Biological Nomenclatures: A Source of Lexical Knowledge and Ambiguity

O. Tuason, L. Chen, H. Liu, J.A Blake, and C. Friedman

Pacific Symposium on Biocomputing 9:238-249(2004)

BIOLOGICAL NOMENCLATURES: A SOURCE OF LEXICAL KNOWLEDGE AND AMBIGUITY

O. TUASON[1], L. CHEN[1], H. LIU[1],
J.A BLAKE[2], C. FRIEDMAN[1]
1. Department of Biomedical Informatics, Columbia University, 622 W 168 St, VC-5, New York, NY 10032
2. The Jackson Laboratory, 600 Main Street, Bar Harbor, ME 04609

There has been increased work in developing automated systems that involve natural language processing (NLP) to recognize and extract genomic information from the literature. Recognition and identification of biological entities is a critical step in this process. NLP systems generally rely on nomenclatures and ontological specifications as resources for determining the names of the entities, assigning semantic categories that are consistent with the corresponding ontology, and assignment of identifiers that map to well-defined entities within a particular nomenclature. Although nomenclatures and ontologies are valuable for text processing systems, they were developed to aid researchers and are heterogeneous in structure and semantics. A uniform resource that is automatically generated from diverse resources, and that is designed for NLP purposes would be a useful tool for the field, and would further database interoperability. This paper presents work towards this goal. We have automatically created lexical resources from four model organism nomenclature systems (mouse, fly, worm, and yeast), and have studied performance of the resources within an existing NLP system, GENIES[1]. Using nomenclatures is not straightforward because issues concerning ambiguity, synonymy, and name variations are quite challenging. In this paper we focus mainly on ambiguity. We determined that the number of ambiguous gene names within the individual nomenclatures, across the four nomenclatures, and with general English ranged from 0%-10.18%, 1.187%-20.30%, and 0%-2.49% respectively. When actually processing text, we found the rate of ambiguous occurrences (not counting ambiguities stemming from English words) to range from 2.4%-32.9% depending on the organisms considered.

1 Introduction

The amount of scientific literature has increased exponentially over the past few years, providing a rich source of genomic information. Recently, there has been increased activity involving exploration of natural language processing (NLP) and information retrieval (IR) methods to help extract, organize and facilitate access to the information. One type of application involves automatic extraction of genomic entities, such as genes and proteins[2, 3, 4, 5, 6]. In order to perform extraction, a system generally requires a resource that specifies and classifies genomic entities, that associates them with normalized terms and also unique identifiers that preferably are identifiers associated with a standardized nomenclature system so that the extracted entities are well-defined. In addition, an automated extraction system must be able to effectively utilize the resources.

Associating terms mentioned in text with specific biological entities is extremely challenging because 1) new genes are continually being named or known ones renamed, 2) the number of biomolecular entities is very large, 3) the nomenclature conventions differ for different organisms, 4) researchers do not

strictly follow standard naming conventions when they write articles, and 5) the names of biological entities are associated with synonymy and ambiguity: a gene can have multiple aliases (synonyms) in addition to its official symbol, and genes that are functionally different across species often have the same name (ambiguity). In addition to ambiguities among gene names, problems also arise when a gene has the same name as an English word, such as the genes named *was*, *nervous*, and *to*.

There are numerous specialized genomic databases, which are invaluable resources that were developed to assist biological researchers. These databases are also valuable for NLP purposes because they publish nomenclature and ontological specifications for biomolecular entities in online databases that are continually updated. Among these are model organism databases, such as Mouse Genome Informatics (*Mus musculus*: http:// www.informatics.jax.org, FlyBase (*Drosophila melanogaster*. http://www.flybase.org), WormBase (*Caenorhabditis elegans*: http://www.wormbase.org), and Saccharomyces (yeast) Genome Database (*Saccharomyces cerevisiae*: http://www.yeastgenome.org). Although these databases are resources for NLP, they were developed for different purposes, and therefore a variety of automated procedures must be developed to use them effectively for NLP. One issue is that they are heterogeneous: the database formats are different, as are the ontological specifications and naming conventions. Obtaining a uniform structure and semantics containing gene names and their unique identifiers is a crucial first step in recognizing and identifying them in the literature. A resource developed specifically for NLP that automatically acquires biological knowledge for NLP purposes from diverse resources, and that provides effective tools for utilizing the knowledge would be of great benefit to the NLP and research community. As a first step towards this goal it is important to study issues that influence the effectiveness of such a resource. The work reported in this paper has several aims. One is to develop a lexicon automatically for NLP use containing gene names from several model organism databases. Later, this will be expanded to other types of entities and organisms. The second is to study aspects of performance, especially ambiguity, when using the lexicon to process abstracts. We performed an experiment to test recall when using an existing NLP system GENIES[1] and the lexical resources that were generated. We analyzed the errors in order to categorize and determine the causes. Additionally, the ambiguous nature of the lexical resource that was created was quantified because ambiguous lexical entries pose difficult problems for NLP systems and lead to decreased precision. Ambiguity within each species, across all four species, and with general English words was measured.

