

## Evaluating Genomics Text Data Mining Systems: Lessons Learned from the KDD Challenge Cup 2002

Alexander Yeh, Lynette Hirschman, Alexander Morgan  
The MITRE Corporation  
Bedford, MA 01730, USA  
{asy,lynette,amorgan}@mitre.org

For the past year, we have been looking at the processing of texts describing biological research, with a specific interest in the evaluation of information retrieval, extraction and question and answering tasks for such texts. This talk will cover our current work in running an evaluation competition on texts dealing with the *Drosophila* (fruitfly) genome. This competition is one of two tasks in this year's KDD Challenge Cup, a competition held in conjunction with the annual ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), July 23–26, 2002.

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 contain 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.

In the KDD Challenge task, we focused on the work performed by Prof. William Gelbart at Harvard and the Flybase Harvard curators (see <http://www.flybase.org/> for information on Flybase, a publicly available database on the genetics and molecular biology of *Drosophila*). In consultation with the Flybase curators, we explored automated aids for curating biomedical databases; after exploring several alternatives, we settled on a fundamental task at the beginning of the Harvard Flybase curation "pipeline", namely identification of the papers to be curated for *Drosophila gene expression information*.

The Flybase criterion for curation for gene expression is:

Does the paper contain *experimental evidence* of interest to these curators for *gene expression*, specifically, information about *the gene products (mRNA transcripts or polypeptides or proteins)* 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 whether the full paper has experimental evidence for gene products (mRNA and/or protein).

Thus for each paper containing experimental evidence of interest on gene expression, we asked that a system 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).

We selected the database curation task for several reasons. First, database curation is a critical challenge to working biologists: keeping these databases up-to-date with respect to the increasing amounts of new research literature is a real problem facing the genomics community. Second, curated databases provide a large annotated training data set for machine learning methods, provided these methods can be adapted to use the "weakly annotated" data available. Specifically, biological databases contain pointers to the articles from which the database entries are derived. However, in general, these annotations are only at the article level. There is no indication of the specific section or paragraph or sentence that provided the basis for the annotation. To date, most linguistically based methods have required much finer-grained annotation (at the phrase level) and have also required completely annotated data, where unannotated material is interpreted as negative examples. However, such "coarsely annotated" data sets are appropriate for data mining and classification techniques, and for machine learning methods that rely on unsupervised methods or weakly labeled training data.

The preparation of the training and test data for the evaluation were completed to fit within the KDD Challenge Cup schedule, which included a 6 week period when the training data were made available, and 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. This allowed us to side-step a very difficult issue related to gene and protein nomenclature. The typical situation is that one gene has many synonyms, and frequently one term may be ambiguous between gene or protein, and on occasion, between multiple genes.<sup>1</sup> The Flybase database provides standardized nomenclature for the genes, and also provides synonym lists for each gene. These resources, along with the set of relevant database entries for each article, were provided as part of the training data. The test set consisted of another 249 articles, with the genes mentioned in each article.

One set of challenges concerned access to the literature for automated text processing. Many of the automated text processing systems are designed to handle plain text. In the literature for this task, things like superscripts, subscripts, Greek letters (in English text), italics and sometimes even figures and tables convey important information. For HTML versions of full text papers, it was necessary to "translate" some of the HTML conventions (subscripts, bold font, Greek letters) into an ASCII representation for further processing. Another complication was that full articles are harder to obtain than abstracts. Full text articles may be unavailable for free download, or may occur in difficult-to-read formats (pdf). By contrast, abstracts are easy to download through Medline. However, the task required the full articles in order to determine whether there was experimental evidence for the findings, since many findings are mentioned in the full articles, but not in the abstracts.

Relying on the existing database for training data presented other challenges. Not surprisingly, curation standards change over time and differ between individuals. 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. Even with this reverse engineering, some residual noise remains in the training data.

