Overview of BioCreAtIvE task 1B: Normalized Gene Lists

Lynette Hirschman[1 §], Marc Colosimo[1], Alexander Morgan[1], Alexander Yeh[1]

[1]The MITRE Corporation, 202 Burlington Road, Bedford, MA 01730, USA

[§]Corresponding author

Email addresses:
LH: lynette@mitre.org
MC: mcolosimo@mitre.org
AM: amorgan@mitre.org
AY: asy@mitre.org

Abstract

Background

Our goal in BioCreAtIve has been to assess the state of the art in text mining, with emphasis on applications that reflect real biological applications. To this end, we have focused on the curation process for model organism databases. This paper summarizes the BioCreative task 1B, the “Gene Identifier List” task, which is inspired by the gene list typically supplied for each curated paper in a model organism database. For the assessment, systems were given a set of abstracts from each of three model organism databases (Yeast, Fly, and Mouse), along with synonym lists for these organisms that define the correspondence between unique gene identifiers and the mentions of these genes and gene products in the curated literature. The systems were evaluated on their ability to produce the correct list of unique gene identifiers for the genes and gene products mentioned in the abstracts for each organism. For the evaluation, we prepared a training data set of 5000 abstracts per organism with (noisy) gene lists derived automatically from the gene lists for the full text articles; a development test data of 100-200 abstracts per organism with hand-corrected gene lists; and a blind test set of 250 abstracts per organism with carefully annotated gene lists.

Results

We report results from 8 groups fielding systems for the three data sets (Yeast, Fly, and Mouse). The results are reported as balance F-measure (the harmonic mean of precision and recall). For Yeast, the top scoring system (out of 15 systems) achieved an F-measure of 0.92; for Mouse and Fly, the task was more difficult, due to larger numbers of genes, and more ambiguity in the gene naming conventions. For Fly, the top F-measure was 0.82 out of 11 systems and for Mouse, it was 0.79 out of 16 systems.

Conclusion

The gene list task 1B builds on findings from BioCreative task 1A, identification of gene mentions in text. Systems achieved scores of over 80% F-measure for task 1A. Task 1B added an additional complication, namely the mapping of a gene name to its unique identifier. Because there is extensive ambiguity in gene names, task 1B required that systems distinguish between common English words and gene names (“dorsal”, “yellow”), as well as between multiple possible identifiers associated with a particular gene name (e.g., actin appears in FlyBase as a synonym for 6 genes). Because the nomenclature for Yeast is fairly unambiguous, systems did well, indicating that this application could be automated for production use in curation. Both Mouse and Fly presented different challenges, namely gene names with greater ambiguity and more complex names; these applications will require further research to reach useful accuracy. However, the major finding is that multiple groups were able to perform a real biological task across a range of organisms, given lexical resources derived from the model organism databases. This holds out great promise for partial automation of the curation process in the next few years, as natural language processing applications refine techniques and increase their ability to be quickly adapted to new model organism curation tasks.

Background

Why Evaluate?

Our goal in organizing BioCreative was to provide a systematic assessment of the state of the art for a set of “building block” biological tasks. There has been increased activity in the field of text mining and information extraction applied to the biological literature. However, each group has tackled a different problem and reported on a different data set [Hirschman 2002a; Hirschman 2002b]. With BioCreative, our goal was to assemble a suite of tasks that would:

- Attract researchers from natural language processing and bioinformatics;
- Address problems of importance to the biology and bioinformatics community;
- Create legacy training and test data suites that could be used for development and benchmarking of future applications.
- Permit the assessment of the state of the art on real biological tasks.

We chose to frame these tasks in terms of aids for the curation of biological databases. This built on earlier work in organizing one of the first challenge evaluations in text mining for biology [Yeh 2003], which also focused on a task related to the curation of biological literature, namely the identification of articles containing experimental evidence for gene products for Flybase [Flybase].

Choice of Evaluation Task

In designing the tasks for BioCreative, we were motivated by several factors: first, the need to define meaningful biological applications; second, the availability of training and "gold standard" test data; third, the need for a simple evaluation procedure; and fourth, the need to attract participants from fields such as natural language processing and text mining, as well as from bioinformatics.