2 Background and Related Work

2.1 Model Organism Databases

The research done here is based on the gene nomenclatures of four model organisms: mouse, fly, worm, and yeast, as mentioned above, because they have

excellent resources that are easy to access through their websites, the organisms are well-studied, much effort has gone into development of their nomenclatures, and their nomenclatures are mature. Their websites specify information needed for NLP such as official gene symbols, locus names, gene synonyms (aliases), unique identifiers, as well as other information, such as mappings to the same entity in other standardized nomenclature systems, such as Gene Ontology (http://www.geneontology.org). Additionally, the websites list associations betweens genes and journal articles, providing a reliable and cost-effective resource that can serve as a gold standard for evaluation.

2.2 Name Recognition Systems

There have been many systems and experiments described in the literature that employ different techniques for biological name recognition. Recognition of gene names is a partial solution: in order to obtain important biological information, identification of the exact gene being referred to is crucial, as the names serve as indices to the literature that contains the knowledge and the results[7]. Fukuda[2] developed the system PROPER which identifies protein names in the literature, using rules based on protein nomenclature. Another system[4] utilizes a name dictionary that contains human symbols and aliases extracted from different databases, such as HUGO, and LocusLink. An algorithm developed by Hanisch[5] uses name tokenization as well as a curated gene symbol dictionary to recognize protein names. Proux[3] also uses both lexical analysis and contextual analysis for recognizing gene symbols and names. In addition to protein and gene names, a system to recognize chemical names has also been developed[6]. Our system GENIES recognizes biological entities and also extracts their relations. GENIES can use either a straightforward lexical lookup method or process text that has already been tagged by a separate module.

Hirschman[7]performed a lexical-based pattern matching experiment for tagging genes using a list of genes symbols and synonyms obtained from FlyBase. A list of known associations between journal articles and gene names contained within each article served as the gold standard, against which the experimental results were compared. For the full text of the articles, this experiment yielded a precision of 2%. This experiment showed that problems in precision were largely due to gene name ambiguities (with each other, between genes and proteins, and with English words).

Our work differs from the above related work on biological entity recognition in that we are focusing on the automated acquisition of a uniform lexical resource for NLP and on issues affecting performance of the resource, whereas the related work focused on development of methods for recognizing gene names. Furthermore, in measuring performance we study issues associated with performance in conjunction with identification and not just recognition. Our work is similar to Hirschman's. However, we experiment with four different organisms and

quantify ambiguity within and across organisms as well as with general English words.

3 Methods

3.1 Creating a Lexical Resource and Measuring Its Ambiguity

We automatically created a lexicon from the four model organism databases. Specific files containing gene information were downloaded from the fly, mouse, and worm websites in January 2003. These files included *FBgn.acode*, *wormpep.93, MRK_List1.sql.rpt, MRK_LocusLink.rpt,,* and *MRK_Synonym.sql.rpt* from Flybase, WormBase and MGI. The file from the yeast database, *registry.genenames.tab*, was downloaded in June 2003. Since the file format for each different organism varied, the files were processed using different Perl scripts to extract gene symbols, aliases, full names, and identifiers, and to map the information to a single uniform format. For each organism, a gene name lexicon was created so that there was one entry per gene name, which contained the **gene name**, **unique database identifier**, and full name, if one exists. Figure 1a shows an example of three entries associated with the same name but denoting different genes. Also for the name of each entry, we kept track of whether it was an official symbol, synonym (alias), or full name, but this is not shown in the figure.

