Analysis of gene and protein name synonyms in Entrez Gene and UniProtKB resources

Thesis by

Basil Arkasosy

In Partial Fulfillment of the Requirements

For the Degree of

Master of Science

King Abdullah University of Science and Technology

Thuwal, Kingdom of Saudi Arabia

May 2013

COMMITTEE APPROVAL FORM

The thesis of Basil Arkasosy is approved by the examination committee.

Committee Chairperson: Professor Vladimir Bajic

Committee Member: Professor Mikhail Moshkov

Committee Member: Professor Xiangliang Zhang

ABSTRACT

Analysis of gene and protein name synonyms in Entrez Gene and UniProtKB resources

BASIL ARKASOSY

Ambiguity in texts is a well-known problem: words can carry several meanings, and hence, can be read and interpreted differently. This is also true in the biological literature; names of biological concepts, such as genes and proteins, might be ambiguous, referring in some cases to more than one gene or one protein, or in others, to both genes and proteins at the same time. Public biological databases give a very useful insight about genes and proteins information, including their names.

In this study, we made a thorough analysis of the nomenclatures of genes and proteins in two data sources and for six different species. We developed an automated process that parses, extracts, processes and stores information available in two major biological databases: Entrez Gene and UniProtKB. We analysed gene and protein synonyms, their types, frequencies, and the ambiguities within a species, in between data sources and cross-species. We found that at least 40% of the cross-species ambiguities are caused by names that are already ambiguous within the species. Our study shows that from the six species we analysed (*Homo Sapiens, Mus Musculus, Arabidopsis Thaliana, Oryza Sativa*, *Bacillus Subtilis* and *Pseudomonas Fluorescens*), rice (*Oriza Sativa*) has the best naming model in Entrez Gene database, with low ambiguities between data sources and cross-species.

ACKNOWLEDGEMENTS

Alhamdulilah (Praise to Allah)

I would like to express my sincere gratitude to my supervisor, Professor Vladimir Bajic, for investing his invaluable support, guidance, feedback, and time throughout the duration of this work. His endless enthusiasm and encouragement always motivated me to carry out my research with a keen sense of duty and commitment. I would like to thank him for being a part of this rich, rewarding journey, and for his generosity in sharing his wisdom, and vast expertise in the field.

I sincerely thank the erudite members of my defence committee, Professor Mikhail Moshcov and Professor Xiangliang Zhang, for their time, insightful feedbacks and guidance.

I deeply appreciate the support of Martin Senger, a Senior Research Scientist at KAUST's CBRC Center. I have used and built upon his software components, I made numerous research discussions with him, and I have largely benefited from his constructive feedback.

I would like to express my profound gratitude to Professor Aiman El-Maleh, my advisor at my alma mater, King Fahd University of Petroleum and Minerals (KFUPM), for his influence in my eventual decision to continue my graduate studies.

I cannot forget to thank my dearest friends, who used to always follow up, and greatly support me during this journey. My special thanks also go to my colleagues here at KAUST; your support, collaboration and late hours study companionships, gave this journey that extra bit of joy that complemented my curricular activities.

To my father-in-law Mohammad and my aunt Nada, your unlimited support and continuous encouragements for graduate studies, served a big contribution to where I am now today. My sincere thanks also go to my uncles and aunts Nabeel, Sabah, Iman, Ghadah, Maher and Samir; with your support and motivation I always feel proud with what I achieve.

To my brothers Fadi, Ahmad, Baraa and Majd, you were always there for me, serving as the solid rocks over which I leaned whenever reaching for the peak of the steep mountain seemed daunting. To my father, Saeed, and my mother, Huda, the greatest parents I could ever wish for, I am very grateful to you. I could not have gotten here if you had not supported me with your care, support, love, and prayers. Simply put, without you, I would not have been.

To the lights of my eyes, my kids, Mohammad and Abdurrahman, even though I had to make the tough but necessary decision to spend time away from you every now and then, I always did so looking forward to the priceless joy that would occupy me on my back home; in the knowledge that you both would never fail to reenergize me with your smiles which it always lightened my long, study nights.

Last but most important, is my grateful appreciation to the best friend I have ever had, my beloved wife Nesreen. The love, support, confidence, and patience you gifted me were pivotal to this success. Words fail me when I need them the most, and thanking you can never sufficiently express my feeling of gratitude. Nothing I do would mean anything without you, this is really our success Nesreen; we have made it together.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

CDA	Cross-species Degree of Ambiguity
DA	Degree of Ambiguity
DB	Database
EG	Entrez Gene Database
ODA	Overlapping Degree of Ambiguity
OFN	Official Full Name of a gene/protein
OFS	Official Symbol of a gene/protein
SDA	Strong Degree of Ambiguity
UPK	UniProtKB Database

LIST OF ILLUSTRATIONS

LIST OF TABLES

Chapter I

Introduction

With the exponential growth of the volume of scientific literature, it is becoming harder to enforce standards used in the research worldwide. Naming conventions for biological concepts such as genes and proteins are among the challenges faced today in the biology domain [1]. Standards for naming genes and proteins might not be well defined, or in some cases, do not exist [2]. In addition, researchers might not adhere to the existing guidelines when they assign names to new genes/proteins, when they use them or when they refer to them in their publications. This adds a significant overhead and many difficulties to the on-going research in these fields, especially when new genes/proteins are discovered and reported on daily basis [1].

I.1 Organisms and Species

An organism refers to any living biological creature such as an animal, a plant or a bacterium. Taxonomy, one of the branches of biology, defines the principles for describing and classifying organisms that share certain characteristics into groups, and assigning names to these groups [3]. Taxonomic classification is hierarchical, where each level is known as a rank, and *kingdoms* represent the largest, most inclusive classification category [4]. The Six-Kingdom system classifies organisms based on their characteristics into the kingdoms: *Bacteria*, *Archaea*, *Protista*, *Fungi*, *Plantae*, and *Animalia [4, 5]*. On the other hand, at the

bottom of the taxonomic hierarchy, the basic unit of the biological classification is known as a species. Generally, species have common names besides their scientific names that are used in the literature. For example, 'Human' is a species with the scientific name '*Homo Sapiens'*. Similarly, 'Mouse' is a common name for a species scientifically known as '*Mus Musculus*'.