By choosing tasks related to the curation process of some of the major biological databases, we guaranteed that the tasks would have biological relevance, since these are tasks that are presently performed by expert human curators. This also meant that there would be "gold standard" annotated data available: annotations produced by experts that could be used to as training data for system development and as an evaluation standard for the blind test data.

Task 1B, the normalized gene list task, is intermediate in the BioCreAtiVe tasks. It builds on task 1A, the gene mention identification task [Yeh 2004], but it is much simpler and requires far less understanding of the underlying biology than task 2, functional annotation from text [Blaschke 2004]. It reflects a step in the curation process for the model organism databases: once an article is selected for curation, an important step is to list those genes discussed in the article that have sufficient experimental evidence to merit curation – see discussion in [Colosimo 2004]. Therefore, we were able to extract the expert-curated gene lists from the model organism databases, to use as training and test data.

By defining the task as the generation the list of unique gene identifiers, we made the evaluation much simpler than for either task 1A or task 2. Task 1A required the comparison of annotated text segments, raising issues of how to annotate and complex gene names (e.g., *TTF-1-binding sites (TBE) 1, 3, and 4*), as well as questions about gene name boundaries. Task 2 required expert human evaluation of whether a text passage constitutes adequate evidence for a particular Gene Ontology annotation. By contrast, the gene list task simply required the comparison of a proposed set of gene identifiers against a "gold standard" list. This made the actual evaluation process very straightforward. Originally we had also wanted evidence for each answer, to provide a safeguard against manual curation of the articles, but our instructions for this were not clear, different people submitted different things and we did not evaluate this.

In order to make the task as accessible as possible, we extracted synonym lists for Fly [FlyBase], Mouse [MGI], and Yeast [SGD]. These consisted of the list of unique gene identifiers and their associated gene symbol and synonym lists. We made these lists available in a simple standard flat file format.

We chose to use abstracts as the basis for the gene list task, rather than full text articles. This simplified the task for the participants, since abstracts are much shorter and easier to process than full text article (because they are around 250 words long and are available in ASCII). The abstracts can also be readily collected and distributed to the participants, unlike the full text articles. However, using abstracts meant that we had to “prune” the gene lists provided by the model organism database, since these were usually based on the full text articles (although withYeast, curators often only read the abstract to create the gene list). We developed an automated pruning procedure to remove genes from the gene list that were not mentioned in the abstract. As discussed in [Colosimo 2004], this was a “noisy” process. We delivered the noisy training data “as is” but we hand corrected the development test data and the blind test data. In addition to pruning the gene lists to reflect the content in the abstracts, we made one additional simplification in the task. The model organism databases do not curate every gene mentioned in a paper – they curate only those genes that meet a set of (organism-specific) criteria, including presentation of experimental evidence related to gene or gene included in the gene list. However, we felt that the abstract might not provide enough context to determine whether a gene had sufficient evidence for curation or was mentioned only in passing, so for the test data sets, the annotators added by hand all genes mentioned in the abstract. This was not done for the automatically generated training data, so the automatically generated training set had significant recall errors (see Tables 2-4).

Task Description

The task is defined as follows: given an abstract from a specific model organism (Fly, Mouse or Yeast), the task is to create the list of unique gene identifiers for the genes that are mentioned in the abstract (see Figure 1). These mentions will include explicit mentions of genes (gene names, e.g., *esterase-6* and gene symbols, e.g., *est-6*, as well as gene mentions implicit in mentions of gene mutants, alleles, and products, e.g., *esterase-6* in the sentence *some allozymes of the enzyme esterase 6 in Drosophila melanogaster*…. Genes must come from the appropriate organism for the specific database (e.g., Drosophila melanogaster for FlyBase, Saccharomyces cerevisiae for Yeast, Mus musculus for Mouse) and must be identified by their unique gene identifier from the database.

Figure 1 shows the gene list for an abstract, along with entries from the FlyBase synonym list. At the top of the figure is the gene list, given as a set of pairs of (abstract number, gene identifier). In the middle is the text of the abstract. At the bottom are the entries for the two Fly base genes on the gene list (FBgn0000592 and FBgn00026412).

There are thus two distinct aspects to the gene list task: finding the gene mentions for the specific organism (similar to Task 1A); and mapping the gene mention to the appropriate unique gene identifier, which requires resolving ambiguities. It is clear from the 18 entries for the gene Est-6 that the synonym list contains many variants – some of which are not obvious (e.g., Est-5 is listed as a synonym of Est-6). In addition, some synonyms are ambiguous: EST also occurs in the genomic literature as an abbreviation for Expressed Sequence Tags.

