Quantitative Assessment of Dictionary-based Protein Named Entity Tagging

Hongfang Liu, PhD, Zhang-Zhi Hu, MD, Manabu Torii, PhD, Cathy Wu, PhD, Carol Friedman, PhD

Abstract

Objective: Natural language processing (NLP) approaches have been explored to manage and mine information recorded in biological literature. A critical step for biological literature mining is biological named entity tagging (BNET) that identifies names mentioned in text and normalizes them with entries in biological databases. The aim of this study was to provide quantitative assessment of the complexity of BNET on protein entities through BioThesaurus, a thesaurus of gene/protein names for UniProt knowledgebase (UniProtKB) entries that was acquired using online resources.

Methods: We evaluated the complexity through several perspectives: ambiguity (i.e., the number of genes/proteins represented by one name), synonymy (i.e., the number of names associated with the same gene/protein), and coverage (i.e., the percentage of gene/protein names in text included in the thesaurus). We also normalized names in BioThesaurus and measures were obtained twice, once before normalization and once after.

Results: The current version of BioThesaurus has over 2.6 million names or 2.1 million normalized names covering more than 1.8 million UniProtKB entries. The average synonymy is 3.53 (2.86 after normalization), ambiguity is 2.31 before normalization and 2.32 after, while the coverage is 94.0% based on the BioCreAtive data set comprising MEDLINE abstracts containing genes/proteins.

Conclusion: The study indicated that names for genes/proteins are highly ambiguous and there are usually multiple names for the same gene or protein. It also demonstrated that most gene/protein names appearing in text can be found in BioThesaurus.

Introduction

Natural language processing (NLP) approaches have been explored to manage and mine information recorded in biological literature. One critical step for the development of NLP applications in the biomedical domain is biological named entity tagging (BNET) that identifies names mentioned in text and normalizes them with entries in biological databases. For example, in EDGAR, which extracted relationships between cancer-related drugs and genes from the literature, terms for genes need to be identified before the extraction of the relationships. In various pathway construction NLP systems, the identification of biological entity terms such as genes/proteins is the crucial component in order to construct molecular pathways from the text.

Methods for biological entity name identification can be categorized in two ways: one is to use a dictionary and a mapping method and the other is to mark up terms in the text according to contextual cues or specific verbs. The performance of systems relying on dictionaries depends on the coverage of the dictionary and a method to disambiguate gene/protein names from other biomedical terms or general English words. Another requirement is the ability to associate these names with corresponding entries in biomedical databases in order to be used by other automated systems for literature mining. This task is called biological entity name normalization, which requires a knowledge base that associates names identified in text with entries in databases. We refer to the task that identifies entity terms in the text and associates them with entries in the database as biological named entity tagging (BNET). BNET is not a trivial task because of several characteristics associated with biological entity names, namely: synonymy (i.e., different terms refer to the same database entry), ambiguity (i.e., one term is associated with different entries), and novelty (i.e., entity terms or entities are not present in databases or knowledge bases). Both tasks (i.e., biological entity name identification and biological entity name normalization) were tackled by the researchers in the community (i.e., Task 1A and Task 1B) in the first BioCreAtive workshop (Critical Assessment of Information Extraction in Biology).
The performance of the systems participating in the workshop was measured using precision, recall, and F-Measure. A system's precision is defined as the ratio of the number of names (or entities) correctly identified (or normalized) to the total number of names (or entities) being identified (or normalized). A system’s recall is defined as the ratio of the number of names (or entities) correctly identified (or normalized) to the total number of names (or entities) in the test set. An F-Measure of a system is defined as the harmonic mean of its precision and recall. In the following, we provide a brief summary of the workshop related to the above two tasks.

For the biological entity name identification task (Task 1A), the data were prepared using a gene/protein name tagger, AbGene, 22 and manually curated by domain experts. 17,24 There were 15 teams in Task 1A and a number of teams achieved F-Measure of 80%. Most systems in Task 1A fell into the following two machine learning approaches: Statistical modeling 25–27 and Support vector machine. 27–29 Besides machine learning approaches, a rule-based system, Text Detective, was using a combination of handcrafted rules with lexical knowledge sources to identify genes mentioned in the text. 30 The teams that achieved an F-Measure of 90% or more tended to use Statistical modeling together with post-processing methods, various features, and external knowledge sources. The data for the normalization task of BioCreAtive (Task 1B) 31 were prepared using a gene list associated with the full journal articles found in the model organism databases, i.e., SGD (yeast), MGD (mouse), and Flybase (fly). 34 As described by Hirschman et al., 14 the normalization task can be divided into several steps: i) identifying gene occurrences in the text; ii) associating gene occurrences to one or more unique gene identifiers; iii) selecting the correct identifier in case of ambiguity; and iv) assembling the final gene list for each abstract. Eight systems participated in the evaluation and a variety of approaches were adapted for the above steps. 19,30,35 Generally, identifying gene occurrences in text can be classified into two groups: i) matching against the lexical resource, 36,37; and ii) using the results obtained in Task 1A. The second step was simply a table look-up. The methods used to select unique identifiers fell into two categories: prune the lexical resource by removing ambiguous lexicon, or perform word sense disambiguation. Most systems used thresholds to select final lists and one system applied a maximum entropy classifier for removing bad matches. The precision and recall rates reported for Task 1B ranged from a maximum of 92% F-Measure for yeast to 79% for mouse. We participated in Task 1B and used an extensive list of synonyms obtained from online resources to perform biological named entity identification and normalization. The system achieved the best recall for mouse and yeast while the precision needs to be improved. Incorporating more synonyms into the system could improve recall while word sense disambiguation would be critical to improve the precision. 12

