Published in final edited form as:
Bioinformatics. 2007 December 1; 23(23): 3232–3240. doi:10.1093/bioinformatics/btm495.

Mining experimental evidence of molecular function claims from the literature

Colleen E. Crangle[1,*], **J. Michael Cherry**[2], **Eurie L. Hong**[2], and **Alex Zbyslaw**[1]
Limsoon Wong

[1]Converspeech LLC, 60 Kirby Place, Palo Alto, CA 94301

[2]Department of Genomics, Stanford University, Stanford, CA 94025, USA

Abstract

Motivation—The rate at which gene-related findings appear in the scientific literature makes it difficult if not impossible for biomedical scientists to keep fully informed and up to date. The importance of these findings argues for the development of automated methods that can find, extract and summarize this information. This article reports on methods for determining the molecular function claims that are being made in a scientific article, specifically those that are backed by experimental evidence.

Results—The most significant result is that for molecular function claims based on direct assays, our methods achieved recall of 70.7% and precision of 65.7%. Furthermore, our methods correctly identified in the text 44.6% of the specific molecular function claims backed up by direct assays, but with a precision of only 0.92%, a disappointing outcome that led to an examination of the different kinds of errors. These results were based on an analysis of 1823 articles from the literature of *Saccharomyces cerevisiae* (budding yeast).

Availability—The annotation files for *S.cerevisiae* are available from ftp://genome-ftp.stanford.edu/pub/yeast/data_download/literature_curation/gene_association.sgd.gz. The draft protocol vocabulary is available by request from the first author.

Contact—crangle@converspeech.com

1 INTRODUCTION

As gene-related information continues to appear in the scientific literature, the need for efficient procedures to extract information from text grows more urgent. What is needed in addition to information extraction, however, is a way for extracted information to be compiled and interpreted. The work we present here addresses the problem of determining what molecular function claims are being made in a scientific article and what type of experimental evidence the article presents for those claims. To come to a conclusion of this kind requires more than information extraction. The relevant information is scattered throughout the document and the context of the information is as important as its content.

For example, the article by Sedman *et al*. (2000) presents experimental evidence for the claim that the Hmi1 protein exhibits ATP-dependent DNA helicase activity. Several passages in the text establish these claims. The following sentence in the Abstract section

*To whom correspondence should be addressed..

Conflict of Interest: none declared.

suggests in its use of the phrase ‘purified recombinant protein’ that a direct assay was performed: ‘The purified recombinant protein can unwind duplex DNA molecules in an ATP-dependent manner’. The title of Figure 3 in the paper—‘The recombinant Hmi1 protein is a DNA helicase’—has the protein name occurring in the same context as the phrase ‘DNA helicase’. Finally, conclusive evidence is given in the paragraph that begins with ‘We overexpressed and purified the Hmi1 protein in *Escherichia coli* by using the pGEX41-based expression system ...’ and continues with ‘The recombinant Hmi1 protein is an ATPase that is stimulated by single-stranded DNA ... The helicase activity of the Hmi1 protein was analyzed by using a partially double-stranded DNA substrate ...’.

This task of identifying gene-related claims and the evidence for those claims in scientific articles is one that curators of model organism databases perform every day. To foster consistent descriptions by curators of gene products across organisms and databases, the Gene Ontology (GO) project developed three structured controlled vocabularies. These describe gene products in terms of their associated biological processes, cellular components and molecular functions (Gene Ontology, 2000). These vocabularies are hierarchical, allowing a gene product to be associated with a high-level description such as signal transduction or lower-level descriptions involving specific receptor tyrosine kinases, for example. The entry in a model organism database that associates a GO term with a gene product is known as an annotation and it includes the source of the information along with an indication of the kind of evidence there is in the source for the association. Each entry in GO includes an identifier, the name of the GO term, a definition and a position within the hierarchy. In making an annotation, the curator takes into account not just the GO term name but also the definition of the GO term. Different kinds of evidence may be available within the same article to support annotating a gene product to different GO terms at different levels. For example, there might be a direct assay showing that a protein localizes to the mitochondrion, and a physical interaction suggesting localization to the mitochondrial matrix. A given experimental procedure may provide one kind of evidence in one context and another in another context. Curators can annotate a gene to both a parent and a child term in the hierarchy, and cite the same or different kinds of evidence for the annotations as appropriate. When curators use GO to annotate a gene product, they are not only extracting relevant pieces of information and assembling them to reach a conclusion, in choosing a GO term, they are performing an act of interpretation that links text in the source article to the language of the GO term and its definition.

Our approach to the problem of identifying gene-related findings in the scientific literature draws on the experience of model-organism database curators. We used the literature of the budding yeast *Saccharomyces cerevisiae* and annotations in the *Saccharomyces* Genome Database (SGD; www.yeast genome.org) to develop and test our approach (Issel-Tarver *et al*., 2000). In this article, we present the results of our analysis of 1823 articles in which over 2500 experimentally based claims were made, more than 840 of which concerned molecular function. These results are compared with results from other work on automated database curation—in particular, the Knowledge Discovery and Data Mining (KDD) challenge (Yeh *et al*., 2003) and the BioCreAtIvE challenge (Critical Assessment of Information Extraction Systems in Biology) (Blaschke *et al*., 2005; Camon *et al*., 2005; Hirschman *et al*., 2005). Although our focus was on molecular function claims and the experimental evidence backing up those claims, the approach we present applies to other kinds of gene-related information and other kinds of evidence.