I.2 Genes Nomenclature

Genes are segments of the DNA from where copies of DNA, transcripts, are produced. These transcripts can perform various functions in cells. They may encode information based on which functional units called proteins are synthesized or they may have regulatory roles [6]. Gene nomenclature refers to the set of standards and conventions used for the scientific naming of genes [7]. Genes and proteins are normally given names relevant to their functionalities [8], leading in many cases to long descriptive names. For example, 'Eye Color 1 (green/blue)', 'hair color 1 (brown) ' and 'hair growth associated' are official names for some genes, whose functionalities are well described by their names. It should be noted that a gene or a protein can have many functions, so not all gene/protein functionalities could be captured in a single name, simply because not all functionalities were known at the time the first gene function was established. Besides official full names, genes and proteins may have shorter official symbols and/or aliases that are also used in the literature.

In most cases, several names, known as synonyms, are used to refer to the same gene and protein in various sources. For example, 'hair color 2 (red)', 'RHC' and 'HCL2' are synonyms that refer to the same gene (Entrez Gene ID 3057) in the Entrez Gene database [9]. Synonyms could make it harder for researchers to link information about particular genes from different

publications where different synonyms are used to refer to the same gene. Homonyms, on the other hand, refer to those names/terms that are denoting two or more biological entities. For example, the symbol 'MPPH' is an alias that refers to two different genes in Entrez Gene database (Entrez Gene IDs: 5296 and 10000). Naming ambiguity occurs because of homonyms; that is when a gene or protein name is used in the text, but it cannot be uniquely associated to a specific gene or protein, or even identified as a gene or as a protein.

I.3 Biological Databases

The huge amount of biological data has motivated researchers to build public databases that store and catalogue biological information. Public biological databases maintain and organize high quality information about genes and proteins. They assign unique identifiers to them, and associate these identifiers with the different names, aliases and biological properties genes and proteins have, such as, for example, gene expressions or protein structures. Most of these databases were constructed using information from literature but also form primary experimental data. When the population of a database resource is based on the literature, the concepts from articles are transferred to the corresponding fields in the database using, controlled vocabulary or gene ontologies [8, 10, 11].

The biological databases differ in their sizes, contents and the objectives for which they were built. Some are organisms specific; they only contain information about a specific species and are usually maintained by research groups that are focused to these particular species, such as TAIR (for *Arabidopsis thaliana*) [12] and Flybase (for the fruit fly, *Drosophila melanogaster*) [13]. Other databases, however, are more general; they contain rich information about many

different kinds of species. Yet, the information contained about specific species in general databases varies largely from a few hundreds to hundreds of thousands of records, depending on the species. In addition to the size and the content, these databases differ in their data structures, the way data have been curated and stored, and, more importantly, the way data is presented to users or made available to programs. Some databases provide data on flat or XML files which are easy to parse and extract, while others provide data in more complicated forms that require extra processing and manipulation [8, 14]. Entrez Gene [9] and UniProtKB [15] are two widely used public biological databases that contain information on genes and proteins, respectively, for many different species.

Entrez Gene is a gene-specific database, hosted at the National Center for Biotechnology Information (NCBI) in the United States. It assigns unique identifiers to genes, and links them with the genes' nomenclature, along with much of the other information. It also provides many useful, easy to use, reporting features [9].

UniProt (Universal Proteins Resource) [15] is a central repository for storing and integrating information on proteins gathered from different resources. At the heart of UniProt, a well-curated database known as UniProt Knowledge Base (UniProtKB) is maintained. UniProtKB is made up of two parts: UniProtKB/Swiss-Prot and UniProtKB/TrEMBL, where the former contains curated protein records, while the later contains computationally analysed high quality records, supported with automatic annotation and classification [15].

I.4 Problem Definition

The wealth of information available in public gene and protein databases makes a very useful

information source about genes and proteins. In this study, we made a thorough analysis of the use of naming of genes and proteins in different data sources, and for several different species with the aim to understand the level of ambiguities that is present. Our study is important to researchers who work in biological data analysis, for example in literature mining or data integration.

We develop and automate a process that parses, extracts, processes and stores information from the data available in two of the most used gene and protein databases: Entrez Gene and UniProtKB. This process will greatly simplify the analysis of the large number of data records, and the hundreds of thousands of synonyms, for different species in these resources. Additionally, we perform various analyses of gene and protein synonyms, their types, frequencies, and the ambiguities within a species, in between data sources for the same species, and in between different species. Finally, we analyse and discuss the results for three pairs of different species: *Homo Sapiens* and *Mus Musculus, Arabidopsis Thaliana* and *Oryza Sativa, Pseudomonas Fluorescens* and *Bacillus Subtilis*, where each pair represents species that belong to the same *kingdom*.

I.5 Thesis Organization

The remainder of this document is organized as follows: Chapter 2 discusses the related work, and highlights a summarized literature review. Chapter 3 explains the methodology and the implementation details. Chapter 4 presents the experiments, their results and the discussion and interpretation of the results, while conclusions are given in Chapter 5. References and Appendices are available at the end of the document.

Chapter II

Literature review

The problems of ambiguities in the biological and medical literature have been well known since many years, and there have been several studies that addressed these problems. Finding abbreviations and expanding them, named entity disambiguation, gene and protein names identification, nomenclature guidelines, building dictionaries and analysing public databases represent different on-going research in the field [14].

Abbreviations, such as gene and protein symbols, are one of the main sources of ambiguity in texts. In order to understand them, different methods and algorithms were developed to expand gene and protein symbols (associate them to their definitions or full forms). Hongfang et al. [16] studied the ambiguity of the three-characters abbreviations in the MEDLINE abstracts and reported that it is possible to automatically expand abbreviations that are frequently associated with their definitions in texts. Yu and his colleagues [17] presented a similar work and developed methods that automatically map abbreviations to their full forms. They found that about (25%) of the abbreviations used in biomedical articles are also defined in the same biomedical text.

