Background and Overview for KDD Cup 2002 Task 1: Information Extraction from Biomedical Articles

Alexander Yeh, Lynette Hirschman, Alexander Morgan
The MITRE Corporation
202 Burlington Road
Bedford, MA 01730, USA
[asy,lynette,amorgan]@mitre.org

ABSTRACT

This paper presents a background and overview for 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. Task 1 dealt with detecting which papers, in a set of fruit fly genetics papers (texts), contained experimental results about gene products (transcripts and proteins), and also within each paper, which genes had experimental results about their products mentioned.[1]

Keywords

KDD Cup, competition, biology, genomics, text mining

1. BACKGROUND

This paper presents a background and overview for 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. Task 1 dealt with text data-mining to provide semi-automated aids for biological database curation. Alexander Yeh was the co-chair for task 1.

Biomedical information exists in both the research literature and various semi-structured databases. The literature is a rich source of information. Abstracts of much of the published literature are easily accessible via PubMED; full text articles have more limited availability, but contain critical information not available in the abstracts. Biological databases serve as repositories and distillations of what is described in the literature. Such databases exist for genes and proteins in general, and also for more specific areas, such as the genome of a specific organism. These databases typically have fields that contain structured entries, e.g., genetic or protein sequences, measurements, or gene or protein or tissue names (in a controlled vocabulary). However, these databases also contain significant amounts of semi-structured information, including summaries, comments, and short descriptive phrases. In addition, biological databases are generally accompanied by rich resources, including nomenclatures or ontologies that specify allowable entries for the database fields.

To keep these databases useful, they need to be curated: that is, these databases need to be kept up-to-date with respect to the increasing volume of new research literature. Currently, this curation is done manually. For example, a curator reading a paper on the Drosophila (fruit fly) genes and proteins may encounter the following passage:[2]

> *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;

The curator recognizes that this passage is describing the steps of an *immunolocalization* assay (use an anti-body to stain some tissue and then look at it). So in the database's assay field for the PHM protein for this paper, the curator enters the controlled vocabulary term *immunolocalization*.

2. THE TASK

In KDD Challenge task 1, we focused on the work performed by Prof. William Gelbart and colleagues at Harvard in connection with FlyBase Harvard (see http://www.flybase.org/ for information on FlyBase, a publicly available database on Drosophila genetics and molecular biology). We discussed 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 papers containing experimental gene expression evidence of interest to the curators, specifically, experimental evidence about the products (mRNA transcripts (TR) or proteins/polypeptides (PP)) associated with a given gene.

To create the KDD Challenge Cup Task, we defined the following task, 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

[2]From: A. Kolhekar *et. al.* Neuropeptide amidation in *drosophila*: Separate genes encode the two enzymes catalyzing amidation. *J. of Neuroscience*, 17:1363–1376, 1997.

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 articles in order of probability of the need for curation, where papers containing experimental evidence of interest rank higher than papers that do not contain such evidence;

2. A yes/no decision on whether to curate each article;

3. For each gene in each article, a yes/no decision about whether the article contained experimental evidence for the gene products (RNA, protein/polypeptide).

Appendix A provides more details.

The KDD Challenge Cup schedule included a 6 week period when the training data was made available, followed by a two week period to complete the running of the test material. The training set consisted of 862 "cleaned" full text articles, of which 283 had been judged to need curation. Each article 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 (in an evidence file) of relevant database entries for each article, were provided as part of the training data. The test set consisted of another 213 articles, together with the genes mentioned in each article.

3. CHALLENGES POSED BY THE TASK

The task presented to the contestants is only a part of what the FlyBase Harvard curators do. But even just this part is of real importance to the curators, because most of the papers (for example, 2/3 of our training papers) given to them contain no results of interest, and filtering out such papers is useful. Following KDD Cup in July, other database curation groups have asked us about our interest in using their databases as test cases for a similar evaluation, because they need these kinds of tools in their daily work.

Even this one "simple" task provides plenty of challenges for the contestants. One challenge is that 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 just apply to laboratory induced mutations.

Another challenge are the multiple names (synonyms) of many genes. This challenge carries over to gene products because the texts usually reference the products via their associated gene. As mentioned above, the contestants were provided with a list of synonyms for the genes. However, the list was probably not complete (there are many typographical variants of names), and an additional complication is that some names can refer to more than one gene. An example is *Clk*, which is both a symbol for the *Clock* gene and also is a synonym for the *period* gene.

