BIOINFORMATICS

Vol. 19 Suppl. 1 2003, pages i331–i339
DOI: 10.1093/bioinformatics/btg1046

Evaluation of text data mining for database curation: lessons learned from the KDD Challenge Cup

Alexander S. Yeh, Lynette Hirschman and Alexander A. Morgan*

The MITRE Corporation, 202 Burlington Road, Bedford, MA 01730, USA

Received on January 6, 2003; accepted on February 20, 2003

ABSTRACT

Motivation: The biological literature is a major repository of knowledge. Many biological databases draw much of their content from a careful curation of this literature. However, as the volume of literature increases, the burden of curation increases. Text mining may provide useful tools to assist in the curation process. To date, the lack of standards has made it impossible to determine whether text mining techniques are sufficiently mature to be useful.

Results: We report on a Challenge Evaluation task that we created for the Knowledge Discovery and Data Mining (KDD) Challenge Cup. We provided a training corpus of 862 articles consisting of journal articles curated in FlyBase, along with the associated lists of genes and gene products, as well as the relevant data fields from FlyBase. For the test, we provided a corpus of 213 new ('blind') articles; the 18 participating groups provided systems that flagged articles for curation, based on whether the article contained experimental evidence for gene expression products. We report on the evaluation results and describe the techniques used by the top performing groups.

Contact: asy@mitre.org

Keywords: text mining, evaluation, curation, genomics, data management

INTRODUCTION

The research literature is a major repository of biological knowledge. To make this knowledge accessible, it is translated by expert curators into entries in biological databases. This serves several purposes: experts consolidate data about a single organism or a single class of entity (e.g. proteins) in one place, often in conjunction with sequence information. Second, this process makes the information searchable through a variety of automated techniques, given that the curators use standardized terminologies or ontologies. However, it is becoming more and more difficult for curators to keep up with the increasing volume of literature, creating a demand for automated curation aids.

There has been a growing volume of work in text mining for biological literature, but until now, there has been no way to compare results of the different systems (Hirschman *et al.*, 2002). Several related fields have addressed this problem by organizing open 'challenge' evaluations, e.g. for protein structure prediction, there have been the successful CASP evaluations (Critical Assessment of Techniques for Protein Structure Prediction, http://predictioncenter.llnl.gov/). For natural language processing, there was the series of Message Understanding Conferences (MUCs) for information extraction on newswire text (Hirschman, 1998). The Text REtrieval Conferences (TREC, http://trec.nist.gov/; (Voorhees and Buckland, 2002)) for information retrieval are ongoing.

The idea behind these series of open evaluations has been to attract teams to work on a problem by providing them with real (or realistic) training and test data, as well as objective evaluation metrics. These data sets are often hard to obtain, and the open evaluation makes it much easier for groups to build systems and compare performance on a common problem. If many teams are involved, the results are a measure of the state-of-the-art for that task. In addition, when the teams share information about their approaches and the evaluations are repeated over time, then the research community can demonstrate measurable forward progress in a field.

For these reasons, we decided that it would be valuable to test whether text mining techniques could help the curation process. To do this, we created and ran a common challenge evaluation (a contest) using data from a biological database and a task performed by curators of biological databases. The contest that we created and ran was Task 1 (of 2 tasks) of the KDD Challenge Cup 2002, a competition held in conjunction with the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), July 23–26, 2002.[†] This contest

*To whom correspondence should be addressed.

[†] See http://www.biostat.wisc.edu/~craven/kddcup/ for a description of the task.

 © Oxford University Press 2003; all rights reserved.

focused on text mining to provide semi-automated aids for biological database curation (Yeh *et al.*, 2003). FlyBase, a publicly available database on *Drosophila* genetics and molecular biology (FlyBase Consortium, 2002), supplied the data and biological expertise. This paper describes the results and lessons that we learned from setting up and running this contest.

METHODS: CONTEST SET-UP

For this contest, we drew on the work performed by Prof. William Gelbart and colleagues at Harvard in connection with FlyBase Harvard. We discussed how to provide automated aids for curating biomedical databases with the FlyBase curators and settled on a fundamental task at the beginning of the FlyBase Harvard curation pipeline, that of identifying the papers to be curated for *Drosophila* gene expression information.

FlyBase Harvard curates journal articles containing experimental gene expression evidence, specifically, experimental evidence about the products—mRNA transcripts (TR) and proteins/polypeptides (PP)—associated with a given gene.

We defined the following task for the contest, based on materials obtained from FlyBase:

- Given a set of papers (full text) on genetics or molecular biology and, for each paper, a list of the genes mentioned in that paper:

- Determine whether the paper meets the FlyBase gene-expression curation criteria, and for each gene, indicate whether the full paper has experimental evidence for gene products (mRNA and/or protein).