Weeber et al. [1] studied the ambiguity related to the use of human gene symbols from LocusLink database (Entrez Gene now) in MEDLINE articles. They explained that even though over (40%) of the symbols appeared in MEDLINE articles, many of them were not

related to genes. Yushida et al. [18] developed a workbench for building a dictionary for protein names abbreviations. Adar [19] implemented a high performance dictionary-building tool to disambiguate abbreviations and symbols in biomedical texts. Zhou et al. [20] built ADAM, a database of abbreviations, both acronym and non-acronym, and their definitions in MEDLINE titles and abstracts. Xu and his colleagues [21] studied the abbreviations in clinical notes and described a model for constructing a database of abbreviations in medical notes.

Another area of research is related to identifying biological concepts in texts. Hirschman et al. [8] summarized the problems related to biological names identifications and explained the challenges experienced with recognizing fly gene names. To further support their analysis, they compared information extraction from news and from biology and explained why tagging named entities is harder for biologists in general. Malik et al. [22] created CONANR, a system that combines different algorithms to tag genes, proteins and biological concepts and link them to MeSH [23] and Gene Ontology. Tanabe and Wilbur [24] proposed a statistical and knowledge-based approach for tagging gene and protein names in biomedical texts. Fukuda et al. [25] developed PROPER, a method that uses proteins nomenclature to extract proteins names that are known or newly defined from medical and biological articles with high accuracy.

Fundel at al. [26] implemented a simple approach for gene and protein identification from free text. They maintained a synonyms list that maps the database identifiers to the different synonyms for each gene and protein. Settles [27] implemented ABNER, an open source software tool, that uses machine learning techniques to automatically tag gene and protein names in texts. BioCreAtIvE was the first assessment of the text mining methods used for

gene and protein names extraction and identification [28, 29]. Different research groups participated in this assessment, which aimed to extract and identify gene and protein names for mouse, fly and yeast species. Similar work related to the extraction and identification of gene and proteins names can be can be found in [30-34].

In efforts to address gene and protein nomenclature problems, research committees for different species set the standards and the guidelines for the gene/protein nomenclatures of these species. The first guidelines for the human genes nomenclature were published at the Edinburgh Human Genome Meeting in 1979 [2, 35], and were updated through years after that. Similar guidelines exist for mouse [36], bacteria [37], rice [38] and many other species. Recently, the first 'Gene Nomenclature Across Species' meeting was held in 2009, and discussed and organized gene naming across vertebrates [39]. The meeting discussed the implementation and the coordination of the gene nomenclature across species in the databases.

One of the important guidelines that could help resolve the ambiguity in gene/protein names is to assign official full names and official symbols to genes and to use these, instead of aliases, in the literature. Nobert and Wain [35] highlighted this point and explained the importance of using official gene names and symbol.

Chen et al [40] reported that authors of biological and biomedical texts tend to use genes/proteins official names at the level of only 7.6% in their publications, while they use symbols at the level of 17.7%, and in the remaining 74.7% aliases are used. They also investigated mouse genes nomenclature and found that the ambiguity of symbols is in 14% of cases, which is low compared to the aliases ambiguity of 85%, supporting the hypothesis that most of the ambiguity is caused by aliases. As potential solutions to the ambiguity problems,

they suggested that authors should strictly follow the rule of only using official names or official symbols in text and avoid using other notation. Some journals such as *Genomics* and *Nature Genetics* already support this suggestion by forcing the rule, and it could be very helpful if other journals adopted it as well. Schumie et al [41] came to similar conclusion when they studied the distribution of information in the abstracts and the full texts of the biomedical publications. They reported that genes official full names are not frequently used in the literature and gene symbols are introduced most of the time without their definitions. They found that only 30% of the time, the full names are used along with their symbols in the biomedical abstracts. They reported, however, that the percentage of having the genes full names expanding their symbols in the full text drops to as low as 18%.

Analysing public biological databases is another active research areas targeting gene nomenclature problems. Tuason and his colleagues [2] conducted the first comparisons of genes nomenclatures across species. They studied the amount of ambiguity for four different species (mouse, worm, fly and yeast) within their respective databases, in between databases and with English words. They found that between 0% to 10.18% of the names within species, are ambiguous and most of the ambiguity is caused by aliases. They also reported that the naming conventions followed by different species committees have an impact on the degree of the ambiguity, where more lenient rules lead to higher percentages of ambiguities.

Fundel and Zimmer [14] compared the degree of ambiguity of the gene and protein names for five different species (human, mouse, rat, fly and yeast). They extracted data from different organisms-specific and public data sources and compared the ambiguity within and in-between these data sources. They found that the degree of ambiguity for the same species

could vary in different data sources, even if these data sources contain similar number of synonyms for the species under study. They also studied the ambiguity across species and found that human, mouse and rat have higher degrees of inter-species ambiguity over fly and yeast. In addition, they analysed the overlap of synonyms between different data sources and found that the overlap varies significantly and ranges from 11% to 83%. They explained that this big range could be related to the differences in the data structure of the data sources and the strategies used to maintain them. They also used this big range to support the hypothesis that combining entries from different data sources is essential to build complete dictionaries.

Chapter III

Methodology

In this study, we developed and automated the process to parse, extract and store information from genes and proteins nomenclature data available in two major public databases. Additionally, the process performs various analyses and generates different results of gene and protein synonyms, their types, frequencies and the ambiguities within a species, in between data sources, and in between different species. Fig. 1 below illustrates the various integrated components that build the process.