2 SYSTEM AND METHODS

2.1 Data

A gene-association file containing 32 708 *S.cerevisiae* annotations was provided by the SGD group (version dated 12 November 2006 of the gene_association.sgd file). These annotations were abstracted by the SGD curators from published articles, which had been selected as being of possible interest for yeast annotation based on a weekly-automated search of MEDLINE citations in the Entrez PubMed database. The annotation file listed 5919 documents in total. Of these, 2193 were available in full-text form, converted automatically from portable document format (pdf) files available from the publishers. In the end, we further restricted ourselves to documents that had both a Results and a Methods section that could be identified using automated heuristics. In total, there were 1834 such full-text documents. Eleven were used as training documents along with three additional papers on yeast, two on *Arabidopsis*, two on *Caenorhabditis elegans*, two on mice and two on humans. The remaining 1823 full-text documents gave rise to 3830 annotations based on evidence from the primary literature (the article itself) as opposed to other secondary sources. Of these 3830 annotations, 2117 were inferred from direct assays, a very reliable form of experimental evidence. Of these 2117 annotations, 607 were for molecular function. These 607 molecular-function direct-assay annotations were produced from 1250 papers. We compiled a list of candidate genes for each of the 1823 full-text articles by identifying in the SGD annotations all genes that had been annotated using evidence from these articles, whatever the kind of evidence or sub-ontology, whether biological process, cellular location or molecular function. These SGD annotations were used as the gold standard against which we measured our system's results.

2.2 Document representation

A careful analysis of the 22 papers used for training led us to represent documents as sets of triples $<s, g, w>$, where s is the name of the section of the article within which word or phrase w appears, g is a gene or gene product name that appears within the same sentence or sentence fragment as w and w is a member of the set of biologically meaningful words and phrases that appear in the document. The section names are specified as follows: $s \in \{T, A, N, R, C, D, M\}$, where T denotes Title, A Abstract, N Introduction, R Results, C Conclusion, D Discussion and M Methods (or sections with similar labels). A word or phrase w is considered biologically meaningful if it can be matched, with some degree of flexibility, against the terms in either a federated biomedical language resource that we have used in prior work or a draft protocol vocabulary designed to capture the words and phrases commonly used to describe relevant experimental procedures (Crangle, 2002; Crangle and Zbyslaw, 2004). The federated biomedical language resource has been used to improve text information access (Biomedical Information Science and Technology Initiative 2003 Symposium, http://www.bisti.nih.gov/2003meeting/abstracts/; GO Users Meeting, Stanford University, January 2004, http://www.geneontology.org/meeting/Stanford_GO_Program2004.html#6) and for concept extraction and synonymy management in biomedical information retrieval (TREC2004 Meeting, November 2004, NIST, http://trec.nist.gov/pubs/trec13/papers/converspeech.geo.pdf). This document representation takes into account the contribution of a given word or phrase, not simply on its own, but in the context of a specific section. We had initially thought that the Methods section would be the source of the most useful information on experimental evidence. Our preliminary investigations, however, suggested that, on the contrary, it is the appearance of a protocol vocabulary term in the Results section that tells us the most about what kind of evidence there is for a molecular function claim.

2.3 Biomedical language resources

Our biomedical language resources (known collectively as BioMedPlus) are modeled on WordNet, a general lexical resource for the English language (Miller *et al*., 1990). BioMedPlus contains over one million unique entries (words or phrases) and is built from publicly available biomedical resources, including the *S.cerevisiae* database of gene and gene product names, LocusLink (now superseded by the Entrez Gene database) and MeSH, the medical subject headings used to index MEDLINE records. A word or sequence of words from an article (loosely called a phrase) is matched to entries in BioMedPlus using a normalizing matching process. The matching process has the following features: word order is ignored; words such as articles and prepositions that are not significant in this context are ignored; hyphens and other non-alphabetic characters such as commas and digits break words up into subordinate words for matching; words are stemmed using a modified Porter stemmer (Porter, 1980) and case is taken into account as follows: a word in BioMedPlus that has any uppercase letters (e.g. the partial gene name 'ACT') will only match words in the text that also contain some uppercase letters (e.g. 'Act' or 'ACT' but not 'act'). This matching strategy is not flawless (it will fail to identify 'act1' as a reference to a recessive mutant to the ACT1 gene, for example) but it does eliminate many false matches to ordinary English words. For the task reported in this article, we did not use the MeSH or LocusLink derived portions of BioMedPlus.

We compiled a protocol vocabulary, currently in draft form, to capture the language that scientific curators look for in an article to determine that there is experimental evidence for a molecular function claim. We focused on direct assay experiments, using several primary sources from which we manually selected vocabulary items. The presence of any of the selected words or phrases in an article suggests that a direct assay has been performed. The sources were as follows: the children of the GO term 'protein complex' in the cellular component sub-ontology; the GO definition of the GO term 'protein complex' and the definitions of all its children; enzyme names, including recommended name, systematic name and synonyms (Bairoch, 2000; Barthelmes *et al*., 2007); the GO definition of the GO term 'enzyme' and headings and subheadings for selected chapters in Current Protocols in Molecular Biology, specifically chapters 9, 10, 12, 14, 16, 17, 18, 20, 21 and 27 (Ausubel *et al*., 2007). There are approximately 550 general entries (words or phrases) in the draft protocol vocabulary Terms from this vocabulary were matched against document text using the same matching process described above. Further work on this vocabulary will draw on the Ontology for Biomedical Investigations (OBI) effort, which seeks to model the design, protocols, instrumentation, material, data and analysis of biomedical investigations (Whetzel *et al*., 2006). Correctly identifying references to genes and gene products is crucial. We were aided in this by the well-regulated gene naming guidelines adopted by the yeast community and maintained by SGD. Gene products are named by capitalizing only the initial letter of the gene name and either appending the letter 'p' or following the gene name with the word 'protein'. For example, the gene *SEC11* has gene product Sec11p or Sec11 protein. Standard gene names and aliases are available in the SGD annotations file. To identify a gene or gene product name in the document text, we used the same matching process previously described.