For each paper, a system needed to return three things:

1. A ranked list of papers in order of probability of the need for curation, where papers containing experimental evidence of interest should rank higher than papers that did not contain such evidence;
2. A yes/no decision on whether to curate each paper;
3. For each gene in each paper, a yes/no decision about whether the paper contained experimental evidence for the gene's products (RNA and protein/polypeptide).

The KDD Challenge Cup schedule included a 6 week period when the training data was made available for the contestants to build and train a system, followed by a two week period to complete the running of the test material. The results were then submitted to MITRE for final scoring.

The training and test data

The training data set consisted of 862 'cleaned' full text papers, of which 283 had been judged to need curation. Each paper came to the Harvard curators with a list of the genes (in a standardized nomenclature) mentioned in the paper. Along with its standardized nomenclature, the FlyBase database provides synonym lists for each gene. These resources, along with the set of relevant FlyBase database entries for each paper, were provided to KDD participants as part of the training data.

The test data set came from papers that had already been curated for genes (so the gene list was available to both the FlyBase gene product curators and the general public), but for which the gene product curation was not public yet. In the end, the test set consisted of 213 papers, together with the genes mentioned in each paper.

For each paper, we 'cleaned' the text by converting non-plain text (superscripts, subscripts, italics, Greek letters) into plain text; this was critical to preserve distinctions in meaning for further processing. For example, in '*Appl*^{d}', the *Appl* in italics indicates that the Appl gene is being mentioned (and not the protein) and the superscript *d* indicates that the Appl gene's *d* allele is being mentioned. The resulting conversion produced the form '@Appl@[d]', using the conventions developed by FlyBase for their gene name lists.

The list of genes for a paper was given in the form of a template in XML that also indicated the yes/no decisions to be made. For the training papers, a filled-out version of this template was also provided. For example, the following template indicated the mention of the `sws` and `Ecol\lacZ` genes in the associated paper:

```
<curate>?</curate>
<gene symbol="sws">
        <tr>?</tr><pp>?</pp></gene>
<gene symbol="Ecol\lacZ">
        <tr>X</tr><pp>X</pp></gene>
```

Systems gave their yes/no (`Y/N`) answers by returning these templates with the `?`'s replaced by `Y` or `N`. For each gene, returning `<pp>Y</pp>` meant that a system found experimental data of interest in the paper for some *protein* of that gene. Returning `<pp>N</pp>` meant that a system did not; `<tr>Y</tr>` and `<tr>N</tr>` indicated analogous findings for that gene's *transcripts*.

Lethal (e.g. `l(2)52A`), foreign (e.g. `Ecol\lacZ`) and anonymous (e.g. `anon-56Cb`) genes were especially hard to handle and were deemed 'special'. Systems did not have to answer `Y/N` for those genes' products. We indicated this by replacing the appropriate `?`'s with `X`'s.

The overall decision on whether a paper had experimental evidence for a product of any gene (*including* 'special' genes) was indicated by changing the `?` in `<curate>?</curate>` into a `Y` for yes and `N` for no.

For the training papers, we also provided the experimental data that FlyBase extracted from that paper. For example, below is relevant evidence from Kolhekar *et al.* (1997):

```
(gene="Phm" product="Phm-P1"
 ptype="pp" evtype="asm"):
                    immunolocalization
```

This indicated that the assay mode (*asm*) was *immunolocalization*, used on the *Phm-P1* protein product (*pp*) of the *Phm* gene.

The data sets presented a number of complications. First, the list of synonyms for the genes provided to the contestants was not complete because of the many typographical variants of names. For example, FlyBase listed the following 7 synonyms for the *Appl* gene:

```
APPL, appl, EG:65F1.5, CG7727, BcDNA:GH04413,
&bgr; amyloid protein precursor-like,
&bgr;-amyloid-protein-precursor-like.
```

But this list did not include the synonym *APP-like* as a gene name, which appears in Rosen *et al.* (1989). In addition, some names are not unique to a particular gene. For example, *Clk* is both a symbol for the *Clock* gene and a synonym for the *period* gene. This meant that it was not trivial to map between the genes listed for an article and their mentions in the text.

The training set came from papers that were already curated and publicly available from FlyBase. One small source of noise in the training data was due to the fact that, not surprisingly, curation standards change over time and differ between individuals. For example, FlyBase is only interested in gene expression results that are applicable to 'regular' flies found in the wild (wild-type), and not in expression results that apply just to laboratory induced mutations. In addition, FlyBase is normally interested in wild-type experimental gene expression results that are repeats of results found in other, earlier papers. However, if the focus of a paper is not on the gene products in wild-type flies, but the paper does have a few experimental results on wild-type flies (usually to serve as controls in an experiment) that have already been seen elsewhere, then FlyBase is *not* interested in that particular paper's gene expression results. The border between a 'few' results (not of interest) and enough results to be of interest is a bit fuzzy. Such borderline papers were removed from the test data, but were left in the training data.

