that contain the 2 biomolecules that are the subject of the table.

(2) Number of sentences describing the interaction

(3) Number of times interaction-indicating terms appear in all sentences

(4) Number of times an interaction-indicating term describes the interaction between the pair of biomolecules of interest in the sentences

(5) Total times of appearance of correct interaction-indicating terms

For each category or form:

(6) Times of appearance of each form or category of the interaction-indicating term in the sentence set, e.g., H177 in the table for myosin and ATP (the same table is assumed for the following examples.)

(7) Times of appearance of each form or category of the correct interaction-indicating word, e.g., H179 in the table

(8) 100%*(7)/(6), which gives the probability of an interaction-indicating term being the correct one as a function of its form or category, e.g., H180

(9) Times of appearance of each form or category of an interaction-indicating term describing the interaction in the sentence, e.g. H182

(10) 100%*(9)/(6), which gives the probability that an interaction-indicating term describes the interaction in a sentence as a function of its form or category, e.g., H183

(11) The number of sentences containing an interaction-indicating term appearing in each form or category, e.g. H185

(12) The number of sentences containing an interaction-indicating term appearing in each form or category and describing an interaction between the two biomolecules, e.g. H186

(13) 100%*(12)/(11), which gives the probability that sentences containing an interaction-indicating term appearing in a specific form or category describe an interaction between the two biomolecules, e.g., H187

(14) The number of phrases containing both the two biomolecules and an interaction-indicating term appearing in a specific form or category, e.g. H188

(15) The number of phrases containing the two biomolecules together with an interaction-indicating term appearing in a specific form or category that describes an interaction between the two biomolecules, e.g. H189

(16) 100%*(15)/(14), which gives the probability that phrases containing the two biomolecules plus an interaction-indicating term appearing in a specific form or category describes an interaction between the two biomolecules, e.g., H190

(17) The number of phrases containing the two biomolecules plus an interaction-indicating term that appears in a specific form or category placed between the two biomolecules, e.g. H191

(18) The number of phrases containing the two biomolecules plus an interaction-indicating term that appears in a specific form or category placed between the two biomolecules, that describe an interaction between the two biomolecules, e.g. H192

(19) 100%*(18)/(17), which gives the probability that phrases containing an interaction-indicating term appearing in a specific form or category between the two biomolecules, describes an interaction between the two biomolecules, e.g. H193

For the position data:

(20) The number of interaction-indicating terms also appearing in the same phrase as the two biomolecules, e.g. Y177

(21) The number of correct interaction-indicating terms appearing in a phrase also containing the two biomolecules, e.g. Y179

(22) 100%*(21)/(20), which gives the probability that an interaction-indicating term appearing in the same phrase as the two biomolecules is a correct interaction word, e.g. Y180

(23) The number of occurrences of interaction-indicating terms appearing in the same phrase as the two biomolecules and appearing between the two biomolecules, e.g., Z177

(24) The number of interaction-indicating term occurrences that are correct interaction descriptions, appearing between the two biomolecules and in the same phrase as the two biomolecules, e.g. Z179

(25) 100%*(24)/(23), which gives the probability that an interaction-indicating term appearing in the same phrase as the two biomolecules and appearing between the two biomolecules is a correct interaction term, e.g. Z180

(26) The number of times interaction-indicating terms appear in the same phrase as the two biomolecules but **not** appearing between the two biomolecules, e.g. Z182

(27) The number of times interaction-indicating terms that are the correct interaction appear in the same phrase as the two biomolecules but **not** between the two biomolecules , e.g. Z184

(28) 100%*(27)/(26), which gives the probability that an interaction-indicating term appearing in the same phrase as the two biomolecules but **not** appearing between the two biomolecules is a correct interaction word, e.g. Z185

(29) The number of interaction-indicating terms **not** appearing in the same phrase as the two biomolecules, which is the difference between the total appearances of interaction-indicating terms and the appearances of interaction-indicating terms appearing in a phrase, e.g. Y182

(30) The number of times of appearance of interaction-indicating terms that are the correct interaction but **not** in the same phrase as the two biomolecules, which is the difference between the total correct interaction-indicating term appearances and the number of the appearances of correct interaction-indicating terms in the phrase, e.g. Y184

(31) 100%*(30)/(29), which gives the probability that an interaction-indicating term **not** appearing in the same phrase as the two biomolecules gives a correct interaction, e.g. Y185

(32) The number of interaction-indicating terms that are the correct interaction and do **not** appear in the same phrase as the two biomolecules but do appear between the two biomolecules, which is the difference between the total interaction-indicating term appearances and the interaction-indicating terms appearing in a phrase, e.g. X182

(33) The number of times of appearances of interaction-indicating terms that are the correct interaction and **not** in the same phrase as the two biomolecules, but appearing between the two biomolecules, which is the difference between the total correct interaction-indicating term appearances and the number of appearances of correct interaction-indicating terms appearing in a phrase, e.g. X184

(34) 100%*(33)/(32), which gives the probability that an interaction-indicating term **not** appearing in the same phrase as the two biomolecules, but between the two biomolecules, represents the correct interaction, e.g. X185.

(35) The number of interaction-indicating terms appearances **not** in the same phrase as the two biomolecules and **not** between the two biomolecules, which is the difference between the total appearances of interaction-indicating terms and the interaction-indicating terms appearing in a phrase, e.g. X187

(36) The number of interaction-indicating term appearances that correctly indicate an interaction **not** appearing in the same phrase as the two biomolecules and **not** appearing between the two biomolecules, which is the difference between the total times of correct interaction-indicating term appearances and the number of appearances of correct interaction-indicating terms appearing in a phrase, e.g. X189

(37) 100%*(36)/(35), which gives the probability that an interaction-indicating term **not** appearing in the same phrase as the two biomolecules and **not** between the two biomolecules, represents the correct interaction between the biomolecules, e.g. X190

Appendix II Results of proportion of different pairs in sentences, phrases, interaction indicating terms investigation

Interaction term appearance	noun	adj	adv	present tense	-ing	PAST	Perfect	Asso-ciation	Modi-fication	negative	positive	Trans-portation	transcript	create	vacuous
ATY MYOSIN	0.46	1		0.28	0.85	0.42		0.74	0.86	0.00	0.18	0.00	0.00	0.00	0.18
cre cytokinin	0.63	0.00		0.30	0.75	0.20		1.00	0.00	0.00	0.11	0.20	0.00	0.00	0.00
nitrite xanthine	0.80	0.00		0.08	0.73	0.39		0.00	0.96	0.00	0.00			0.64	0.22
glucose-6-p starch	0.74	0.25		0.50	0.25	0.22		0.88	0.25	0.00	0.09	0.00	0.00	0.85	0.00
glucose-starch	0.53			0.43	0.00	0.37		0.00	0.96	0.00	0.00		0.00	0.87	0.00
glucose pyruvate	0.45	0.60		0.24	0.33	0.13		0.00	0.13	0.00	0.00	0.00	0.00	0.90	0.75
acetyl-coa leucine	0.72	1.00		0.17	0.00	0.65		0.00	0.65	0.14	0.00	0.00		1.00	0.89
indole acetic acid starch	0.00	0.00			0.00	0.00		0.00	0.00	0.00	0.00	0.00		0.00	0.00
carotenoid ipp	0.40			0.40	0.00	0.14		0.80		0.00	0.00		0.00	0.20	0.00
pyruvate dehydrogenase phosphofructokinase	0.03	0.00		0.00	0.00	0.00		0.00	0.00	0.00	0.00	0.00		0.00	
Sentence describing interaction															
ATY MYOSIN	0.69			0.67	0.52	0.63		0.62	0.68	0.33	0.40	0.71	0.00	1.00	0.77
cre cytokinin	0.89	0.50		0.71	0.33	0.81		0.91	0.00	0.78	0.75	1.00	1.00	1.00	1.00
nitrite xanthine	0.69	0.00		0.75	0.70	0.60		1.00	0.72	0.40	0.68			0.50	0.71
glucose-6-p starch	0.74	1.00		1.00	0.75	0.69		0.57	1.00	0.71	0.45	1.00		0.84	1.00
glucose-starch	0.48			0.64	0.80	0.43		0.00	0.80	0.38	0.27		0.00	0.62	0.20
glucose pyruvate	0.42	0.33		0.60	0.40	0.33		0.00	0.33	0.33	0.20	0.33	1.00	0.50	0.57
acetyl-coa leucine	0.61	0.50		0.67	0.40	0.63		0.38	0.73	0.33	0.44	0.00		0.74	0.64
indole acetic acid starch	0.43	0.50		1.00	0.50	0.40			0.67	0.60	0.00	0.00		0.50	0.33
carotenoid ipp	0.50			1.00	0.50	0.50		0.75		1.00	0.33		0.67	0.67	0.00
pyruvate dehydrogenase phosphofructokinase	0.05	0.00		0.00	0.13	0.00		0.00	0.00	0.06	0.08	0.00		0.00	0.00

Phrase describing interaction	noun	adj	adv	present tense	-ing	PAST	Perfect	Asso-ciation	Modi-fication	negative	positive	Trans-portation	transcript	create	vacuous
ATY MYOSIN	0.69			0.83	0.50	0.67		0.64	0.68	0.50	0.44	0.57		1.00	0.80
Cre cytokinin	0.83	0.00		0.50	0.00	0.80		0.90	0.00	0.75	0.56	0.50			1.00
nitrite xanthine	0.78			0.71	0.50	0.69			0.83	0.33	0.80			0.56	0.67
glucose-6-p starch	0.78					0.80		0.60	1.00	0.83	0.50	1.00		0.82	1.00
glucose-starch	0.50			0.71	1.00	0.71			1.00	0.55	0.50			0.86	0.33
glucose pyruvate	0.52	0.50		0.56	0.60	0.33		0.00	0.57	0.33	0.25	0.00	1.00	0.67	0.62
acetyl-coa leucine	0.58	0.50		1.00		0.67		0.75	0.86	0.50	0.50			0.57	0.67
indole acetic acid starch	1.00			1.00					1.00	1.00					
carotenoid ipp	0.00			1.00				1.00			1.00		1.00		
pyruvate dehydrogenase phosphofructokinase	0.13				1.00	0.00			0.00	0.17	1.00				

1. N/A means no data available

Appendix III Interaction-Indicating Term List

verbID	verb	verb_group	verb_category_id	#_trioccurrences
1100001	abolish	negative regulation	3	55
1100002	accompany	association	1	62
1100003	accumulate	positive regulation	4	1
1100004	acetylate	modification	2	22
1100005	activate	positive regulation	4	380
1100006	aggregate	association	1	78
1100007	agonize	positive regulation	4	116
1100008	amplify	positive regulation	4	6
1100009	antagonize	negative regulation	3	139
1100010	associate	association	1	280
1100011	attenuate	negative regulation	3	29
1100012	augment	positive regulation	4	36
1100013	bind	association	1	810
1100014	block	negative regulation	3	249
1100015	blunt	negative regulation	3	11
1100016	cocluster	association	1	0
1100017	coexpress	positive regulation	4	15
1100018	combine	association	1	176
1100019	complex	association	1	289
1100020	conjugate	association	1	62
1100021	convert	vague	8	3
1100022	cooverexpress	positive regulation	4	2
1100023	costimulate	positive regulation	4	4
1100024	counteract	negative regulation	3	9
1100025	couple	association	1	112
1100026	decrease	negative regulation	3	576
1100027	degradate	modification	2	78
1100028	deplete	negative regulation	3	88
1100029	depress	negative regulation	3	76
1100030	derive	create	6	1
1100031	destabilize	modification	2	5
1100032	detoxify	modification	2	3
1100033	diminish	negative regulation	3	38
1100034	diminute	negative regulation	3	6
1100035	disrupt	negative regulation	3	23

