BIOINFORMATICS ORIGINAL PAPER

Vol. 21 no. 2 2005, pages 248–256
doi:10.1093/bioinformatics/bth496

Gene name ambiguity of eukaryotic nomenclatures

Lifeng Chen[1,], Hongfang Liu[2] and Carol Friedman[1]*

[1]Department of BioMedical Informatics, Columbia University, New York, NY 10032, USA and [2]Department of Information Systems, University of Maryland, Baltimore County, Baltimore, MD 21250, USA

Received on April 11, 2004; revised on August 19, 2004; accepted on August 20, 2004
Advance Access publication August 27, 2004

ABSTRACT

Motivation: With more and more scientific literature published online, the effective management and reuse of this knowledge has become problematic. Natural language processing (NLP) may be a potential solution by extracting, structuring and organizing biomedical information in online literature in a timely manner. One essential task is to recognize and identify genomic entities in text. 'Recognition' can be accomplished using pattern matching and machine learning. But for 'identification' these techniques are not adequate. In order to identify genomic entities, NLP needs a comprehensive resource that specifies and classifies genomic entities as they occur in text and that associates them with normalized terms and also unique identifiers so that the extracted entities are well defined. Online organism databases are an excellent resource to create such a lexical resource. However, gene name ambiguity is a serious problem because it affects the appropriate identification of gene entities. In this paper, we explore the extent of the problem and suggest ways to address it.

Results: We obtained gene information from 21 organisms and quantified naming ambiguities within species, across species, with English words and with medical terms. When the case (of letters) was retained, official symbols displayed negligible intra-species ambiguity (0.02%) and modest ambiguities with general English words (0.57%) and medical terms (1.01%). In contrast, the across-species ambiguity was high (14.20%). The inclusion of gene synonyms increased intra-species ambiguity substantially and full names contributed greatly to gene-medical-term ambiguity. A comprehensive lexical resource that covers gene information for the 21 organisms was then created and used to identify gene names by using a straightforward string matching program to process 45 000 abstracts associated with the mouse model organism while ignoring case and gene names that were also English words. We found that 85.1% of correctly retrieved mouse genes were ambiguous with other gene names. When gene names that were also English words were included, 233% additional 'gene' instances were retrieved, most of which were false positives. We also found that authors prefer to use synonyms (74.7%) to official symbols (17.7%) or full names (7.6%) in their publications.

Contact: lifeng.chen@dbmi.columbia.edu

INTRODUCTION

Owing to the advancement of high-throughput technologies, such as genomic sequencing, DNA chips and powerful computers, the amount of scientific literature has been increasing exponentially in the past decade. Much of the literature has been published online as well as in journal articles. The plethora of the online literature is an invaluable resource for researchers if used effectively. However, the information is stored in heterogeneous formats by different groups of researchers and used for various purposes. This makes management and reuse of these resources problematic. First, the large volume of information makes it impractical for manual identification and entry of relevant information into databases. Second, when searching for materials that may be of interest, the vast amount of information can easily dumbfound researchers. Fortunately, automated processes such as natural language processing (NLP) may shed light on this problem by extracting and structuring relevant textual information from online literature in a timely manner (for review see Hirschman *et al*., 2002b).

Recently, there have been multiple NLP applications developed in medical (Christensen *et al*., 2002; Sager *et al*., 1995; Friedman *et al*., 1994) as well as biological domains (Fukuda *et al*., 1998; Jenssen and Vinterbo, 2000; Proux *et al*., 1998; Hanisch *et al*., 2003; Narayanaswamy *et al*., 2003; Friedman *et al*., 2001) that are able to extract and encode clinical or biological terms from clinical reports or biological literature. In biomedical literature, genomic entities such as proteins and genes consist of an important type of information and are currently a main focus of NLP researchers. Fukuda *et al*. (1998) developed the system PROPER which extracts protein names in the literature using rules

*To whom correspondence should be addressed.

based on protein nomenclature. Another system developed by Jenssen and Vinterbo (2000) utilizes a name dictionary containing human gene symbols extracted from different databases, such as HUGO and LocusLink. Hanisch *et al*. (2003) uses name tokenization along with a curated gene symbol dictionary to recognize protein names. Some systems use support vector machines and hidden Markov models to recognize entities in the biomedical domain (Shen *et al*., 2003), while other methods include the use of morphological analysis of entity names, classification techniques and collocation (Yamamoto *et al*., 2003). These systems focused on recognition of biological entity names. But identification is also important and is required for effective utilization. To accomplish this, NLP needs a comprehensive resource that specifies and classifies genomic entities and that associates them with normalized terms and unique identifiers assigned by a standardized nomenclature system so that the extracted entities are well defined.

Since journal articles are associated with a broad range of organisms, it would be invaluable for the advancement of NLP if a lexical resource covering gene information for all organisms were built. To be useful, such a resource must be comprehensive, accurate and up-to-date. Many model organisms have their own specialized online databases that are valuable for this purpose because they have mature nomenclatures and ontological specification for biological entities, for instance, the Human Genome Nomenclature Committee Database (Genew) (Wain *et al*., 2002a), the Mouse Genome Informatics (MGI) (Blake *et al*., 2003), FlyBase (The FlyBase Consortium, 2003), WormBase (Harris *et al*., 2004), the Rat Genome Database (RGD) (Steen *et al*., 1999), the Zebrafish Information Network (ZIN) (Sprague *et al*., 2001), the Saccharomyces (yeast) Genome Database (SGD) (Cherry *et al*., 1998), DogMap (Dolf, 1999), BovMap and ARKdb, which includes cat, chicken, cow, deer, horse, sheep, pig, salmon, Tilapia and turkey (Hu *et al*., 2001). LocusLink is also an excellent resource for gene information (Pruitt and Maglott, 2001), and is the result of a substantial effort that involves gathering and organizing gene, sequence, protein and homolog information for multiple species. Currently the scope of LocusLink is *Caenorhabditis elegans*, chicken, cow, fruit fly, human, HIV-1, mouse, pig, rat, sea urchin, *Xenopus laevis*, *Xenopus tropicalis* and zebrafish. However, LocusLink does not have gene information for all species, nor does it have all the gene information for the species it covers.