Also, it took significant reverse engineering to determine how the experimental evidence was encoded in the database, and exactly what kinds of information constituted experimental evidence. This reverse engineering was not perfect, and the imperfections form another source of residual noise. Many FlyBase transcript and protein data fields contain experimental data, such as *transcript length* and *assay mode*, but many others do not, including the fields *public protein symbol* and *synonyms for transcript symbol*. There are also fields that contain data that used to be of interest to the gene product curators, but are no longer, because another group is now curating these fields. Examples are *protein domains* and *protein characteristics*. The *comment* field is a special case by itself. It usually contains experimental results of interest to the gene product curators, so we included its existence as an indicator of a training paper having results of interest for an associated gene's transcript or protein. However, the field is used for any information that will not fit anywhere else in FlyBase, and so the field sometimes contains material that is either not of interest or of borderline interest.

We originally wanted a contestant's system to provide evidence for its response, by indicating a passage in the text describing relevant experimental results. When using a system to aid in curation, providing such a passage would give a person checking the system a basis on which to accept or reject that finding. But while FlyBase stores the results of interest found in a paper, it does not indicate which passage(s) in that paper support or describe those results. Furthermore, the entry in FlyBase often uses wording that is very different from what is explicitly stated in the passage(s). For example, FlyBase's assay field for the PHM protein in the paper (Kolhekar *et al.*, 1997) uses the controlled vocabulary term *immunolocalization*. In that paper, there is *no* mention of the term 'immunolocalization' (or any similar term) in the text. Instead, the supporting text describes the various steps taken to perform an *immunolocalization* assay (use an anti-body to stain some tissue and then look at it), as illustrated in this figure caption excerpt from Kolhekar *et al.* (1997):

> *Figure 12.* **Top.** Whole-mount tissue staining using an affinity-purified anti-PHM antibody in the CNS and in non-neural tissues. *A,* The third instar larval CNS exhibits distributed cell body and neuropilar staining. This view displays only a portion of the CNS; ...

Another example is that for the paper Tingvall *et al.* (2001), FlyBase records that mRNA transcripts of certain reporter constructs (a construct is a combination of a reporter gene and a gene of interest) appear in certain parts of the body. The paper itself never explicitly mentions any transcript. Instead, the supporting text mentions where the associated reporter protein is detected. The FlyBase curators infer the transcripts' locations from the places where the protein is detected. Manually finding such 'evidence' passages for use in training a system would have been both time consuming and difficult[†]. So we dropped the passage finding requirement.

[†] Especially since it can require a lot of biology knowledge and our contestants had more of a data-mining background than a biology background.

We also originally wanted the participating systems to generate the names of the gene product(s) that had experimental results in a paper. However, different proteins in the FlyBase database are named using different conventions (and likewise for transcripts). For example, FlyBase lists 5 different forms of proteins for the **Doa** gene, which are named using 4 different conventions: **Doa$^+$P105kD** is named after the form's size (105 kilodaltons). **Doa$^+$P517** is named after the form's length (517 amino acids). **Doa-P1** and **Doa-P2** are named using more recent naming conventions for distinct forms of **Doa** protein found in the literature. **Doa$^+$P** is a name used for results that apply to one or more forms of **Doa** protein, but the curator cannot tell from the paper which specific form(s).

Furthermore, the product names used in the papers do not always match the corresponding FlyBase names. Determining the correspondences may not be so difficult with FlyBase names that contain some product property like size or length, for example '105-kD protein' or '105-kD DOA isoform' in (Yun *et al.*, 2000), which are mentions of **Doa$^+$P105kD**. However, determining the correspondences to other FlyBase product names is difficult. For example, in the same paper, phrases like '55-kD DOA protein' and '55-kD isoform' are recorded in FlyBase as **Doa-P2**. Also in that paper, phrases like 'protein kinase', 'DOA kinase', 'DOA protein', and 'DOA' can either refer to all forms of DOA protein, the two forms studied in detail in that paper (**Doa$^+$P105kD** and **Doa-P2**) or to one or more forms, but the paper is unclear (to a biologically-trained curator) as to which, leading the curator to use **Doa$^+$P**.