1100036	dissociate	association	1	77
1100037	distribute	transportation	5	125
1100038	downregulate	negative regulation	3	5
1100039	elevate	positive regulation	4	130
1100040	elicit	positive regulation	4	4
1100041	eliminate	negative regulation	3	44
1100042	enhance	positive regulation	4	2
1100043	evoke	positive regulation	4	44
1100044	form	create	6	2
1100045	hydrolyze	modification	2	146
1100046	improve	positive regulation	4	35
1100047	incorporate	association	1	266
1100048	increase	positive regulation	4	2
1100049	induce	positive regulation	4	2
1100050	inhibit	negative regulation	3	1316
1100051	invoke	positive regulation	4	1
1100052	link	association	1	144
1100053	lower	negative regulation	3	181
1100054	lytic	modification	2	5
1100055	maintain	positive regulation	4	70
1100056	mediate	vague	8	221
1100057	mobilize	transportation	5	34
1100058	modify	modification	2	187
1100059	modulate	positive regulation	4	67
1100060	obliterate	negative regulation	3	2
1100061	overexpress	negative regulation	3	136
1100062	oxidate	modification	2	1
1100063	phosphorylate	modification	2	183
1100064	potentiate	positive regulation	4	50
1100065	produce	create	6	3
1100066	promote	positive regulation	4	41
1100067	provoke	positive regulation	4	10
1100068	raise	positive regulation	4	53
1100069	redistribute	transportation	5	5
1100070	reduce	negative regulation	3	1
1100071	regulate	vague	8	3
1100072	release	transportation	5	426
1100073	repress	negative regulation	3	3
1100074	rise	positive regulation	4	78
1100075	stabilize	positive regulation	4	2

1100076	stimulate	positive regulation	4	2
1100077	suppress	negative regulation	3	1
1100078	transactivate	positive regulation	4	24
1100079	transform	modification	2	81
1100080	translocate	transportation	5	31
1100081	underexpress	negative regulation	3	0
1100082	upregulate	positive regulation	4	86
1100083	synthesize	create	6	634
1100085	remove	negative regulation	3	0
1100086	affect	vague	8	1
1100087	transport	transportation	5	247
1100088	limit	negative regulation	3	88
1100089	metabolize	vague	8	445
1100090	control	vague	8	1
1100091	change	vague	8	461
1100092	encode	transcription	7	33
1100093	influence	vague	8	136
1100094	catalyze	positive regulation	4	5
1100095	catalyze	positive regulation	4	2
1100096	generate	create	6	110
1100097	reconstruct	create	6	1
1100098	decarboxylate	modification	2	14
1100099	oxidize	modification	2	3
1100100	catabolize	modification	2	9
1100101	accept	association	1	1
1100102	reduce	modification	2	306
1100103	subtract	negative regulation	3	3
1100104	dehydrogenate	modification	2	2
1100105	carboxylate	modification	2	11
1100106	import	transportation	5	1
1100107	convey	transportation	5	1
1100108	utilize	vague	8	85
1100109	alter	vague	8	231
1100110	disassemble	modification	2	3
1100111	depolymerize	modification	2	1
1100112	remove	negative regulation	3	110
1100113	nitrate	modification	2	18
1100114	perceive	association	1	2
1100115	receive	association	1	499
1100116	dephosphorylate	modification	2	34

1100117	transmit	transportation	5	21
1100118	deregulate	negative regulation	3	1
1100119	interact	vague	8	219
1100120	transduce	transportation	5	5
1100121	consume	vague	8	13
1100122	breakdown	modification	2	22
1100123	disassociate	association	1	1
1100124	stabilize	positive regulation	4	34
1100125	moderate	vague	8	20
1100126	delay	negative regulation	3	34

Appendix IV Previous Research

Text Empirics-Based Mining of Biomolecular Interactions from Texts: Theory and Application

Daniel Berleant[1], Lifeng Zhang[2], Jing Ding[3], Tuan Cao[2],
Daniel Nettleton[2], Jun Xu[4], Andy Fulmer[4], and Eve Wurtele[2]
[1]Dept. of Information Science, University of Arkansas at Little Rock,
jdberleant@ualr.edu,
501-569-3448;
[2]Iowa State University, Ames; [3]Ohio State University Medical Center;
[4]Procter & Gamble Co., Miami, Ohio

This is online at:
http://ifsc.ualr.edu/jdberleant/papers/LifengZhangDissertationAppendixIV.pdf

Appendix V 300 sentences

http://ifsc.ualr.edu/jdberleant/papers/LifengZhangDissertationAppendixV.txt

Appendix VI Sentences’ score by three different combination methods

Interaction described?	ALL	PRI	ORI	Interaction described?	ALL	PRI	ORI
1	0.425	0.891	0.795	0	0.328	0.831	0.735
1	0.322	0.783638	0.688	0	0.359	0.867	0.794
1	0.336	0.805	0.706	1	0.661	0.976	0.968
0	0.58	0.965	0.936	0	0.681	0.977	0.971
0	0.633	0.973	0.958	0	0.661	0.976	0.968
1	0.449	0.909	0.835	1	0.659	0.976	0.966
1	0.429	0.924	0.857	0	0.361	0.834	0.75
1	0.429	0.924	0.857	0	0.401	0.907	0.818
1	0.59	0.967	0.94	1	0.401	0.907	0.818
1	0.498	0.947	0.891	1	0.303	0.764	0.644
0	0.485	0.943	0.883	1	0.556	0.962	0.929
1	0.367	0.884	0.769	1	0.62	0.972	0.954
0	0.4	0.907	0.834	1	0.569	0.963	0.931
1	0.405	0.938	0.865	1	0.406	0.881	0.776
0	0.475	0.921	0.854	1	0.61	0.97	0.95
0	0.544	0.96	0.924	0	0.606	0.973	0.958
1	0.622	0.973	0.957	1	0.446	0.907	0.832
1	0.62	0.972	0.954	0	0.476	0.921	0.845
1	0.633	0.973	0.958	0	0.44	0.902	0.825
1	0.633	0.973	0.958	1	0.437	0.948	0.892
0	0.622	0.972	0.954	0	0.676	0.977	0.97
1	0.64	0.974	0.96	1	0.696	0.978	0.973
0	0.465	0.916	0.833	1	0.558	0.961	0.925
1	0.668	0.977	0.969	1	0.506	0.935	0.878
0	0.4	0.906	0.816	1	0.437	0.948	0.892
1	0.62	0.972	0.954	1	0.652	0.976	0.966
1	0.709	0.978	0.975	1	0.602	0.969	0.946
1	0.659	0.976	0.966	1	0.705	0.978	0.974
1	0.401	0.907	0.818	1	0.705	0.978	0.974
1	0.686	0.977	0.971	1	0.705	0.978	0.975
1	0.401	0.907	0.818	1	0.095	0.456	0.23
1	0.461	0.913	0.829	0	0.437	0.948	0.892
1	0.463	0.955	0.913	1	0.458	0.915	0.839
0	0.437	0.948	0.892	1	0.458	0.915	0.839
1	0.44	0.902	0.825	0	0.676	0.977	0.97
1	0.7	0.978	0.973	1	0.346	0.902	0.811

0	0.51	0.951	0.899
1	0.495	0.93	0.869
0	0.381	0.894	0.789
1	0.475	0.921	0.854
1	0.703	0.968	0.954
1	0.033	0.033	0.971
1	0.441	0.948	0.892
0	0.401	0.907	0.818
1	0.666	0.976	0.968
1	0.437	0.948	0.892
1	0.651	0.976	0.966
1	0.477	0.925	0.861
1	0.652	0.976	0.966
1	0.651	0.976	0.966
1	0.48	0.927	0.864
1	0.47	0.922	0.855
1	0.579	0.965	0.935
0	0.4	0.906	0.816
1	0.686	0.977	0.971
1	0.419	0.888	0.797
0	0.578	0.966	0.938
1	0.448	0.909	0.834
1	0.634	0.974	0.962
1	0.591	0.967	0.941
1	0.033	0.033	0.078
0	0.472	0.939	0.874
1	0.487	0.932	0.879
1	0.726	0.979	0.978
1	0.518	0.963	0.93
1	0.501	0.932	0.869
1	0.558	0.961	0.925
1	0.49	0.934	0.882
1	0.328	0.831	0.735
1	0.611	0.973	0.957
1	0.4018	0.9065	0.8183
1	0.633	0.973	0.958
1	0.4018	0.9065	0.8183
1	0.62	0.972	0.954
0	0.558	0.961	0.925
1	0.561	0.968	0.944
1	0.643	0.974	0.962
1	0.321	0.776	0.683
1	0.587	0.971	0.951
1	0.727	0.979	0.978
1	0.437	0.948	0.892
1	0.451	0.911	0.831
0	0.523	0.954	0.906
1	0.455	0.913	0.835
0	0.436	0.906	0.822
0	0.535	0.956	0.913
1	0.472	0.939	0.874
0	0.612	0.97	0.95
0	0.514	0.938	0.884
0	0.405	0.876	0.774
0	0.503	0.934	0.875
0	0.322	0.789	0.677
0	0.4	0.906	0.816
0	0.364	0.838	0.753
0	0.42	0.888	0.796
0	0.341	0.811	0.711
0	0.31	0.775	0.653
0	0.429	0.924	0.857
0	0.485	0.925	0.856
0	0.033	0.033	0.033
0	0.44	0.902	0.825
0	0.405	0.876	0.774
0	0.429	0.924	0.857
0	0.367	0.841	0.756
0	0.405	0.876	0.774
0	0.402	0.907	0.819
0	0.343	0.813	0.713
0	0.401	0.907	0.818
0	0.477	0.921	0.849
0	0.344	0.815	0.715
0	0.354	0.826	0.743
0	0.493	0.929	0.863
0	0.364	0.838	0.753
0	0.402	0.907	0.819
1	0.591	0.967	0.941
0	0.484	0.943	0.883
0	0.361	0.835	0.749
0	0.4018	0.9065	0.8183

1	0.4018	0.9065	0.8183	1	0.4018	0.9065	0.8183
0	0.506	0.935	0.878	0	0.534	0.956	0.913
0	0.42	0.888	0.796	0	0.336	0.805	0.706
0	0.483	0.925	0.796	0	0.361	0.834	0.75
0	0.457	0.912	0.825	0	0.643	0.974	0.962
1	0.612	0.97	0.95	1	0.429	0.924	0.857
0	0.324	0.791	0.679	0	0.424	0.892	0.811
0	0.42	0.888	0.796	1	0.429	0.924	0.857
1	0.659	0.976	0.966	0	0.508	0.952	0.906
0	0.133	0.574	0.325	0	0.623	0.972	0.954
0	0.343	0.813	0.713	0	0.344	0.815	0.715
1	0.593	0.972	0.955	0	0.457	0.911	0.831
0	0.448	0.906	0.823	0	0.333	0.8	0.702
0	0.513	0.937	0.879	0	0.469	0.917	0.842
1	0.476	0.921	0.845	1	0.643	0.974	0.962
1	0.4	0.906	0.816	1	0.47	0.922	0.855
1	0.726	0.979	0.978	0	0.4	0.906	0.816
0	0.58	0.965	0.936	1	0.429	0.924	0.857
0	0.63	0.973	0.957	0	0.599	0.972	0.954
0	0.6	0.969	0.946	0	0.62	0.972	0.954
0	0.429	0.897	0.805	1	0.676	0.978	0.974
0	0.62	0.972	0.954	0	0.533	0.958	0.918
0	0.612	0.97	0.95	0	0.405	0.876	0.774
1	0.675	0.978	0.974	1	0.643	0.974	0.962
0	0.401	0.907	0.818	1	0.546	0.959	0.919
0	0.477	0.921	0.849	0	0.429	0.924	0.857
1	0.467	0.917	0.848	0	0.6	0.969	0.946
0	0.484	0.925	0.852	0	0.643	0.974	0.962
0	0.546	0.959	0.919	0	0.633	0.973	0.958
1	0.587	0.971	0.951	1	0.649	0.975	0.964
1	0.401	0.907	0.818	1	0.696	0.978	0.973
1	0.429	0.924	0.857	0	0.479	0.923	0.857
1	0.659	0.976	0.966	1	0.633	0.973	0.958
1	0.659	0.976	0.966	1	0.612	0.97	0.95
1	0.659	0.976	0.966	1	0.401	0.907	0.818
1	0.659	0.976	0.966	1	0.62	0.972	0.954
0	0.361	0.834	0.75	0	0.62	0.972	0.954
1	0.6	0.969	0.946	0	0.62	0.972	0.954
1	0.612	0.97	0.95	0	0.64	0.974	0.96