2.4 GO term representation

Identifying GO descriptions of molecular function in text depends crucially on how GO terms are represented. It is this representation that determines what automated methods look for when they are trying to find a GO description of a molecular function in an article. A basic representation takes the individual words that make up a GO term, along with the individual words that make up its synonyms, and puts them together to form an unordered list. We extend this basic representation by additionally including in the list biologically

related words and phrases of the GO term that are provided by a number of so-called 'GO mappings'. These mappings, available at http://www.geneontology.org/external2go/, list terms that, in some context or another, are used in a roughly synonymous way with their associated GO term. Because these terms are not part of the controlled vocabulary of GO, they are not given as synonyms in GO. The following mappings were used: Swiss-Prot keywords; InterPro; MultiFun Classifications; EGAD and COG Functional Categories. GO terms were found in text using the normalizing matching process described earlier, in which word order is not significant, words are stemmed before matching, case is significant but does not require exact matching and hyphens and other non-alphabetic characters demarcate the most basic units for matching. In addition, several words such as 'activity' were considered optional in that they were not required in the document text for a match to be made.

2.5 Approach

The problem of finding molecular function claims backed by experimental evidence in a scientific article was broken down into the following two subtasks:

Subtask I: determine whether or not there is experimental evidence for a molecular function claim about a given gene, without reference to any specific molecular function.

Subtask II: identify potential molecular function terms in GO that may be associated with the gene for which experimental evidence has been found.

2.5.1 Subtask I: finding experimental evidence—This analysis was conducted for each article and for each candidate gene identified in Section 2.1.

Find all occurrences of the given gene or gene product name within the Results section.

Then find all terms from the draft protocol vocabulary that occur in the same sentence or sentence fragment as this gene or gene product name within the Results section.

For each protocol term so found, produce the following 6-tuple, where PMID is the PubMed identifier of the full-text document being analyzed; Gene is the name of the gene; P_ID identifies the cooccurring term from the draft protocol vocabulary; Co_Occur gives the number of times both the gene name and protocol term appear in the same sentence; Sec_Size gives the number of sentences in the Results section; and Score is the ratio of Co_Occur to Sec_Size.

<PMID; Gene; P_ID; Co_Occur; Sec_Size; Score>

As long as there is at least one such 6-tuple for a given gene, it is assumed that there is experimental evidence for a molecular function claim about that gene in that article.

The information given in these 6-tuples is over-specified. For example, for the gene *SEC15* and the paper with PMID 10022848, over a dozen 6-tuples were produced, including these:

<10022848, SEC15, PL:0000029, 1, 376, 0.00266>
<10022848, SEC15, PL:0000075, 1, 376, 0.00266>
<10022848, SEC15, PL:0000092, 3, 376, 0.00798>
<10022848, SEC15, PL:0000095, 13, 376, 0.03457>

That is, over a dozen different protocol terms occurred in the context of 'SEC15' within the Results section in this article. For the current task, it does not matter which protocol terms were found. All the 6-tuples for *SEC15* and the paper with PMID 10022848 can therefore be merged, with the values for Co_Occur and Score respectively being added up to produce totals for one 6-tuple for the given article and gene. Several scoring heuristics were tried. For example, we experimented to see what would happen if we ignored those 6-tuples whose score was less than 0.01. It turned out, however, that none of the scoring heuristics made any appreciable difference and they were discarded for the current task.

2.5.2 Subtask II: finding molecular function GO terms—Our approach to this second, more difficult, subtask was as follows. Subtask I established that there (purportedly) was experimental evidence in the given article for a molecular function annotation of some kind for the given gene. For each 6-tuple generated in subtask I, our system found those GO terms that occurred in the same sentence or sentence fragment as the gene or gene product name in the 6-tuple, but only in sections other than Methods and References. That is, for every 6-tuple,

<PMID; Gene; P_ID; Co_Occur; Sec_Size; Score>,

one or more 4-tuples of the following form were generated for each co-occurring GO term:

<PMID; Gene; P_ID; GO_ID, Excerpt>,

where GO_ID is the identifier of the GO term, and Excerpt is the sentence or sentence fragment in which both the GO term and the gene name occurred. The GO term in every such 4-tuple counted as a proposed molecular function description for the given gene.

3 IMPLEMENTATION

3.1 System implementation and evaluation

Our methods for the two subtasks were implemented in the Perl programming language, running on a variety of PCs using both Fedora Linux and FreeBSD. For the biomedical language resources, we used the open-source Berkeley DB database library.