In addition, difficulties in determining the correspondences can lead to difficulties in determining when a transcript or protein described in a paper is actually new to FlyBase, and has yet to be listed in the database. For these reasons, we avoided the issue of naming gene expression products; we simply required the systems to provide a 'yes/no' answer for whether a paper had experimental results for that gene's transcripts and proteins.[†]

Scoring measures

The contest task was divided into 3 sub-tasks. The ranked-list and 'yes/no curate paper' sub-tasks are two possible ways to help a curator with filtering out the papers that have no information of interest. The ranked-list can help by providing an ordering on the relative likelihood of a paper being of interest. If accurate, the 'yes/no curate paper' decisions are direct indicators of what papers to concentrate on. The third sub-task ('yes/no' for products of each gene) is a way to tell a curator what gene(s) to concentrate on in a paper.

After defining the task and preparing the training and test data, we developed a simple scoring method for each of the three sub-tasks.

For the ranked-list sub-task, we used as a metric the area under the receiver operating characteristic curve (AROC); the ROC curve (Duda *et al.*, 2001, Section 2.8.3) measures the trade-off between sensitivity (recall) and the probability of a false alarm. As the area under the curve increases, a system will on average be more sensitive for the same false alarm rate.

For the yes/no curation decisions for the set of papers, we used the standard balanced F measure, which is a combination (the harmonic mean) of *recall* and *precision*.[‡] *Recall* is the percentage of the correct 'yes' decisions that are actually returned by the system. This measures how sensitive a system is in finding what it should find. *Precision* is the percentage of the 'yes' decisions returned by the system that are actually correct. This measures how specific a system is in finding just what it should find.

We also used the balanced F measure for the yes/no decisions on experimental evidence for products of the genes mentioned in the papers. The sum of these three scores (equally weighted) was used to provide an overall system score.

Table 1. Results of the 32 submissions

Sub-Task	Best	1st Quartile	Median	Low
Ranked-list:	84%	81%	69%	35%
Y/N paper:	78%	61%	58%	32%
Y/N products:	67%	47%	35%	8%
Overall:	76%	61%	55%	32%

RESULTS

Overall, 18 teams returned 32 separate submissions for evaluation (up to 3 per team). There were eight countries represented, including Japan, Taiwan, Singapore, India, Israel, UK, Portugal and USA. There were groups from industry, academia and government laboratories, often teamed. The top performing team, ClearForest and Celera, obtained both the highest overall score and the highest score on the each sub-task. The results of the 32 submissions for the three metrics and the overall score (normalized to 100%) are given in Table 1. The top 5 teams for the ranked-list sub-task all had close scores for this sub-task (81–84%).

[†] Even with this simplification, products of 'special' (lethal, foreign and anonymous) genes can be hard to handle, so we added the further simplification that contestants did not need to make 'yes/no' decisions about these products, as mentioned earlier.

[‡] The balanced F measure is (2**precision***recall*)/(*precision* + *recall*).

High-performing teams

We declared a winning team and three honorable mention teams. The teams used a variety of approaches. The winning team (Regev *et al.*, 2003) used an information extraction approach with manually constructed rules that matched against patterns deemed of interest. A focus was finding patterns in figure captions. These often involved linguistic constructs, such as noun phrases (e.g. 'the developing midgut') and verb phrases (e.g. 'does not antagonize'). The output of the rules was combined to produce scores at both the document and gene level.

One honorable mention team was from three Singapore-based organizations (Shi *et al.*, 2003). Their system looked for the presence of certain manually chosen 'keywords'.[†] Within each paragraph of a paper, it computed the distance (measured as the number of sentence boundaries crossed) between each keyword mention and each mention of a gene name or synonym. For each gene and keyword pair, the minimum distance was noted, as was the number of occurrences with that minimum distance. The effects of different keywords on decisions about a gene's products were combined using Naïve Bayes (Duda *et al.*, 2001, Section 2.11).

The honorable mention team from Imperial College and Inforsense (Ghanem *et al.*, 2003) had a system that used regular expressions[‡] to find particular patterns of words. It automatically extracted these patterns from sentences in the training corpus. The patterns were restricted to be within a sentence or neighboring sentences, and to contain gene name(s) or keyword(s) that appeared in the experimental database fields from FlyBase associated with the training papers. When searching for the products of a particular gene, only sentences related to that gene were examined. The patterns served as features to be combined by a support vector machine (SVM) classifier (Duda *et al.*, 2001, Section 5.11) (http://svmlight.joachims.org), which made the final decisions.

The honorable mention team from Verity and Exelixis (B.Chen, personal communication) also had a system that used regular expressions and SVMs (two types: transductive and inductive). The system ignored certain sections of papers.

One thing these highly-ranked teams had in common is that they all moved away from the 'bag of words' approach common in text classification and information retrieval. This approach represents a document as an unordered bag of words, thus losing any grammatical relations among words. The words are then weighted by frequency to create a vector for each document. These vectors are then compared to find similar documents or passages.[§] One group (Ghanem *et al.*, 2003) in fact tried this approach at first, but found that the resulting system did not perform well. In general, use of pattern matching and local context seemed to work better, probably because it was important to associate experimental results with specific relevant genes; document level association may simply be too weak for this set of tasks.