fbp1 MGI:109606^formin binding protein 1 fbp1 MGI:95492^fructose bisphosphatase 1 fbp1 MGI:95568^folate receptor 1 (adult)	fbp1 MGI:109606^forming binding protein 1+MGI:95492^fructose bisphos phatase 1+MGI:95568^folate receptor 1 (adult)
Figure 1a - Ambiguous gene name entries created from the MGI nomenclature. The name *fbp1* refers to three distinct genes, one corresponding to an official symbol and the other two to aliases**.**	**Figure 1b -** The merged lexical entry for *fbp1*. The target forms in **1a** were combined by concatenating the individual target forms, and a '+' was used to separate them.

After the initial lexical entries were created, an automated program merged all entries associated with the same name that had different target forms, so that all of the entries were combined into a single entry with a single target form consisting of the union of the individual target forms. Figure 1b shows an example of the merged entry. After merging entries, the number of ambiguities in the lexicons was counted.

Once the individual lexicons were created, we explored resources to use for identifying English words so that we could identify gene names that were ambiguous with general English. We explored three different resources, analyzed their effectiveness, and chose the best. We considered a resource effective if it did not intentionally contain genes names. The three sources were: 1) a list of English words obtained from the Moby lexicon project website

(http://www.dcs.shef.ac.uk/research/ilash/Moby/mwords.html) containing 74,550 English words, which occurred in two or more published dictionaries, 2) words obtained from the Wall Street Journal (WSJ) corpus, and 3) a list of words that occurred in Medline abstracts from 1969-2002. The WSJ corpus consisted of one million words selected from samples of articles appearing in 1988 and 1989. The words in the WSJ corpus were tagged with parts of speech; we eliminated words from the corpus that were tagged as proper nouns because they were not general English words. As a result of manual analysis of the different lists, we determined that the Moby lexicon was the most appropriate to use. Based on the list of English words in the Moby lexicon, we identified gene names that were also English words and computed how often they occurred within each organism database. They were then removed from each of the four lexicons. In addition, we found that the majority of single and double letter names, such as *a, a1,* and *to* were highly ambiguous even if some were not general English words, and also removed them from the lexicons, thereby creating individual lexicons MB (mouse), FB (fly), WB (worm), and SC (yeast). Therefore for each lexicon, each unique name had only one entry, which was not an English word.

The entries in each of the four lexicons were then combined to create a combined lexicon. Using the same merging process described above, entries for ambiguous names were merged. In the resulting lexicon each entry corresponded to a unique name, creating lexicon ALL4. Target forms for ambiguous gene names were combined as before except they could also consist of a union of identifiers associated with the four different nomenclatures. Ambiguous gene names were then quantified across species.

3.2 Evaluating Recall and Ambiguity

In independent runs, we used GENIES with each of the five lexicons to study performance, and to analyze problems. One aspect measured recall. This was accomplished by using the lexical lookup method in GENIES, which is a straightforward string matching procedure that finds the longest match. We used GENIES to capture other information as well, but for this work we only focus on gene name recognition. We realized that straightforward string matching was not an ideal method, but our aim was to perform an analysis of the instances where genes were missed, and to categorize and quantify the reasons, as a preliminary step to refining lexical lookup. Our intention was to determine the tools that would be most useful to accompany the lexicon. We focused on the mouse model organism (lexicon MB), and automatically obtained a gold standard set of abstracts by downloading a file named **MRK_Reference.rpt** from the MGI website, which listed MGI gene identifiers and the corresponding Medline abstracts containing those genes. This correspondence was established manually by curators[8], thus serving as an independent and accurate gold standard (ftp://ftp.informatics.jax.org/pub/reports/index.html). Based on this file, 45,000

Medline abstracts were obtained that contained at least one curated MGI gene. The abstracts were divided into two groups according to the number of MGI genes they were associated with: Group I had 26,000 abstracts that each corresponded to only one curated MGI gene, whereas Group II had 19,000 abstracts associated with two or more genes.