In this paper, we present a quantitative assessment of the complexity of dictionary-based protein named entity tagging where a protein named entity thesaurus, BioThesaurus, was assessed on synonymy, ambiguity and coverage. In the following, we first present background and related work. The assessment method is presented next. We then provide a discussion and conclude our work.

Background and Related Work

In this section, we first present a brief description of BioThesaurus. Background information about synonymy, ambiguity, and coverage is described next. We then provide background description of online resources used in the study. Examples of synonymy and ambiguity as well as related work are then given next.

BioThesaurus

BioThesaurus 38 is a Web-based system designed to map a comprehensive collection of protein and gene names to protein entries in the UniProt Knowledgebase (UniProtKB), a knowledge base of protein sequence and function created by UniProt—the Universal Protein Resource. 39 UniProtKB combined information contained in three databases: Swiss-Prot (a curated protein sequence database), TrEMBL (a computer-annotated supplement of Swiss-Prot, translated from nucleotide sequence database EMBL), and PIR-PSD (PIR-International Protein Sequence Database). Currently covering more than two million proteins, BioThesaurus consists of over 2.8 million names extracted from multiple molecular biological databases according to the database cross-references in iProClass, an integrated database that provides rich links with executive summaries to over 90 biological databases. 40 An overview of BioThesaurus construction is shown in Figure 1. The thesaurus was designed to provide comprehensive gene/protein names for all protein entries in UniProtKB. An extensive collection of names was extracted from many online resources according to database cross-references in iProClass. A parser was developed to automatically gather names from the underlying sources and parse individual names from annotation fields that contain multiple names separated by parentheses or other delimiters such as semicolons or commas. A raw thesaurus was then compiled, associating names with the corresponding UniProtKB entries. The raw thesaurus was further filtered to remove highly ambiguous and nonsensical names. The “name filter” was compiled based on frequency counts of names in UniProtKB entries and by curator judgment as “nonsensical.” Examples of filtered names include “novel protein,” “fragment,” and “hypothetical protein.” We also mapped names in the thesaurus to concepts in the UMLS (Unified Medical Language System), 41 a biomedical knowledge resource. Note that mapping names to UMLS concepts can be helpful for further curating the thesaurus. Some usages of semicolons and parentheses in annotation fields were not for separating synonyms but for comments, while our parser could not distinguish them and extract those comments as names. For example, the usages of parentheses in the description field for UniProtKB entry Q727V5, “precorrin-6Y C5,15-methyltransferase (decarboxylating)” and that for entry Q5JBL9, “TraA fimbrial protein precursor (pilin)” are different where the latter one is used to denote a synonym while the former one is not.

Characteristics Associated with Biological Entity Names

There are several characteristics associated with biological entity names: synonymy (also referred to as alias), ambigui-
ity (or referred to as homonymy) and coverage (or novelty), which are detailed in the following.

**Synonymy**—Different terms can refer to the same entity. The synonymous pairs can be classified into two types: derivable and non-derivable. A pair is considered derivable if one can be derived from the other algorithmically (i.e., can be unified through ignoring cases or punctuations, normalizing morphological difference, or ignoring word order). Another type of derivable is abbreviation-derivable. For example, the pair (“cAMP receptor protein,” “cyclic AMP receptor protein”) is derivable because “cAMP” is an abbreviation for “cyclic AMP.” In contrast, the synonymous pair (“cyclic AMP receptor protein,” “catabolite activator protein”) is non-derivable. Different methods have been used to recognize synonyms. One way to obtain synonyms is to use various online knowledge resources. For example, synonymous relations that are specified in the UMLS can be used to recognize synonym pairs (e.g., “cyclic AMP receptor protein” and “catabolite activator protein”) were associated with the same UMLS Concept Unique Identifier, C0007364). For abbreviation-derivable pairs, methods such as mapping abbreviations to full names using contextual cues (e.g., parentheses) have been developed. Wren et al. provided a review of a list of abbreviation databases obtained by extracting full names for abbreviations from text using either patter-based methods or supervised machine learning methods that are available to the public. Non-derivable synonymous pairs can be detected using contextual cues, such as “also called” or “also known as.” For example, the synonymous pair “IL-8,” “neutrophil-activating peptide 1” can be detected in the sentence “IL-8 (also known as neutrophil-activating peptide 1) is recognized as a potent effector of neutrophil functions.”