Results

Tables 2-4 show the scores from each participating system, by group and run (each run was considered a system) for Yeast (Table 2), Fly (Table 3) and Mouse (Table 4). Each user was allowed to submit up to three systems for each organism. The systems were scored against the manually created “gold standard” for each abstract in the test set (250 abstracts per organism). The results are presented in terms of the following metrics:

True Positives:		Number of correctly detected genes

False Positives:		Number of genes incorrectly marked as being present

Misses:			Number of genes NOT detected by the system

Precision: True Positives / (True Positives + False Positives)

Recall: True Positives/ (True Positives + Misses)

F-measure: Balanced precision/recall computed as 2*P*R/(P+R)

The first two rows of each table show first the **Gold Standard** compared to itself, which always yields a score of 100% or 1. The second line, **AutoFound,** shows the results of comparing the test data run through the “automatic cleaning” procedure and compared to the Gold Standard. This provides an estimate of the quality of the automatically generated training data.

Next, for each organism, we show High, Median and Low scores for each of these quantities, followed by the scores of each user by run.

In addition to the tables, Figure 2 shows a composite graph of precision versus recall for all systems and all organisms. This graph also shows the estimates of training data quality (marked as Yeast Train, Fly Train and MouseTrain in the legend and in solid symbols on the graph). The diagonal line indicates balanced precision versus recall.

The results demonstrate several things, in particular that there are significant differences among organisms.

1. Yeast is the easiest. The F-measures of the systems tended to be high, with several groups achieving an F-measure of over 0.90, and a median F-measure of 0.86. Also, the quality of the training data was high (F-measure 0.92).
2. Fly was harder than Yeast: the high F-measure was 0.82, and there was much greater variability in performance (median F-measure was 0.66). The training data quality for Fly was significantly lower than for Yeast (0.83). Fly was hard because there are many ambiguous terms, and also extensive overlap between Fly gene names and abbreviations and English words, as in “not”, “period”, “was”, etc.

3. Mouse was the hardest, as measured by system performance (best F-measure 0.79), although the median system performance for Mouse was better than for Fly (0.74). The training data quality was the lowest (F-measure of 0.71). The poor training data quality was related to the stringent Mouse curation criteria. As a result, there were relatively many more genes that were "mentioned in passing" and needed to be added back into the Gold Standard. These genes were not included in the automatically generated training data (hence the low recall and low F-measure for the training data). Indeed, for Mouse, the median system F-measure was actually higher than the training data F-measure, indicating that the systems did a good job in generalizing away from the noise.

A second observation is that systems may have been limited by the quality of the noisy training data. For both Yeast and Fly, the estimated training data quality was just a shade higher than the final top performing systems.

Methods

This section discusses the methods used to prepare the evaluation materials.

Data Preparation

In order to evaluate the performance of the systems, the organizers prepared a hand-coded gold standard, as described in [Colosimo 2004]. First, each abstract was associated with the gene ID list from the appropriate model organism database. Since we were using abstracts rather than full text, the gene list from the model organism database then had to be "adjusted" to conform to the names mentioned in the abstract. This was done in several steps, as follows:

- Removing gene IDs that were not found in the abstract, but were found in the underlying full text article. This was done automatically, using the synonym list, to generate large quantities of “noisy” training data. This corresponds to the **Auto Found** column on the tables for the model organism performance data.

- Hand checking to make sure that the automatic procedure did not eliminate genes that were present in the abstract (development test set and blind test set only). This could occur if, for example, the mention in the text was a variant of the synonyms provided in the lexical resource, e.g., “polgamma B” versus “polgamma 2”.
- Adding in any additional genes mentioned “in passing” in the abstract (development test set and blind test set only). This was necessary because each model organism database curated genes according to a certain set of criteria, so not all genes mentioned were necessarily on the gene list. There might, for example, be additional genes mentioned “in passing” such as genes located near a gene of interest, or possible homologues etc.

Overall, we estimate that it took between 1-2 staff weeks of time from an experienced curator to edit and check a 250 abstract test set. The checking was particularly important because we detected significant interannotator variability, particularly for the Mouse annotations – see [Colosimo 2004] for a detailed discussion of the data preparation and interannotation agreement studies.