All 45,000 abstracts were parsed using GENIES with lexicon MB. The output that was generated for each abstract contained the target forms as specified in lexicon MB (see Figure 1b). For each abstract, the MGI identifiers obtained were compared with the MGI identifiers that were associated with that abstract in the gold standard. A true positive instance was considered to be one where the output contained at least one instance of the appropriate MGI identifier. Recall was calculated for each group and an overall average was computed. Recall was computed as the ratio of the number of appropriate MGI identifiers that were found divided by the total number that should have been found. In order to determine the cause of recall errors, a random sample of 100 abstracts associated with errors from each group (200 abstracts in all) was chosen, and an error analysis was performed by one of the co-authors (LC) who has a background in biology. When manual analysis of the abstract failed to detect the appropriate gene, the complete article was retrieved and examined to see whether mention of the gene occurred somewhere else in the article other than in the abstract.

In the next step, we determined the number of ambiguities in the output that contained the appropriately retrieved MGI identifiers. This phase consisted of three parts: one part involved using lexicon MB to determine occurrences of ambiguity within the mouse nomenclature, the second part involved using lexicon ALL4 to determine occurrences of ambiguities when considering all four nomenclatures, and the third, lexicon MBE, also included ambiguities with English words. To create lexicon MBE, gene names were added to MB that were previously removed because they were English words. Lexicons MB, ALL4, and MBE were each used by GENIES in three separate runs to process the set of abstracts. For each run, the number of MGI genes that were appropriately identified and that had more than one target form was counted and compared to the number of MGI genes that were appropriately retrieved.

4 Results

4.1 Ambiguity of the Lexical Resources

Table 1 shows the amount of ambiguity within each database, across all databases, and with English words for gene symbols as well as for all names, which includes gene symbols, full names, and aliases as listed in the individual nomenclature databases. In the mouse database, only 43 out of the 19,175 gene symbols (0.22% of all gene symbols) had ambiguities *with other gene symbols* and only 948 out of the 55,795 names (1.69% of all names) had ambiguities with other

names in the mouse database. The other databases also exhibited a very low rate of ambiguity within the same organism, except for Flybase, which had a rate of .68% for symbols, but 10.18% when considering all names.

		Ambiguities in Database	Ambiguities with English words	Ambiguities across databases
MOUSE	symbols (19,175)	43 (0.22%)	307 (1.60%)	1585 (8.27%)
	all names* (55,795)	948 (1.69%)	846 (1.52%)	3693 (6.62%)
WORM	symbols (3,221)	0 (0%)	0 (0%)	205 (6.36%)
	all names* (27,268)	0 (0%)	0 (0%)	511 (1.87%)
FLY	symbols (43,394)	296 (0.68%)	731 (1.68%)	1668 (3.84%)
	all names* (82,553)	8407 (10.18%)	1985 (2.40%)	3279 (3.97%)
YEAST	symbols(5,117)	0 (0%)	3 (0.06%)	1039 (20.30%)
	all names* (7,264)	120 (1.65%)	5 (0.07%)	1372 (18.89%)

Table 1. Results quantifying ambiguities of model organism gene names within each respective database, with English words, and across databases. * The category "all names" in the table comprises gene symbols together with synonyms (aliases).

We determined that Moby was the best of the three resources that we experimented with for identifying English words. The Wall Street Journal corpus did not have broad enough coverage of English. The Medline articles did have good coverage of English, but the corpus also contained many gene names and symbols, and therefore was inappropriate. Based on the Moby list of words, 307 (1.60%) of the mouse gene symbols were found to be ambiguous with English words, and an additional 539 mouse names (1.52%) were ambiguous with English words. Flybase exhibited the largest amount of ambiguities with English (2.40%), and WormBase exhibited none. Results for the Yeast were similar to those for the Worm.

Not surprisingly, in all the databases, the amount of ambiguities increased substantially when considering all four nomenclatures. When compared to the other three organisms 1,585 mouse gene symbols (8.27%) were ambiguous, and 3,963 mouse names (6.62%) were ambiguous with the other three organisms. The Yeast database exhibited the largest rate of ambiguity with other databases, amounting to 20.3% for gene symbols and 18.89% for all the names.