Online organism databases are not suitable for direct use by NLP systems because (1) different databases have very different formats and ontological specifications, making it difficult to utilize them directly; (2) new genes are discovered and added to databases constantly while some others are retired and modified, making it difficult for a lexical resource to keep the database current; and (3) for specific species, multiple databases usually exist and cover different scopes of gene information. For example, for the species cow, there is BovMap in France, Bovine ARKdb in the UK and Cattle Genome Database (CGD) in Australia. To be comprehensive, the information from different resources must be combined but extra care must be taken to avoid conflicts and redundancy.

Even if we successfully build a uniform lexical resource for NLP purposes, to associate terms appearing in text with specific biological entities is extremely challenging since a name (1) could refer to several different genetic entities, either from the same species or from other organisms, e.g. the string '*CAT*' represents different genes in cow, chicken, fly, human, mouse, pig, deer and sheep; (2) could refer to another type of biological entity, such as a protein or phenotype, e.g. the mouse gene '*hair loss*'; (3) could be other types of concepts in closely related domains, such as the clinical field, e.g. the mouse gene '*diabetes*'; and (4) could be the same as common English words, e.g. fly genes '*can*' and '*lie*'.

The ambiguity problem (one name referring to different entities) must be addressed. Failure to do so may significantly affect the usefulness of NLP systems for biological applications. Many NLP engines have reported promising results in recognizing gene or protein names, with precision and recall rates each ranging from 60 to 80% (for review see Hirschman *et al*., 2002b). However, the situation is different for identification. A previous paper by Hirschman *et al*. (2002a), in an effort to identify fly genes, yielded a much lower rate of precision (2%) when processing full-text articles using a straightforward string-matching method due to gene name ambiguity. A similar study by Tuason *et al*. (2004) found that one out of three mouse gene names were ambiguous with other gene names, either from the species mouse or from yeast, fly or worm. Another work on extraction of gene names and establishment of human gene name network from PubMed also displayed high error rates (up to 40%) in retrieved 'gene pairs' due to abbreviations or words being mistaken for gene symbols (Jenssen *et al*., 2001).

An important step in automated gene identification would be to understand the extent of the naming ambiguity problem. The research presented here extends our previous work in studying gene name ambiguities for four model organisms (Tuason *et al*., 2004). However, this study is significantly different from the previous one in that we (1) used a much larger and more comprehensive dataset by combining and obtaining gene information for 21 model organisms from heterogeneous databases; (2) quantified ambiguities for all gene names, including official symbols, synonyms and full names; (3) measured ambiguities considering both upper and lower cases of letters; and (4) expanded the study to include gene name ambiguity with a closely related domain—the medical domain by acquiring abbreviations (Liu *et al*., 2001) and concepts from the Unified Medical Language System (UMLS) (Lindberg *et al*., 1993), a comprehensive nomenclature system containing medical and biological terms with corresponding

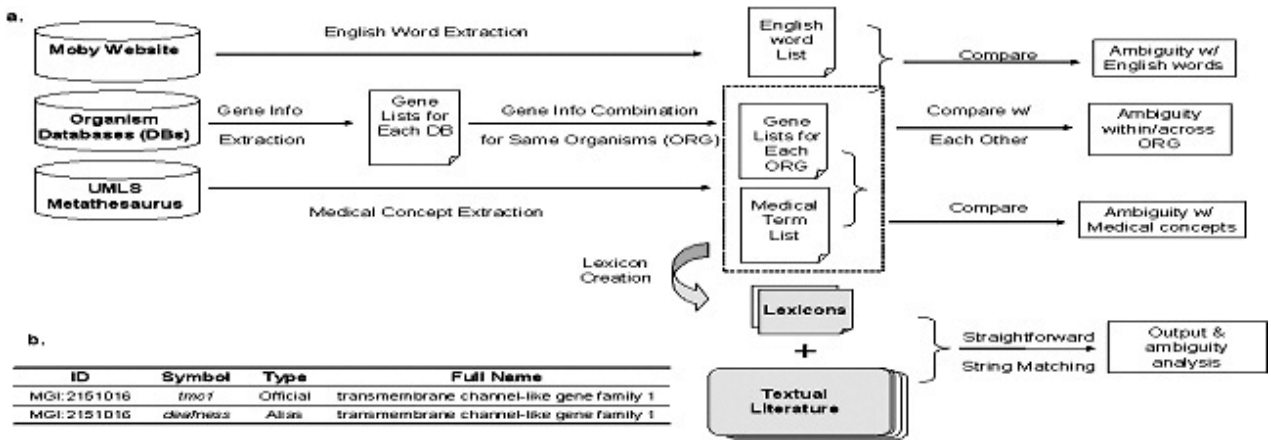

Fig. 1. (**a**) Methods flowchart; (**b**) sample entries from the mouse gene list.