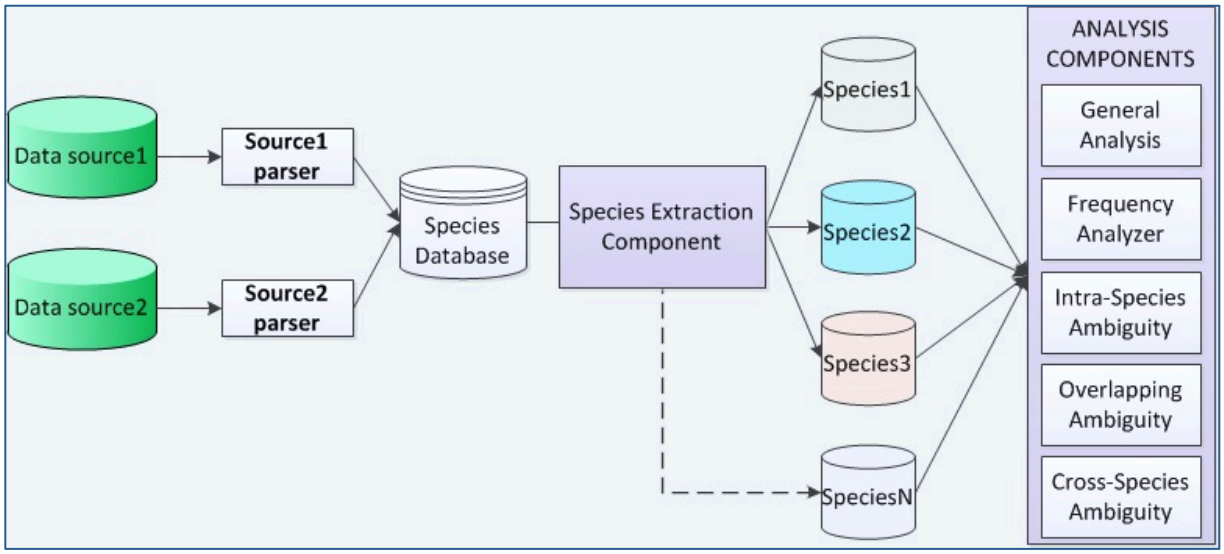

Figure 1. Automated process components to analyze genes/proteins nomeclatures. *

The first part of the automated process is related to data gathering from different data sources,

* Martin Senger implemented the essential components of this process. These are: data parsers and species extraction components. Some were co-developed such as the frequency and intra-species ambiguity components.

parsing and storage for all species. After that, data can be filtered with a species-extraction component that extracts and stores all the records of particular species. Once species records are separated into smaller, separated databases, different components can execute different procedures and generate various types of results.

III.1 Data Sources

We are analysing genes and proteins datasets available in Entrez Gene and UnirpotKB (both parts: Swiss-Prot and TrEMBL) public databases. We use data from both sources as of 29th January 2013.

III.2 Data Parsing and Extraction

Data provided by different sources have different structures and formats. Therefore, a source-specific parser is needed for each data source. We have implemented two parsers to parse and extract data from the flat files provided by Entrez Gene and UniProtKB.

Since we are interested in the different synonyms that genes and proteins have, we mark each synonym along with its type. A synonym could be an *Official Full Name*, an *Official Symbol* or an *Alias*. Entrez Gene has additional fields other than the official name and symbol, such as A*lternate_Name* and an *Alias*, but we unify these and consider them as synonyms of type *Alias*. Locus_tag is another field in Entrez Gene which is a systematic identifier assigned to genes. Therefore, we will not take it into consideration when computing synonyms and ambiguities. Similarly, UniProtKB contains different fields, out of which we will mark the official names and official symbols from the *RecName* fields, as they represent the

recommended names by the UniProt Consortium, and we consider other fields as aliases. Entry_Name field is the unique identifier given to protein records in UniProtKB, and we will not consider it our study.

In order to ease our computations, we combine and store the extracted data records from both sources in a MySQL database in a synonym-oriented format. Table 1 below shows the main fields of the table that stores the data records and their descriptions.

Table 1. Main fields and descriptions of the table that stores data records.

Field Name	Description
ID	Auto generated ID
Source	Data source name
SID	Unique ID provided by the data source [Entrez Gene ID or UniProtKB ID]
Synonym	Synonym
Type	Synonym type (All synonyms, Official Name, Official Symbol, or Alias)

Table 2 demonstrates samples of data records from different data sources. For example, the second row represents a synonym ‘sqs1’ of the type ‘official symbol’, for the gene in Entrez Gene database (Entrez Gene ID: 541614) and the official full name ‘squalene synthase1’.

Table 2. Sample data records parsed, extracted and stored in MySQL database.

ID	Source	SID	Synonym	Type
1	Entrez Gene	541614	squalene synthase1	Official_Full_Name
2	Entrez Gene	541614	sqs1	Official_Symbol
3	Swis-Prot	A0AEM0	Arginine deiminase	RN_Full
4	Swiss-Prot	A0AEM0	ADI	RN_Short
5	TrEMBL	B2YI80	Elongation factor P	RN_Full
6	TrEMBL	B2YI80	EF-P	RN_Short

III.3 Species Extraction

In order to conduct our analysis, it is important to extract the records for species we intend to study. Each species of interest will be separated and stored in its own MySQL database in the same structures and formats as shown above in Tables 1, 2.

III.4 Frequency Computation

Different species have different distributions of synonyms over data records. Genes and proteins may have one or more synonym from the same type or from different types, and the ranges of the number of synonyms for genes/proteins vary largely between species. In order to study these variations, we define the frequency of synonyms, $Freq_i(T)$, as the percentage of data records that have ***i*** number of synonyms of type T, as shown in Equation 3.1.

$$Freq_i(T)\% = \frac{X_i\ (T)}{Total\ number\ of\ data\ records}\ X\ 100\% \qquad (3.1)$$

Where T is the synonym type, T ∈ (All Synonyms, Official Full Name, Official Symbols, Aliases) and X_i (T) is the number of records X that have ***i*** number of synonyms of type T.

Computing the frequencies may take long time, depending on the size of data records, and hence once we compute these, we materialize and store them in a separate frequency table in the species-specific MySQL database. Table 3 below shows the main fields of the frequency table and their description.

Table 3. Main fields and descriptions of the materialized frequency tables.

Field Name	Description
ID	Auto generated ID
Source	Data source name
Type	Synonym type
Num of Syns	Number of synonyms a gene/protein may have
Count of records	Number of data records that have this number of synonyms

Table 4 demonstrates samples of the computed the frequency records for the *Mus Musculus* species. The first row, for example, indicates that out of the *Mus Musculus* records in Entrez Gene database, 2167 records have a single synonym. Row number 4, however, indicates that only 159 of the *Mus Musculus* records in Entrez Gene have exactly two synonyms.