A third set of challenges comes from a mismatch between natural language processing (NLP) systems and curation for FlyBase. NLP systems are mainly designed to find/extract information that is explicitly

---

<sup>1</sup> The difficulty related to naming genes, transcripts and proteins prevented us from defining a more conventional extraction task. We determined that naming conventions for newly discovered transcripts and proteins was quite complex. For example, an initial mention might consist of a description ("a 145-kDa membrane-associated precursor"). In the database, these are generally just numbered, e.g., Appl-P1 (Appl gene, first protein mentioned). This kind of naming would have been virtually impossible for participants to reproduce, so we abandoned the idea of specific extraction and moved to a classification (yes/no) approach for the task.

mentioned in the text (text strings), with perhaps some limited normalization or stemmed involved, while FlyBase curation often involves finding information that is deduced fairly indirectly from what the text actually states. One example of inference is that curators often conclude an experiment is using the immunolocalization method without seeing any mention of the term "immunolocalization" (or any similar term) in the text. Instead, the curators conclude this by reading (in the text) descriptions of the various steps taken to perform an immunolocalization analysis. Another example is that curators may conclude the existence of mRNA transcripts in certain locations without seeing any mention of transcripts in the text. Instead, the text describes an association of a reporter protein with the gene for the transcript, and then reports where the reporter protein is detected.

Because of these issues, we simplified the task as presented to the KDD contestants. We performed the simplification by taking out many parts of the original task involving extraction and inference and putting more emphasis on the part of just making a yes/no decision on which papers contain any useful information. Even with this simplification, the task remains one of real importance to the curators, because most of the papers given to them contain no information of interest, and filtering out such papers is useful. (We have since been approached by other database curation groups, asking if we would be interested in using their databases as test cases for a similar evaluation, because they need these kinds of tools in their daily work).

After defining the task and preparing the training and test data, we developed three simple scoring methods for each of the three subtasks. For the ranking task, we used as a metric the area under the receiver operating characteristic curve (AROC); the ROC curve measures the trade-off between specificity (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<sup>2</sup>; 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.

We had over 40 groups who downloaded the training data. Of these, 18 teams obtained the test data and submitted 32 separate results for evaluation (up to 3 per team). There were seven geographic regions/countries represented, including Japan, Taiwan, Singapore, India, UK, Portugal, USA. There were groups from industry, academia and government laboratories, often teamed. The top performing team (ClearForest teamed with Celera) will present a talk on their approach at the KDD Conference; this team obtained both the highest overall score and the highest scores on the individual tasks. The results for the three metrics are shown below.

Evaluation Task	Best	Median
Ranked-list:	84%	69%
Yes/No curate paper:	78%	58%
Yes/No gene products:	67%	35%

Once one of the issues of concern to us, in creating the evaluation, has been to determine some way to measure the success of the evaluation itself -- a kind of meta-evaluation. Some obvious criteria are:

- Did we get participants? Yes -- many groups expressed interest in using the data after the test was completed, saying that the timing of the Challenge Cup itself was bad, but they are interested.
- Did we make the materials accessible to people from multiple fields? Yes, although having an interdisciplinary team was probably very useful.
- Was the task tractable? Yes, apparently at least the curation/no-curate decision at the paper level was a reasonable task; clearly, finer-grained extraction at the individual gene level was much harder.
- Was the cost of the evaluation reasonable? We estimate that it took about a staff year of effort overall: 9 staff months of our time, and probably 3 staff months from Harvard FlyBase. If we were

---

<sup>2</sup> The balanced F measure is  $(2 * \text{precision} * \text{recall}) / (\text{precision} + \text{recall})$ , where recall is number of correct "yes" decisions returned by the system over the number of possibly correct "yes" decisions; precision is the number of correct "yes" decisions returned by the system over the total number of "yes" decisions returned.

to do this again, however, the cost would be significantly lower, since part of the cost was the learning curve associated with FlyBase, part had to do with KDD standards, and part had to do with learning and understanding enough biology to set up the evaluation.

- Is this a useful task to biologists, such that groups would be interested in repeating variants of this task? The verdict is still out, but there seem to be many databases, and the curators for these databases have similar problems. If we could demonstrate that the tools are close to being usable, then there would be significant interest in this class of task.

In conclusion, we believe that this has been a highly educational experience for us, and we hope an equally profitable one for our team member, Harvard Flybase, as well as all the participants in the KDD Challenge Cup. We believe that a text data mining approach can leverage existing technologies and provide value-added to working biologists, while also pushing the state of the art in text data mining and information extraction.