To evaluate the system, we compared our results against the gold standard provided by the SGD annotations. We used the standard measures of recall and precision and the harmonic mean F of recall and precision. Recall is the ratio of the number of correct answers returned by the system to the actual number of correct answers. Precision is the ratio of the number of correct answers returned by the system to the total number of answers returned by the system. The 'answer' may be a yes/no decision on the existence of experimental evidence or it may be a GO term. Recall measures how sensitive a system is in finding what it should find. Precision measures how specific a system is in finding just what it should find. Recall and precision are typically given as percentages. The *F* measure is defined as follows:

$$F = \frac{(2 \times \text{precision} \times \text{recall})}{(\text{precision} + \text{recall})}$$

The SGD annotations each contain an evidence code, which could be any one of the following defined by the GO Consortium: Inferred from direct assay (IDA); inferred from physical interaction (IPI); inferred from expression pattern (IEP); inferred from genetic interaction (IGI); inferred from mutant phenotype (IMP); inferred from sequence or

structural similarity (ISS); inferred by curator (IC); inferred from electronic annotation (IEA); inferred from genomic context (IGC); inferred from reviewed computational analysis (RCA); traceable author statement (TAS) and non-traceable author statement (NAS). We had a particular interest in the IDA evidence code because a direct assay is considered a very reliable form of experimental evidence. However, inferences from expression pattern, genetic interaction, mutant phenotype and physical interaction are also based on experimental evidence. Furthermore, inferences from sequence or structural similarity are made on the basis of information in the primary literature, namely the article itself, as opposed to information in some secondary source. This kind of evidence was therefore also of interest to us.

3.2 Results for subtask I

For subtask I, we analyzed each final 6-tuple as follows. If the SGD annotations contained an entry that had the given gene and PubMed identifier and the appropriate kind of evidence, then the 6-tuple was counted as a correct answer. So for example, if we were considering the evidence codes IDA, IEP, IGI, IMP and IPI as representing a claim of experimental evidence, then any annotation with the given gene name and PubMed identifier and any of IDA, IEP, IGI, IMP or IPI as evidence code would give rise to a correct answer for our results. If we restricted our analysis to direct assays (IDA code), then only annotations with the given gene name and PubMed identifier and an IDA evidence code would give rise to a correct answer.

Correct answers, also known as true positives, can be compared to incorrect answers of two different kinds—false positives and false negatives. While a true positive is one for which the claimed annotation does exist—that is, the given document was used to support a molecular function annotation for the given gene and it provided experimental evidence of the relevant kind—a false positive is one in which there was no molecular function annotation supported by the relevant kind of evidence in the gold-standard annotations. A false negative is one for which a molecular function annotation for the given gene does exist and was derived by the curators from the given article, but our system failed to generate a 6-tuple containing the gene name and the PubMed identifier.

To interpret the results correctly we had to filter the false positives by removing all 6-tuples for which the given gene did have an experimentally based molecular function annotation, as confirmed by an entry in the gold-standard annotations, even though that annotation cited a document with a different PubMed identifier. This filtering was necessary because SGD provides only one article for an annotation. (Other model organism databases provide all articles found with information for a given gene.) Consequently, our system's claim may be correct; this article may simply be one of the papers the curators put aside during triage in relation to the given gene and GO term because a relevant annotation already existed. The annotations removed during filtering cannot automatically be counted as true positives, however. The particular document that we claim supports an experimentally based annotation for the given gene may or may not actually do so; there is simply no way of telling without having a scientific curator analyze the paper itself.

We computed recall, precision and *F* scores for several combinations of GO evidence codes: IDA (direct assay) only; IDA and IPI (physical interaction) and IDA, IPI, IEP, IGI, IMP and ISS—that is, any of the six primary literature evidence codes, omitting IC, IEA, IGC, RCA, TAS and NAS. Results are summarized in Table 1. For the most difficult task—identifying evidence from direct assays only (IDA evidence code)—our methods achieved recall of 70.7% and precision of 65.7%. When physical interactions (IPI evidence code) were also included, our methods achieved recall of 70.0% and precision of 77.0%. When all direct sources of evidence in the text were included, our methods achieved recall of 69.7% and

precision of 97.7%. Our third set of results included annotations inferred from sequence similarity (ISS evidence code), a method not always considered experimental. However, in the complete set of SGD documents, a full 77% of the articles that contributed an ISS annotation also contributed an experiment-based annotation, and only 11% of those articles also produced a non-experiment-based annotation. There is thus a high level of agreement between articles with claims based on sequence similarity and articles with claims based on more direct experimental evidence.

3.3 Results for subtask II

As we did for subtask I, we calculated filtered results for the molecular function annotations returned by the system for subtask II. Because SGD provides only one article for an annotation, when our system proposes a GO annotation for a gene based on a particular article, it may be correct even if there is no gold-standard annotation for that gene and GO term that contains the PubMed identifier for that particular article. This article may simply be one the curators did not include because a relevant annotation had already been established using another article. It may also be that this article was used by the curators to assign a molecular function term at a higher or lower level in the GO hierarchy, or with a non-experimental evidence code, and the GO term or evidence code suggested by our system is preferable. But these possibilities were simply counted as false positives.

Results for subtask II are summarized in Table 2. A true positive for subtask II is an answer given by our system that is in alignment with the gold-standard annotations; that is, the molecular function description (the GO term) that our system identified for the given gene *was* derived from the given article by the curators. A false positive for subtask II is an answer given by our system that proposed a molecular function description (a GO term) for the given gene that the curators did *not* derive, at least not using the given article. (For the filtered results, neither the given article nor any other was used to derive the molecular function description for the given gene.) A false negative for subtask II is given by a gold-standard annotation with a molecular function description (GO term) for the given gene that our system failed to identify. That is, our system failed to generate even one 4-tuple containing the given gene name and GO term and PubMed identifier found in the gold-standard annotation.

For direct-assay annotations, our methods correctly identified 44.6% of the molecular function descriptions associated with a given gene, an encouraging result. However, precision for this recall result was only 0.92%. When physical interactions (IPI evidence code) were also counted, recall dropped to 37.9% and precision rose slightly to 1.10%. When all direct sources of evidence in the text were counted, recall dropped further to 28.1% and precision again rose slightly to 1.49%.