Lexical Resources

An analysis of the lexical resources provides insight into the differences in difficulty observed for the three organisms. Table 5 shows the number of unique identifiers (IDs), the number of synonyms and the average number of synonyms per identifier for each organism. We can see that the Yeast resources are the most parsimonious (1.9 synonyms per ID), and Fly the richest (2.9 synonyms per ID). In addition, the last column shows the average length (in words) for the synonyms. Again, Yeast is very compact, with barely over one word per synonym; however, Mouse has the longest synonyms on average, at 2.77 words per synonym. This may have contributed to recall problems in identifying Mouse gene mentions, since longer names tend to be more descriptive and therefore, to show significant syntactic variation. Also, longer names are more difficult to identify (see [Yeh 2004], this volume, for a discussion of these issues in the context of task 1A).

The resources for these organisms also differ in amount of ambiguity among the synonyms. This is shown in Table 6. The 4th column of this table lists the absolute number of synonyms that were associated with multiple gene identifiers. Some of these were associated with many identifiers – see Figure 3 for the distribution of synonyms associated with multiple gene identifiers. Again we observe that Yeast is the least ambiguous (168 synonyms and an average of 1.013 identifiers per synonym, column 5), while Fly, with the most synonyms on average per gene, is also the most ambiguous, at 1.085 gene identifiers per synonym. The 6th column shows the absolute number of synonyms that overlap with the 5000 most common English words, and the last column shows the average number of ambiguities with English words per synonym. While this is very low for Yeast (0.00014), it is over 10 fold higher for Mouse(0.00171) and about four-fold higher than that for Fly (0.0065). These figures correlate with the differences in difficulty between Yeast, Fly and Mouse. Yeast was relatively easy, with few problems of ambiguity; Fly and Mouse were both significantly harder, for slightly different reasons. The Fly lexical resources were the richest, and as a result, the most ambiguous (with respect to gene identifiers and also with respect to overlap with regular English words). Mouse, on the other hand, had longer names and a more limited set of synonyms. This required more complex gene mention and gene ID matching procedures; this may have been offset by having reduced problems with ambiguity. In addition, Mouse had the noisiest training data, which may have contributed to the difficulty of the task. The high scores for Mouse and Fly were quite similar: for Fly, the high recall was 0.841, precision 0.831 and F-measure of 0.815 (all these scores were from the same group [Hanisch 2004]); for Mouse, high recall was 0.898, precision 0.828, and F-measure 0.791; for Mouse, these three high scores came from three different groups.

Discussion

There were eight groups participating in task 1B; 7 groups submitted 15 systems for Yeast; 6 groups submitted 11 systems for Fly; and 7 groups submitted 16 systems for Mouse.

Of the eight participating groups, two groups did not submit extended write-ups and are not discussed in detail here. Four systems are documented in articles in this issue [Crim 2004; Fundel 2004; Hanisch 2004; Tamames 2004]. For descriptions of the other two systems, see [Hachey 2004; Liu 2004] in the BioCreative Workshop Handout. The remainder of this section discusses the challenges presented by task 1B and how the participating systems approached these challenges.

Technical Challenges for Task 1B

The requirements for task 1B can be divided into four inter-dependent steps:

- Identifying gene mentions in the text
- Associating gene mentions to one or more unique gene identifiers
- Selecting the correct gene identifier in cases of ambiguity
- Assembling the final gene list for each abstract

These steps were highly interdependent. There are complex recall/precision trade-offs that occur in capturing candidate gene mentions and in assigning a unique (and correct) gene identifier to these mentions. This is because of significant ambiguity among gene synonyms (one word might be a synonym for multiple genes) and also because of significant overlap between gene synonyms (“white”, “dorsal”) and English vocabulary. At the same time, the synonym lists provided by the model organism databases, while extensive, were by no means exhaustive; also in some cases, very confusable synonyms were entered, such as the synonym “A” for “abnormal abdomen” in FlyBase. As noted above, the lexical resources differed in number of synonyms per gene identifier and in ambiguity of terms within the resource.