**Ambiguity**—One term can be associated with different entities as well as other concepts. The ambiguity of biological entity terms can be classified into four different types:

- **Systematic ambiguity**—terms that represent concepts that are closely related. For example, gene products (e.g., mRNA or proteins) are usually represented by names that refer to genes (note that such ambiguity has been referred to as class ambiguity in the literature). Homologous genes are usually represented by the same name. For example, “CAP” refers to rat cystinyl aminopeptidase protein and also human cystinyl aminopeptidase protein.
- **Entity-specific ambiguity**—terms that represent multiple unrelated proteins/genes. For example, the term “CAP” refers to biological entities, such as capsid protein, cystine aminopeptidase, catabolite gene-activator protein, cyclase-associated protein, and calcium activated protease.
- **Cross-medicine ambiguity**—terms that refer to clinical terms as well as protein/gene terms. For example, the term “CAP” also refers to the following clinical concepts: community acquired pneumonia, congenital alveolar proteinosis, cochlear action potential, carotid artery pressure, and compound action potentials.
- **Cross-general ambiguity**—symbols that are also general English words. For example, the following are common English words and also biological entity terms (“not,” “can,” “bad,” and “for”).

Several researchers have started to investigate the ambiguity issues. Hatzivassiloglou et al. considered systematic ambiguity (i.e., classifying names to three different classes: gene, protein, or mRNA) and applied machine learning tech-
Examples of Ambiguity and Synonymy in BioThesaurus

Figure 2 shows 17 synonymous protein names from multiple data sources for the human PDZ and LIM domain protein 1 (UniProt entry O00151). The names and synonyms can be normalized based on case and morphological variations and the number of synonyms decreases to 11 after normalization. Figure 3 shows both entity-specific and systematic ambiguities of protein names. The eight UniProtKB entries associated with “CLIM1” demonstrate both systematic ambiguity resulting from protein homology and name overloading from gene symbols that represent different proteins. Indeed there are three heterogeneous groups of proteins as indicated by their different assignments in UniRef90/50 sequence clusters as well as in distinct families in PIRSF (PIR SuperFamily), a protein family classification system at PIR that classifies proteins based on global sequence similarities of full-length proteins to reflect their evolutionary relationships.56 The “CLIM1” example represents entity-specific ambiguity among the three PIRSF families as well as systematic ambiguity within each PIRSF family. Examination of the source of protein name ambiguities revealed that “CLIM1” was derived from “human 36-Kda carboxyl terminal LIM domain protein (hCLIM1),” a cytoskeleton regulator.53 Independently, “cofactors of LIM homeomain proteins,” transcriptional activators associated with the LIM homeoproteins, were abbreviated as “CLIM,” and two forms of CLIMs were named as “CLIM1” and “CLIM2,” respectively.54 Clearly, the ambiguity could arise from independent naming of different proteins. However, it is not clear how “CLIM1” was annotated as a synonym of “polymeric immunoglobulin receptor” found in MGD.

Related Work

In Chen et al.,55 gene information associated with 21 organisms was obtained and quantified, naming ambiguity with species, across species, with English words and with medical terms, was measured. The purpose of Chen’s work was to study gene name ambiguity associated with eukaryotic nomenclatures that were obtained from each specific organism database, where the assessment of ambiguity did not distinguish systematic ambiguity from other types of ambiguity.

Here, we assessed the complexity of dictionary-based biological named entity tagging through several aspects related to biological entity names. According to our knowledge, there is no related work that provides a detailed analysis of synonymy, ambiguity and coverage of a thesaurus for all protein records in UniProtKB with respect to biological named entity tagging even though most researchers are aware of them.6

Methods

In the study, we first generated the gene/protein name thesaurus using the iProClass release 2.67. Noticing the ambiguous usages of parentheses and semicolons in annotation fields of online resources, we post-processed BioThesaurus to remove obvious non-protein names obtained from the online resources such as species names using the UMLS that is described in the following. We...
then evaluated BioThesaurus with respect to three characteristics: synonymy, ambiguity, and coverage.