Table 4. Sample data records of the materialized frequency tables for *Mus Musculus*.

ID	Source	Type	Num of Syns	Count of records
1	Entrez Gene	All Synonyms	1	2167
2	Swiss-Prot	All Synonyms	1	4074
3	TrEMBL	All Synonyms	1	875
4	Entrez Gene	All Synonyms	2	159
5	Swiss-Prot	All Synonyms	2	3552
6	TrEMBL	All Synonyms	2	721

III.5 Ambiguity

A synonym is ambiguous if it refers to more than one gene or protein name. Many gene and protein names that are extracted from a single data source for a specific species are ambiguous. Another type of ambiguity can be caused by the overlapping gene names from different data sources. Moreover, different species could share the same names for genes and proteins, even though these genes/proteins may not have the same functionalities. In this

work, we analyze and study these different types of ambiguities.

III.5.1 Intra-species Ambiguity

Intra-species ambiguity is caused by ambiguous synonyms of genes and proteins, which are extracted from the same data source and for the same species. Different species are expected to have different range of ambiguities, depending on the naming conventions and standards used by those species research committees. We compute the degree of intra-species ambiguity for a synonym type T as the quotient of the ambiguous T synonyms over the total number of T synonyms. Recall that T represents the synonym type; for example, we might be interested to find the number ambiguous official name synonyms for a species. The degree of ambiguity $DA(T)$ for synonyms of type T is denoted in Equation 3.2

$$DA(T) = \frac{\#\text{AmbigSyn(T)}}{\#\text{TotalSyn(T)}} X\ 100\% \qquad (3.2)$$

Where $\#\text{AmbigSyn(T)}$ represents the number of ambiguous T synonyms, and $\#\text{TotalSyn(T)}$ represents the total number of synonyms. Additionally, we define the Level (L) of ambiguity of a synonym as the number of entities that share that synonym. For example, a synonym that has an ambiguity level (L=3) indicates that this synonym is shared between three gene or protein names. As the distribution of the ambiguity levels and the degrees of ambiguity for these levels vary between different species, we compute the degree of ambiguity for synonyms of type T and level L by finding proportions of ambiguous T synonyms that have ambiguity level L, $\#Syn_L(T)$, relative to the unique set of all T synonyms as shown in Equation 3.3.

$$DA_L(T) = \frac{\#Syn_L(T)}{\#\text{UniqSyn(T)}} X\ 100\% \quad (3.3)$$

Computing these degrees for different levels might be computationally expensive, and hence, we materialize and store them in a separate ambiguity table in the species-specific database. Table 5 below shows the main fields of the ambiguity table and their descriptions.

Table 5. Main fields and descriptions of the materialized ambiguity tables.

Field Name	Description
ID	Auto generated ID
Source	Data source name
Type	Synonym type
Level	The level of ambiguity
Count of Synonyms	Number of synonyms that have this level of ambiguity

Table 6 demonstrates samples of the computed ambiguity records. For example, the first row indicates that there are 14368 *Mus Musculus* synonyms that have level L=2, i.e. that are shared between two records.

Table 6. Sample data records of the materialized frequency tables for *Mus Musculus*.

ID	Source	Type	Level	Count of Synonyms
1	Entrez Gene	All Synonyms	2	14368
2	UniProtKB	All Synonyms	2	732
3	Entrez Gene	All Synonyms	3	2501
4	UniProtKB	All Synonyms	3	209

III.5.2 Overlapping Ambiguity

Overlapping ambiguity is caused by ambiguous synonyms of genes and proteins that exist in two different data sources for the same species. For this type of ambiguity, we analyze the distribution and the degree of overlapping ambiguity, with reference to each data source separately. First, we find the overlapping synonyms by extracting the set of unique synonyms from both data sources, and then compare and match these sets. The synonyms that are found in both sets are said to be overlapping ambiguous synonyms. Then, we find the distribution of these overlapped ambiguous synonyms in each data source by analyzing the proportions they represent out of the unique set of synonyms in each data source. The pseudo code to find the overlapped synonyms is shown in Algorithm 1 below:

Algorithm 1: FINDING DATA SOURCES OVERLAPPING SYNONYMS

```
1 Extract the unique synonyms from data source1 and store them in hash set S1
2 Extract the unique synonyms from data source2 and store them in hash set S2
3 Sort S2
4 for each synonym S in S1: do
5 |   Apply binary search for (S) in (S2)
6 |   if S is found in S2 then
7 |   |   matchedCount++
8 return matchedCount
```

We compute the degree of overlapping ambiguity, for synonyms of type T, between the data sources S1 and S2 relative to S1 ($ODA^{S1}_{S1,S2}(T)$) as follows

$$ODA^{S1}_{S1,S2}(T) = \frac{\#\ Matched\ T\ Synonyms}{\#\ of\ Unique\ T\ Synonyms\ in\ S1}\ X\ 100\% \qquad (3.4)$$

III.5.3 Cross-species Ambiguity

We refer to the ambiguity caused by synonyms shared between different species, as the cross-

species ambiguity. In this study we analyze the cross-species ambiguity between groups of species, two at a time, within the same data source. First, we find the cross-species ambiguous synonyms by extracting the set of unique synonyms of the two species from the same data source, and then comparing and matching these sets. The synonyms that are found in both sets are said to be cross-species ambiguous synonyms. Then, we find the distribution of these synonyms relative to each species by computing their degree of ambiguity.

The relationship between cross-species and intra-species ambiguities could give interesting highlights about the ambiguous synonyms. Here, we study this relationship by defining the strongly ambiguous synonyms. A synonym is strongly ambiguous if it causes a cross-species and intra-species ambiguities. In order to find the strongly ambiguous synonyms with respect to a species, we first extract the unique set of ambiguous synonym from that species, and then compare and match this set with the set of cross-species ambiguous synonyms that was extracted before. The synonyms that are found in both sets are said to be strongly ambiguous synonyms. We define the degree of strong ambiguity with respect to a species as the quotient of the unique set of strongly ambiguous synonyms over the set of cross-species ambiguous synonyms. We execute all of these procedures for case-sensitive and case-insensitive name matching, separately. The pseudo-code to find the cross-species ambiguous synonyms and the strongly ambiguous synonyms is shown in Algorithm 2.