Precision errors could be caused by:

- False alarms for gene mentions (for example, taking an English word to be a gene name)
- Incorrect disambiguation of ambiguous gene names
- Assignment of gene identifiers to genes from non-relevant organisms (e.g., human genes are often discussed in Mouse abstracts, but should be entered into the gene list)

Recall errors could be caused by:

- Failure to recognize a gene mention (perhaps due to mismatch with the organism-specific synonym list)
- Incorrect disambiguation of ambiguous gene names

Finding Gene Mentions

The participating groups took a variety of approaches to these challenges. For gene mentions, the approaches fell into roughly two groups:

- Matching against the lexical resource; in many cases, an approximate matching approach was used. For example, [Crim 2004] used exhaustive pattern matching against the synonym lists to generate a high recall system (91% for fly; 79% for mouse; and 90% for Yeast), but with very low precision, similar to [Morgan 2004]. The approach described in [Liu 2004] also used an enriched lexical resource to achieve high recall (but lower precision) results for Mouse and Yeast.

- Gene mention identification as done for task 1A, adapted to the three specific organisms in 1B [Hachey 2004]. To do this, Hachey et al used a technique to generate "noisy" training data similar to that described in [Morgan 2004].

Association with Unique Gene Identifier

The second stage, association with a unique identifier, was essentially a table look-up. For groups that used a task 1A-type gene mention tagger, they were then able to use the table look up to filter out erroneous gene mention candidates. However, recall at this step was limited by the completeness of the synonym list from the model organism database. While the synonyms contained many variant forms (see the example with *Est-6* in Figure 1), there were still more variations that had to be handled. The incompleteness of the lexical resources could lead to recall errors. This was also the stage at which ambiguity was flagged, since some synonyms could refer to multiple genes (see Table 5). A number of groups chose to edit the lexical resources, removing highly ambiguous or uninformative terms and adding additional variants or descriptions [Crim 2004; Fundel 2004; Hanisch 2004; Tamames 2004]. The systematic editing and expansion of the underlying lexical resources was at the core of two high performing systems [Hanisch 2004; Fundel 2004]. Both Tamames [2004] and Liu [2004] used the same tokenization for the lexicon as was used for the gene mention identification and also used stemming to improve the matching between lexicon terms and candidate gene names in the text.

For several groups, the gene mention tagging, gene identifier look-up and disambiguation were interleaved; for example, Hanisch et al [2004] accrued evidence during the process of identifying candidate gene mentions that was then used to disambiguate the gene mention to a specific gene identifier. For Tamames [2004], these stages were also combined.

Disambiguation

The final stage, disambiguation for gene synonyms associated with multiple identifiers, turned out to be the most interesting feature of this task. The extensive ambiguity of gene names, particularly for Fly and to a lesser extent, for Mouse (see Figure 3), required that systems develop techniques for disambiguation. These included pruning the lexicon or accumulating multiple sources of contextual evidence for use in a classifier. Hanisch et al [2004] used a multi-stage process that included correlating abbreviations with their long forms and also a filter for abstracts based on organism specificity. Liu [2004] used features derived from rich lexical resources to create feature vectors used in word sense disambiguation. Crim [2004] followed their high recall pattern matching system with a maximum entropy classifier trained to distinguish correct matches from bad matches. Hachey et al [2004] used information retrieval techniques to associate candidate gene identifiers with term frequencies in a document. They used this to create a pool of candidate gene identifiers for a given abstract, based on its similarity to abstracts in the training data.

Generating the Final Gene List

Once these stages were completed, the systems assembled the final gene list for each abstract as output. For some groups, this stage was parameterized in terms of a certainty threshold. Increasing the threshold traded recall for precision, e.g., in [Hanisch 2004] and [Liu 2004]. One group [Crim 2004] was able to achieve reasonable performance (well above the median of the reported systems) using a single approach across all three organisms, based on high recall pattern matching, followed by a maximum entropy classifier for remove bad matches. Many groups found that it was possible to use much simpler techniques for Yeast than for Mouse or Fly, due to the more tightly constrained nomenclature.