0	0.401	0.907	0.818	0	0.401	0.907	0.818
1	0.429	0.924	0.857	1	0.444	0.907	0.818
1	0.465	0.915	0.838	1	0.416	0.915	0.831
0	0.333	0.8	0.702	1	0.612	0.97	0.95
0	0.417	0.887	0.805	1	0.589	0.967	0.942
1	0.344	0.815	0.715	0	0.437	0.948	0.892
1	0.659	0.976	0.966	0	0.401	0.907	0.819
0	0.491	0.929	0.866	1	0.7	0.978	0.973
0	0.506	0.935	0.878	0	0.612	0.97	0.95
0	0.498	0.945	0.86	1	0.499	0.932	0.872
1	0.453	0.908	0.827	1	0.401	0.907	0.819
1	0.485	0.87	0.766	1	0.61	0.97	0.95
0	0.396	0.925	0.856	1	0.467	0.917	0.848
0	0.126	0.558	0.312	0	0.666	0.976	0.968
1	0.649	0.975	0.964	0	0.401	0.907	0.818
0	0.485	0.925	0.856	0	0.367	0.841	0.756
1	0.487	0.927	0.863	1	0.635	0.974	0.962
1	0.48	0.927	0.864	0	0.453	0.932	0.852
0	0.483	0.945	0.892	1	0.651	0.976	0.966
1	0.581	0.971	0.952	1	0.448	0.906	0.823
0	0.426	0.895	0.801	1	0.643	0.974	0.962
1	0.495	0.93	0.869	0	0.429	0.924	0.857
1	0.633	0.973	0.958	1	0.557	0.961	0.924
1	0.623	0.972	0.954	1	0.484	0.943	0.882
1	0.58	0.965	0.936	1	0.501	0.932	0.869
1	0.429	0.924	0.857	1	0.357	0.83	0.747
1	0.666	0.976	0.968	0	0.471	0.919	0.851
1	0.643	0.974	0.962	1	0.612	0.97	0.95
0	0.406	0.877	0.782	1	0.429	0.924	0.857
1	0.591	0.967	0.941	0	0.459	0.915	0.844
1	0.525	0.943	0.892	0	0.269	0.596	0.596
1	0.506	0.935	0.878	0	0.499	0.932	0.872
0	0.556	0.962	0.929	0	0.421	0.888	0.791
1	0.508	0.952	0.906	1	0.401	0.907	0.818
1	0.612	0.97	0.95	1	0.509	0.936	0.876
1	0.599	0.972	0.954	1	0.488	0.927	0.856

1	0.492	0.929	0.859	1	0.462	0.918	0.843
0	0.434	0.907	0.821	0	0.401	0.907	0.818
0	0.477	0.921	0.849	0	0.465	0.915	0.838
1	0.643	0.974	0.962	1	0.696	0.978	0.973
1	0.401	0.873	0.776	1	0.696	0.978	0.973
0	0.475	0.921	0.854	1	0.612	0.97	0.95
1	0.606	0.973	0.958	1	0.457	0.911	0.831
0	0.569	0.963	0.931	1	0.116	0.531	0.288

Appendix VII The data about relationship between distance of IIT from biomolecules and IIT’s correctness

Far correct IITs distance	pyruvate dehydrogenase phosphofructinase	glucose pyruvate	atp myosin	g-6-p starch	iaa starch	glucose starch	cre cytokinin	carotenoid ipp	acetylcoa	nitrite xanthine	New verbs	Total
0	0	1	1	1	0	3	4	0	0	1	6	17
1	0	2	5	1	0	3	1	0	4	2	5	23
2	0	2	8	6	0	4	4	0	3	11	4	42
3	0	11	5	2	0	3	5	0	7	3	6	42
4	0	1	4	1	0	4	4	1	4	3	7	29
5	0	1	0	3	0	1	4	0	3	7	7	26
6	0	2	3	2	0	2	3	0	0	4	6	22
7	0	1	5	2	0	2	2	0	4	7	12	35
8	0	2	4	2	0	1	4	0	6	0	4	23
9	0	1	4	1	0	4	4	0	4	4	4	26
10	0	1	4	0	0	1	1	1	2	3	5	18
11	1	0	3	3	0	1	1	0	1	2	9	21
12	0	1	1	0	0	1	1	1	2	2	1	10
13	0	1	4	1	0	0	0	1	2	2	3	14
14	0	1	3	2	0	1	1	0	3	1	2	14
15	0	4	0	2	0	0	0	0	2	3	1	12
16	0	1	2	0	0	0	0	0	3	2	1	9
17	0	0	1	0	0	0	1	0	2	0	2	6
18	0	0	1	1	0	0	0	0	1	0	1	4
19	0	0	0	0	0	0	0	0	1	1	1	3
20	0	1	0	0	0	2	0	0	1	0	3	7
21	0	0	0	1	0	1	0	0	0	0	0	2

22	0	4	0	0	0	0	0	0	1	0	0	5
23	0	0	0	0	0	0	1	0	0	1	0	2
24	0	1	2	0	0	0	0	0	0	1	4	8
25	0	0	0	0	0	0	0	0	0	0	1	1
26	0	0	0	0	0	0	0	0	0	1	0	1
27	0	0	0	0	0	0	0	0	0	3	1	4
28	0	0	0	0	0	0	0	0	0	0	0	0
29	0	0	0	0	0	1	0	0	0	0	0	1
30	0	0	1	0	0	0	0	0	2	0	0	3
31	0	0	0	0	0	0	0	0	1	0	1	2
32	0	0	0	0	0	0	0	1	0	0	0	1
33	0	0	0	0	0	0	1	0	0	0	0	1
34	0	0	0	0	0	0	0	0	0	0	1	1
35	0	0	0	0	0	0	0	0	0	0	0	0
36	0	0	0	0	0	0	0	0	0	0	0	0
37	0	0	0	0	0	0	0	0	1	0	0	1
38	0	0	0	0	0	0	0	0	0	0	0	0

Far IITs distance	pyruvate dehydrogenase phosphofructinase	glucose pyruvate	atp myosin	g-6-p starch	iaa starch	glucose starch	cre cytokinin	carotenoid ipp	acetylcoa	nitrite xanthine	New verbs	Total	odds
0	0	1	1	1	0	3	4	0	0	3	6	19	8.5
1	0	2	8	1	1	4	3	1	4	2	7	33	2.3
2	1	9	10	7	1	5	6	0	3	13	8	63	2
3	5	13	7	2	1	4	6	1	7	8	11	65	1.826087
4	3	10	6	2	0	10	7	1	9	9	21	78	0.591837
5	8	4	6	6	4	6	8	1	4	14	19	80	0.481481
6	2	4	6	6	2	3	8	0	3	7	13	54	0.6875
7	4	6	8	6	1	5	4	0	5	10	22	71	0.972222
8	5	10	8	3	0	4	5	1	8	1	18	63	0.575
9	3	5	8	2	2	4	5	0	7	8	13	57	0.83871
10	5	7	8	2	3	4	1	3	5	3	11	52	0.529412
11	4	6	5	7	2	2	3	1	4	2	17	53	0.65625
12	2	3	2	2	1	4	1	2	2	4	9	32	0.454545

13	6	4	5	4	0	1	1	2	2	3	10	38	0.583333
14	2	1	7	3	1	2	5	1	5	3	15	45	0.451613
15	2	6	0	3	0	1	2	0	3	3	5	25	0.923077
16	0	2	4	0	0	2	2	1	4	2	6	23	0.642857
17	2	3	2	0	2	3	3	0	3	3	9	30	0.25
18	0	1	3	3	1	0	0	1	2	0	4	15	0.363636
19	2	2	0	0	0	2	3	1	1	1	4	16	0.230769
20	0	2	1	0	1	3	0	0	1	2	11	21	0.5
21	1	0	1	1	0	1	2	0	1	0	5	12	0.2
22	1	4	2	0	0	1	0	0	4	0	2	14	0.555556
23	1	1	0	1	0	0	2	0	0	1	4	10	0.25
24	0	1	2	1	0	0	1	1	1	2	7	16	1
25	1	0	0	0	0	0	0	1	0	1	4	7	0.166667
26	1	1	0	1	0	1	1	0	1	1	6	13	0.083333
27	0	1	0	0	0	2	0	0	0	3	1	7	1.333333
28	0	1	2	0	0	0	0	0	0	0	2	5	0
29	0	0	0	0	0	1	1	0	0	1	0	3	0.5
30	0	0	1	1	0	1	0	0	2	1	1	7	0.75
31	0	0	0	0	0	0	0	0	1	0	3	4	1
32	0	0	0	0	0	0	0	1	0	0	0	1	#DIV/0!
33	0	0	0	0	0	0	1	0	0	0	0	1	#DIV/0!
34	1	0	0	0	1	0	0	0	0	0	1	3	0.5
35	0	0	0	0	0	0	0	0	0	0	1	1	0
36	0	0	0	0	0	0	1	0	0	0	1	2	0
37	0	0	0	0	0	0	0	0	1	0	0	1	#DIV/0!
38	1	0	0	0	0	0	0	0	0	0	0	1	0

Nearest correct IITs distance	pyruvate dehydrogenase phosphofructinase	glucose pyruvate	atp myosin	g-6-p starch	iaa starch	glucose starch	cre cytokinin	carotenoid ipp	acetylcoa	nitrite xanthine	New verbs	Total
0	0	16	28	26	0	13	30	1	18	19	40	191
1	0	8	14	0	0	5	5	2	14	11	17	76
2	0	6	6	0	0	4	1	0	8	7	6	38
3	0	2	2	1	0	3	3	0	5	11	15	42
4	0	3	4	0	0	3	0	0	0	4	7	21
5	1	0	0	0	0	1	3	0	4	5	3	17
6	0	0	0	1	0	0	0	0	1	1	1	4
7	0	2	3	1	0	3	0	1	1	1	1	13
8	0	0	3	0	0	0	0	0	0	1	1	5
9	0	0	0	1	0	0	0	0	1	1	1	4
10	0	0	0	0	0	0	0	0	0	0	2	2
11	0	0	0	1	0	0	0	0	0	0	1	2
12	0	0	0	0	0	0	0	0	0	1	0	1
13	0	0	0	0	0	0	0	0	3	0	2	5
14	0	0	0	0	0	0	0	0	0	0	0	0
15	0	1	0	0	0	0	0	0	0	0	1	2
16	0	0	0	0	0	1	0	1	1	0	0	3
17	0	0	0	0	0	1	0	0	0	1	0	2
18	0	0	1	0	0	0	0	0	1	0	1	3
19	0	0	0	0	0	0	0	0	0	0	0	0
20	0	1	0	0	0	1	0	0	1	1	0	4
21	0	0	0	0	0	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	0	1	0	1	2
24	0	0	0	0	0	0	0	0	0	0	0	0
25	0	0	0	0	0	0	0	0	0	0	0	0
26	0	0	0	0	0	0	0	0	0	0	0	0
27	0	0	0	0	0	0	0	0	0	0	0	0
28	0	0	0	0	0	0	0	0	0	0	0	0
29	0	0	0	0	0	0	0	0	0	0	0	0
30	0	0	0	0	0	0	0	0	0	0	0	0