Data Preparation
A version of BioThesaurus was constructed using the iProClass release 2.67 (May 15, 2005). As we have presented earlier in this paper, BioThesaurus associates terms derived from various online resources with protein records in UniProtKB. To prepare data for our analysis, BioThesaurus was computationally curated using the UMLS to remove terms that are obviously not genes/proteins such as cells, small molecules, or organisms (Figure 1; also see Background section for rationale). To do this, we first identified the UMLS semantic categories that correspond to genes/proteins. For each UMLS semantic category, we counted the number of terms in the raw dictionary with seven or more letters that can be found in the UMLS using strict string matching while ignoring care differences. The reason for restricting to terms with seven or more letters in the process is that abbreviations especially with less than seven letters are used frequently for medical concepts as well as for genes/proteins; including them in the identification process will favor categories that contain many abbreviations (e.g., body part). Based on the acquired frequency information and the UMLS semantic category definition, we derived a list of UMLS semantic categories containing genes/proteins. We considered terms with a semantic category not in this list as ones representing other types of concepts instead of genes/proteins. After removing terms representing non-gene or non-protein concepts and those that are nonsensical, we acquired the final thesaurus for assessment.

Assessment
Gene/protein names in BioThesaurus were assessed according to the three major characteristics of biological entity names—synonymy, ambiguity, and coverage. The assessment was conducted under two circumstances: i) using names before or after normalization, and ii) including or removing systematic ambiguity. Name normalization is to convert names to lower cases, to remove
punctuation marks, and to unify morphological variants. The normalization was done because most mapping methods used by biological named entity tagging systems can handle derivable synonyms, such as case variants (e.g., “arsC” vs. “ARSC”), punctuation variants (e.g., “IL12A” vs. “IL-12A”), and morphological variants (e.g., “superoxide dismutase” vs. “superoxide dismutases”).

Removing systematic ambiguity (the same name for homologous proteins across species) can help estimate the extent of name ambiguity for entities representing different functions (entity-specific) and concepts from different domains (cross-medicine) in the thesaurus. Systematic ambiguity can be removed based on a full-scale family classification system with curated families of homologous proteins. While the PIRSF family classification provides manually curated families of full-length proteins, the curated families of the system have not covered the complete UniProtKB (coverage ~50%–75%). Instead, we used the UniRef clusters as a way for removing systematic ambiguity. UniRef clusters provide full-scale automatic clustering of all UniProtKB entries based on sequence identity, and the clusters even at 50% sequence identity level (UniRef50) were found to well correspond to one PIRSF family, even to a subfamily within the PIRSF family. To remove systematic ambiguity, we grouped protein entries belonging to the same UniRef90 or UniRef50 clusters, where proteins shared at least 90% or 50% sequence identity, respectively. We manually inspected several UniRef50 clusters and found all proteins in the same cluster belong to the same PIRSF family.

The synonymy is a measure of the number of names per UniProtKB entry (i.e., entity). The average synonymy was measured by the total number of entity-name pairs (EID, NAME) in BioThesaurus divided by the total number of entities with one or more names, where EID is UniProtKB entry ID and Name is gene/protein name. We counted the number of entities in eight synonymy ranges: 1, 2, 3–4, 5–8, 9–16, 17–32, 33–64, and >64 (i.e., [2^n - 1, 2^n] for n from 1 to 5). We also measured the synonymy considering systematic ambiguity by using UniRef90 and UniRef50 to group UniProtKB entries.

The ambiguity is a measure of the number of UniProtKB entries per name. The average ambiguity was measured by the total number of entity-name pairs (EID, NAME) in BioThesaurus divided by the total number of names. We counted the number of names in eight ambiguity ranges: 1, 2, 3–4, 5–8, 9–16, 17–32, 33–64, and >64. The ambiguity of BioThesaurus with biomedical concepts was assessed based on names that were mapped to the UMLS concepts with semantic categories that were not highly related to gene or proteins. The ambiguity with general English words was assessed by matching single-word names in BioThesaurus with a common English word list we assembled using frequent word lists generated by Waring (http://www1.harenet.ne.jp/~waring/vocab/wordlists/vocfreq.html).

We evaluated the term coverage using the test set of Task 1B provided by the 2003 BioCreAtive Workshop for three model organisms (yeast, mouse, and fly). The following is a brief summary of the test set used in BioCreAtive for Task1B in detail. In the test set, genes that are described in the associated papers were identified and mapped to corresponding identifiers in the associated model organism databases by human experts with a total of 3,375 pairs of (GENE, PAPER). Among them, the number of pairs
identified by human experts as GENE that could be inferred explicitly from the abstract of PAPER was 1,593. The remaining 1,782 pairs that were judged as either GENEs were inferred explicitly from full paper or were inferred implicitly from the paper. Note that systems participating in Task1B of the 2003 BioCreAtive Workshop were evaluated using the 1,593 pairs only. When assessing the coverage, we also eliminated 250 pairs from the 1,593 because GENEs were not mapped directly to UniProtKB protein entries, thus had no corresponding BioThesaurus records. Note that when considering the whole 3,375 pairs, a total of 905 pairs did not have matching BioThesaurus records. The final identifier list consisted of 1,343 identifiers. For each gene GENE, names associated with GENE were extracted from BioThesaurus and, if there was at least one occurrence of any of those names in the associated abstract, we considered GENE was covered by BioThesaurus. Note that when identifying the occurrences of those terms, we used the normalized form of the terms. We computed the percentage of identifiers for each model organism that were covered by BioThesaurus.