4.2 Recall

In a total of 25,804 abstracts from Group I that were processed, 7,899 did not result in identification of the appropriate MGI gene, yielding a recall of 69.4%. Of the 96,712 genes that were associated with 18,636 abstracts from Group II (two or more genes per abstract according to the gold standard), 70,305 genes were not recognized, resulting in a much lower recall of 27.3%. Overall a recall of 36.2% was achieved. An analysis of the failures identified seven primary reasons, which are shown in Table 2. The most frequent cause in Group 1 was due to simple name

variation between names in the abstracts and in the lexical entries. This can be further divided into more specific categories: a) punctuation variations (*bmp-4*, *bmp4*); b) numerical variations, (*syt4*, *syt iv*), c) variations of Greek letters (*iga*, *ig alpha*), and d) word order differences (*integrin alpha 4*, *alpha4 integrin*). Gene name variations accounted for 79% of the failures in Group I, and for 22% in Group II. However, when errors in only abstracts were considered, the error rate in Group II became 61%. A significant source of error occurred in Group II (58%) because we processed only abstracts, but the curated genes appeared in the full text only and not in the abstracts. In contrast, only 2% of Group I errors were due to this reason. Another substantial source of error was due to partial matches (*trophoblast specific protein alpha*, *trophoblast specific transcription factor*), which accounted for 14% in Group I. Smaller amounts of error were due to several other reasons: a) a gene name was not found in either the abstract or in the full text, b) a gene name was the same as an English word, which was deliberately removed from the lexicon, c) gene names only appeared in the reference section but not in the text of the article, and d) the original abstract was missing from the Medline database.

Reasons for failure	**Group I**	**Group II**	**Total**
1. Gene name variations	79	22 (8)*	101 (8)*
2. In full text only	2	58	60
3. Partial match	14	14 (6)*	28 (6)*
4. Not in article at all	1	4	5
5. Same as English word	3	1	4
6. In reference section only	0	1	1
7. Abstract missing	1	0	1
Total	100	100	200

Table 2. Reasons for recall failures based on analysis of 100 abstracts in Group I and 100 genes in Group II. *Numbers in parentheses signify the error occurred in the full text and not in the abstract.

4.3 Ambiguities in the Output

We determined ambiguous occurrences when using the MB lexicon. For Group I abstracts, 1,557 (8.7%) out of 17,891 MGI genes that were recognized appropriately by the straightforward lookup method had an ambiguous target form. Similar results were found for Group II abstracts, where 2,073 (7.9%) of the 26,378 genes had multiple target forms; the rates of the two groups did not differ significantly. Similar results were obtained for Group I and II abstracts, and therefore we combined the results. In total, 43,721 MGI genes were recognized correctly; of those 10% (4,389) had multiple mouse gene targets, and 24.7% (10,818) of the MGI genes had identical symbols with one or more fly genes. The ambiguities with *C. elegans* and yeast were 2.4% and 4.2%, respectively. Overall 32.9% of the curated MGI genes shared the same name with other genes, either within MGI database or across the species we examined (see Table 3).

The last question we addressed was ambiguity with English words. With lexicon MB (which had English words removed), roughly 328,000 gene symbols were recognized by the lexical lookup method. When Moby English words were

included and lexicon MBE was used, about 149,000 additional MGI IDs (a 45% increase) were obtained when processing the same set of abstracts, bringing the total to 477,000.

Ambiguities	**Number**
Within MGI	10.0% (4,389)
With Flybase	24.7% (10,818)
With WormBase	2.4% (1,045)
With Yeast	4.2% (1,798)
Total	32.9% (14,373)

Table 3. Occurrences of ambiguities of MGI gene names within MGI and across species. These were obtained as a result of processing a set of 45,000 abstracts. Note that the total number of ambiguities is not the simple sum of individual ambiguities since many overlap.

5 Discussion

Results showed that the number of ambiguous names within each species varied from 0%-10.18%, with the number per name ranging from 2 to over 100; most ambiguities were caused by gene synonyms or aliases. For example, *fbp1*, as shown in Figure 1, corresponded to 3 different genes, but 2 of the names were aliases. One factor contributing to ambiguity is that some nomenclatures intentionally include broader terms in their synonym lists to facilitate access to gene data because authors do not always use the appropriate names. One potential solution would involve refinement of the synonym lists provided by the databases to include a separate category for terms that are broader. Another factor that contributes to ambiguity is the gene nomenclature rules established by the organism databases themselves. The four different species have various naming rules that specify how researchers should name their genes. For example, the rule for naming yeast genes states that the gene name symbol should consist of three letters (the gene symbol) followed by an integer (e.g. ADE12), and also requires that the name symbol be unique within that nomenclature. The low percentage of ambiguities within the yeast database, as well as with English words shows that this rule was effective in avoiding problems due to ambiguity. Similarly, the standards for naming genes in WormBase follow this pattern: “A Predicted Gene: A dot name, such as F59E12.2; A Named Gene: A three letter name, such as zyg-1.” This also accounts for the low percentage of ambiguities within the worm database and with English words, but there still exists ambiguity with names in other organisms as shown in Table 1.