31	0	0	0	0	0	0	0	0	0	0	0	0
32	0	0	0	0	0	0	0	0	0	0	0	0
33	0	0	0	0	0	0	0	0	0	0	0	0
34	0	0	0	0	0	0	0	0	1	0	0	1
35	0	0	0	0	0	0	0	0	0	0	0	0
36	0	0	0	0	0	0	0	0	0	0	0	0
37	0	0	0	0	0	0	0	0	0	0	0	0
38	0	0	0	0	0	0	0	0	0	0	0	0

Nearest IITs distance	pyruvate dehydrogenase phosphofructinase	glucose pyruvate	atp myosin	g-6-p starch	iaa starch	Glucose starch	cre cytokinin	carotenoid ipp	acetylcoa	nitrite xanthine	New verbs	Total	odds
0	2	36	40	31	3	18	38	5	22	34	73	302	1.720721
1	10	19	22	2	4	9	15	3	18	19	47	168	0.826087
2	6	12	7	3	3	10	4	1	15	9	29	99	0.622951
3	5	11	12	5	1	9	6	1	5	19	32	106	0.65625
4	2	7	7	6	5	6	4	1	5	10	20	73	0.403846
5	5	5	5	2	2	5	10	0	4	8	11	57	0.425
6	8	4	2	3	2	3	2	2	3	1	8	38	0.117647
7	4	4	6	3	0	6	2	2	4	1	12	44	0.419355
8	3	2	5	1	0	1	0	1	1	2	5	21	0.3125
9	1	2	1	2	0	1	2	1	4	1	5	20	0.25
10	3	1	0	0	1	0	1	0	0	0	7	13	0.181818
11	5	1	0	2	0	1	0	0	0	1	4	14	0.166667
12	1	0	1	1	3	1	0	1	2	1	4	15	0.071429
13	2	1	1	1	0	1	1	0	4	0	5	16	0.454545
14	1	0	0	0	0	2	0	0	0	0	0	3	0
15	0	1	0	0	0	0	0	1	0	1	3	6	0.5
16	1	1	2	0	0	1	0	1	1	0	1	8	0.6
17	1	1	1	1	0	1	1	0	0	2	2	10	0.25
18	0	0	1	0	0	0	0	0	1	0	4	6	1
19	1	0	0	1	0	0	1	0	0	0	0	3	0
20	0	1	0	0	0	2	0	0	1	1	0	5	4

21	1	1	0	1	0	1	0	0	1	1	1	7	0
22	0	0	0	0	0	0	0	0	0	0	1	1	0
23	0	0	0	0	0	0	0	0	1	0	2	3	2
24	0	0	0	0	0	1	0	0	0	0	1	2	0
25	0	0	0	0	0	0	0	0	0	0	0	0	#DIV/0!
26	0	0	0	0	0	0	0	0	0	0	0	0	#DIV/0!
27	0	0	0	0	0	0	0	0	0	0	2	2	0
28	1	0	0	0	0	0	0	0	0	1	0	2	0
29	0	0	0	0	0	0	0	0	0	0	0	0	#DIV/0!
30	0	0	0	0	0	0	0	0	0	0	0	0	#DIV/0!
31	0	0	0	0	0	0	0	0	0	0	0	0	#DIV/0!
32	0	0	0	0	0	0	0	0	0	0	0	0	#DIV/0!
33	0	0	0	0	0	0	0	0	0	0	0	0	#DIV/0!
34	0	0	0	0	0	0	0	0	1	0	0	1	#DIV/0!
35	0	0	0	0	0	0	0	0	0	0	0	0	#DIV/0!
36	1	0	0	0	0	0	0	0	0	0	0	1	0
37	1	0	0	0	0	0	0	0	0	0	0	1	0
38	1	0	0	0	0	0	0	0	0	0	0	1	0

Appendix VIII Acronyms List

MEDLINE	Medical Literature Analysis and Retrieval System (MEDLARS) Online
MIPS	Munich Information center for Protein Sequences
KEGG	Kyoto Encyclopedia of Genes and Genomes
BIND	Biomolecular Interaction Network Database
IR	Information Retrieval
POS	Part-Of-Speech
NLP	Natural Language Processing
DPBE	Dragon Plant Biology Explorer
ACS	Associative Concept Space
LMMA	Literature Mining and Microarray Analysis
tf	Term Frequency
idf	Inversed Document Frequency
BEV	Bird's Eye View
GO	Gene ontology
SLIM	Slider Interface for MEDLINE/PubMed searches
BLAST	Basic Local Alignment Search Tool
SGO	Semantic Gene Organizer
GENIES	GENomics Information Extraction System
GIS	Gene Information System
GIFT	Gene Interactions Finder in Text
BN	Bayesian Network
CRF	Conditional Rrandom Field

Appendix VIIII Birds Eye View

1. Introduction

As biotechnology rapidly progresses, more and more experimental data are becoming available. Extraction of knowledge and conclusions from those data are requirements of biological research. However, it can be hard to investigate so much data manually, to find useful genes from long lists of numbers and to extract useful biological knowledge. Here we introduce a software application to help biologists analyze voluminous experiment data, such as microarray data. It gives biologists a graphical “Birds Eye View” of experimental data, hence its name, “BirdsEyeView.” A lot of work has been done to visualize biological information like biocyc (Karp et al., 2002), reactome (Matthews et al., 2008) and FCModeler (Dickerson et al., 2003). Different from them, which directly draw the information, BEV maps information onto preexisting biological figures.

BirdsEyeView (BEV) is a part of the MetNet platform. At the center of the MetNet platform is the MetNet database (MetNetDB). This database has biochemical interactions and pathways, genomic, gene ontology, transcriptomic, proteomic, and metabolomic data on Arabidopsis thaliana and other plants.

BEV uses the information stored in MetNetDB and data in the experiment data file to give users overviews of the experimental data. It shows a unique view of experimental results by mapping the experimental entities (genes or others biomolecules) to different views based on those entities’ profile information in MetNetDB. In addition, it graphically (by color and distribution of entities among compartments) indicates experiment differences among entities across different compartments in different views. Currently, BEV provides cellular view, pathway view, Gene Ontology function view, and Gene Ontology process view functions, in which cell view is on a fixed cellular picture, pathway view mapping is dynamically implemented by Treemap (Bederson et al., 2002), and Gene Ontology view is on a series of hierarchy rectangles representing Gene Ontology groups. BEV accepts input data and then maps the data in the file to those four views to help users find interesting results about cellular compartments, pathways, and biological processes and functions.

2. System features

2.1 Program overview

BEV is a stand-alone Java application that accepts users' input data and maps the data to different views to give users a "bird's-eye view" of their input data. BEV reads the list of biological entity IDs in an input file and finds the corresponding biological information from MetNetDB including name, location, pathway, Gene Ontology information, and so on. Based on this information, the program shows the biomolecular entities in the input file, e.g., genes, as small graphical icons (circles or rectangles) in different places in the different views. In addition, the program adjusts the icons' colors based on the average value of entities represented by the icons. Then patterns of experimental input will be shown in different views so users can view useful information such as significant genes, compartments, pathways, Gene Ontology Processes, and so on. Users can switch among views and experiments to see different biomolecule's distributions.

2.2 Input data

The general use of BEV is to show experimental result, like microarray result data. For displaying microarray results, the input could be in several files: a chip list file containing a list of Affymetrix probe ID (chip probes) in the experiments, a file containing the list of experiments, or a file of numbers describing the value of experimental results. Alternatively the input could be one file containing probes, experiments, and numbers together. Besides probe ID's microarray data input above, BEV could accept other similar format input files as long as it provides a list of entities IDs like locus IDs acknowledgeable by BEV.

2.3 User Interface

The main interface is divided into 4 panels (Figure 1): the info panel, view panel, legend panel, and annotation panel. In the info panel, users can use buttons and drop-down lists to switch experiments and views, load files, and do other view-related operations.

The view panel is the largest panel, and holds the different views (displayed one at a time by default) including biomolecular entities in the input file shown as icons inside the views.

The legend panel shows the legend of the view panel and gives user options for the view panel, including color range values, cut values for showing biomolecular entities whose data values are over or under this value, number of entities per icon for icons representing multiple entities, and the maximum number of entities to show in the annotation panel before opening a new frame. The number of entities per icon has a minimum value based on number of entities in cell compartments and sizes of the compartments. Users can adjust the icons to make one icon represent more or less entities as long as it is over the minimum value.

The annotation panel is at the bottom. It shows detailed information about entities represented by a graphical icon the user selects in the view panel.

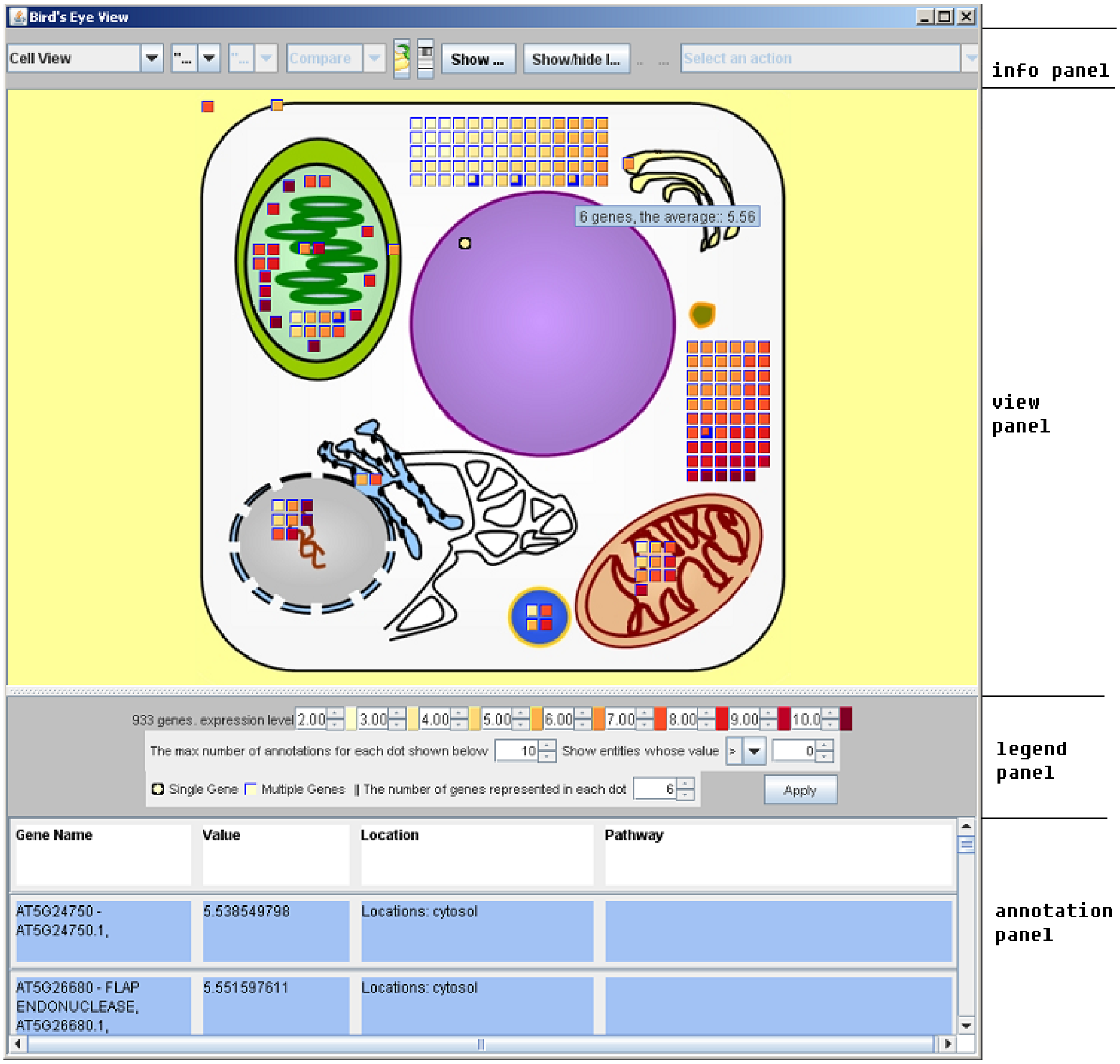
2.4 Cell View

In this view, an image showing the structure of a plant cell is displayed. The main compartments, nucleus, plastid, vacuole, mitochondrion, golgi apparatus, endoplasmic reticulum, apoplast, thylakoid, chloroplast, cytosol, and membranes of some compartments are shown with their cellular location relationships.