Conclusions

Overall, BioCreative demonstrated the ability of automated systems to do gene normalization for a range of organisms, given a simple lexical resource consisting of the set of unique gene identifiers and their names and synonyms. The actual performance depended more on the organism than on the kind of system. Factors included the number of genes, the number of synonyms per gene identifier, the consistency of naming conventions, the length and complexity of names, and the degree of ambiguity in the naming conventions. The more ambiguity (among genes, between genes and English) and the more complex the names (descriptions versus simple gene symbols), the harder the problem. Yeast naming is relatively simple and regular -- and good performance could be achieved with relatively simple methods (such as expanded lexical look-up). Fly is hard because of ambiguity of short names, both with English words and among gene names; the Flybase lexicon is quite large, with many synonyms per gene; for this task, editing the synonym lists turned out to be a useful technique for reducing ambiguity. Mouse is hard because names are often long and descriptive, subject to many variants (grammatical as well as syntactic and typographic). Mouse was also harder because of the our decision to simplify that task to include all gene mentions; this required that the annotators add many genes in by hand, which made training and test data preparation difficult (and somewhat less reliable than other organisms).

It is always important to evaluate the evaluation. The success of an evaluation can be gauged by several criteria:

- The level of participation: did the evaluation attract good researchers from diverse groups and backgrounds?
- The results: was the task sufficiently challenging, but not too easy?
- The research: does the task raise important and interesting research questions?
- The relevance of the application: does the evaluation task have applicability to some application that users care about?
- The data: was there sufficient training and test data? Will these resources be available to the larger research community after the evaluation, for further benchmarking?
- Repeatability: Would people want to do this again?

By all the these criteria, the BioCreative task 1B evaluation was a success.

Participation

We attracted 8 groups from five countries and from some of the major groups involved in information extraction in biology

Results

The results were promising – with the results for Yeast sufficiently good to think about insertion into a production system. The results for Fly and Mouse were lower (around 80% F-measure), and it is not clear that this would be good enough for a production quality system – so there is more to do.

Research

The task raised three interesting research questions. (1) How to achieve high recall (achieving high precision seems relatively easy, but only one system achieved high recall, at the expense of precision). (2) How to disambiguate ambiguous synonyms, including both abbreviations or short forms of gene names, and longer forms. This problem requires word sense disambiguation, but this is a new way of framing the problem that should provide an interesting testing ground for various approaches to the problem. (3) How to do rapid adaptation to different task domains, given appropriate lexical resources (synonym list for the organism gene identifiers). Some of the successful systems found that the different organisms benefited from somewhat different approaches. And several systems made use of additional lexical resources.

Relevance

The task of listing gene names is a task that is currently performed (manually) by curators for various model organism databases. In addition, ability to identify and map gene names to normalized gene identifiers would have great applicability for search and indexing operations.

Data

We were able to use our "noisy" training data, though the noisy data may have imposed limitations on system performance. The cost of preparing the training and test sets was greater than we expected: 1-2 person weeks of expert annotator time for a 250 abstract test set. And the difficulties of achieving reliable interannotator agreement were greater than we expected. The training and test data are now available for other groups to use in further experiments.

Repeatability

The participants at the workshop seemed interested to repeat the evaluation. For task 1B in particular, we need to consider several questions. First, is the task realistic enough? The real task that curators perform uses full text articles (not abstracts, although the Yeast curators do curate from abstracts part of the time). Furthermore, the real task involves a biologically complex set of criteria about which genes to list and which genes that fall outside the scope of what is curated (for example, they belong to another organism, or they are only mentioned in passing). It would be far easier for the organizers to prepare "real" data sets, because it would require none of the editing that was performed for this year's BioCreative task 1B. On the other hand, it would be harder for the participants, because they would have to handle full text and they would have to replicate biological decisions in terms of which genes to list. In conclusion, we look forward to receiving feedback from the participants and to defining a follow-on task for the next BioCreative evaluation.

Acknowledgements
This paper reports on work done in part at the MITRE Corporation under the support of the MITRE Sponsored Research Program and the National Science Foundation (contract number EIA-0326404). Copyright © 2004 the MITRE Corporation. All rights reserved.