However, the other two species (mouse and fly) have a higher percentage of ambiguities within their respective databases. The naming conventions for these two are more lenient. For the fly nomenclature, the names should be concise, should allude to the gene's function, mutant phenotype or other relevant characteristic, and should not have been previously used for a Drosophila gene. This more general rule does not place too many restrictions on the format of the gene names, and thus more ambiguities tend to arise. For example, *alp* is a symbol for the *abnormal leg pattern* gene and is also a synonym for *activin like protein at 23B*. For the mouse gene

nomenclature the names of genes and loci should be brief, should convey accurate information, and should begin with a letter. This less stringent rule may lead to ambiguities within the database, as well as with English words. One potential solution would involve refinement of the naming conventions.

Not surprisingly, the ambiguities of gene names across the four nomenclatures seemed to be more severe, ranging from 1.19%-20.30%. Factors similar to the ones we discussed above for a single nomenclature apply in this situation also. However, another factor is that historically, different nomenclatures tend to name orthologs (the same genes in different species that typically have the same functions) with the same name. For example, MGI has curated 9,981 mouse/human ortholog relationships, and only 2,329 of them differ in their names. Using the same names for orthologs makes intuitive sense and facilitates user comprehension, but is confounding for NLP applications.

When we employed the NLP engine to recognize gene names in Medline abstracts using the lexicon ALL4, the ambiguity problem was exacerbated. Overall, 33% of the mouse genes that were extracted shared a name in common with other genes, either within the mouse database or across databases for different organisms. This shows that the ambiguity problem is a serious one for NLP and that more research in this area is needed. This problem may be reduced significantly if the appropriate organism-specific lexicon is used to process an article. This suggests that a method that first identifies the applicable organism(s) for each article would help alleviate the ambiguity problem somewhat.

The information presented in this research may not be complete in a number of respects. We found that the worm data that we collected did not contain any aliases. It may be because only the official symbols are used in the literature, or because information concerning aliases were not available on the website. If aliases are used in journal articles in place of the official worm symbols, they would cause more ambiguities than we determined. This also raises the issue of completeness. We obtained the names of synonyms (aliases) from the websites and considered ambiguity in gene names and in the number of ambiguous occurrences based on that information. However, the ambiguity problem would be worse if the information we obtained was incomplete. We also did not account for ambiguities with other organisms, or with biomedical terms, such as drug names, diseases, clinical procedures, or symptoms. In particular, we did not consider ambiguities between gene and protein names, which is a serious problem also. Therefore it is likely that the quantities we obtained for ambiguities and ambiguous occurrences substantially underestimated the problem if the broader biomedical domain were to be considered. Future work will involve expanding our study of ambiguity to include more organisms and also terms in the Unified Medical Language System [9], a comprehensive nomenclature system containing medical and biological terms. Because gene-disease and gene-drug relationships contain important genomic information, gene names that are ambiguous with clinical terms will be worthwhile to study further.

The amount of gene names ambiguous with English words ranges from 0-2.4% of gene names in all four organisms. However, this seemingly low percentage of genes could cause substantial difficulties for NLP and IR systems. If a gene name is a frequent English word, such as *was*, and *to* it will occur frequently in the articles, and cause a large decrease in precision unless special disambiguation procedures are available. Such a situation is consistent with the results reported by Hirschman and colleagues[7]. Their results showed an extremely low rate of precision (2%) because English words were included in their list of gene names. We found that when we used a lexicon containing English words, an additional 149,000 (45% increase) "genes" were extracted. The percent of these that occurred as actual gene names is currently under study, but the rate is probably low as suggested by the low rate of recall error that occurred when we removed genes that were English words from the lexicon (2%). It appears that sacrificing genes that are English words is likely to result in a small drop in recall and substantial increase in precision. It will also be interesting to see in what form common English words actually appear in the articles. For example, we did not find any occurrences of a gene named *Was* in the 45,000 articles.