After loading the input files, the program displays entities from the list of input files in appropriate compartments by drawing graphical icons representing entities in compartments based on location information about each entity loaded from MetNetDB.

Color-coding the icons

An important function is finding entities that are significant relative to other entities. The experimental data is used to color the circle and rectangular icons. The program uses the YIOrRd (light yellow to orange to dark red) color system of the RColorbrewer color combination strategy (http://casoilresource.lawr.ucdavis.edu/drupal/node/192) (Brewer et al., 2002) to indicate different values (such as expression levels in microarray data). Higher values usually get darker colors. The program automatically adjusts the color distribution based on the current value range of the entities, so users can more readily notice differences between low and high values. However, users can adjust the color range to explore entity value distributions over a narrower value range. In addition, all entity icons in a compartment are arranged from low value (light colors) to high value (dark colors). Using this coloring strategy, users can easily notice significant entities or an interesting distribution of entities in a particular compartment.



933 genes. expression level 2.00 3.00 4.00 5.00 6.00 7.00 8.00 9.00 10.0

The max number of annotations for each dot shown below 10 Show entities whose value > 0

○ Single Gene □ Multiple Genes || The number of genes represented in each dot 6 Apply

Gene Name	Value	Location	Pathway
AT5G24750 - AT5G24750.1,	5.538549798	Locations: cytosol	
AT5G26680 - FLAP ENDONUCLEASE, AT5G26680.1,	5.551597611	Locations: cytosol	

Figure 29. Main interface

Automatic scaling

The size of each compartment space in the image of cell view and the icon size is used to calculate how many entity icons it can accommodate. When loading a large file holding a long list of entities, the number of entities in some compartments may exceed the maximum number of icons the area of the compartment in the image can accommodate. In this situation, the program will use one icon to represent multiple entities. The numbers of entities per icon is determined by the number of entities that need to be indicated and the space available in the image of the cell compartment. In this way, the application can accept inputs with any number of entities. When a unit represents several entities, its shape is a rectangle, rather than a circle, which is used for an icon representing only one entity. This information is noted in the legend panel.

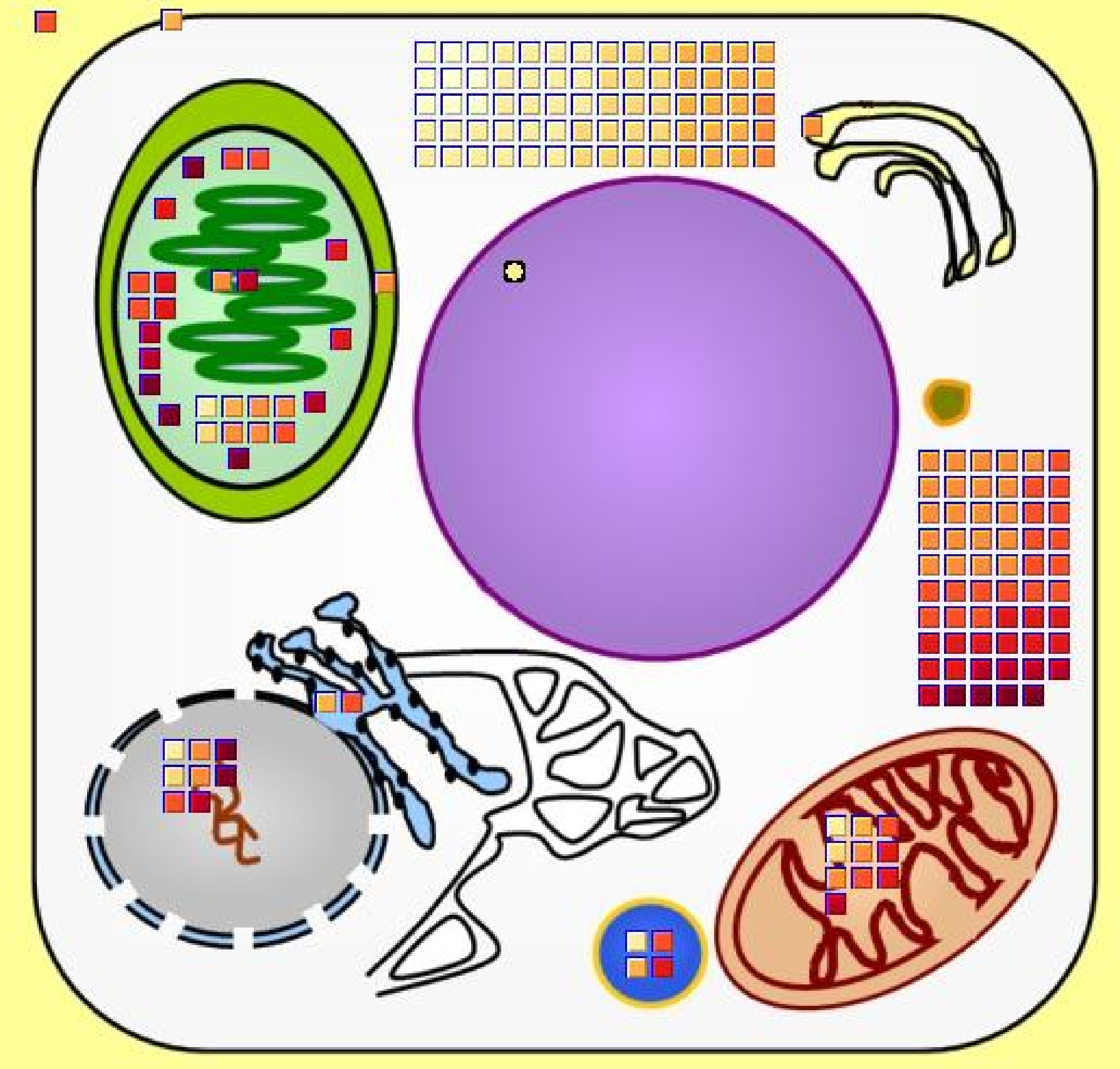


Figure 30. Cell view

Calling up detailed annotations

When an icon only represents one entity, if the user moves the mouse over it, basic information about the entity will be shown in a pop-up "tooltip" temporary window beside it. But when an icon represents multiple entities, the pop-up window only shows number of entities and their average value. To see more detailed information, the icon may be clicked. Then a list of rows, each containing the name, value, location, and pathway of an entity belonging to the icon, is shown in the bottom annotation panel.

Meanwhile, the icon will be highlighted by surrounding it with a small blue circle. When the user moves the mouse over the highlighted icon, the annotation panel will automatically scroll to rows with detailed information about the entities represented by the icon, if there are multiple highlighted icons and hence multiple groups of rows of detailed information. If the highlighted icon is clicked again, the highlighting and the icon's rows in the annotation panel will disappear. To remove all highlight statuses at once, the right mouse key may be clicked on any highlighted icon and the first item on the resulting pop-up menu selected. The second item on that menu is for viewing more detailed information. A new web browser will open to show information provided by another application in MetNet, AtGeneSearch (http://metnet.vrac.iastate.edu/MetNet_atGeneSearch.htm). This will display all information related to this entity stored in MetNetDB.

When an icon represents many entities, the list of rows in the annotation panel will be very long. Showing a list of too many rows and scrolling to some positions in such a list can tend to make users lose focus. To avoid that, the program will open a frame to show a list of rows of entity information when more than 10 entities belong to the icon. This number is adjustable in the legend panel. (If adjusted to 0, then the new frame will be always open when an icon is clicked.)

2.5 Pathway View

The pathway view shows the entities in the pathways in which they participate and represents pathways as nested rectangles. The pathways and their corresponding cellular compartments information are preloaded by the program from its configuration file and MetNetDB, and are independent from the input file. For the natural hierarchy relationships between cellular compartments and the pathways involved in, and the dynamic entities input, we used the Treemap (Bederson et al., 2002) method to display the results of mapping biomolecule entities onto different pathways as nested rectangles with entities icons. The application of Treemap here chooses the pattern of nested rectangles based on input file and preloaded pathway information. In general, the more entities a pathway contains, the larger size of the rectangle representing this pathway compared to other pathways' rectangles involved in the same cellular compartment. The program divides a pathway view rectangle into different smaller rectangles in two levels.

All these rectangles are contained in a larger rectangle for the overall pathway view. Within this encompassing rectangle, cell compartments are second-level rectangles. These in turn contain the third-level rectangles representing the pathways. Pathway rectangles in a cellular compartment rectangle have the same color as the color of the compartment in the cell view, so users can easily see the correspondence between pathways and compartments. Many entities participate in multiple pathways, and many pathways are in multiple compartments, so they will appear multiple times in this display. As in the cell view, the arrangement of entity icons into pathway rectangles is based on the pathway information about each entity loaded from MetNetDB. Users can do the same operations on the entities icons that were described earlier for the cell view.

2.6 GO Biological Process and Molecular Function Views

BEV provides a Biological Process and Molecular Function view in addition to cellular components, which can be viewed from the cell view. The GO Biological Process and Molecular Function view displays and maps entities based on their Gene Ontology information stored in MetNetDB. The Gene Ontologies are structured as directed acyclic graphs. These can be organized as hierarchies, and one entity may belong to several hierarchies.

The GO view shows each biomolecule entity in a Gene Ontology categories node in the categories graph as a small square (Fig. 32). The Gene Ontology (http://www.geneontology.org) contains three structured controlled vocabularies (ontologies): biological processes, cellular components and molecular functions.