Table 1. Lexicons created in this study

Lexicons	Contents	Sample lexical entries
LEX1	Gene information from the 21 organisms.	tmc1 **gene** GDB:636178^transmembrane channel-like 1 tmc1 **gene** MGI:2151016^transmembrane channel-like gene family 1 deafness **gene** MGI:2151016^transmembrane channel-like gene family 1
LEX2	Same as LEX1 but target forms for the same gene names were combined	tmc1 **gene** GDB:636178 transmembrane channel-like 1+MGI:2151016^transmembrane channel-like gene family 1 deafness **gene** MGI:2151016^transmembrane channel-like gene family 1
LEX3	Entries of LEX2 that were general English words removed	tmc1 **gene** GDB:636178^transmembrane channel-like 1+MGI:2151016^transmembrane channel-like gene family 1
LEX4	Medical concepts from UMLS	deafness **symptom** C0581883^deafness
LEX5	Combination of LEX2 and LEX4, consisting of both **medical** and **gene** information. Genes named as English words were retained	tmc1 **gene** GDB:636178^transmembrane channel-like 1+MGI:2151016^transmembrane channel-like gene family 1 deafness **symptom** C0581883^deafness deafness **gene** MGI:2151016 transmembrane channel-like gene family 1
LEX6	Combination of LEX3 and LEX4, consisting of both **medical** and **gene** information without genes named as English words	tmc1 **gene** GDB:636178^transmembrane channel-like 1+MGI:2151016^transmembrane channel-like gene family 1 deafness **symptom** C0581883^deafness

A prefix of GDB is used for the human and MGI is used for the mouse.

unique identifiers and synonym information; (5) investigated authors' preference in using official symbols, aliases or full names in abstracts.

METHODS

Overview

Gene name ambiguity. The outline of the methods is shown in Figure 1a. A detailed description of methods is given below. Gene information was obtained from each database. Then the information from different databases for the same species was combined to generate uniformly structured gene lists where **gene ID**, **symbol**, **type** (official or alias) and **full name** were retained. Figure 1b shows two entries from the mouse list. A list of general English words was obtained from the Moby project website (see Methods below), and a medical term list was obtained from the UMLS. Gene lists were then compared with each other and with lists of English words and medical terms to evaluate ambiguities.

An automated process was performed to create lexicons from the lists of genes and medical terms. A total of six lexicons were created (Table 1), two of which—LEX5 and LEX6—were used to identify gene names in a set of PubMed abstracts using straightforward string matching (Fig. 1a). Both lexicons contain information on genes and medical terms, but, in LEX6, gene names that were English words were removed, while in LEX5 they were retained (Table 1).

Generating gene lists for the organisms

Obtaining gene information from each database. Gene information for the 21 organisms was obtained from their

databases. There were several criteria in choosing organism databases:

(1) The model organism is of great research interest, being well studied and has sufficient information for this study.

(2) Gene name information is publicly available and free to download.

(3) Viruses, prokaryotes and plants were excluded although they may be investigated in the future.

Files containing gene name information were downloaded from the corresponding websites in January 2004. Original files from each organism's database varied tremendously in format. They were processed using different Perl scripts to extract and to map the relevant information to a single uniform representation. The raw files include *MRK_LocusLink.rpt* from MGI, *zebrafish_genetic_markers* and *zebrafish_marker_alias* from ZIN, *searchdata.txt* from the Genew website, *FBgn.acode* from FlyBase, *wormpep115* from WormBase, *ratgene.txt* from RGD, *registry.genenames.tab* from SGD, *LL_tmp1* from LocusLink and all results obtained by searching for the marker type 'GENE' for all species in ARKdb. The list of databases used in this study is not complete but substantial. The main data obtained were official the gene symbol, aliases, a unique identifier and an official full name, if one was available. These fields are important for creating lexicons where gene symbols are associated with a standardized target form consisting of the unique identifier and the official full name.

Nonetheless, not all databases contain all the information that was relevant. For instance, cat, horse, deer, salmon, sheep, tilapia, turkey and dog do not have information on gene aliases. Also several species, such as worm and yeast, do not have information on the full names of genes, but only have brief descriptions of gene functions. Most of the databases have unique identifiers for their genes. However, some do not provide one, for instance BovMap and Arkdb. For the purpose of this study, a unique ID was arbitrarily assigned to each of the genes that did not have a unique identifier and was kept consistent throughout the study. For example, bovine gene '*AANAT*' was assigned an ID 'BovMap3'.

Combining information on same species. The second step involved combining information for the same organism from different sources. For instance, human genes from both LocusLink and Genew were pooled together for that species. Extra care was taken in combining gene lists. If the same symbol points to different genes, e.g. the symbol '*VIP*' referred to two human genes—'vasoactive intestinal peptide' and 'alpha-2 macroglobulin family protein VIP', both records were kept in the merged list. In contrast, if both records refer to the same gene, only the one from LocusLink was kept to avoid redundancy. After this step, each organism had only one gene list.

Gene name ambiguity

After gene lists for the 21 organisms were generated, in order to study intra-species ambiguity, official symbols from each organism were compared to each other. For across-species ambiguity, official symbols from a specific organism were compared to the entire body of official symbols except those from the organism itself. Aliases were then added to measure ambiguity for all gene symbols (official and aliases). Finally, full name information was added to see its effect on naming ambiguities. In independent runs, all gene symbols/names were transformed to lower case and compared.

In order to investigate gene-symbol–general-English-word ambiguity, a list of 74 550 common English words was obtained from the Moby lexicon project website (http://www.dcs.shef.ac.uk/research/ilash/Moby/mwords.html). This list contains words common to two or more published dictionaries. Gene symbols/full names were then compared to the English list in case-sensitive and case-insensitive manners in separate runs.