While examining the gene names, we converted all the symbols to non-italicized lowercase letters. Although this helped to discover gene name ambiguities within and across species, gene names in the literature can usually be identified by features such as capitalization (either the first letter or all of the letters), being italicized, or being surrounded by quotation marks or parenthesis. Ambiguities between genes of the same string but with different cases may be easy to resolve, as would genes that are completely or partially upper case when occurring in the middle of sentences. A pre-processor could easily tag these situations and avoid ambiguities with English words. However, this process would still not be straightforward since different organisms may have different conventions. A pre-processor would have to determine which rules to use based on which organism the article discussed.

In this study, Group I abstracts showed a substantially higher rate of recall (69.4%) than that for Group II (27.3%). In Hirschman's work, the full text articles resulted in a much higher recall rate than the abstracts. Not surprisingly, in the random sample for Group II, 58% of missed MGI genes were not in the abstracts. Interestingly, Group I only had only 2% in this category. Presumably, this was because these articles were identified by the curators as containing only one primary gene, and therefore that gene was likely to be in the abstract. For abstracts with multiple genes, many of the genes may not be considered primary findings and therefore do not occur in the abstract, but possibly the most important ones do. The largest recall problems occurred because of simple gene name variants. A refinement of the string matching algorithm could alleviate this problem, and bring about increased recall, although it could also possibly result in decreased precision.

6 Conclusions

Identifying gene names is crucial for database interoperability and for development of automated techniques that extract important genomic information from the biomedical literature. Nomenclature databases are invaluable resources for identifying gene names, but these resources are heterogeneous. Combining the nomenclature information into one resource could facilitate interoperability and also benefit NLP systems in this domain. However, a uniform resource alone is not enough. Our results show that the ambiguous nature of gene names within and across model organism databases presents a significant roadblock to reliable gene identification. More work on disambiguation by the NLP community is needed to address ways to resolve this problem. Furthermore, quantification of the ambiguities and their detrimental effect on the precision of automated text processing systems may provide useful feedback to the model organism communities. NLP methods could be cost-effective tools to assist in the curation process, but the ambiguous nature of the names and the lack of standard conventions across organisms are serious obstacles for NLP.

Acknowledgements
This work was supported in part by grant EIA-031 from the National Science Foundation and LM06274 and LM7659 from the National Library of Medicine.

References

(1) Friedman C, Liu H, Shagina L, Johnson SB, Hripcsak G. Evaluating the UMLS as a source of lexical knowledge for medical language processing. Baaken S, editor Phila: Hanley&Belfus, 2001:189-194

(2) Fukuda K, Tsunoda T, Tamura A, Takagi T. Information extraction: identifying protein names from biological papers. Hawaii: 1998:707-718

(3) Proux D, Rechenmann F, Julliard L, Pillet V, Jacq B. Detecting Gene Symbols and Names in Biological Texts: A First Step toward Pertinent Information Extraction. Genome Inform Ser Workshop Genome Inform 1998; 9:72-80.

(4) Jenssen TK, Laegreid A, Komorowski J, Hovig E. A literature network of human genes for high-throughput analysis of gene expression. Nat Genet 2001; 28(1):21-28.

(5) Hanisch D, Fluck J, Mevissen HT, Zimmer R. Playing biology's name game: identifying protein names in scientific text. Pac Symp Biocomput 2003;403-414.

(6) Narayanaswamy M, Ravikumar KE, Vijay-Shanker K. A biological named entity recognizer. Pac Symp Biocomput 2003;427-438.

(7) Hirschman L, Morgan AA, Yeh AS. Rutabaga by any other name: extracting biological names. J Biomed Inf:2002;35(4):247-59.

(8) Blake JA, Richardson JE, Bult CJ, Kadin JA, Eppig JT. MGD: the Mouse Genome Database. Nucleic Acids Res 2003; 31(1):193-195.

(9) Lindberg D, Humphreys B, McCray AT. The Unified Medical Language System. Meth Inform Med 1993; 32:281-291.