3.4 Comparison with results from other database curation endeavors

Several groups have developed systems to discover gene-related findings in text (Chiang and Yu, 2003; Couto *et al*., 2005; Hu *et al*., 2005; Kim and Park, 2004; Koike *et al*., 2005; Pérez *et al*., 2004; Rebholz-Schuhmann *et al*., 2006). Most work only on MEDLINE data, which includes the title and abstract, not full-text articles. There have been two significant world-wide efforts focused on full-text data mining in biomedicine, both concerned with the problem of automating human and model-organism database curation: the KDD (Yeh *et al*., 2003) and BioCreAtIvE (Blaschke *et al*., 2005; Camon *et al*., 2005; Hirschman *et al*., 2005) challenges. These provide results for comparison with ours. However, it should be noted that the KDD corpus and the BioCreAtIvE corpus (for the comparative task) consisted of only 213 and 99 papers, respectively, whereas ours consisted of 1834 papers.

The KDD challenge sought methods of automating the curation of articles in FlyBase, the model organism database for *Drosophila* (Flybase Consortium, 2003). Given a set of full-text papers on the genetics or molecular biology of *Drosophila* and, for each paper, a list of the genes mentioned in that paper, participants were asked to develop a system that determined whether or not the paper should be curated based on the existence of experimental evidence in the paper and then, for each of the genes listed for the paper, whether or not the paper contained experimental evidence for that gene's expression. For KDD, 18 teams submitted 32 separate results for evaluation. The *F* measure was used to evaluate the yes/no curation decisions and the yes/no decisions for specific gene products. The best and the median *F* scores in the KDD competition were:

Yes/No experimental evidence in general: Best: 78%; Median: 58%.

Yes/No experimental evidence for gene products: Best: 67%; Median: 35%.

The training corpus made available to participants consisted of 862 articles curated in FlyBase together with the associated lists of genes and gene products. The supporting text that FlyBase had extracted from each paper was also provided. It consisted of words and phrases from the text accompanied by explanations of the evidentiary value of these extracts. The test corpus consisted of 213 new articles; 91 papers (43%) of the 213 test papers had results of interest.

There are several points of difference between the KDD effort and ours. Our *F* scores were computed on annotations not papers. Had we computed our scores on papers (number of papers we claimed had experimental evidence versus number of papers with actual experimental evidence, for example), we could not have made a comparison with KDD since our test set included only papers that had indeed been curated by SGD, none that had been put aside in the triage process as being of lower priority, perhaps not offering curatable material. Of the 213 documents in the KDD test set, 57% had no data of interest. The KDD task could therefore be seen to have provided more opportunities for false positives. However, of the 1823 documents in our test set, a full 56% gave rise to non-experimental annotations, regardless of whether or not they also gave rise to experimental annotations. Our test documents therefore gave ample and roughly equivalent opportunity for generating false positives on annotations, the basis of our measurements.

The KDD task comparable to ours determined whether or not specific genes had experimental data in a given paper. For that task, the best *F* score was 67%, the median 35%. Our second set of results (for IDA and IPI evidence codes) provides the fairest comparison. There our *F* score was 73.3%.

There were several areas of similarity between the successful approaches in KDD and our approach—for example, the attention paid to document structure, to phrases rather than individual words, and to figure captions. The most successful approach (Regev *et al*., 2002) focused on figure captions, looking for linguistic patterns containing words and phrases commonly used to describe relevant experimental procedures. This approach also used specific vocabulary items to rule out unwanted experimental evidence, such as from mutant phenotypes. We compiled a much richer vocabulary of experimental terms and, although we did not focus on figure captions, we specifically included them. Many of the sentence or sentence fragments that gave rise to the 6-tuples were from figure captions. Our protocol vocabulary seems to have provided considerable leverage in identifying experimental procedures of interest. A problem cited in the KDD report is that for a procedure such as immunolocalization, the text may nowhere use that term itself but instead contain a description of the specific steps taken to perform a particular immunolocalization assay, as in this figure caption: ‘Whole-mount tissue staining using an affinity-purified anti-Phm antibody in the CNS and in non-neural tissues. ... The third instar larval CNS exhibits

distributed cell body and neuropilar staining. This view displays only a portion of the CNS;...' Our draft protocol vocabulary includes the term 'affinity purification', which also matches against linguistic variations of the term, such as 'affinity-purified' appearing in the figure caption. We would therefore have identified a relevant experimental procedure from this piece of text.

BioCreAtIvE focused on the automatic assignment of GO annotations to human proteins using full-text articles. Task 2.2 of BioCreAtIvE is most relevant to the work reported here. For this task, the systems were given the number of GO terms there were for a given protein in a paper. Systems were to identify the correct GO terms assigned to the protein and provide supporting text passages. There were seven participants with 18 submissions in total. The training corpus consisted of 803 papers. The test corpus consisted of 99 papers, which gave rise to 1227 annotations. In terms of identifying the correct GO term, the highest recall, computed from Table 5 in Blaschke *et al*. (2005), was 6.4%, with precision of 12% (reported as 12.30% or 78/634 in Table 5), giving an *F* score of 8.4%—the highest obtained overall (Ehrler *et al*., 2005). The highest precision was reported as 34.62% or 9/26 in Table 5 (but given the size of the numerator and denominator is more precisely stated as 30%) with recall of 1% (Chiang and Yu, 2004).