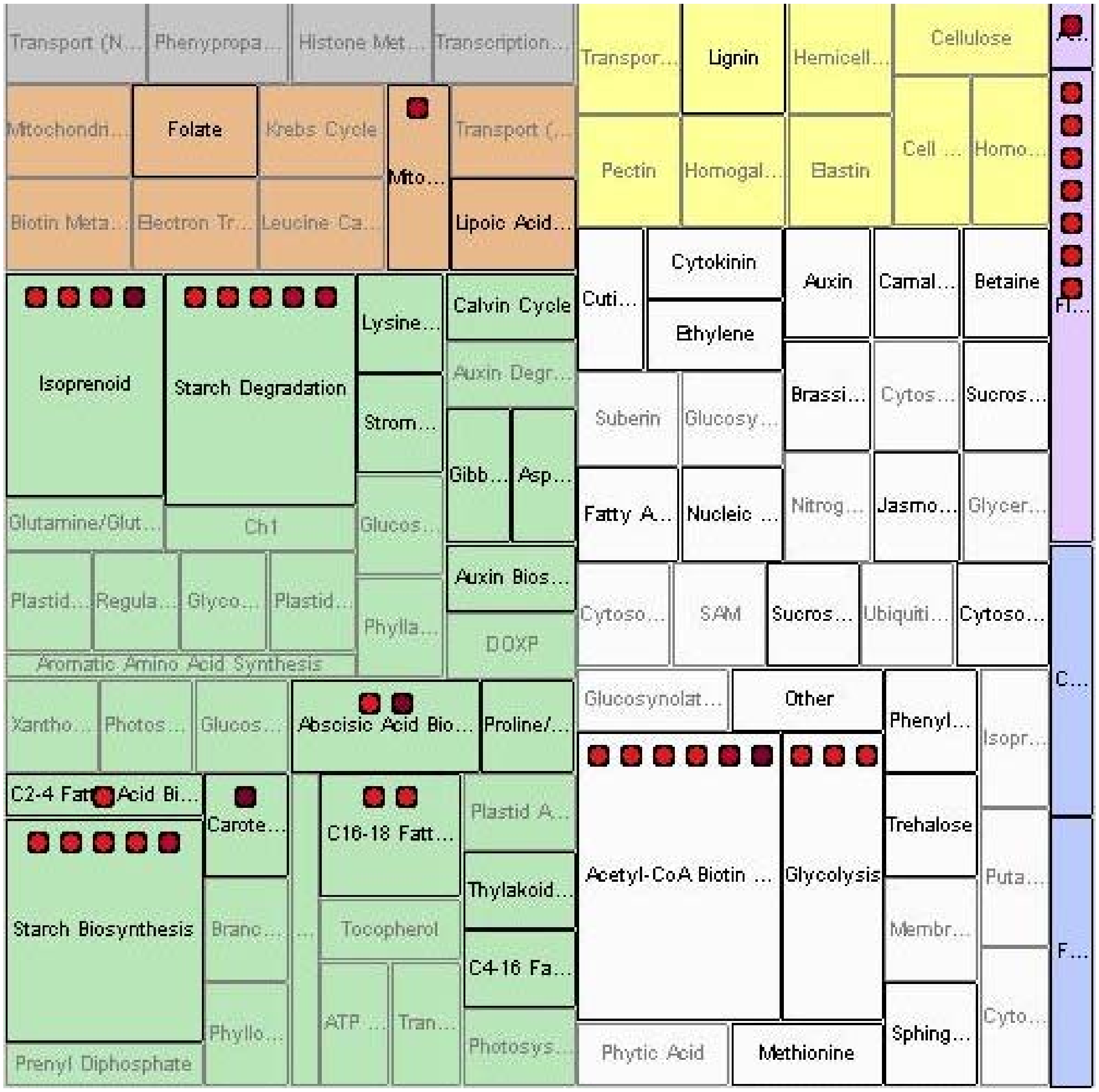


Figure 31. Pathway view

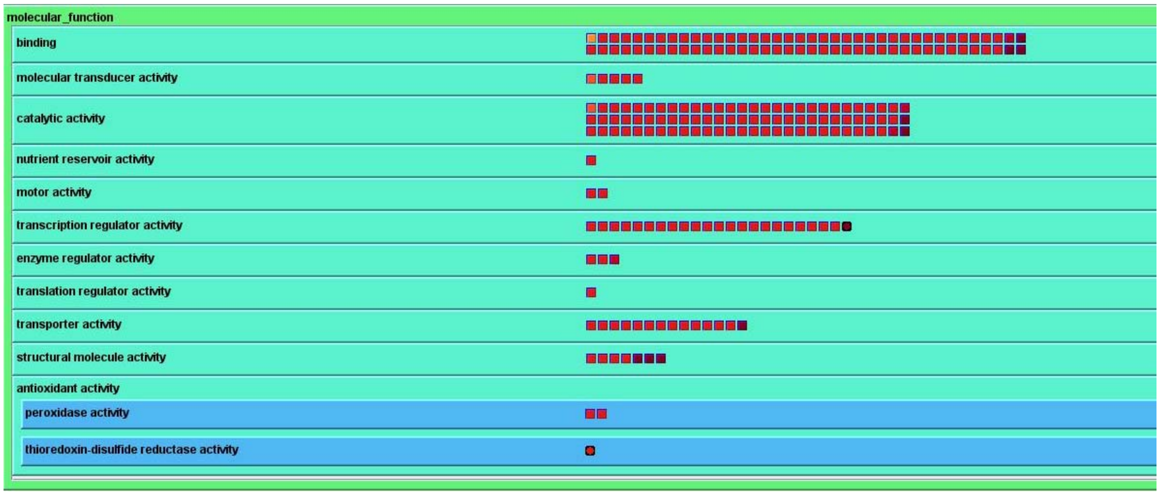

molecular_function
binding
molecular transducer activity
catalytic activity
nutrient reservoir activity
motor activity
transcription regulator activity
enzyme regulator activity
translation regulator activity
transporter activity
structural molecule activity
antioxidant activity
peroxidase activity
thioredoxin-disulfide reductase activity

Figure 32. Gene Ontology (GO) view

BEV first shows all inputted entities as graphical icons in one big rectangle. From there, users can click the "low" button in the info panel, or click the GO terms of a rectangle, hierarchy, etc., to show lower-level GO categories as smaller sub-rectangles holding entity icons. This process can be repeated, subdividing rectangles until the deepest (lowest) GO level. Users also can click the name of an open category, or the "up" button, to delete the sub-rectangles in the current rectangle and thus close the current category display. The colors of rectangles in different levels are different from each other and also different from entity icons' colors, so users can easily see the different GO hierarchy levels and the entity icons.

The design of entity icons in the GO view is the same as in the cell view: an icon can represent one or more entities, basic annotation information can be called up in pop-ups with more detail in the annotation panel, and fully detailed annotations can be called up in a new web browser window.

2.7 Comparison function and linked view display

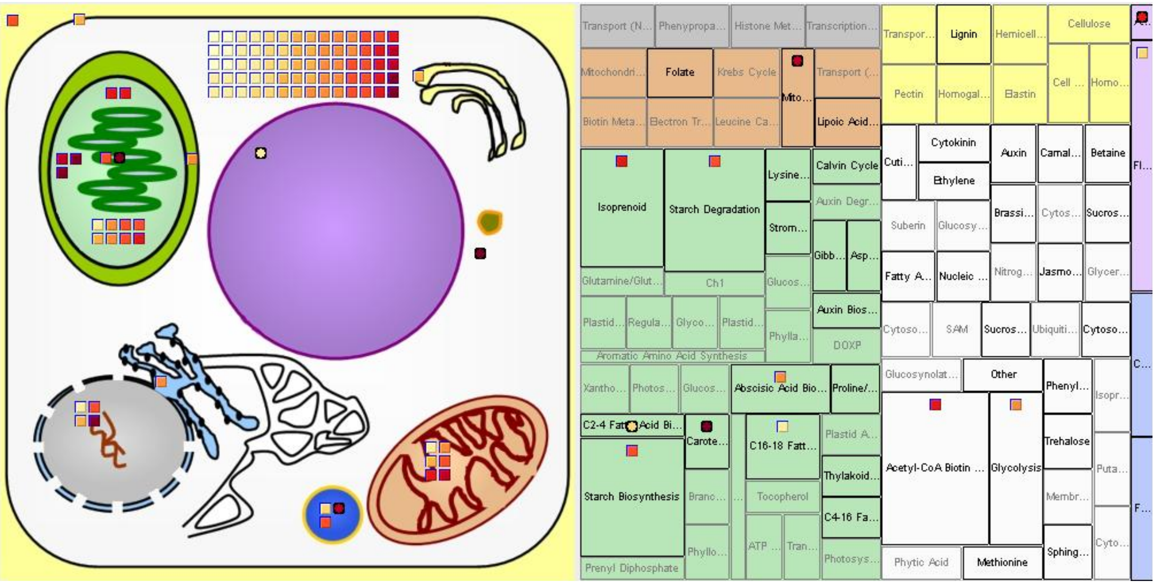


Figure 33. Two views together

Besides displaying one sample one time, BEV supports comparison of two samples. Users can see result value differences, folds (the first experiment's values divided by the second), and difference folds (result value differences divided by the second experiment's values) for the same entities in two different experiments. The display technique is similar to that used for a single experiment. The user simply needs to choose the two experiments and the comparison desired.

BEV also can show two views together. This allows users to track entities across different views. Users may choose this option from the view options box and then choose the two views to be shown. Then the entities in the input file will be displayed in two views at the same time. Users can do the same operations on the entity icons as in the single view. However, because each view in a two-view display has less screen space, some details of a view might not be the same as when only a single view is displayed. For example, in Figure 5, when cell view and pathway view are shown together, entity icons in the pathway view become multiple-entity rectangles instead of the larger number of circles that appear when displaying just the pathway view. By comparing views side by side, some cell, pathway, and GO Biological Process or Molecular Function properties of interesting entities can be shown simultaneously.

3. BirdsEyeView, PathBinder and MetNet

As mentioned before, BirdsEyeView (BEV) is a part of the MetNet platform and the MetNet bioinformatics platform is a suite of software applications designed for bioinformatics analysis. Each part of the suite provides different functions for biologists' needs. For example, PubMed Assistant provides an interface that help biologists search PUBMED more easily. They are all managed and supported by the MetNet platform, and can communicate with each other inside the platform, which means one can call another to provides the user with the function of the called component. For example, if BEV users find two interesting genes in their graphical view, they can ask BEV to send information about these two genes to PathBinder through MetNet platform. PathBinder will return extracted interaction descriptions to BEV and then to the users who called it to help explore the related literature about these two genes to users. In the reverse direction, PathBinder can ask BEV to show genes graphically if PathBinder's users need a graphical view for two genes. All software inside MetNet platform communicate and combine together to give biologists a comprehensive bioinformatics software suite.

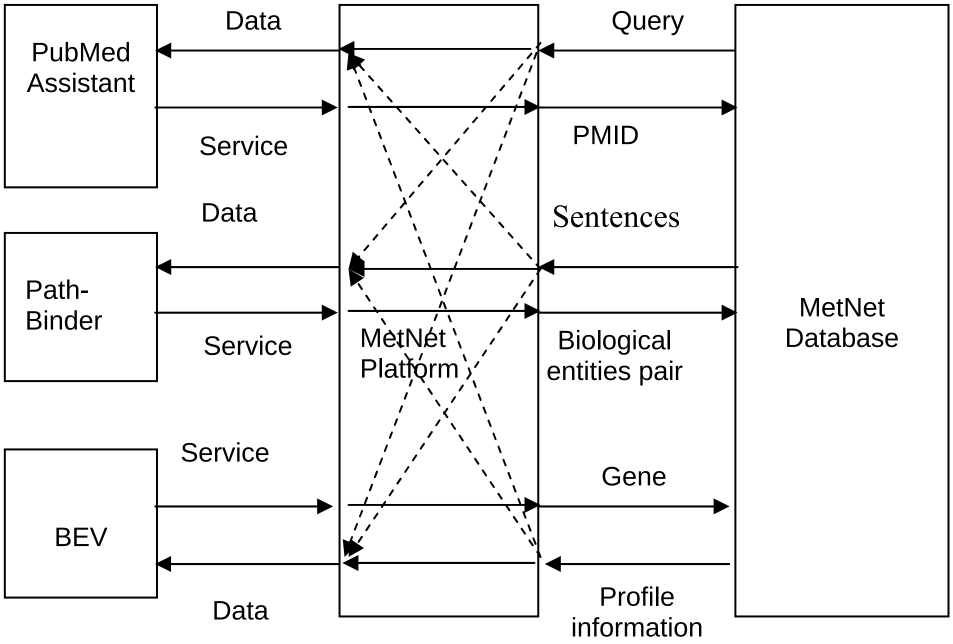


Figure 34. BEV, PathBinder, and MetNet

4. Conclusion

BEV provides cell, pathway, GO Biological Process, and GO Molecular Function views for experimental or similar data. Users can easily see important biological information about each entity in the input both graphically and with annotations. Users can flexibly zero in on interesting entities, cell compartments, pathways, biological processes, and molecular functions relevant to their experiments based on the locations and colors of the entity icons. These capabilities help support the needs of systems biologists to view experimental data both in broaden contexts as well as in considerable detail.