References (convert [ ]'s to numbers later)

1. [BioCreAtIvE 2004].
2. *BioCreAtIvE Workshop Handouts*, Granada, Spain, March 2004. http://www.pdg.cnb.uam.es/BioLINK/workshop_BioCreAtIvE_04/handout/index.html
3. [Blaschke 2004]
4. Blaschke C, Valencia A. BioCreative Task 2 Overview. **BMC Bioinformatics** this issue.
5. [Colosimo 2004]
6. Colosimo, M, Morgan, A, Yeh, A, Colombe, J, Hirschman, L. Data Preparation and Interannotator Agreement: BioCreAtIvE Task 1B. **BMC Bioinformatics** this issue.
7. [Crim 2004].
8. Crim J, McDonald R, Pereira F: Automatically Annotating Documents with Normalized Gene Lists. **BMC Bioinformatics** this issue.
9. [FlyBase]
10. **The FlyBase Database** [http://flybase.org/]
11. [Fundel 2004]
12. Fundel K, Güttler D, Zimmer R, Apostolakis J. A simple approach for protein name identification: prospects and limits. **BMC Bioinformatics** this issue.

13. [Hachey 2004].
14. Hachey B, Nguyen H, Nissim M, Alex B, Grover C: Grounding Gene Mentions with Respect to Gene Database Identifiers. *BioCreAtIvE Workshop Handouts*, Granada, Spain, March 2004.
15. [Hanisch 2004]
16. Hanisch D, Fundel K, Mevissen H-T, Zimmer R, Fluck J. ProMiner: Organism-specific protein name detection using approximate string matching. **BMC Bioinformatics** this issue.
17. [Hirschman 2002a].
18. Hirschman L, Park JC, Tsujii J, Wong L, Wu CH: Accomplishments and challenges in literature data mining for biology. *Bioinformatics* 2002, **18:**1553-1561.
19. [Hirschman 2002b].
20. Hirschman L, Morgan A, Yeh A: Rutabaga by any other name: extracting biological names. *J. of Biomedical Informatics* 2002, **35:**247-259.
21. [Liu 2004]
22. Liu H. BioTagger: A Biological Entity Tagging System. BioCreative Workshop Handouts, http://www.pdg.cnb.uam.es/BioLINK/workshop_BioCreAtIvE_04/handout/index.html
23. [Morgan 2004]
24. Morgan A, Hirschman L, Colosimo M, Yeh A, Colombe J. Gene Name Identification and Normalization Using a Model Organism Database. **J Biomedical Informatics** to appear.
25. [MGI]
26. **The Mouse Genome Database** [http://www.informatics.jax.org]
27. [SGD]
28. **Saccharomyces Genome Database** [http://www.yeastgenome.org]
29. [Tamames 2004]
30. Tamames J: **Text Detective:** Text Dectective: A rule-based system for gene annotation in biomedical texts. **BMC Bioinformatics** this issue.
31. [Yeh 2003].
32. Yeh AS, Hirschman L, Morgan AA: **The Evaluation of text data mining for database curation: lessons learned from the KDD challenge cup.** *Bioinformatics* 2003, **19:**i331-i339.
33. [Yeh 2004]
34. Yeh, AS, Hirschman, L, Morgan, A, Colosimo, M. BioCreAtIvE task 1A: gene mention finding evaluation. **BMC** this issue.

Figures

Figure 1 Abstract with gene list and synonym list excerpt

(see pdf)

Figure 2 Task 1B results for all organisms: precision vs. recall

(see pdf)

Figure 3 Distribution of ambiguous synonyms in Fly, Mouse and Yeast task 1B lexical resources

(see pdf)

Tables

Table 1: Task 1B training and test data sets

Abstracts	Yeast	Fly	Mouse
Training (noisy annotation)	5000	5000	5000
Development test (hand corrected)	108	110	250
Blind Test (extensively corrected)	250	250	250

Table 2: Task 1B results on Yeast gene list task

YEAST	**F-measure**	**Precision**	**Recall**	**True Positives**	**False Positives**	**Missed**
Gold Standard	**1**	**1**	**1**	**613**	**0**	**0**
AutoFound	**0.918**	**0.985**	**0.86**	**527**	**8**	**86**
Hi	**0.921**	**0.969**	**0.962**	**590**	**329**	**171**
Low	**0.763**	**0.642**	**0.721**	**442**	**15**	**23**
Median	**0.858**	**0.94**	**0.848**	**520**	**34**	**93**
user5_1B_1	0.819	0.948	0.721	442	24	171
user5_1B_2	0.848	0.915	0.79	484	45	129
user5_1B_3	0.848	0.969	0.754	462	15	151
user6_1B_1	0.857	0.912	0.809	496	48	117
user6_1B_2	0.858	0.907	0.814	499	51	114
user8_1B_1	0.921	0.95	0.894	548	29	65
user8_1B_2	0.91	0.95	0.873	535	28	78
user16_1B_1	0.897	0.951	0.848	520	27	93
user16_1B_2	0.899	0.966	0.84	515	18	98
user16_1B_3	0.897	0.951	0.848	520	27	93
user18_1B_1	0.904	0.94	0.871	534	34	79
user19_1B_1	0.773	0.646	0.962	590	324	23
user19_1B_2	0.77	0.642	0.962	590	329	23
user19_1B_3	0.763	0.661	0.902	553	284	60
user24_1B_1	0.897	0.917	0.878	538	49	75