To study gene-name–medical-term ambiguity, we acquired medical concepts from *MRCON.txt*, a UMLS metathesaurus file that consists of medical and biological concepts with unique identifiers and synonyms from the 2003 AA version of the UMLS (Lindberg *et al*., 1993). We also obtained a list of UMLS abbreviations extracted by Liu *et al*. (2001) in an independent study. These two lists were merged to give a list of 1 990 195 UMLS terms. Since, in this step, we were concerned with gene name ambiguity associated with non-biomolecular entities in the UMLS, a total of 261 740 terms associated with the UMLS semantic categories *Gene or Genome*, *Biologically Active Substance*, *Amino Acid, Peptide or Protein*, *Enzyme*, *Immunologic Factor and Receptor* were identified and excluded before the gene names were compared. The resulting list contained 1 728 455 medical terms. Again, both case-sensitive and case-insensitive situations were studied.

Creating lexicons for text processing

Various lexicons were then created automatically from the gene lists. Each lexical entry contained an official symbol, alias or full name, followed by the standard target form. All target forms were assigned at least the corresponding unique identifier. Information of the corresponding full gene name was included in the target form if available. For example, the *tmc1* gene in LEX1 (Table 1) has standard target forms 'MGI:2151016 ^transmembrane channel-like gene family 1' and '*GDB:636178^transmembrane channel-like 1*', denoting a mouse and a human gene, respectively. We also recorded whether the gene symbol was an official symbol, an alias or a full name (data not shown) (Table 1).

After the initial gene lexical entries were created, an automated program merged all entries associated with the same symbol/full name that had different target forms, so that all of them were combined into a single entry with a single target form consisting of the union of the individual target forms.

For example, in Table 1 the gene '*tmc1*' in LEX2 has a merged target form of '*MGI:2151016^transmembrane channel-like gene family 1 + GDB:636178^transmembrane channel-like 1*'. During the processing of journal articles, when a string in the literature matches a lexical entry, the output generated will be the target form for that gene name, which consists of all possible gene entities the name refers to.

Since gene names that were general English words are highly ambiguous, we identified those entries and removed them, creating a lexicon where each unique gene name had only one entry, which was not an English word. Note that the mouse gene '*deafness*' was removed in LEX3 because it matches a general English word.

A lexicon (LEX4) consisting of medical terms in the UMLS was also created, and the lexicons containing medical (LEX4) or gene information (LEX2) were combined to create LEX5. All lexical entries from both resources were kept separately even if they shared the same name. This is because medical terms and gene names belong to completely different semantic categories and thus could not be combined for NLP use. For example, the mouse **gene** '*deafness*' is also a **symptom** as shown in LEX5 (Table 1). Finally, LEX 3, where gene names that were English words were removed was combined with LEX4 to create LEX6, a lexicon consisting of both medical and gene information without genes named as English words.

Ambiguities in actual text processing

We evaluated the effect of using the lexicons that were created as resources in actual text processing. We focused on the mouse organism and obtained 45 000 abstracts from PubMed according to the *MRK_Reference.rpt* file from the MGI site (ftp://ftp.informatics.jax.org/pub/reports/index.html). This file associates MGI gene identifiers with PubMed abstracts where the associations were established by human curators, and thus were used to serve as a reference standard in the following evaluation. Since we were not evaluating system performance in this study, precision and recall were not measured. Instead, we addressed issues only associated with the different types of gene name ambiguity.

Two lexicons, LEX6, consisting of gene and medical information without genes named as English words, and LEX5, which included genes that were also English words were used in separate runs to identify genes by processing the abstracts using a straightforward string matching program in a case-insensitive manner. The case of the letters was ignored because authors do not always follow gene nomenclature conventions with regard to the use of capitalization. Genes were considered identified if the output generated as a result of text processing contained the appropriate MGI identifiers whose association was previously established by curators.

Authors' use of gene names

In order to estimate authors' preference in choosing gene symbols and full names, 50 abstracts were randomly chosen from the same set of PubMed abstracts. The first author of this paper (L.C.) manually read through the abstracts and decided which gene names were used, and whether they were official symbols, aliases or full names.

Finally, the effect of Gene–UMLS ambiguity in actual text processing cannot simply be estimated by straightforward string matching because contextual information is critical for determining whether a term is used in the abstract as a gene or as a medical term. In order to investigate this question, 100 abstracts containing one or more genes that share their names with UMLS terms were randomly selected. Again, we went through them to identify concepts and to decide if the phrases were used as gene names or phenotypes.

RESULTS AND DISCUSSION

Ambiguities within resources

Official symbols A total of 149 805 official symbols were obtained from the 21 organisms. When official symbols only were considered in a case-sensitive manner, the intra-species ambiguities were negligible. Overall, only 25 (0.02%) official symbols were ambiguous within the organisms. However, when official symbols from all 21 organisms were combined, the ambiguity increased substantially. A total of 21 279 (14.2%) official symbols were found to be ambiguous with genes from another organism. The ambiguities with general English words and UMLS concepts were only 0.57 and 1.02%, respectively. Figure 2 is a graphic representation of ambiguities for organisms with more than 1000 official gene symbols.

When gene symbols were compared ignoring the case of letters, most organisms showed just slight increase for the intra-species ambiguity. In contrast, the across-species ambiguities showed a great increase. A total of 37 757, or every one out of four genes, were found to be ambiguous across species. Ambiguities with general English words and UMLS concepts were increased to 1.26 and 2.36%, respectively.

As shown above, official gene symbols generally are not ambiguous within a particular organism. In contrast, the across-species ambiguity is high. Although some nomenclatures such as zebrafish recommend not using names identical to those used in other species (Sprague *et al*., 2001), potential ambiguity across organisms is not completely avoidable. And our list of organisms was not complete. It is likely that when more organisms are considered, the ambiguity problem will worsen.