Algorithm 2: FINDING CROSS SPECIES AMBIGUITY

```
1  Extract the unique synonyms for first species and store them in set UniqueS1
2  Extract the unique ambiguous synonyms for first species and store them in set AmbigS1
3  Extract the unique synonyms for second species and store them in set UniqueS2
4  Extract the unique ambiguous synonyms for second species and store them in set AmbigS2
5  Sort UniqueS2
6  for each synonym S in UniqueS1: do
7      Apply binary search for (S) in (UniqueS2)
8      if S is found in UniqueS2 then
9          add S to CrossSpeciesSet crossSpecisAmbi++;
10 for each synonym S in CrossSpeciesSet: do
11     if S is found in AmbigS1 then
12         strongAmbigInS1++
13     if S is found in AmbigS2 then
14         strongAmbigInS2++
```

The degree of cross-species ambiguity ($CDA_{S1,S2}^{S1}$) between the set of synonyms of the first species S1, and the set of synonyms of the second species S2, relative to S1 is defined in Equation 3.5 below.

$$CDA_{S1,S2}^{S1} = \frac{\#\ Cross\ Species\ Match\ Synonyms}{\#\ of\ Unique\ Synonyms\ in\ S1}\ X\ 100\% \tag{3.6}$$

We also denote the degree of strong ambiguity between the set of synonyms of the first species S1, and the set of synonyms of the second species S2, relative to S1 as follows:

$$SDA_{S1,S2}^{S1} = \frac{\#\ Strong\ Ambig\ Synonyms\ in\ S1}{\#\ Cross\ Species\ Matched\ Synonyms}\ X\ 100\% \tag{3.7}$$

Chapter IV

Experimental Results and Discussion

We analyze and compare the gene and protein nomenclatures for six different species grouped into three pairs of species, where each pair represents two species that belong to the same kingdoms. These are: *Kingdom Animalia:* human (*Homo Sapiens*) and mouse (*Mus Musculus*), *Kingdom Plantae*: arabidopsis (*Arabidopsis Thaliana*) and rice (*Oryza Sativa*), and *Kingdom Bacteria*: *Pseudomonas Fluorescens* and *Bacillus Subtilis*.

For each of these species, we use different components of the process to generate relevant results:

1. General Analysis: this component provides three types of results. The first is related to the distribution of the records and synonyms of the species. The second describes the distribution of different synonyms types (*Official Full Names*, *Official Symbols*, and *Aliases*) for the species per data source. Third, analysis on how official names and official symbols are distributed over records is provided. Some records use both official names and symbols, others contain at least one of them and some contains no official synonyms or no aliases.
2. Frequency Analyser: provides frequency distributions for different synonyms types. For all species, the frequency tables and curves are available in the appendix.
3. Intra-species Ambiguity: provides analysis on the distribution of ambiguous synonyms

and the degrees of ambiguity within the species. The ambiguity tables are available in the appendix.

4. Ambiguity Between Data Sources: provides analysis on the percentage of synonyms, for a given species, that are ambiguous between the two databases, with respect to each database.
5. Cross-Species Ambiguity: For each species, we compute the degrees of ambiguity between this species and the remaining five species, as well as the percentage of strongly ambiguous synonyms. This is computed considering the case sensitive/insensitive matching.

IV.1 General Analysis

IV.1.1 Data Records Distribution

The distributions of data records and synonyms for each of the six species in Entrez Gene and UniProtKB databases are summarized below in Table 7 and Table 8, respectively. Each table shows the species, number of records in the resource that corresponds to this species, the percentage of this species records in the resource, the number of found synonyms in the records of this species in the resource, and the percentage that the synonyms of this species represent, out of the total number of synonyms in the resources.

Table 7. Data records and synonyms distribution for the six species in Entrez Gene DB.

Species	# records	% records in	# synonyms	% synonyms
Homo Sapiens	43894	1.30%	396029	6.57%
Mus Musculus	58084	1.71%	334670	5.55%
Arabidopsis Thaliana	24992	0.74%	93846	1.56%
Oryza Sativa	30621	0.90%	33265	0.55%
Pseudomonas Fluorescens	5525	0.16%	5636	0.09%
Bacillus Subtilis	22070	0.65%	22080	0.37%

Table 8. Data records and synonyms distribution for the six species in UniProtKB DB.

Species	# records	% records in	# synonyms	% synonyms
Homo Sapiens	26836	0.59%	77451	0.72%
Mus Musculus	18280	0.40%	54709	0.51%
Arabidopsis Thaliana	12588	0.28%	34117	0.32%
Oryza Sativa	6356	0.14%	15319	0.14%
Pseudomonas Fluorescens	8224	0.18%	21217	0.20%
Bacillus Subtilis	10408	0.23%	23861	0.22%

IV.1.2 Synonym Types Distribution

The distribution of different synonyms types for the six species in Entrez Gene and UniProtKB are shown in Fig. 2 and Fig. 3 respectively. In Entrez Gene database, we can observe that only *Homo Sapiens* and *Mus Musculus* have synonyms tagged as official full names and official symbols, while other species do not, and for which we consider synonyms as aliases. However, the percentage of official synonyms is only about 17% for *Homo Sapiens* and about 33% for *Mus Musculus*.