Many of the submissions came from teams, and these teams often included biologists in the role of 'domain expert'. The domain experts seemed to be most useful for these teams near the start of the contest. This was the indication that we got in talking to a member of the winning team. The two honorable mention teams who wrote descriptions of their work, (Shi *et al.*, 2003) and (Ghanem *et al.*, 2003), both mention using domain experts to produce some of the feature lists that they used in their experiments. However, one thing to keep in mind is that as mentioned in **The training and test data** Section, we made several simplifications to this competition to make it less dependent on domain knowledge.

Test-set paper analysis

In our post-competition analysis, we looked at several factors that might have contributed to overall task difficulty. The first factor was how well the training data and test data sets were matched. The training data had 33% of articles that were judged to contain curatable experimental evidence for gene products. By contrast, the test set had a statistically significantly higher percentage:[¶] 91 papers (43%) of the 213 test papers had results of interest.

This led us to look at whether systems had been overly conservative in marking a paper as containing evidence for curation; we concluded that they had been. Overall, 26 (81%) of the 32 submissions marked less than 91 test papers with 'yes, curate'.[‖]

We also tried to characterize what made a curation decision harder for an individual paper. To do this, we counted how many of the 32 submissions made the correct

[†] A 'keyword' could actually be more a single word, e.g. *northern blot*.

[‡] Many text pattern matching systems use regular expressions to define the patterns, including the *Perl* programming language and the *Unix grep* utility.

[§] The SMART information retrieval system uses this 'bag of words' approach (Salton and McGill, 1983, Ch. 4). Often in this approach, words are stemmed (e.g. removal of plural *s*) and stop words are removed, e.g. *a, of, the, on, in*. Then each document or passage is represented as a vector of words, generally using a variant of the 'tf-idf' scheme, which weights words (terms) by their frequency within a document and by the inverse of the number of documents containing that word (inverse document frequency). Two documents are often compared by taking an inner (dot) product of their vectors, also known as a cosine measure.

[¶] Significant at the 0.005 level using a single-sided equal-variance *t* test.

[‖] This is statistically significantly higher (at the 0.015 level) than 50%, the highest expected figure if overall, the submissions were not conservative. A 1-sided test with a Normal approximation of a binomial distribution plus the Yates correction was used. This statistical significance holds even if one assumes only 18 of the submissions are independent (1 independent submission per team).

Table 2. 'no' versus 'yes' papers

Paper Type	average r'	Fraction with $r' > 50\%$
'no'	24.3 (76%)	93% (114 of 122 papers)
'yes'	17.6 (55%)	54% (49 of 91 papers)

Table 3. *both* versus *either* papers

Paper Type	average r'	Fraction with $r' > 50\%$
both	74%	85% (41 of 48 papers)
either (but not both)	35%	19% (8 of 42 papers)

'Y/N curate' decision for a given paper; we call this number the r' value for the paper.

Given the conservativeness observed above, it is not surprising that papers which had no results of interest (correct answer marked 'no') tended to be easier than papers with results (correct answer marked 'yes'). The 'no' papers had a higher average r' ($\overline{r'}$) than the 'yes' papers.[†] Another way to view this is that a larger fraction of the 'no' papers were correctly marked by over half the submissions (had $r' > 50\%$) than the 'yes' papers. See Table 2.

We did a further analysis to see if we could determine what made the 'yes-curate' papers hard. We noted that all but one of the 'yes' papers (90/91) had results of interest for at least one 'regular'[‡] gene product. These 90 papers could be divided into two groups.[§] Papers in the first group had results of interest on *both* transcripts and proteins. All test set papers of this type also had at least one 'regular' gene for which both transcript and protein results were present in the paper. Papers of the second type had results of interest on *either* transcripts or proteins, but **not** both. The *both* papers were easier to identify than the *either* papers, with the former having a higher $\overline{r'}$ than the latter,[¶] as shown in Table 3. Another way to view this is that a higher fraction of the *both* papers have $r' > 50\%$ than the *either* papers.

The *either* papers may be harder because they seem more likely to also have experimental results that only apply to laboratory-produced mutants (results not of interest), which can obscure the results that are of interest (wild-type).

[†] The r' standard deviations ($sd(r')$) are 14 and 26%, respectively. The difference in the averages is statistically significant (at the 0.0005 level) using a single-sided equal-variance t test.

[‡] A 'regular' gene is one that is not 'special' (anonymous, lethal or foreign) as mentioned in **The training and test data** Section.