A third set of challenges comes from a mismatch between natural language processing (NLP) systems and the training data as provided by the FlyBase database. NLP systems are mainly designed to find/extract explicit mentions of information in the text (text strings), with perhaps some limited normalization or stemming involved. FlyBase stores what results of interest were found in a paper, but

1. It does not indicate which passage(s) in that paper support or describe those results and

2. The entry in FlyBase may use wording that is very different from what is explicitly stated in the passage(s).

An example is the curation example in Section 1. The database entry of *immunolocalization* does not indicate what text in the paper supports this entry. Also, the curator concluded that an *immunolocalization* assay is mentioned in the paper without seeing any 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.

4. RESULTS

After defining the task and preparing the training and test data, we developed a simple scoring method for each of the three subtasks. For the ranked-list sub-task, we used as a metric the area under the receiver operating characteristic curve (AROC); the ROC curve measures the trade-off between sensitivity (recall) and the probability of a false alarm. For the yes/no curation decisions for the set of papers, we used the standard F measure[3]; we also used F measure for the yes/no decisions on whether there was experimental evidence for gene products for each gene mentioned in every paper. The sum of these three scores (equally weighted) was used to provide an overall system score.

18 teams submitted 32 separate results for evaluation (up to 3 per team). There were eight countries represented, including Japan, Taiwan, Singapore, India, Israel, UK, Portugal, 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 for the three metrics and the overall score (normalized to a maximum of 100%) are ("1Q" is first quartile):

Evaluation Sub-Task	Best	1Q	Median	Low
Ranked-list:	84%	81%	69%	35%
Yes/No curate paper:	78%	61%	58%	32%
Yes/No gene products:	67%	47%	35%	8%
Overall:	76%	61%	55%	32%

The top 5 teams for the ranked-list sub-task all had close scores for this sub-task (81%-84%).

In this issue are three other articles on this task: one article by the winning team and one article each by two of the three honorable mentions (teams listed east to west):

- Design Technology Institute Ltd., the Mechanical Engineering Dept. at the National University of Singapore and the Genome Institute of Singapore
- Data mining group at Imperial College (UK) and Inforsense Ltd.
- Verity, Inc. and Exelixis, Inc. (no separate article): used inductive support vector machines. Gene names were found with regular expressions. Also, certain sections of papers were excluded from consideration.

[3]The balanced F measure is (2**precision***recall*) divided by (*precision* + *recall*), where *recall* is the percentage of the correct "yes" decisions that are actually returned by the system; *precision* is the percentage of the "yes" decisions returned by the system that are actually correct.

5. 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 worked to make this KDD Cup task possible, especially William Gelbart, Beverly Matthews, Leyla Bayraktaroglu, David Emmert and Don Gilbert.

APPENDIX

A. SOME TASK DETAILS

For each paper, we provide the text of the paper in which parts of the paper that are beyond plain English text (superscripts, italics, Greek letters, etc.) have been converted into a representation in plain English text. We also provide a template in XML for each paper, which both lists the genes mentioned in that paper and also indicates the yes/no decisions to be made. The template for the paper from the Section 1 *immunolocalization* example is shown here (actually, only 4 of the 12 listed genes are shown here):

```
<article file="R171" pubmedid="9006979">
<curate>?</curate>
<gene symbol="l(2)05006"><tr>X</tr><pp>X</pp></gene>
...
<gene symbol="Ecol\lacZ"><tr>X</tr><pp>X</pp></gene>
<gene symbol="Phm"><tr>?</tr><pp>?</pp></gene>
...
<gene symbol="Thiolase"><tr>?</tr><pp>?</pp></gene>
</article>
```

This view of the template indicates some mention of the `l(2)05006`, `Ecol\lacZ`, `Phm` and `Thiolase` genes in the paper.

The contestants give their yes/no (`Y/N`) answers by returning these templates with the `?`'s replaced by `Y` or `N` as appropriate. For each gene, returning `<pp>Y</pp>` means that a system found experimental evidence of interest in the paper for some polypeptide/protein of that gene. Returning `<pp>N</pp>` means that a system did not find such evidence. Returning `<tr>Y</tr>` or `<tr>N</tr>` indicates analogous findings for that gene's transcripts. An example of such evidence is the passage in that *immunolocalization* example, which mentions experimental evidence for some `Phm` polypeptide(s). This passage describes an assay which finds where `Phm` polypeptide(s) are present in a fruit fly.

Lethal (e.g., `l(2)05006`), foreign (e.g., `Ecol\lacZ`) and anonymous genes are particularly hard to handle, so the contestants did not have to answer `Y/N` for those genes' products. We indicate this by having an `X` where an `?` would normally be found in a template.

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