One factor that contributes to the across-organism ambiguity is the naming convention of homologous genes. At the Gene Mapping Workshop held during the 22nd Conference of the International Society for Animal Genetics (ISAG) in East Lansing, MI, it was resolved unanimously that animal gene nomenclature should 'follow the rules for human gene nomenclature, including the use of identical symbols for homologous genes and the reservation of human symbols

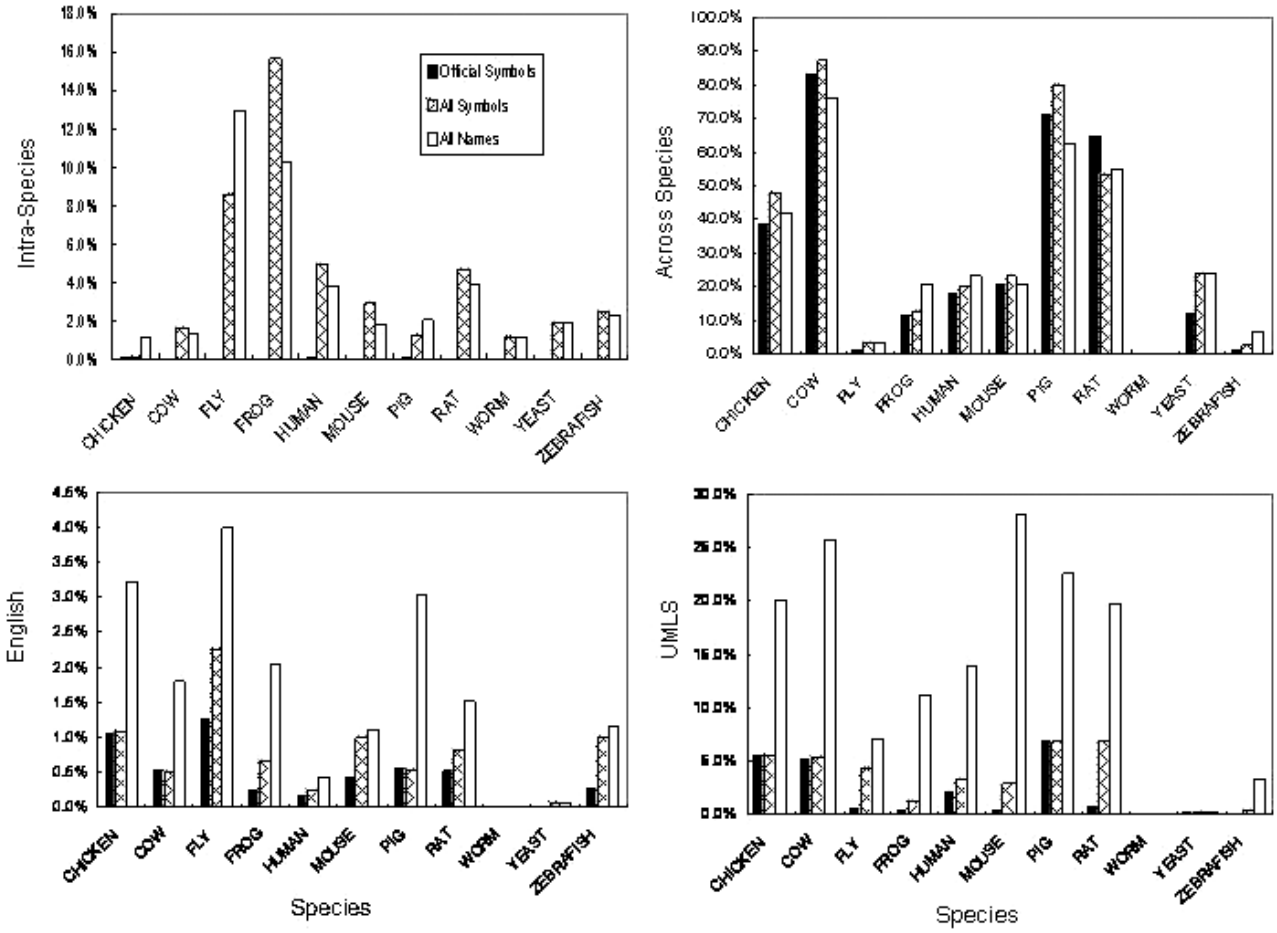

Fig. 2. Graphic representation of ambiguities of gene names for organisms with more than 1000 official symbols. For every organism, solid bar, official symbols only; shaded bar, all symbols, including official and aliases; hollow bar, all names, including all symbols and official full names. Ambiguities were measured when with cases of letters retained. Note different scales for different types of ambiguity.

for yet unidentified animals genes'. As a result, identical homologous names are ubiquitous. For example, the gene symbol *brca1* appears in human, cow, chicken, pig, frog, rat and mouse. In this study, we did not quantify the exact amount of ambiguity caused by homologous genes because of the complex nature in determining homology. But a random sample of 20 ambiguous instances showed that 16 of them were caused by apparent homology (data not shown), indicating that it contributes significantly to the across-organism ambiguity.

All gene symbols (official and aliases) When alias information was considered, a total of 331 287 symbols (a 120% increase in terms of number of official symbols) were obtained. Not surprisingly, every type of ambiguity increased as well. For instance, the total intra-species ambiguity was increased from 0.02 to 5.02%, i.e. over 16 000 gene symbols/aliases were found to be ambiguous even within their own species, compared to 25 that were found ambiguous when aliases were not considered. The ambiguities across the species, with general English words and with UMLS concepts were 13.43, 1.10 and 2.99%, respectively (Fig. 3).