[§] Products of 'special' genes were ignored in the determination of the groups' members.

[¶] $sd(r')$ is 20 and 16% respectively. The difference in the averages is statistically significant (at the 0.0005 level) using a single-sided equal-variance t test.

DISCUSSION: LESSONS LEARNED

One lesson we learned from running this contest is that access to the literature itself is a complex matter. Abstracts of papers are fairly easy to obtain via PubMed/Medline (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi). However, many of the results of interest to the FlyBase curators are only described in the full paper, and not in the abstract. As an example, for the protein **Appl$^+$P145kD**, FlyBase records that (Torroja *et al.*, 1996) finds 17 expression patterns relating to *when* (in the life cycle) and *where* (in the body) that protein is found. Only 2 (12%) of these patterns (an adult's brain and an adult's mushroom body) are mentioned in that paper's abstract. The other 15 (88%) patterns (for example: a larva's photoreceptor cell and a pupa's lobula) are only mentioned in the full paper—see Figure 1.

Using full papers introduces complications. One complication is that easily accessible electronic versions of some papers do not exist. Other papers could be obtained in PDF format, but they were not suitable for processing by most text mining systems. A subset of the papers were available in HTML format; however, this HTML version needed to be freely available to the public. For the contest, we began with a list of 7100 possible papers, but were able to obtain only 1118 freely available papers in HTML, from which both the training and test papers were drawn.

HTML has its own challenges. Publishers set up many of the HTML versions of the papers so that the main file and the directory to the linked files have the same path name. The linked files include most figures and also some figure captions and tables. So in a straightforward download, one cannot get both the main file and the linked files. Either one downloads the main file first and then replaces it with the directory when a linked file is downloaded, or vice-versa. We chose to keep the main files and leave out the directory and associated linked files. FlyBase curators have mentioned that many of the experimental results are presented in figures and their captions (B.Matthews, personal communication). Fortunately, most captions were not left out, and the captions typically described what was of interest in the actual figure. Also, most text processing systems cannot actually handle the figures (images) themselves.

Another complication was that many automated text processing systems have been designed to handle plain text, but, as mentioned in **The training and test data** Section, full papers of interest to FlyBase often have information expressed in typesetting conventions, such as superscripts, subscripts, italics and Greek letters. It

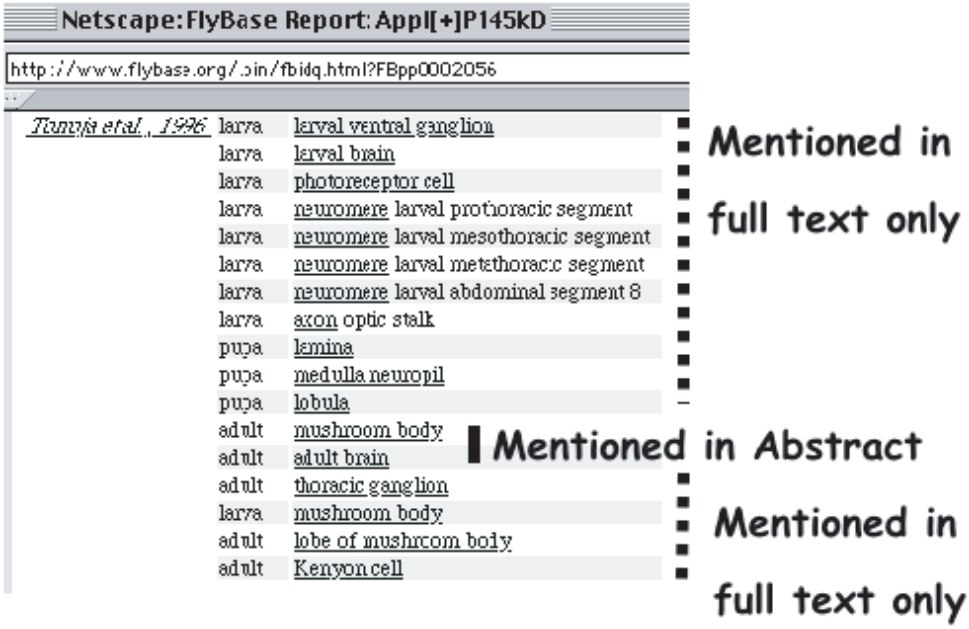

Fig. 1. Expression patterns found in full text versus abstract.

was necessary to apply conversion routines to produce versions that translated typographic conventions into plain text corresponding to the FlyBase conventions.

A second lesson is that the information of interest can differ quite a bit in appearance between the paper and the corresponding curated database entries, for example, in gene product naming (see **The training and test data** Section). FlyBase does not store pointers to the specific passage(s) that support a database entry. As a result, finding the evidence for a given entry may require significant biology expertise and sometimes, also expertise in FlyBase conventions.