To map to the correct entity, biological named entity tagging systems need to resolve ambiguity of terms that occur in the text. We assessed the ambiguity using the following methods. For each pair (GENE, ABSTRACT), there might be multiple terms associated with GENE occurring in ABSTRACT, and they may have various ambiguities. Tagging systems need to at least resolve the ambiguity of the least ambiguous term (that is, the term with the least number of gene identifiers associated with it) in order to map to the right entity. We considered the ambiguity of the least ambiguous term as the ambiguity that BNET systems need to resolve. We measured the ambiguity that BNET systems need to resolve twice: once considering systematic ambiguity and once ignoring systematic ambiguity according to UniRef90.

**Results**

The version of BioThesaurus used here consisted of 1.8 million UniProtKB entries with a total of 3.2 million names. A total of 34,598 names in BioThesaurus were mapped to concepts from 120 (out of 134) UMLS semantic categories when ignoring cases. Among them, we considered 18 categories highly related to genes/proteins based on the UMLS semantic category definitions. Table 1 lists these 18 categories with the number of mapped names. The top seven UMLS semantic categories: Amino Acid, Peptide, or Protein (T116), Gene or Genome (T028), Enzyme (T126), Biological Active Substance (T123), Receptor (T192), Immunological Factor (T129), and Pharmacologic Substance (T121), each had over one thousand matched names. Note that some categories listed in Table 1 are not concepts for genes/proteins per se but are highly related. For example, disease or syndrome (T047) and lab procedure (T059) are highly related to genes/proteins because disease terms are often used to name disease genes/proteins (e.g., disease name “Wolman disease” used to name the associated protein with UniProtKB identifier Q5T073) and lab procedures involving genes/proteins may be named after them (e.g., enzyme name Leucine aminopeptidase used to name the associated lab procedure with UMLS identifier C0202118). Several UMLS semantic categories had more than 100 matched names in BioThesaurus but were not deemed highly related to genes/proteins because disease terms are often used to name disease genes/proteins (e.g., disease name “Wolman disease” used to name the associated protein with UniProtKB identifier Q5T073) and lab procedures involving genes/proteins may be named after them (e.g., enzyme name Leucine aminopeptidase used to name the associated lab procedure with UMLS identifier C0202118). Several UMLS semantic categories had more than 100 matched names in BioThesaurus but were not deemed highly related to genes/proteins based on their definitions, such as Plant (T002), cell component (T026), Body Part, Organ or Organ Component (T023), and a few other categories (e.g., Quantitative Concept (T081) and Intellectual Product (T170)).

As shown in Table 2, the average synonymy of each BioThesaurus record is 3.53 (i.e., one protein entry is represented by an average of 3.53 names). After normalization, there were 2,106,222 names with an average synonymy 2.32. After grouping homologous entries and combining their names to remove systematic ambiguity according to UniRef90 and UniRef50, the number of entries is reduced to 1,015,207 and 580,956 respectively, and the synonymy was increased to 4.41 and 6.13 respectively before normalization and 3.61 and 4.90 respectively after.

The ambiguity result of BioThesaurus shows that the majority of names (83.7%) were associated with a single
UniProtKB entry, with an average ambiguity of 2.31 for the 2,665,968 names (Table 3). After normalization, the average ambiguity of the 2,106,222 names is slightly increased to 2.32. When systematic ambiguity was removed according to UniRef90 and UniRef50, the ambiguity dropped to 1.77 and 1.46 respectively before normalization and 1.79 and 1.49 respectively after.

The cross-biomedical domain ambiguity was evaluated based on BioThesaurus names that were mapped to terms in UMLS semantic categories using exact string match. Among 34,598 matched BioThesaurus names, only 1,483 names (4.3%) were mapped to UMLS semantic categories not considered highly related to genes/proteins (i.e., those not listed in Table 1). After normalization, 2,143 (4.0%) out of 53,052 matched normalized names were mapped to concepts with categories not considered highly related to genes/proteins. Examples include “Juvenile,” “minute,” and “purple.” The common English collection