The problem we solved was more difficult than that solved by the BioCreAtIvE teams in two respects. We were not given specific genes for each paper but a set of candidate genes, some of which did have experimentally based molecular function annotations derived from the given paper and some not. Second, true positive GO term identifications in BioCreAtIvE included not only exact matches but also close matches that the human judges considered acceptable. As reported in Camon *et al*. (2005), true positive, otherwise known as perfect, predictions for task 2.2 were those that were evaluated as 'high' for both the GO term and the protein identification. The GO term assignment was deemed to be 'high' if it was correct or close to what a curator would choose given the evidence text. (Other possibilities were 'general' if the GO term was in the correct lineage given the evidence text but was at too high a level or was too specific, and 'low' if it was simply wrong.)

Our task was made somewhat easier, however, by the well-regulated gene naming guidelines for yeast, which simplified the task of identifying references to yeast genes and proteins in text. For task 1B of BioCreAtIvE, for example, the top score for identifying yeast names in abstracts was over 90%, while for fly (*Drosophila*) and mouse it was around 80%. Our results clearly favored recall over precision. However, the precision we obtained undeniably needs improvement. One general explanation for the relatively poor results achieved to date in GO term identification is that the information to be extracted is complex, spread over several sentences or located in different places altogether in the text. What is noteworthy, however, is that of the participating groups in BioCreAtIvE, only two made use of the external knowledge resources of MeSH and the HUGO database of human gene names and symbols, and these groups obtained the highest *F* scores of 8.4% (Ehrler *et al*., 2005) and 7.0% (Ray and Craven, 2005). Similarly, as pointed out in the comparison of our results with KDD, we made extensive use of external knowledge resources, which appears to have contributed to our high recall results.

4 DISCUSSION

Greater insight into the data mining methods might be gained if scientific curators could render a judgment on each sentence that contributed to the Score in a 6-tuple for subtask I or appeared in an Excerpt in a 4-tuple for subtask II. If the curators asserted that a sentence supported the claim of experimental evidence or the proposed GO term for the given gene, then that sentence would be considered valid. If the curators did not think the sentence

offered such support, then it would be considered spurious. However, it is difficult and time consuming to assess the evidentiary value of thousands of pieces of text, as the organizers of BioCreAtIvE confirmed in their report. In addition, our experience suggests that the existence of a spurious sentence (or even several) does not invalidate the conclusions drawn from the 6-tuples and perhaps also from the 4-tuples. However, a more complete evaluation of the evidentiary passages produced by our system would provide a deeper understanding of how to improve our results, particularly with respect to precision.

It is unlikely now or in the near future that database curation will be fully automated. What is needed is a tool scientific curators can use to make their task easier and more efficient. Even a highlighting tool would make a difference, one that showed all mentions of a gene or protein and all its synonyms or aliases within an article, and also indicated any terms used to describe relevant protocols. For example, Table 3 shows several passages from the article with PubMed ID 10207049 (Lin and Lis, 1999) that provided experimental evidence for the annotation of the *GAC1* gene to the GO term ‘heat shock protein binding’. The passages on the left were provided by a SGD curator for comparison with the passages identified by our system, shown on the right. The passages identified by our system seem not only to produce reasonable results but also to be potentially useful as a screening tool for curators in the analysis of new papers.

In addition to these passages, our system also picked out a number of irrelevant passages. Experiments with human subjects are needed to see at what point the distraction of the irrelevant passages overcomes the usefulness of the relevant passages.

It is instructive to examine the different kinds of errors our system is making. One kind of error lies in extending an annotation to additional genes in a family when it is not warranted. For example, a paper that provided experimental evidence for the assignment of the GO term ‘6-phosphogluconolactonase activity’ to genes *SOL3* and *SOL4* was thought by our system also to provide support for assigning this term to *SOL1* and *SOL2*. This kind of error produces false positives, lowering precision. Conversely, there is the mistake of failing to apply a particular annotation to additional genes in a family, as when the GO term ‘acetyltransferase activity’ was correctly assigned by our system to *SAS2* and *SAS3* but not to *SAS4* and *SAS5*, when it should have been. This kind of error produces false negatives, lowering recall and precision.

One kind of error in particular produces thousands of false positives in our runs. This error occurs when a paper is correctly identified as providing experimental evidence for the assignment of a given GO term to a specific gene, but then many other unrelated GO terms are also thought to apply to the gene. For example, the paper with PMID 15889139 (Chang *et al*., 2005) mentions 20 different genes. But, based on direct assays described in the paper, it supports the assignment of only the following GO terms to the genes *RMI1, SGS1* and *TOP3*:

GO term ‘four-way junction DNA binding’ assigned to *RMI1*.

GO term ‘single-stranded DNA binding’ assigned to *RMI1*.

GO term ‘ATP-dependent DNA helicase activity’ assigned to *SGS1*.

GO term ‘DNA topoisomerase type I activity’ assigned to *TOP3*.

Our system correctly identified only *RMI1*, *SGS1* and *TOP3* as having molecular function claims supported by direct-assay evidence in the paper, and it correctly assigned the above three GO terms to the correct genes. However, our system also incorrectly assigned nine other GO terms to *RMI1*, eight to *SGS1* and seven to *TOP3*. Of these 24 additional GO assignments, 6 have terms that are in the same lineage as the correct GO terms, just higher in

the GO hierarchy. For *RMI1*, our system proposed the two GO terms of ‘DNA binding’ and ‘nucleic acid binding’. Although ‘nucleic acid binding’ may be too far up in the hierarchy to be useful, ‘DNA binding’, an ancestor of both ‘four-way junction DNA binding’ and ‘single-stranded DNA binding’ may be judged useful in directing curator attention to a particular part of the GO hierarchy. Similarly, for *SGS1*, our system proposed ‘DNA helicase activity’, ‘catalytic activity’ and ‘helicase activity’, and for *TOP3* our system proposed ‘DNA topoisomerase activity’, all of which are more general than the correct GO terms but in the correct lineage. A systematic study of incorrect additional GO assignments could provide dramatic improvement in precision.