This competition was held at a data-mining conference, so not surprisingly, many contestants made use of statistical, automated learning and/or automated weighting techniques, including Naïve Bayes, SVMs, Widrow Hoff linear classifiers, linear regression and the 'tf-idf' weighting scheme. However, such techniques were not enough to do well. The contestants also needed to either manually determine what features to look for and/or where to look for them. Examples of features included keywords, patterns of words and types of patterns. The winning team used manually determined patterns, while the honorable mention teams used mixtures of manually determined items and items gleaned from statistics or automated learning.

It was also important to know *where* to look for features or patterns. The winning system made good use of information in figure captions. A number of groups looked only at sentences containing gene name(s) or certain keyword(s). Some groups made use of the document structure, preferring to look in certain sections (e.g. 'Results' or 'Methods') and avoiding other sections. The 'References' section is one to especially avoid, as it contains citations that included names of genes not discussed in the paper.

The third sub-task was the hardest and the most fine-grained. This sub-task required determining which genes in a paper had experimental data on their wild-type (non-mutant) products, as opposed to just making an overall determination for the paper. So especially for this sub-task, a contestant's system needed to do more than look for good indicators of experimental results and good indicators of results for wild-type versus mutant genes. The system also needed to associate indicator terms relating to experimental findings (e.g. 'Northern blot' or 'Western blot') with particular genes. Some of the high performing systems handled this by looking for particular patterns of words that would associate an indicator with a particular gene, with the patterns often being contained within a sentence or two. Another system handled this by measuring how close (in sentences) an indicator (feature) was to a gene name and restricting the measurements to occurrences of gene and indicator mentions within the same paragraph.

A common feature of these approaches was that they used information about the document structure and linguistic structure of a paper, e.g. sections, paragraphs, sentences, and phrases. This is in contrast to the information

retrieval approach of treating a paper as just an unstructured set of words. We expect that systems will need to make more extensive use of linguistic and document structure to achieve better results and to accommodate more realistic tasks. For example, linguistic structure may provide critical clues once the simplifications mentioned in **The training and test data** Section are removed, including requiring systems to handle mentions of foreign genes, lethal genes or anonymous genes. Similarly, if the list of genes is not provided in advance for each paper, this makes the task of identifying the set of genes discussed in an article more difficult. The system would have to determine when a new name refers to a new gene and when it is a synonym for something already known. In this case too, both linguistic structures and document structures can provide critical information

One of our goals in running this evaluation was to evaluate the evaluation. For this, we defined three criteria:

- The evaluation should be repeatable and affordable. This should include a reusable training data set, cost-effective preparation of 'gold standard' data for test and repeatable scoring procedures that are easy to run and easy to understand.
- The evaluation must attract participants. This means that it needs to be a problem of importance to biologists, but also accessible to researchers in text mining teamed with biologists.
- The task must be tractable, but should also push the state of the art. If the task is well chosen, groups will demonstrate that they are on the path to the development of a useful capability.

Our assessment of the KDD evaluation was that it was successful along all of these dimensions. It was affordable. We estimate that it took us approximately 9 staff months of time to complete the tasks associated with setting up and running the KDD evaluation, including: (1) defining the task; (2) obtaining and normalizing the texts; (3) preparing and packaging the training data; (4) releasing the training data and answering questions; (5) developing and explaining the scoring routines; and (6) scoring the test results. In addition to our time, it took 2 staff months of time from the FlyBase curators to curate the test set and answer questions (both our questions and those of the participants).

We were able to create a reusable training corpus, which we will continue to make available.[†]

We were able to attract a reasonable number of participating groups (18) from a wide range of countries. However, because of the venue (KDD), we attracted mostly researchers in text mining, rather than biologists. We would like to attract more participation from the biology community.

The task we chose is one of real importance to curators responsible for maintaining biological databases. We believe that there are many other text mining tasks that could be of great potential utility to biological database curators.

CONCLUSIONS

We successfully organized an initial evaluation on text mining systems to aid biological database curation, as part of the KDD Challenge Cup 2002. Many teams took part in the evaluation, and their results indicate that curated data from a biological database can be used to train text mining systems to perform a potentially useful task.

The task that we presented to the contestants is only a small part of what the FlyBase Harvard curators do. But even this limited task is of real importance to the curators, because most of the papers (for example, 2/3 of our training papers) given to the curators contain no results of interest, and filtering out such papers is useful. The results from the ranked-list sub-task look especially promising (the best teams were 81–84%). But we need to perform further experiments to determine whether the resulting lists will actually help the curators with filtering papers.