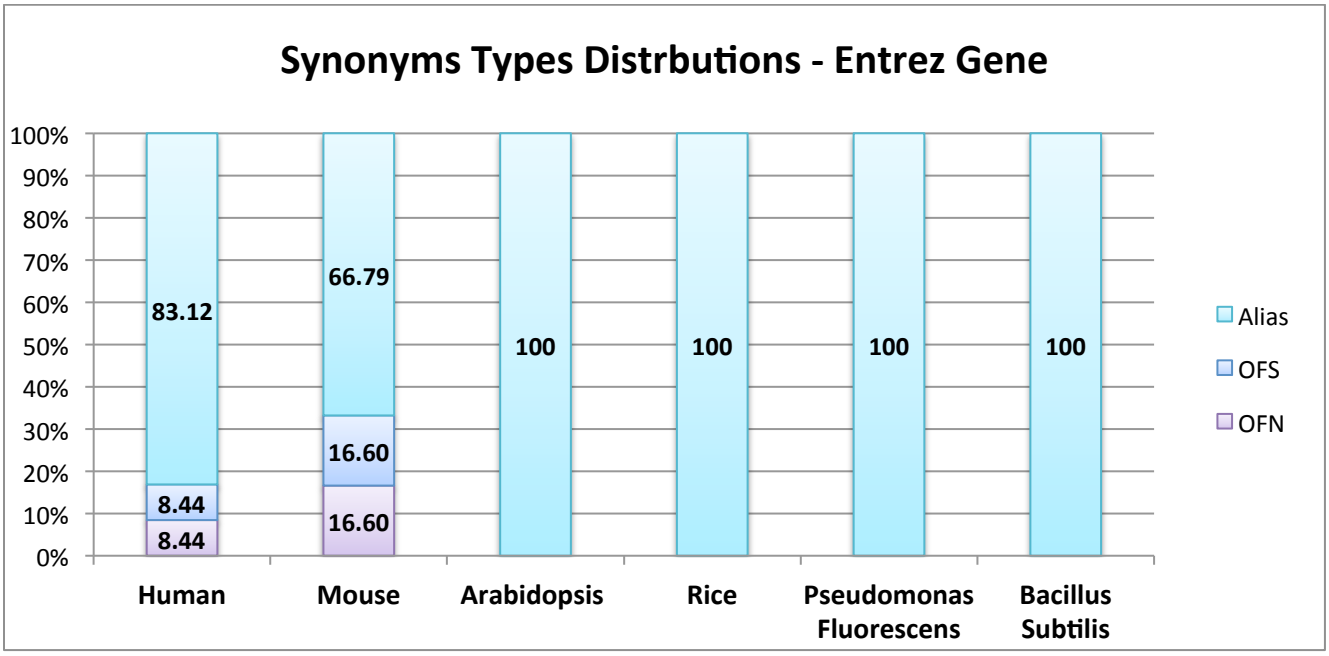

Figure 2 Synonyms types distributions in Entrez Gene DB.

In UniProtKB on the other hand, all species have official synonyms. We notice that the official synonyms are distributed at relatively close ratios for all species; records with the official full names make 33.41% to 43.62% of all records in a database, and records with the official symbols range from 7.57% to 9.23%.

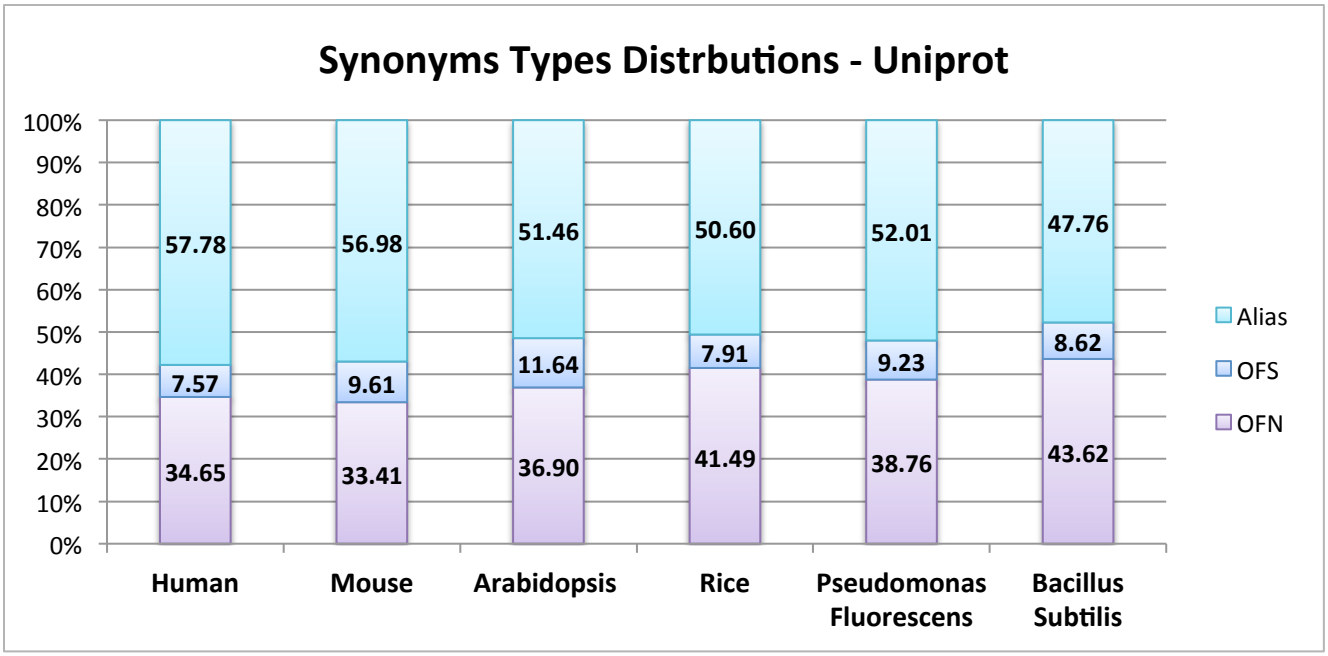

Figure 3. Synonyms types distributions in UniProtKB.