One straightforward way to increase the number of true positives is to pick up acronyms and abbreviations given within the article itself. For example, the article featured in Table 3 defined ‘HSF’ as an acronym for ‘heat shock factor’, but our system does not yet pick up definitions given in the text. Consequently, out methods failed to select the title as a text passage of interest even though the phrase ‘heat shock factor’ appears in it. Methods such as those presented by Okazaki and Ananiadou (2006) and Yu *et al*. (2007) could be used. The title of the article in Table 3 also does not contain any reference to an appropriate experimental method or any recognizable reference to the GAC1 gene. The linguistic relation between ‘glycogen synthase phosphatase’ in the title and the gene name ‘GAC1’ is not straightforward. Although our language resources know that ‘GLyCogen’ is the description for the standard name ‘GLC7’, the fact that Gac1 is a regulatory subunit of Glc7 phosphatase is not known to our system. This fact is stated simply and clearly in the text, however, which suggests that it could be productive to find ways to mine such linguistic information from the text itself.

Our biggest challenge is to improve precision by reducing the number of false positives, but without significantly harming recall. An analysis of the 29 363 false positives and the 271 true positives for GO annotations based on direct assays reveals a useful screening tool that does just that. This analysis shows that we can find the GO term, using our normalized matching procedure, in the title and abstract of 90% of the true positives but only 40% of the false positives. Thus, while a normalized match to the GO term in the title and abstract is neither a necessary nor sufficient condition for a correct molecular function annotation, it can serve as a useful screen. If we eliminate all the annotations that fail to produce this kind of normalized match, for direct assay-based annotations we eliminate 17 617 false positives but just 27 true positives, lowering recall just a little to 40.2% and raising precision to 2.08%. While the increase in precision may seem small, the practical benefit of a curator's not having to check 17 617 false positives is huge.

Our approach to finding molecular function claims backed by experimental evidence in text can be characterized by two features that distinguish it from other large-scale efforts. First, we made extensive use of external knowledge, specifically a federated biomedical language resource, publicly available GO mappings and a vocabulary of protocol terms. Second, we relied on any one of multiple text passages generated in support of a claim of experimental evidence or a specific molecular function description. The relatively poor results from all efforts to identify GO terms in text testify to the difficulty of matching a given molecular function description with text that in some complex way expresses ‘the same notion’. While our recall results are very encouraging, there is clearly room for improvement, particularly with regard to precision.

Acknowledgments

This work was supported by grant number #R43CAHG003600-01 from the National Human Genome Research Institute (NHGRI) at the US National Institutes of Health. The *Saccharomyces* Genome Database project is

supported by NHGRI as a Genome Research Resource [grant number #2P41HG001315] and the Gene Ontology Consortium project is supported by grant number #P41HG02273 from the NHGRI.

REFERENCES

Ausubel, FM., et al. Current Protocols in Molecular Biology. John Wiley and Sons, Inc.; 2007.

Bairoch A. The ENZYME database in 2000. Nucleic Acids Res 2000;28:304–305. [PubMed: 10592255]

Barthelmes J, et al. BRENDA, AMENDA and FRENDA: the enzyme information system in 2007. Nucleic Acids Res 2007;35:511–514.

Blaschke C, et al. Evaluation of BioCreAtIvE assessment of task 2. BMC Bioinformatics 2005;6(Suppl. 1):S16. [PubMed: 15960828]

Camon E, et al. An evaluation of GO annotation retrieval for BioCreAtIvE and GOA. BMC Bioinformatics 2005;6:S1.7.

Chang M, et al. RMI1/NCE4, a suppressor of genome instability, encodes a member of the RecQ helicase/Topo III complex. EMBO J 2005;24:2024–2033. [PubMed: 15889139]

Chiang J, Yu H. MeKE: discovering the functions of gene products from biomedical literature via sentence alignment. Bioinformatics 2003;19:1417–1422. [PubMed: 12874055]

Chiang, J.; Yu, H. Extracting Functional Annotations of Proteins Based on Hybrid Text Mining Approaches.. In Proceedings of the BioCreAtIvE Challenge Evaluation Workshop; 2004; 2004.

Couto FM, et al. Finding genomic ontology terms in text using evidence content. BMC Bioinformatics 2005;6(Suppl. 1):S21. [PubMed: 15960834]

Crangle, CE. Text summarization in data mining.. In: Bustard, D., et al., editors. Soft-Ware 2002, LNCS 2311. Springer-Verlag; Berlin, Heidelberg: 2002. p. 332-347.2002Proceedings of Computing in an Imperfect World, First International Conference; Belfast, Northern Ireland. April 2002; p. 332-347.

Crangle CE, Zbyslaw A. Identifying gene ontology concepts in natural-language text. EMBC 2004. Conference Proceedings of the 26th Annual International Conference of the Engineering in Medicine and Biology Society 2004;4:2821–2823.