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>154,444</td>
<td>30,263</td>
<td>30,263</td>
<td>82,167</td>
<td>143,594</td>
<td>50,049</td>
<td>70,018</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>523,523</td>
<td>270,976</td>
<td>141,410</td>
<td>432,730</td>
<td>216,890</td>
<td>171,166</td>
<td>92,761</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3–4</td>
<td>808,714</td>
<td>373,322</td>
<td>216,730</td>
<td>426,266</td>
<td>252,340</td>
<td>177,196</td>
<td>118,454</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5–8</td>
<td>222,083</td>
<td>182,712</td>
<td>146,388</td>
<td>183,691</td>
<td>130,275</td>
<td>102,232</td>
<td>72,611</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9–16</td>
<td>76,031</td>
<td>83,169</td>
<td>61,331</td>
<td>113,695</td>
<td>65,823</td>
<td>55,209</td>
<td>39,366</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17–32</td>
<td>1,471</td>
<td>20,234</td>
<td>28,828</td>
<td>13,046</td>
<td>20,784</td>
<td>17,971</td>
<td>20,784</td>
<td></td>
<td></td>
</tr>
<tr>
<td>33–64</td>
<td>908</td>
<td>2,387</td>
<td>8,971</td>
<td>1,271</td>
<td>5,606</td>
<td>5,606</td>
<td>5,606</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&gt;64</td>
<td>41</td>
<td>240</td>
<td>2,447</td>
<td>117</td>
<td>1,280</td>
<td>1,280</td>
<td>1,280</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total entities</td>
<td>1,800,505</td>
<td>1,015,207</td>
<td>580,956</td>
<td>1,015,207</td>
<td>580,956</td>
<td>580,956</td>
<td>580,956</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average synonymy</td>
<td>3.33</td>
<td>2.86</td>
<td>4.41</td>
<td>3.64</td>
<td>4.90</td>
<td>6.13</td>
<td>4.90</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

UniProtKB entry, with an average ambiguity of 2.31 for the 2,665,968 names (Table 3). After normalization, the average ambiguity of the 2,106,222 names is slightly increased to 2.32. When systematic ambiguity was removed according to UniRef90 and UniRef50, the ambiguity dropped to 1.77 and 1.46 respectively before normalization and 1.79 and 1.49 respectively after.

The cross-biomedical domain ambiguity was evaluated based on BioThesaurus names that were mapped to terms in UMLS semantic categories using exact string match. Among 34,598 matched BioThesaurus names, only 1,483 names (4.3%) were mapped to UMLS semantic categories not considered highly related to genes/proteins (i.e., those not listed in Table 1). After normalization, 2,143 (4.0%) out of 53,052 matched normalized names were mapped to concepts with categories not considered highly related to genes/proteins. Examples include “Juvenile,” “minute,” and “purple.” The common English collection

Table 2 - The synonymy assessment result of BioThesaurus, where Original denotes measures before removing systematic ambiguity, UniRef90 shows measures after grouping entities according to UniRef90, and UniRef50 for measures after grouping entities according to UniRef50. We measured the synonymy using eight ranges (1, 2, 3–4, 5–8, 9–16, 17–32, 33–64, and >64). The average synonymy as well as the total number of entities in the dictionary are also provided. The percentage in parentheses shows the percentage of entities with synonymy in the range to the total number of entities.

<table>
<thead>
<tr>
<th>Ambiguity</th>
<th>Original</th>
<th>UniRef90</th>
<th>UniRef50</th>
<th>Before Normalization</th>
<th>After Normalization</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2,233,471</td>
<td>2,233,471</td>
<td>2,233,471</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>223,698</td>
<td>270,826</td>
<td>303,400</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3–4</td>
<td>109,704</td>
<td>94,164</td>
<td>80,654</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5–8</td>
<td>53,535</td>
<td>35,258</td>
<td>26,968</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9–16</td>
<td>21,936</td>
<td>15,757</td>
<td>11,113</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17–32</td>
<td>10,841</td>
<td>7,579</td>
<td>5,579</td>
<td></td>
<td></td>
</tr>
<tr>
<td>33–64</td>
<td>5,646</td>
<td>4,183</td>
<td>3,134</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&gt;64</td>
<td>7,140</td>
<td>4,460</td>
<td>1,649</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total names</td>
<td>2,665,968</td>
<td>2,106,222</td>
<td>1,46</td>
<td>1.46</td>
<td>1.49</td>
</tr>
</tbody>
</table>

Average ambiguity 2.31 1.77 1.46 2.32 1.79 1.49

Table 3 - The ambiguity assessment result of BioThesaurus. Refer to Table 2 for notations. The ambiguity was measured using eight ranges (1, 2, 3–4, 5–8, 9–16, 17–32, 33–64, and >64). The total number of terms and the average ambiguity in the dictionary are also provided. The percentage in parentheses shows the percentage of terms with ambiguity in the range to the total number of terms.
We used the curated classification systems such as PIRSF, we used because of the incomplete coverage of all UniProtKB records as SGD and MGD, cross-referenced by iProClass. Tailored for systems that use records in other databases, such as BioThesaurus, entities (i.e., concepts) in the dictionary can be constructed links terms with protein entries in UniProtKB. Using the comprehensive cross-reference information stored in iProClass, entities (i.e., concepts) in the dictionary can be tailored for systems that use records in other databases, such as SGD and MGD, cross-referenced by iProClass.