6. References

[1] Affymetrix (2002), Affymetrix Microarray Suite User Guide, Version 5 ed., Affymetrix, Santa Clara, CA.

[2] S. Albert, S. Gaudan, H. Knigge, A. Raetsch, A. Delgado, B. Huhse, H. Kirsch, M. Albers, D. Rebholz-Schuhmann, and M. Koegl (2003), Computer-assisted generation of a protein-interaction database for nuclear receptors, *Mol. Endocrinol.* 17:1555-1567.

[3] G. D. Bader, D. Betel, C. W. Hogue (2003), BIND: the Biomolecular Interaction Network Database, *Nucleic Acids Research* 31(1):248-250.

[4] G. D. Bader, I. Donaldson, C. Wolting, B. F. Francis Ouellette, T. Pawson, and W. V. Hogue (2001), BIND—The Biomolecular Interaction Network Database, *Nucleic Acids Research* 29:242-245.

[5] V. B. Bajic, M. Veronika, P. S. Veladandi, A. Meka, M. W. Heng, K. Rajaraman, H. Pan, and S. Swarup (2005), Dragon Plant Biology Explorer: A text-mining tool for integrating associations between genetic and biochemical entities with genome annotation and biochemical terms lists, *Plant Physiol* 138(4):1914-1925.

[6] B. B. Bederson, B. Shneiderman, and M. Wattenberg (2002), Ordered and Quantum Treemaps: making effective use of 2D space to display hierarchies, *ACM Transactions on Graphics (TOG)* 21(4)(October):833-854.

[7] D. Berleant (2004), Combining evidence: the Naïve Bayes model vs. semi-naïve evidence combination, Technical Report SARD04-11.

[8] R. Bowater, M. R. Webb, and M. A. Ferenczi (1989), Measurement of the reversibility of ATP binding to myosin in calcium-activated skinned fibers from rabbit skeletal muscle, *Journal of Biological Chemistry* 264:7193-7201.

[9] C. A. Brewer, W. H. Geoffrey, and A. H. Mark (2003), ColorBrewer in Print: a catalog of color schemes for maps, *Cartography and Geographic Information Science* 30(1):5-32.

[10] M. Bundschus, M. Dejori, M. Stetter, V. Tresp, and H. P. Kriegel (2008), Extraction of semantic biomedical relations from text using conditional random fields. *BMC Bioinformatics* 9:207-221.

[11] R. Bunescu, R. Ge, R. J. Kate, E. M. Marcotte, R. J. Mooney, A. K. Ramani, and Y. W. Wong (2005), Comparative experiments on learning information extractors for proteins and their interactions, *Artificial Intelligence in Medicine (special issue on Summarization and Information Extraction from Medical Documents)*, 33(2):139-155.

[12] J. Chaudière (1986), Possible role of glutathione peroxidase in the regulation of collagenase activity, *Ann Biol Clin* 44:181-7.

[13] J. Chiang, H. Yu, and H. Hsu (2004), GIS: a biomedical text-mining system for gene information discovery, *Bioinformatics* 20: 120-121.

[14] R. Chowdhary, J. Zhang, and J. S. Liu (2009), Bayesian inference of protein-protein interactions from biological literature, *Bioinformatics* 15:1536-1542.

[15] A.M. Cohen and W.R. Hersh (2005), A survey of current work in biomedical text mining, *Briefings in Bioinformatics,* 6:57-71.

[16] A. M. Cohen, W. R. Hersh, C. Dubay, and K. Spackman (2005), Using co-occurrence network structure to extract synonymous gene and protein names from MEDLINE abstracts, *BMC Bioinformatics* 6:103-118.

[17] N. Daraselia, S. Egorov, A. Yazhuk, S. Novichkova, A. Yuryev, and I. Mazo (2004), Extracting protein function information from MEDLINE using a full-sentence parser, *Proceeding of the Second European Workshop on Data Mining and Text Mining for Bioinformatics*, 11-18.

[18] N. Daraselia, A. Yuryev, S. Egorov, S. Novichkova, A. Nikitin, and I. Mazo (2003), Extracting human protein interactions from MEDLINE using a full-sentence parser, *Bioinformatics* 19:1-8.

[19] E. Davis (1990), *Representations of Commonsense Knowledge*, Morgan Kaufmann.

[20] J. A. Dickerson, D. Berleant, P. Du, J. Ding, C. M. Foster, L. Li, and E. S. Wurtele (2005), Creating, modeling, and visualizing metabolic networks, chap. 17 in H. Chen, S. S. Fuller, C. Friedman, and W. Hersh, eds., *Medical Informatics: Knowledge Management and Data Mining in Biomedicine*, Springer, 491-518.

[21] J. A. Dickerson and Z. Cox (2003), Using fuzzy measures to group cycles in metabolic networks, *North American Fuzzy Information Processing Society (NAFIPS) Annual Meeting*, Chicago, IL.

[22] J. Ding, D. Berleant, D. Nettleton, and E. Wurtele (2002), Mining MEDLINE: abstracts, sentences, or phrases? *Pacific Symposium on Biocomputing 7*, Kaua'i, Hawaii, 326-337.

[23] N. Domedel-Puig and L. Wernisch (2005), Applying GIFT, a Gene Interactions Finder in Text, to fly literature, *Bioinformatics* 21:3582-3583.

[24] T. Fayruzov, M. D. Cock, C. Cornelis, and V. Hoste (2009), Linguistic feature analysis for protein interaction extraction, *BMC Bioinformatics* 10:374-391.

[25] C. Friedman, P. Kra, H. Yu, M. Krauthammer, and A. Rzhetsky (2001), GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles, *Comput. Appl. Biosci.* 17: S74-S82.

[26] K. Fundel, R. Küffner, and R. Zimmer (2007), RelEx—Relation extraction using dependency parse trees, *Bioinformatics* 23(3):365-371.

[27] B. Gafurov, Y. D. Chen, and J. M. Chalovic (2004), Ca2+ and ionic strength dependencies of S1-ADP binding to actin-tropomyosin-troponin: regulatory implications, *Biophysical Journal* 87:1825–1835.

[28] T. Goetz and C. von der Lieth (2005), PubFinder: a tool for improving retrieval rate of relevant PubMed abstracts, *Nucleic Acids Res.* 33:W774-W778.

[29] G. R. Grimes, T. Q. Wen, M. Mewissen, R. M. Baxter, S. Moodie, J. S. Beattie, and P. Ghazal (2005), PDQ Wizard: automated prioritization and characterization of gene and protein lists using biomedical literature, *Bioinformatics*, 22:2055-2057.

[30] Y. Hao, X. Zhu, M. Huang, and M. Li (2005), Discovering patterns to extract protein–protein interactions from the literature: Part II, *Bioinformatics* 21: 3294-3300.

[31] W. Hersh (2005), Evaluation of biomedical text-mining systems: lessons learned from information retrieval, *Briefings in Bioinformatics* 6:344-356.

[32] L. Hirschman, J. C. Park, J. Tsujii, L. Wong, and C. H. Wu (2002), Accomplishments and challenges in literature data mining for biology, Bioinformatics, 18:1553-1561.

[33] O. Hofmann and D. Schomburg (2005), Concept-based annotation of enzyme classes, *Bioinformatics* 21:2059-2066.

[34] R. Homayouni, K. Heinrich, L. Wei, and M. W. Berry (2005), Gene clustering by Latent Semantic Indexing of MEDLINE abstracts, *Bioinformatics* 21 :104-115.

[35] M. Huang, X. Zhu, Y. Hao, D. G. Payan, K. Qu, and M. Li (2004), Discovering patterns to extract protein–protein interactions from full texts, *Bioinformatics* 20:3604-3612.

[36] L. Hunter, Z. Lu, J. Firby, W. A. Baumgartner Jr., H. L. Johnson, P. V. Ogren, K. B. Cohen (2008), OpenDMAP: an open source, ontology-driven concept analysis engine, with applications to capturing knowledge regarding protein transport, protein interactions and cell-type-specific gene expression, *BMC Bioinformatics* 9:78-89.

[37] R. Jelier, G. Jenster, L. C. J. Dorssers, C. C. van der Eijk, E. M. van Mulligen, B. Mons, and J. A. Kors (2005), Co-occurrence based meta-analysis of scientific texts: retrieving biological relationships between genes, *Bioinformatics* 21:2049-2058.

[38] R. Johnson (2005), J2EE development frameworks, *IEEE Computer* 38(1), Jan., 107-110.

[39] R. Kabiljo, A. B. Clegg, and A. J. Shepherd (2009), A realistic assessment of methods for extracting gene/protein interactions from free text, *BMC Bioinformatics* 10:233-245.

[40] M. Kanehisa, S. Goto, M. Hattori, K. F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, and M. Hirakawa (2006), From genomics to chemical genomics: new developments in KEGG, *Nucleic Acids Research* 34:D354-D357.

[41] P. D. Karp, C. A. Ouzounis, C. Moore-Kochlacs, L. Goldovsky, P. Kaipa, D. Ahrén, S. Tsokal, N. Darzentas, V. Kunin, and N. López-Bigas (2005), Expansion of the BioCyc collection of pathway/genome databases to 160 genomes, *Nucleic Acids Research* 33(19):6083-6089.

[42] P. Karp, D. Paley, and P. Romero (2002), The Pathway Tools Software, *Bioinformatics* 18:S225-32.

[43] J.-D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii (2003), GENIA corpus—a semantically annotated corpus for bio-textmining, *Bioinformatics* 19:i180-i182.

[44] J. J. Kim, Z. Zhang, J. C. Park, and S. K. Ng (2006), BioContrasts: extracting and exploiting protein-protein contrastive relations from biomedical literature, *Bioinformatics* 22(5):597-605.

[45] S. Kim, J. Yoon, J. Yang, and S. Park (2010), Walk-weighted subsequence kernels for protein-protein interaction extraction, *BMC Bioinformatics* 11:107-128.

[46] A. Koike, Y. Niwa, and T. Takagi (2005), Automatic extraction of gene/protein biological functions from biomedical text, *Bioinformatics* 21(7):1227-1236.

[47] J. Lafferty, A. McCallum, F. Pereira (2001), Conditional random fields: probabilistic models for segmenting and labeling sequence data, in *Proceedings of the Eighteenth International Conference on Machine Learning (ICML-2001)*, 2001.

[48] J. Lewis, S. Ossowski, J. Hicks, M. Errami, and H. R. Garner (2006), Text similarity: an alternative way to search MEDLINE, *Bioinformatics* 22: 2298-2304.

[49] S. Li, L. Wu, and Z. Zhang (2006), Constructing biological networks through combined literature mining and microarray analysis: a LMMA approach, *Bioinformatics* 22(17):2143-2150.

[50] Y. Li, X. Hu, H. Lin, and Z. Yang (2010), Learning an enriched representation from unlabeled data for protein-protein interaction extraction, *BMC Bioinformatics* 11 (Suppl 2):S7.