Ehrler F, et al. Data-poor categorization and passage retrieval for gene ontology annotation in Swiss-Prot. BMC Bioinformatics 2005;6(Suppl. 1):S23. [PubMed: 15960836]

Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet 2000;25:25–29. [PubMed: 10802651]

Hirschman L, et al. Overview of BioCreAtIvE: critical assessment of information extraction for biology. BMC Bioinformatics 2005;6(Suppl. 1):S1. [PubMed: 15960821]

Hu ZZ, et al. Literature mining and database annotation of protein phosphorylation using a rule-based system. Bioinformatics 2005;21:2759–2765. [PubMed: 15814565]

Issel-Tarver L, et al. Saccharomyces Genome Database. Meth. Enzymol 2002;350:329–346. [PubMed: 12073322]

Kim J, Park J. BioIE: retargetable information extraction and ontological annotation of biological interactions from literature. J. Bioinform. Comput. Biol 2004;2:551–568. [PubMed: 15359426]

Koike A, et al. Automatic extraction of gene/protein biological functions from biomedical text. Bioinformatics 2005;21:1227–1236. [PubMed: 15509601]

Lin JT, Lis JT. Glycogen synthase phosphatase interacts with heat shock factor to activate CUP1 gene transcription in Saccharomyces cerevisiae. Mol. Cell. Biol 1999;19:3245–3247.

Miller GA, et al. Introduction to WordNet: an on-line lexical database. Int. J. Lexicogr 1990;3:235–244.

Okazaki N, Ananiadou S. Building an abbreviation dictionary using a term recognition approach. Bioinformatics 2006;22:3089–3095. [PubMed: 17050571]

Pérez A, et al. Gene annotation from scientific literature using mappings between keyword systems. Bioinformatics 2004;20:2084–2091. [PubMed: 15059832]

Porter MF. An algorithm for suffix stripping. Program 1980;14:130–137.

Ray S, Craven M. Learning statistical models for annotating proteins with function information using biomedical text. BMC Bioinformatics 2005;6(Suppl. 1):S18. [PubMed: 15960830]

Rebholz-Schuhmann D, et al. Protein annotation by EBIMed. Nat. Biotechnol 2006;24:902–903. [PubMed: 16900125]

Regev Y, et al. Rule-based extraction of experimental evidence in the biomedical domain – the Kdd Cup (Task 1). SIGKDD Explor 2002:4.

Sedman T, et al. A DNA helicase required for maintenance of the functional mitochondrial genome in Saccharomyces cerevisiae. Mol. Cell. Biol 2000;20:1816–1824. [PubMed: 10669756]

The FlyBase Consortium. The FlyBase database of the Drosophila genome projects and community literature. Nucleic Acids Res 2003;31:172–175. [PubMed: 12519974]

Whetzel PL, et al. FuGO working group. Development of FuGO: an ontology for functional genomics investigations. OMICS 2006;10:199–204. [PubMed: 16901226]

Yeh AS, et al. Evaluation of text data mining for database curation: lessons learned from the KDD Challenge Cup. Bioinformatics 2003;19(Suppl. 1):i331–i339. [PubMed: 12855478]

Yu H, et al. Using MEDLINE as a knowledge source for disambiguating abbreviations and acronyms in full-text biomedical journal articles. J. Biomed. Inform 2007;40:150–159. [PubMed: 16843731]

Table 1

Results for determining whether or not there is experimental evidence of the given type for a given gene in a given article, without reference to any specific molecular function

	Evidence code IDA	Evidence codes IDA and IPI	Evidence codes IDA, IEP, IGI, IMP, ISS, IPI
Experiment-based annotations	2117	2592	3830
Filtered annotations returned by system	2280	2355	2734
True positives (correct answers)	1497	1814	2671
False positives	784	542	64
False negatives	614	772	1153
Recall	70.7%	70.0%	69.7%
Precision	65.7%	77.0%	97.7%
F score	68.1%	73.3%	81.4%

Table 2

Results for determining the specific molecular function (GO term) associated with a given gene for which there is experimental or primary source evidence

	Evidence code IDA	Evidence codes IDA and IPI	Evidence codes IDA, IEP, IGI, IMP, ISS, IPI
Molecular function annotations	607	846	1560
Filtered annotations returned by system	29 363	29 292	29 380
True positives (correct answers)	271	321	438
False positives	29 111	28 990	28 960
False negatives	317	506	1103
Recall	44.6%	37.9%	28.1%
Precision	0.92%	1.10%	1.49%
F score	1.80%	2.14%	2.83%

Table 3

Supporting text passages from article with PubMed ID 10207049 for the annotation of the GAC1 gene to the GO term ‘heat shock protein binding’ based on experimental evidence

Text passages identified by curators	Text passages identified by system
In Title or introductory section without section heading . . .	
Lin JT, Lis JT (1999) Glycogen synthase phosphatase interacts with heat shock factor to activate *CUP1* gene transcription in Saccharomyces cerevisiae. Mol Cell Biol 19(5):3237-45	Here, we used the phage display system to isolate proteins that interact with HSFrr.
In this article, we describe the use of the phage display system to identify an HSFrr-interacting protein, Gac1.	Second, the targeted modification of HSF appears to play * [truncated sentence]
	The phage display system allows the rapid selection and cloning of specific proteins that interact directly with a target
In Results section . . .	
HSFrr and Gac1 proteins can physically interact. A simple pull-down binding assay confirmed the direct physical interaction between HSFrr and Gac1 proteins *in vitro*.	
Purified MBP-HSFrr was mixed with either GST or GST-Gac1(162n406), both of which were bound to glutathione-agarose resin.	
FIG. 3. Physical interaction between HSFrr and Gac1	
We also used an immunoprecipitation analysis to demonstrate that Gac1 and HSFrr interact in a yeast extract.	We also used an immunoprecipitation analysis to demonstrate that Gac1 and HSFrr interact in a yeast extract
In summary, these results demonstrate that Gac1 protein interacts with HSFrr both as purified recombinant proteins and as proteins in crude yeast extracts.	
FIG. 4. Physical interaction between HSFrr and Gac1(130n502) in an immunoprecipitation assay.	Physical interaction between HSFrr and Gac1(130–502) in an immunoprecipitation assay.