The incomplete coverage of all UniProtKB records by the curated classification systems such as PIRSF, we used contained 8,090 words, all in lower case. Among them, 367 were also single-word names in BioThesaurus; when case difference was ignored, the number increased to 594 words. Examples of gene/protein names in common English words include “dare,” “air,” “all” and “ago.” The coverage result is shown in Table 4. The coverage of BioThesaurus is 94.0%, i.e., there are 1,262 out of 1,343 pairs of (GENE, ABSTRACT) where at least one name associated with GENE in BioThesaurus occurred in ABSTRACT. Additionally, 119 out of 1,127 (i.e., 3,375–905–1,343) pairs of (GENE, ABSTRACT), which were originally curated as a GENE that was not explicitly mentioned in ABSTRACT, were found to be errors since names associated with GENE in BioThesaurus were actually found to occur in ABSTRACT.

The coverage was assessed using the test set of the BioCreAtive workshop. The percentage in the first column shows the percentage of terms in BioCreAtive text present in BioThesaurus. The percentages in the second, third, and last columns were acquired by inverting the ambiguity which refer to the precisions of a base system that randomly picks an associated entity for those ambiguous terms.

<table>
<thead>
<tr>
<th>Organisms</th>
<th>Mentioned in Abstracts</th>
<th>The Coverage of Genes in the Evaluation Set Present in BioThesaurus</th>
<th>The Ambiguity of Matched Terms in BioThesaurus</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Including Systematic Ambiguity</td>
</tr>
<tr>
<td>Yeast</td>
<td>557/595 (93.6%)</td>
<td></td>
<td>55.7 (1.8%)</td>
</tr>
<tr>
<td>Mouse</td>
<td>346/378 (91.5%)</td>
<td></td>
<td>28.0 (3.6%)</td>
</tr>
<tr>
<td>Fly</td>
<td>359/370 (98.1%)</td>
<td></td>
<td>50.7 (2.0%)</td>
</tr>
<tr>
<td>Total</td>
<td>1,262/1,343 (94.0%)</td>
<td></td>
<td>46.9 (2.1%)</td>
</tr>
</tbody>
</table>

The ambiguity that a dictionary-based BNET system needs to resolve according to BioCreAtive text was in the range of 28–35 when including systematic ambiguity. It dropped to the range of 13–23 when ignoring systematic ambiguity. If we ignore other types of ambiguity (i.e., cross-general ambiguity and cross-medicine ambiguity), a base system which randomly picks an entity for those ambiguous terms could achieve a precision in the range of 1.8%–3.4% when including systematic ambiguity and of 4.5%–7.5% when ignoring systematic ambiguity (i.e., when we only consider entity-specific ambiguity, a base dictionary-based BNET system achieved a precision in the range of 4.5%–7.5% and a recall in the range of 91.5%–98% for BioCreAtive Task1B test set).

**Discussion and Conclusion**

In this paper we have assessed BioThesaurus regarding its ambiguity, synonymy, and coverage to reveal the complexity of dictionary-based biological named entity tagging. Most existing NLP systems currently do not associate terms appearing in the text with database entries and they do not need to resolve entity-specific ambiguity (i.e., a term represents multiple database entries). The dictionary we constructed links terms with protein entries in UniProtKB. Using the comprehensive cross-reference information stored in iProClass, entities (i.e., concepts) in the dictionary can be tailored for systems that use records in other databases, such as SGD and MGD, cross-referenced by iProClass.

Because of the incomplete coverage of all UniProtKB records by the curated classification systems such as PIRSF, we used UniRef clusters to assess the synonymy and ambiguity when ignoring systematic ambiguity. In fact, UniRef clusters provide tighter sequence groupings than PIRSF, i.e., one PIRSF family can map to one or more UniRef50 clusters. It is known that functions of some homologous proteins can be well conserved from higher to lower organisms, while those of others may only be conserved in close lineage of organisms (e.g., among mammals). Moreover, even homologous proteins including paralogs sharing high degree of sequence identity may differ in functions due to changes of key residues such as active sites (including gaining or losing functions). It is not known the percentage of such proteins in UniRef clusters, which is beyond the scope of this paper. Nonetheless, we feel that using UniRef90 and 50 is still a reasonable way to estimate the level of name ambiguity when disregarding systematic ambiguity of homologous proteins, albeit the true level of this ambiguity could be higher. One advantage of estimating the name ambiguity by removing the systematic ambiguity is that it could help identify protein name misnomers and name annotation inconsistency in databases. For example, human EGFR (epidermal growth factor receptor) was incorrectly assigned as gene name of the human EGF entry (UniProtKB: Q8NDU8), thus, giving the name “EGFR” higher ambiguity.