We found that authors seem to prefer synonyms to official symbols or full names. For instance, the full name and official symbol for the mouse gene '*transformation related protein 53*' and '*Trp53*' were never used by authors in our random sample of 50 PubMed abstracts. Instead the synonym '*p53*' was used. Among the random sample, a total of 565 gene symbols/ synonyms/full names were used. Of these, 100 (17.7%) were official symbols, 43 (7.6%) were full names and 422 (74.7%) were gene synonyms. The use of aliases, although convenient for human beings, has increased the overall intra-species ambiguity of gene symbols significantly, presenting a much more difficult situation for automated processes.

All gene symbols and full gene names When full names were included, a total of 426 380 gene symbols/full names were obtained. For most types of ambiguities, there was only a slight increase. Specifically, the intra-species, across-species

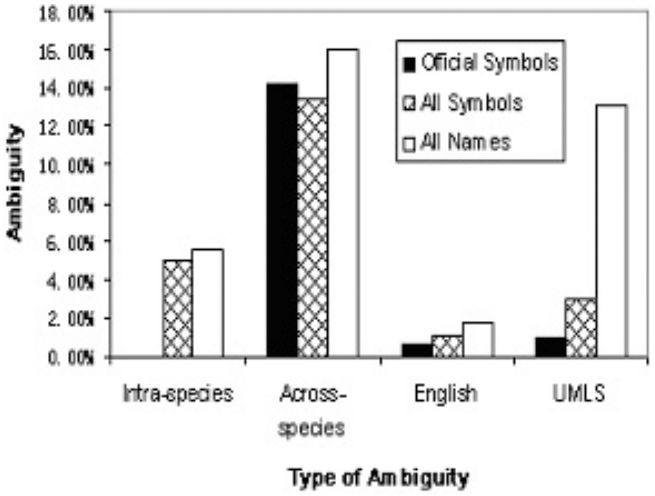

Fig. 3. Graphic representation of ambiguities for all gene names for 21 organisms when cases of letters were considered.

and English ambiguities were 5.59, 16.01 and 1.77%. However, the ambiguity between gene names and UMLS terms increased dramatically from the previous 3.95 to 13.10% when the cases of letters were retained (Fig. 3). One reason is that full names, usually multi-term phrases, simply have more of a chance to match medical terms. For example, the gene '*limb deformity*' could be found ambiguous with the corresponding phenotype 'limb deformity' (C0239337) or a combination of 'limb' (C0015385) and 'deformity' (C0000768).

Identification ambiguity

The ambiguity problem was exacerbated when the lexical resource was used for actual processing. A total of 589 062 'gene' instances were retrieved by the straightforward string matching algorithm for the 45 000 mouse abstracts. A total of 147 233 genes instances were extracted correctly according to the reference standard, but 125 261 (85.1%) were ambiguous with other gene names. The correctness of the remaining instances was not determined because of lack of a reference standard. However, this does not mean that over 400 000 genes were false positives. Although the list of gene–abstract associations was established by curators, it is not a complete list. Genes that occur in an article may not be curated because the gene is not the focus of the article. Manual examination of a random example of 100 instances by the first author showed that about half (56/100) of the extracted genes that were not annotated by curators were reasonable. Thus, this method possibly can be used to help curators refine their list of gene–article associations.

Gene–English ambiguity Although only 2.93% of the gene names were named as general English words in the lexical resource, this small percentage of ambiguity could be devastating for automated processes when attempting to identify genes. In this study, when these genes were included, 1 371 584 (233%) additional 'genes' were recognized for the same set of abstracts, most of which were false positives. For example, the English words 'was' and 'if', which are also the same as mouse gene names, essentially occur in every publication. The large amount of false positives will likely affect the precision and usefulness of automated applications in biological text parsing, even if methods more complex than straightforward string matching are used. Fortunately, some gene names are so ambiguous that authors seem to be reluctant to use them even if they are official symbols. For example, we did not find a single abstract in the 45 000 studied using the official symbol '*was*' as a gene. Instead, a less frequent synonym '*wasp*' (another English word unfortunately) was used. Thus, it may be worthwhile to sacrifice a little recall to achieve a much higher precision rate by ignoring genes that are common English words until improved disambiguation methods are developed. On the other hand, some words determined to be English words by Moby were not actually general English words, for instance, 'catalase' or 'deoxyribonuclease'. This may have biased the results. However, the task of automatically identifying English words is not trivial.

Gene–UMLS ambiguity Many nomenclatures recommend their genes to be named after functions, mutant phenotypes or other characteristics the genes are responsible for. Thus it is common that a gene may be given the same name as a medical concept, e.g. 'limb deformity'. Among the random sample of 100 abstracts known to have at least one phrase that could be either a mutant phenotype or the gene named after that phenotype, a total of 149 such phrases were identified. The majority (122/149, or 81.9%) of phrases occurred as phenotypes/diseases while the rest (27/149 or 18.1%) occurred as genes. There was no abstract in the random sample where the same term denoted both the mutant phenotype and the gene at the same time, and therefore all ambiguous terms had one sense per abstract. Table 2 shows the phrases identified in the random sample.

However, the gene–medical-term ambiguity is impossible to quantify using a straightforward string-matching algorithm. As we may see from Table 2, depending on different abstracts, many phrases can represent either phenotypes/diseases or genes. For instance, promyelocytic leukemia refers to a disease, in Hup-4 is probably expressed in the human uterus and in HL-60 (human promyelocytic leukemia) cells, (PMID:8078484) or a gene in promyelocytic leukemia gene product (PML) is seen aggregating at later points in infection (PMID: 11498534). Although it seems more intuitive that these phrases are phenotypes, 18% of time the phrases were used to represent genes. Note that the phrases listed in Table 2 as 'used only as a gene' or 'phenotype' are limited to our random sample. Chances are that these phrases can be used for both types as well.