[51] C. D. Manning, R. Raghavan, and H. Schütze (2008), *Introduction to Information Retrieval*, Cambridge University Press.

[52] E.M. Marcotte, I. Xenarios, and D. Eisenberg (2001), Mining literature for protein-protein interactions, *Bioinformatics* 17(4):359-63.

[53] L. Matthews, G. Gopinath, M. Gillespie, M. Caudy, D. Croft, B. de Bono, P. Garapati, J. Hemish, H. Hermjakob, B. Jassal, A. Kanapin, S. Lewis, S. Mahajan, B. May, E. Schmidt, I. Vastrik, G. Wu, E. Birney, L. Stein, and P. D'Eustachio (2008), Reactome knowledge base of biological pathways and processes, *Nucleic Acids Res*. 37:D619-D622.

[54] D. M. McDonald, H. Chen, H. Su, and B. B. Marshall (2004), Extracting gene pathway relations using a hybrid grammar: the Arizona Relation Parser, *Bioinformatics* 20:3370-3378.

[55] MEDLINE, http://www.nlm.hih.gov/pubs/factsheets/MEDLINE.html.

[56] H. W. Mewes, C. Amid, R. Arnold, D. Frishman, U. Güldener, G. Mannhaupt, M. Münsterkötter, P. Page, N. Strack, V. Stümpflen, J. Warfsmann, and A. Ruepp (2004), MIPS: analysis and annotation of proteins from whole genomes, *Nucleic Acids Research* 32:D41-D44.

[57] M. Muin, P. Fontelo, F. Liu, and M. Ackerman (2005), SLIM: an alternative Web interface for MEDLINE/PubMed searches – a preliminary study, BMC *Medical Informatics and Decision Making* 5:37-46.

[58] H. M. Muller, E. E. Kenny, and P. W. Sternberg (2004), Textpresso: an ontology-based information retrieval and extraction system for biological literature, *PLoS Biol.* 2(11):1984-1998.

[59] J. Natarajan, D. Berrar, W. Dubitzky, C. Hack, Y. Zhang, C. DeSesa, J. R. Van Brocklyn, E. G. Bremer (2006), Text mining of full-text journal articles combined with gene expression analysis reveals a relationship between

sphingosine-1-phosphate and invasiveness of a glioblastoma cell line, *BMC Bioinformatics* 7:373-289.

[60] G. Nenadic, I. Spasic, and S. Ananiadou (2003), Terminology-driven mining of biomedical literature, *Bioinformatics*, 19:938-943.

[61] Y. Niu, D. Otasek, and I. Jurisica (2010), Evaluation of linguistic features useful in extraction of interactions from PubMed; application to annotating known, high-throughput and predicted interactions in I^{2}D, *Bioinformatics* 26(1):111-119.

[62] S. Novichkova, S. Egorov, and N. Daraselia, MedScan (2003), a natural language processing engine for MEDLINE abstracts, *Bioinformatics* 19:1699-1706.

[63] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa (1999), KEGG: Kyoto Encyclopedia of Genes and Genomes, *Nucleic Acids Research*, Jan 1; 27(1):29-34.

[64] T. Ono, H. Hishigaki, A. Tanigami, and T. Takagi (2001), Automated extraction of information on protein–protein interactions from the biological literature, *Bioinformatics* 17:155-161.

[65] P. Pagel, S. Kovac, M. Oesterheld, B. Brauner, I. Dunger-Kaltenbach, G. Frishman, C. Montrone, P. Mark, V. Stümpflen, H.-W. Mewes, A. Ruepp, and D. Frishman (2005), The MIPS mammalian protein–protein interaction database, *Bioinformatics* 21: 832-834.

[66] C. Perez-Iratxeta, A.J. Pérez, P. Bork, and M. A. Andrade (2002), Computing fuzzy associations for the analysis of biological literature, *Biotechniques* 32:1380-1385.

[67] C. Perez-Iratxeta, A.J. Pérez, P. Bork, and M.A. Andrade (2003), Update on XplorMed: a web server for exploring scientific literature, *Nucleic Acids Res.* 31(13):3866-3868.

[68] C. Perez-Iratxeta, A.J. Pérez, P. Bork, and M.A. Andrade (2001), XplorMed: a tool for exploring MEDLINE abstracts, *Trends in Biochemical Sciences* 26: 573-575.

[69] A. Pertsemlidis and H. R. Garner (2004), Text comparison based on dynamic programming, *IEEE Eng Med Biol Mag*, 23(6):66-71.

[70] PubMed, http://www.ncbi.nlm.nih.gov/entrez/query.fcgi.

[71] S. Ramachandran and D. D. Thomas (1999), Rotational dynamics of the regulatory light chain in scallop muscle detected by time-resolved phosphorescence anisotropy, *Biochemistry* 38:9097-104.

[72] A. K. Ramani, R. C. Bunescu, R. J. Mooney, and E. M. Marcotte (2005), Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome, *Genome Biology* 6:R40.

[73] A. Ramani, E. Marcotte, R. Bunescu, and R.J. Mooney (2005), Using biomedical literature mining to consolidate the set of known human protein-protein interactions, *Proceedings of the ACL-ISMB Workshop on Linking*

Biological Literature, Ontologies and Databases: Mining Biological Semantics, 46-53.

[74] D. Rebholz-Schuhmann, M. Arregui, S. Gaudan, H. Kirsch, and A. Jimeno (2008), Text processing through Web services: calling Whatizit, *Bioinformatics* 24:296-298.

[75] F. Rinaldi, G. Schneider, K. Kaljurand, M. Hess, C. Andronis, O. Konstandi, and A. Persidis (2007), Mining of relations between proteins over biomedical scientific literature using a deep-linguistic approach, *Artificial Intelligence in Medicine* 39:127-136.

[76] T. C. Rindflesch, B. Libbus, D. Hristovski, A. R. Aronson, and H. Kilicoglu (2003), Semantic relations asserting the etiology of genetic diseases, *AMIA 2003 Symposium Proceedings*, 554-558.

[77] B. Rosario and M. Hearst (2005), Multi-way relation classification: application to protein-protein interactions, *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 732-739.

[78] D. L. Rubin, C. F. Thorn, K. E. Klein, and R. B. Altman (2005), A statistical approach to scanning the biomedical literature for pharmacogenetics, *Journal of the American Medical Informatics Association*, 121-129

[79] R. Sætre, S. Kenji, and J. Tsujii (2007), Syntactic features for protein-protein interaction extraction, *Proceedings of the 2nd International Symposium on Languages in Biology and Medicine (LBM 2007)*, 6.1–6.14.

[80] L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie, and D. Eisenberg (2004), The Database of Interacting Proteins: 2004 update, *Nucleic Acids Research*, 32:D449-D451.

[81] C. Santos, D. Eggle, and D.J. States (2005), Wnt pathway curation using automated natural language processing: combining statistical methods with partial and full parse for knowledge extraction, *Bioinformatics* 21:1653-1658.

[82] C. Schnee, T. G. Köllner, M. Held, T. C. Turlings, J. Gershenzon, and J. Degenhardt (2006), The products of a single maize sesquiterpene synthase form a volatile defense signal that attracts natural enemies of maize herbivores, *Proc Natl Acad Sci (USA)*, 103:1129-34.

[83] H. Schoof, R. Ernst, V. Nazarov, L. Pfeifer, H. W. Mewes, and K. F. X. Mayer (2004), MIPS Arabidopsis thaliana Database (MAtDB): an integrated biological knowledge resource for plant genomics, *Nucleic Acids Research* 32, Database issue: D373-D376.

[84] B. J. Stapley, G. Benoit (2000), Bibliometrics: information retrieval and visualization from co-occurrence of gene names in Medline abstracts, *Pacific Symposium on Biocomputing* 5:526-537.

[85] L. Tanabe, U. Scherf, L. H. Smith, J. K. Lee, L. Hunter, and J. N. Weinstein (1999), MedMiner: an internet text-mining tool for biomedical information, with application to gene expression profiling, *BioTechniques* 27:1210-1217.

[86] L. Tanabe and W. J. Wilbur (2002), Tagging gene and protein names in biomedical text, *Bioinformatics* 18:1124-1132.

[87] J. M. Temkin and M. R. Gilder (2003), Extraction of protein interaction information from unstructured text using a context-free grammar, *Bioinformatics* 19:2046-2053.

[88] T. Theodosiou, L. Angelis, A. Vakalia, and G. N. Thomopoulos (2007), Gene functional annotation by statistical analysis of biomedical articles, *International Journal of Medical Informatics* 76:601-613.

[89] C. C. Van der Eijk, E. M. van Mulligen, J. A. Kors, B. Mons, J. van den Berg (2004), Constructing an associative concept space for literature-based discovery, *Journal of the American Society for Information Science and Technology* 55:436-444.

[90] V. N. Vapnik (1998) *Statistical Learning Theory,* John Wiley & Sons.

[91] J. D. Wren and H.R. Garner (2004), Shared relationship analysis: ranking set cohesion and commonalities within a literature-derived relationship network, *Bioinformatics*, 20(2):191-198.

[92] J. D. Wren, R. Bekeredjian, J. A. Stewart, R. V. Shohet, and H. R. Garner (2004), Knowledge discovery by automated identification and ranking of implicit relationships, *Bioinformatics* 20: 389-398.

[93] I. Xenarios, L. Salwínski, X. Duan, P. Higney, S. Kim, and D. Eisenberg (2002), DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions, *Nucleic Acids Research* 30:303-305.

[94] A. Yakushiji, Y. Tateisi, Y. Miyao, and J. Tsujii (2001), Event extraction from biomedical papers using a full parser in biocomputing, *Proceedings of the Pacific Symposium*, 408-419.

[95] D. Yao, J. Wang, Y. Lu, N. Noble, H. Sun, X. Zhu, D. G. Payan, M. Li, K. Qu (2004), PathwayFinder: paving the way towards automatic pathway extraction, *Proceedings of the Second Conference on Asia-Pacific Bioinformatics* 29:53-62.

[96] A. S. Yeh, L. Hirschman, and A. A. Morgan (2003), Evaluation of text data mining for database curation: lessons learned from the KDD Challenge Cup; *Bioinformatics* 19:i331-i339.

[97] Y. T. Yen, B. Chen, H. W. Chiu, Y. C. Lee, Y. C. Li, C. Y. Hsu (2006), Developing an NLP and IR-based algorithm for analyzing gene-disease relationships, *Methods Inf. Med*. 45:321-329.

[98] H. Yu and E. Agichtein (2003), Extracting synonymous gene and protein terms from biological literature, *Bioinformatics* 19:i340-i349.

[99] H. Yu, V. Hatzivassiloglou, C. Friedman, A. Rzhetsky, and W. J. Wilbur (2002), Automatic extraction of gene and protein synonyms from MEDLINE and journal articles, *Proc. AMIA Symp,* 919-923.

[100] A. Yuryev , Z. Mulyukov, E. Kotelnikova, S. Maslov , S. Egorov , A. Nikitin , N. Daraselia, and I. Mazo (2006), Automatic pathway building in biological association networks, *BMC Bioinformatics* 7:171-184.

[101] L. Zhang, D. Berleant, J. Ding, T. Cao, and E. Wurtele (2009), PathBinder—text empirics and automatic extraction of biomolecular interactions, *BMC Bioinformatics* 10(Suppl 11)**:**S18.

[102] S. W. Zhang, Y. J. Li, L. Xia, and Q. Pan (2010), PPLook: an automated data mining tool for protein-protein interaction, *BMC Bioinformatics* 11:326-332.

[103] D. Zhou and Y. He (2008), Extracting protein-protein interactions from MEDLINE using the Hidden Vector State model, *Intentional Journal of Bioinformatics Research and Applications* 4:64-80.