The assessment of synonymy and ambiguity provided in the result section was derived disregarding the fact that some terms may never appear in literature. For terms that actually occur in the text, the average ambiguity for them tends to be higher. We could not assess them directly. We plan to explore ways to assess the real ambiguity in the text. Even though there are not many terms in BioThesaurus that are common English words, but they occur frequently in the literature where most of these occurrences hold meanings other than biological entities (e.g., all or to). We plan to investigate further for disambiguating these terms in text. In the assessment of synonymy, we did not include abbreviation-derivable in the assessment. It is because most gene/protein symbols are actually abbreviation-derivable. For example in Figure 2, the symbol “CLIM1” is abbreviated from “C-TERMINAL LIM DOMAIN PROTEIN 1.” However, it is not easy to handle abbreviation-derivable synonyms in entity tagging systems because of their ambiguities. We plan to further study abbreviation-derivable synonyms.

Table 4: The coverage assessment for BioThesaurus using the test set of the BioCreAtive workshop. The percentage in the first column shows the percentage of terms in BioCreAtive text present in BioThesaurus. The percentages in the second, third, and last columns were acquired by inverting the ambiguity which refer to the precisions of a base system that randomly picks an associated entity for those ambiguous terms.
During the coverage assessment, we removed 250 pairs of (GENE, ABSTRACT) with no corresponding UniProtKB records associated with them. Most of them were from fly (with 63 pairs) and mouse (with 168 pairs). Among these pairs, almost all were associated with records of non-coding genomic regions or gene clusters/complexes. For example, there are over 60 pairs of (GENE, ABSTRACT) in MGI having a type as DNA segment (e.g., “D2Mit6” with identifier MGI:92196), over 10 pairs having a type as complex (e.g., “Hbb” with identifier MGI:96020). Most of such pairs in fly were associated with transposable elements instead of coding genes. Pairs in yeast that were not associated with UniProtKB records are also caused by non-coding gene records. For example, “HSX1” with identifier S00006707 in SGD has a type tRNA.

During the assessment, we also found that UniProtKB has finer granularity than model organism databases for genes that were cross-referenced by iProClass. For example, 25,514 UniProtKB entries were associated with 16,553 FlyBase records. One reason is that while entries in model organism databases are gene-centric, multiple protein products (isoforms) from alternative splicing of the same gene can be represented by multiple entries in UniProtKB. Another factor is that multiple protein fragments of the same gene could be represented in UniProtKB, which may have derived from independent labs, and sometimes may be given different names. For example, there are two UniProtKB entries (i.e., O76923 for isoform A and O96378 for isoform B) associated with the Ap-2 gene (FBgn00023417). In mouse, there are eight UniProtKB entries (e.g., Q08501, Q8C7G1, Q3UPQ1), including spliced variants and fragments, associated with one mouse gene Prlr (MGI:97763). Because of this many-to-one mapping of UniProtKB to the model organism databases, utilizing the rich cross-reference information supplied by iProClass, BioThesaurus can be tailored to dictionaries for entities in other databases (e.g., MGD) by providing a richer source of protein names and synonyms.

During the coverage assessment study, most terms (94.0%) identified by human experts in abstract of the BioCreAtIvE test set were included in BioThesaurus. It indicates that a BNET system which simply annotates a term with all associated entities can achieve a high recall (as high as 94.0%). However, the precision of such system would be very low since there are a lot of false positives caused by the ambiguities (as low as 2.1% when considering systematic ambiguity and entity-specific ambiguity and 5.3% when only considering entity-specific ambiguity). If we already know the corresponding organism for the paper, we can limit entities to a specific organism. The last column of Table 4 shows the result of the ambiguity when we tailored BioThesaurus to specific organisms. From Table 4, the ambiguity (including systematic ambiguity and entity-specific ambiguity) for mouse (or yeast) decreased to 1.23 (or 1.12) which implies that the precision for a BNET system which randomly picks an entity from entities associated with an ambiguous term was 81.3% (or 88.5%). This feature is helpful in literature-based protein database annotation and indeed is already used for semi-automatic mapping of gene/protein names to UniProtKB entries from a rule-based text mining system RLIMS-P for protein phosphorylation.57 However, the real precision measure for such a system would be relatively lower because there are problems with tokenization, cross-general ambiguities, and cross-medicine ambiguities.

In conclusion, biological entity tagging is an essential task for NLP systems in the biomedical domain. We assessed BioThesaurus through several perspectives: ambiguity, synonymy, and coverage. Based on the assessment, we quantified the complexity of dictionary-based biological named entity tagging systems. The study demonstrated that most gene/protein names appearing in text can be found in BioThesaurus. The study confirmed that biological named entity tagging is a non-trivial task because of the high ambiguity of gene and protein names as well as multiple names corresponding to the same gene/protein. While classifying ambiguity into different categories and applying different approaches to resolve these ambiguities, we believe the performance of dictionary-based biological named entity tagging could be improved.

References