We are now involved in planning a larger competition, together with A. Valencia and C. Blaschke (CNB-Madrid), under the umbrella of the ISCB BioLINK Special Interest Group for Text Data Mining (see http://www.pdg.cnb.uam.es/BioLINK/). We are planning two tasks; the first is the extraction of gene or protein names from text, so that we can evaluate the current state of the art in biological entity extraction across systems that have been reported in the literature over the past few years. The second task will require systems to associate Gene Ontology (GO) terms with mentions of proteins in articles curated in the SWISS-PROT database. Our experience in organizing the KDD competition leads us to believe that by using data from curated databases and focusing on tasks of immediate utility both to database curators and to researchers, we can define a good challenge evaluation for text data mining systems.

ACKNOWLEDGEMENTS

This paper reports on work done in part at the MITRE Corporation under the support of the MITRE Sponsored Research Program. In addition, many people at FlyBase have contributed to the KDD Cup task, especially William Gelbart, Beverly Matthews, Leyla Bayraktaroglu, David Emmert and Don Gilbert.

[†] To obtain the training corpus, send e-mail to Alex Yeh, asy@mitre.org

REFERENCES

Duda,R.O., Hart,P.E. and Stork,D.G. (2001) *Pattern Classification*, 2nd edition, Wiley, New York.

FlyBase Consortium (2002) The flybase database of the *Drosophila* genome projects and community literature. *Nucleic Acids Res.*, **30**, 106–108. http://flybase.org/.

Ghanem,M.M., Guo,Y., Lodhi,H. and Zhang,Y. (2003) Automatic scientific text classification using local patterns: KDD Cup 2002 (task 1). *SIGKDD Exploration* newsletter, **4(2)**, 95–96. http://www.acm.org/sigkdd/explorations/.

Hirschman,L. (1998) The evolution of evaluation: Lessons from the message understanding conferences. *Computer Speech and Language*, **12**, 281–305.

Hirschman,L., Park,J.C., Tsujii,J., Wong,L. and Wu,C.H. (2002) Accomplishments and challenges in literature data mining for biology. *Bioinformatics*, **18**, 1553–1561.

Kolhekar,A.S., Roberts,M.S., Jiang,N., Johnson,R.C., Mains,R.E., Eipper,B.A. and Taghert,P.H. (1997) Neuropeptide amidation in *Drosophila*: Separate genes encode the two enzymes catalyzing amidation. *J. Neurosci.*, **17**, 1363–1376.

Regev,Y., Finkelstein-Landau,M. and Feldman,R. (2003) Rule-based extraction of experimental evidence in the biomedical domain—the KDD Cup 2002 (task 1). *SIGKDD Explorations* newsletter, **4(2)**, 90–92. http://www.acm.org/sigkdd/explorations/.

Rosen,D.R., Martin-Morris,L., Luo,L. and White,K. (1989) A *Drosophila* gene encoding a protein resembling the human β-amyloid protein precursor. *Proc. Natl Acad. Sci. USA*, **86**, 2478–2482.

Salton,G. and McGill,M.J. (1983) *Introduction to Modern Information Retrieval*. McGraw-Hill, New York.

Shi,M., Edwin,D.S., Menon,R., Shen,L., Lim,J.Y.K. and Loh,H.T. (2003) A machine learning approach for the curation of biomedical literature–KDD Cup 2002 (task 1). *SIGKDD Explorations* newsletter, **4(2)**, 93–94. http://www.acm.org/sigkdd/explorations/.

Tingvall,T.Ö., Roos,E. and Engström,Y. (2001) The GATA factor Serpent is required for the onset of the humoral immune response in *Drosophila* embryos. *Proc. Natl Acad. Sci. USA*, **98**, 3884–3888.

Torroja,L., Luo,L. and White,K. (1996) APPL, the *Drosophila* member of the APP-family, exhibits differential trafficking and processing in CNS neurons. *J. Neurosci.*, **16**, 4638–4650.

Voorhees,E.M. and Buckland,L.P. (eds) (2002) *The Eleventh Text REtrieval Conference (TREC 2002)*. NIST Special Publication 500-XXX, Gaithersburg, Maryland, http://trec.nist.gov/pubs/trec11/t11_proceedings.html.

Yeh,A., Hirschman,L. and Morgan,A. (2003) Background and overview for KDD Cup 2002 task 1: Information extraction from biomedical articles. *SIGKDD Explorations* newsletter, **4(2)**, 87–89. http://www.acm.org/sigkdd/explorations/.

Yun,B., Lee,K., Farkas,R., Hitte,C. and Rabinow,L. (2000) The LAMMER protein kinase encoded by the *Doa* locus of Drosophila is required in both somatic and germline cells and is expressed as both nuclear and cytoplasmic isoforms throughout development. *Genetics*, **156**, 749–761.