Table 3: Task 1B Results on Fly gene list task

FLY	F-measure	Precision	Recall	True Positives	False Positives	Missed
Gold Standard	**1**	**1**	**1**	**429**	**0**	**0**
AutoFound	**0.834**	**0.863**	**0.807**	**346**	**55**	**83**
Hi	**0.815**	**0.831**	**0.841**	**361**	**684**	**266**
Low	**0.284**	**0.224**	**0.38**	**163**	**70**	**68**
Median	**0.661**	**0.659**	**0.732**	**314**	**146**	**115**
user5_1B_1	0.661	0.592	0.748	321	221	108
user5_1B_2	0.612	0.659	0.571	245	127	184
user5_1B_3	0.602	0.693	0.531	228	101	201
user8_1B_1	0.665	0.638	0.695	298	169	131
user8_1B_2	0.726	0.692	0.765	328	146	101
user16_1B_1	0.781	0.728	0.841	361	135	68
user16_1B_2	0.815	0.831	0.8	343	70	86
user16_1B_3	0.787	0.744	0.834	358	123	71
user18_1B_1	0.417	0.463	0.38	163	189	266
user19_1B_1	0.284	0.224	0.389	167	580	262
user23_1B_1	0.44	0.315	0.732	314	684	115

Table 4: Task 1B results on Mouse gene list task

MOUSE	F-measure	Precision	Recall	True Positives	False Positives	Missed
Gold Standard	**1**	**1**	**1**	**540**	**0**	**0**
AutoFound	**0.709**	**0.99**	**0.552**	**298**	**3**	**242**
Hi	**0.791**	**0.828**	**0.898**	**485**	**674**	**267**
Low	**0.571**	**0.418**	**0.506**	**273**	**69**	**55**
Median	**0.738**	**0.765**	**0.730**	**394**	**131**	**146**
user5_1B_1	0.672	0.767	0.598	323	98	217
user5_1B_2	0.737	0.811	0.676	365	85	175
user5_1B_3	0.619	0.798	0.506	273	69	267
user6_1B_1	0.739	0.813	0.678	366	84	174
user6_1B_2	0.745	0.785	0.709	383	105	157
user8_1B_1	0.744	0.828	0.676	365	76	175
user8_1B_2	0.661	0.635	0.689	372	214	168
user16_1B_1	0.772	0.75	0.794	429	143	111
user16_1B_2	0.777	0.807	0.75	405	97	135
user16_1B_3	0.791	0.765	0.819	442	136	98
user18_1B_1	0.686	0.728	0.648	350	131	190
user19_1B_1	0.58	0.428	0.898	485	648	55
user19_1B_2	0.571	0.418	0.898	485	674	55
user19_1B_3	0.606	0.489	0.798	431	451	109
user24_1B_1	0.767	0.735	0.802	433	156	107
user24_1B_2	0.776	0.764	0.787	425	131	115

Table 5: Lexical Resources: synonymy for Yeast, Mouse, Fly

	# ID	# Synonym	Synonym per ID	Avg Length (wds) per Synonym
Yeast	7,928	14,756	1.861	1.001
Mouse	52,594	130,548	2.482	2.772
Fly	27,749	81,711	2.944	1.470

Table 6: Lexical resources for Yeast, Fly and Mouse: identifiers, synonyms, and ambiguity

	# IDs	# Synonyms	Ambiguous Synonyms	Avg # IDs per Synonym	# Synonyms Overlap w English	Avg Eng Amb per Synonym
Yeast	7,928	14,756	168	1.013	2	0.00014
Mouse	52,594	130,548	1919	1.017	205	0.00171
Fly	27,749	81,711	2736	1.085	396	0.00650