Table 2. Use of phrases that could be either gene names or phenotypes/diseases in 100 randomly selected mouse abstracts. Numbers in parentheses denote numbers of instances. For the third category, the first number indicates instances when phrases were used as phenotypes and the second number indicates instances of genes

Used only as a phenotype/disease
Breast carcinoma (2), generalized neuroaxonal dystrophy (2), kidney disease (16), peroneal muscular atrophy (2), mediterranean fever (2), hair loss (5), growth retarded (5), hemolytic anemia, motor neuron disease (5), cleidocranial dysplasia (1), congenital hydrocephalus (2), Friedreich ataxia (3), neuronal ceroid lipofuscinosis (3), obstructive hydrocephalus (1), osteogenesis imperfecta (5)
Used only as a gene name
Limb deformity (6), Neurofibromatosis 2 (1), lens opacity—4 (2) congenital goiter (2), chronic multifocal osteomyelitis (2) hairy ears (2), fused toes (1), head tilt (1)
Used as a phenotype/disease or gene name
Duchenne muscular dystrophy (29-3), polycystic kidney disease (9-4), promyelocytic leukemia (6-3), severe combined immunodeficiency (13-1), tight skin (3-1)

Potential solutions to reduce gene name ambiguity

The gene name ambiguity problem is a very serious one for NLP and other automated processes. However, it is doubtful that this problem can be solved by any individual group. Instead, it would be beneficial if the broad community is aware of this problem and works synergistically toward the goal of reducing the problem because then the precision of automated text processing methods would be substantially increased, the complexity decreased, and high-throughput text-mining techniques with satisfactory performance would be of significant benefit to the biology community. Specifically, we suggest the following points.

(1) Authors use only official symbols in their publications, and avoid using aliases whenever possible especially symbols that coincide with general English words. If the publication concerns a newly discovered gene, the author should restrain from naming it as a general English word. When naming genes, a combination of letters and numbers should be used to reduce ambiguity associated with English words. Depending on the journal, authors may include organisms studied in a keyword list. This would be helpful for curators as well as NLP systems. The best way to identify a gene uniquely is to include the identifiers that are associated with a standard nomenclature for the gene discussed, particularly in the abstracts since these are most frequently used by researchers in data mining. These efforts are not difficult, would be of great benefit for NLP systems as well as the entire community but require a change in behavior. If authors do provide a way to identify the relevant genes in their articles it should be of significant benefit to them also because it would facilitate access by the research community and ensure their articles are read.

(2) Publishers can help by enforcing the use of official symbols. Some journals such as *Genomics* and *Nature Genetics*, do insist on the use of an approved nomenclature, a requirement that is useful if it were adopted by other journals.

(3) Naming conventions can be revised so that a uniform nomenclature is achieved for all species. Although this is a difficult task, enhancements can still be made to help disambiguate homologous genes in different species. A prefix or suffix that denotes which species a homolog is from could be implemented. Sometimes, when discussing homologous genes, an author would add a prefix for clarity. For example, one article that discusses both mouse and human gene *Hox 2.1* uses *hHox 2.1*, which is not a recorded symbol, to denote the human *Hox 2.1* (PMID: 1355360). However, this is done informally so that an automated system currently may be likely to miss the gene name. There are also nomenclatures trying to add a prefix to indicate the species of origin when describing a gene from other species, e.g. the FlyBase, which provides a list of valid species abbreviations (The FlyBase Consortium, 2003). However, the introduction of species codes by different nomenclatures may present a difficult situation particularly by increasing the amount of gene synonyms simply because of different gene prefixes and suffixes. HGNC suggests distinguishing animal homologous genes when necessary by the use of the established letter-based species codes from SWISS-PROT (Wain *et al*., 2002b). These codes, if recommended by different organism communities, should help solve the homology ambiguity problem without adding to the synonymy problem.

(4) Finally, it would be beneficial if the NLP and text mining community develop new methods to categorize articles based on domains or species, and thus help reduce ambiguity to a much smaller extent. MESH headings, which were assigned manually by librarians to index papers, represent key information of the papers and can certainly be of help in this task. However, assignment of MESH headings requires significant human effort. Other features, such as title, author or keywords can also be useful. Liu *et al*. (2004) studied a combination of these features in text categorization and achieved an F-measure of 94.1% when using a supervised machine learning technique.

CONCLUSIONS

Identification of gene names is an essential task for biomedical knowledge extraction and reuse. Automated techniques such as NLP show promise in extracting important gene

information from the biomedical literature. Various nomenclature databases are invaluable resources for NLP especially when they are combined into one uniform resource. However, the ambiguous nature of gene names presents a significant roadblock. It would be extremely beneficial if the NLP, biologists and publishing communities work together to help solve the gene name ambiguity problem.

ACKNOWLEDGEMENTS

We would like to thank Dr Judy Blake for her insight concerning the nomenclature for the mouse model and gene naming conventions. The work was supported by grants LM7659 from the NLM and EIA-031 from the NSF.

REFERENCES

Blake,J.A., Richardson,J.E., Bult,C.J., Kadin,J.A. and Eppig,J.T. (2003) MGD: the Mouse Genome Database. *Nucleic Acids Res.*, **31**, 193–195.

Cherry,J.M., Adler,C., Ball,C., Chervitz,S.A., Dwight,S.S., Hester,E.T., Jia,Y., Juvik,G., Roe,T., Schroeder,M., Weng,S. and Botstein,D. (1998) SGD: Saccharomyces Genome Database. *Nucleic Acids Res.*, **26**, 73–79.

Christensen,L., Haug,P. and Fiszman,M. (2002) MPLUS: a probabilistic medical language understanding system. In *Proceedings of ACL Workshop in Natural Language Processing.* Philadelphia, PA, July 2002, pp. 29–36.

Dolf,G. (1999) DogMap: an international collaboration toward a low-resolution canine genetic marker map. *J. Hered.*, **90**, 3–6.

Friedman,C., Alderson,P.O., Austin,J., Cimino,J.J. and Johnson,S.B. (1994) A general natural language text processor for clinical radiology. *JAMIA*, **1**, 161–174.

Friedman,C., Kra,P., Yu,H., Krauthammer,M. and Rzhetsky,A. (2001) GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics*, **17**(Suppl. 1), S74–S82.

Fukuda,K., Tsunoda,T., Tamura,A. and Takagi,T. (1998) Information extraction: identifying protein names from biological papers. In *Proceedings of the Pacific Symposium on Biocomputing'98 (PSB'98)*, Hawaii, January 1998, pp. 707–718.

Hanisch,D., Fluck,J., Mevissen,H.T. and Zimmer,R. (2003) Playing biology's name game: identifying protein names in scientific text. *Pac. Symp. Biocomput.*, Kavai, HI, 403–414.

Harris,T.W., Chen,N., Cunningham,F., Tello-Ruiz,M., Antoshechkin,I., Bastiani,C., Bieri,T., Blasiar,D., Bradnam,K., Chan,J. *et al*. (2004) WormBase: a multi-species resource for nematode biology and genomics. *Nucleic Acids Res.*, **32**, D411–D417.

Hirschman,L., Morgan,A.A. and Yeh,A.S. (2002a) Rutabaga by any other name: extracting biological names. *J. Biomed. Inform.*, **35**, 247–259.

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

Hu,J., Mungall,C., Law,A., Papworth,R., Nelson,J.P., Brown,A., Simpson,I., Leckie,S., Burt,D.W., Hillyard,A.L. and Archibald,A.L. (2001) The ARKdb: genome databases for farmed and other animals. *Nucleic Acids Res.*, **29**, 106–110.

Jenssen,T. and Vinterbo,S.A. (2000) A set-covering approach to specific search for literature about human genes. In *Proceedings of the AMIA Symposium*, Los Angeles, CA, October 2000, pp. 384–388.

Jenssen,T.-K., Laegreid,A., Komorowski,J. and Hovig,E. (2001) A literature network of human genes for high-throughput analysis of gene expression. *Nat. Genet.*, **28**, 21–28.

Lindberg,D., Humphreys,B. and McCray,A.T. (1993) The Unified Medical Language System. *Meth. Inform. Med.*, **32**, 281–291.

Liu,H., Lussier,Y. and Friedman,C. (2001) A study of abbreviations in the UMLS. In *Proceedings of the AMIA Symposium*, Hanley&Belfus, Philadelphia, PA, pp. 393–397.

Liu,H. and Wu,C (2004) A study of text categorization for model organism databases. In *Proceedings of NAACLIHLT 2004*, Boston, MA, pp. 25–32.

Narayanaswamy,M., Ravikumar,K.E. and Vijay-Shanker,K. (2003) A biological named entity recognizer. *Pac. Symp. Biocomput.*, Kavai, HI, 427–438.

Proux,D., Rechenmann,F., Julliard,L., Pillet,V. and Jacq,B. (1998) Detecting gene symbols and names in biological texts: a first step toward pertinent information extraction. *Genome Inform. Ser Workshop Genome Inform.*, **9**, 72–80.

Pruitt,K.D. and Maglott,D.R. (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.*, **29**, 137–140.

Sager,N., Lyman,M., Nhan,N.T. and TIck,L.J. (1995) Medical language processing: applications to patient data representation and automatic encoding. *Meth. Inform. Med.*, **34**, 140–146.

Shen,D., Zhang,J., Zhou,G., Su,J. and an,C. (2003) Effective adaptation of a hidden Markov model-based named entity recognizer for biomedical domain. In *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine*, Sapparo, Japan, pp. 49–56.

Sprague,J., Doerry,E., Douglas,S. and Westerfield,M. (2001) The Zebrafish Information Network (ZFIN): a resource for genetic, genomic and developmental research. *Nucleic Acids Res.*, **29**, 87–90.

Steen,R.G., Kwitek-Black,A.E., Glenn,C., Gullings-Handley,J., Van Etten,W., Atkinson,O.S., Appel,D., Twigger,S., Muir,M., Mull,T. *et al*. (1999) A high-density integrated genetic linkage and radiation hybrid map of the laboratory rat. *Genome Res.*, **9**, 793.

The FlyBase Consortium (2003) The FlyBase database of the *Drosophila* genome projects and community literature. *Nucleic Acids Res.*, **31**, 172–175.

Tuason,O., Chen,L., Liu,H., Blake,J. and Friedman,C. (2004) Acquisition of lexical knowledge using biological nomenclatures. *Pac. Symp. Biocomput.*, 238–249.

Wain,H.M., Lush,M., Ducluzeau,F. and Povey,S. (2002a) Genew: the human gene nomenclature database. *Nucleic Acids Res.*, **30**, 169–171.

Wain,H.M., Bruford,E.A., Lovering,R.C., Lush,M.J., Wright,M.W. and Povey,S. (2002b) Guidelines for Human Gene Nomenclature. *Genomics*, **79**, 464–470.

Yamamoto,K., Kudo,T., Konagaya,A. and Matsumoto,Y. (2003) Protein Name Tagging for Biomedical Anonotation in Text. In *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine*, Sapparo, Japan, pp. 65–72.