IV.1.3 Official Synonyms Distribution

The distributions of synonyms over data records are shown in Table 9 and Table 10 below. The tables show the species, the percentage of records for the species that have both an official name synonym and an official symbol synonym, the percentage of records for the species that have any official name synonym or official symbol synonym, the percentage of records for the species that have no aliases, and the percentage of records for the species that have no official name synonym nor official symbol synonym. In Entrez Gene database, we can observe that out of all *Homo Sapiens* records, 76.16% have both an official name synonym and an official symbol synonym, and 76.16% have at least an official synonym. This indicates that each record that has an official name also has an official symbol and vice versa. This is also true for *Mus Musculus*; most of the records, about 95.67%, have both an official

name and an official symbol. Since the other species do not have synonyms tagged as official full names or official symbols, the records that have only aliases represent 100%. On the other hand, Each record in UniProtKB has at least an official synonym, either an official full name or an official symbol. In addition, for all species, about 16% to 26% of the records have both official full names and official symbols at the same time.

Table 9. Data records and synonyms distribution for the six species in Entrez Gene DB.

Species	Both Officials	Either Officials	No Aliases	Only Aliases
Homo Sapiens	76.16%	76.16%	0%	23.84%
Mus Musculus	95.67%	95.67%	0%	4.33%
Arabidopsis Thaliana	0%	0%	0%	100%
Oryza Sativa	0%	0%	0%	100%
Pseudomonas Fluorescens	0%	0%	0%	100%
Bacillus Subtilis	0%	0%	0%	100%

Table 10. Data records and synonyms distribution for the six species in UniProtKB DB.

Species	Both Officials	Either Officials	No Aliases	Only Aliases
Homo Sapiens	16.53%	100%	35.12%	0%
Mus Musculus	21.58%	100%	33.70%	0%
Arabidopsis Thaliana	26.23%	100%	36.65%	0%
Oryza Sativa	17.35%	100%	30.71%	0%
Pseudomonas Fluorescens	18.71%	100%	28.31%	0%
Bacillus Subtilis	15.88%	100%	41.86%	0%

IV.2 Frequency

From the frequency analysis, we find that *Homo Sapiens* and *Mus Musculus* records in Entrez Gene can have at most one official name, and at most one official symbol. The other species do not have synonyms tagged as official full name or official symbols in Entrez Gene. In UniProtKB, however, all records have only one official name, but it is possible to have more than one official symbol.

Table 11 below displays the highest frequency values for the six species. Recall that if (i) represents the number of synonyms that an entity has, $Freq_i\%$ represents the percentage of data records that has (i) synonyms. The table shows that 27.6% of the *Mus Musculus* records in Entrez Gene have 3 synonyms, which is the highest frequency for this species. For example, the record (Entrez Gene ID: 101154638) has the synonyms: 'predicted gene 5121', 'Gm5121' and 'EG330948'. In UniProtKB, however, 27.07% of the *Mus Musculus* records have a single synonym. The frequency diagrams for the six species, considering the number of synonym from 1 to 5, are available in the appendix of this document (Figures 16 - 27).

Table 11. Highest frequency values for the six species in both data sources.

Species	Highest frequency in Enrez Gene		Highest frequency in UniprotKB	
	i	**$Freq_i\%$**	**i**	**$Freq_i\%$**
Homo Sapiens	1	16.21	1	31.34
Mus Musculus	3	27.6	1	27.07
Arabidopsis Thaliana	2	48.34	1	29.06
Oryza Sativa	1	99.32	2	39.99
Pseudomonas Fluorescens	1	98.26	2	27.95
Bacillus Subtilis	1	99.95	1	39.86

IV.3 Homo Sapiens

IV.3.1 Intra-species Ambiguity

The ambiguity distributions for the different *Homo Sapiens* synonym types (All synonyms, *Official Full Names*, *Official Symbols* and *Aliases*) in Entrez Gene and UniprotKB databases are shown in Fig. 4 and Fig. 5 respectively. In both figures, we demonstrate the ambiguity at levels 2, 3, 4, 5, and 6. For example, the first column in Fig. 4 indicates that about 8% of the synonyms in Entrez Gene are ambiguous, as they are shared between two gene names; all of them are aliases. An example to that is the gene alias "NAT-1" which is shared/ambiguous between the two genes (Entrez Gene IDs: 9 and 145389).

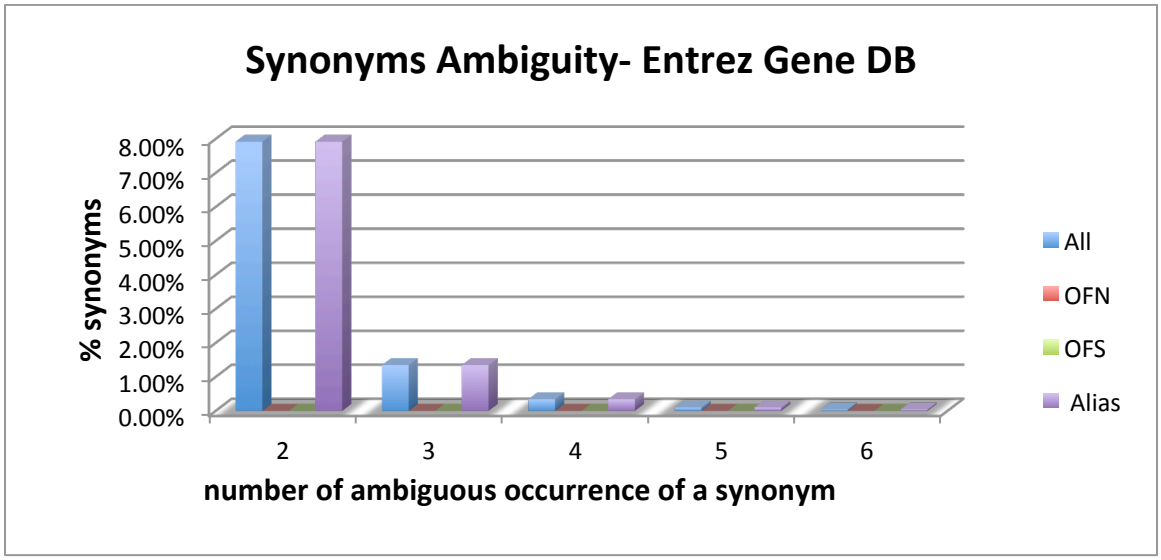

Figure 4. Synonyms ambiguity in Entrez Gene for *Homo Sapiens*.

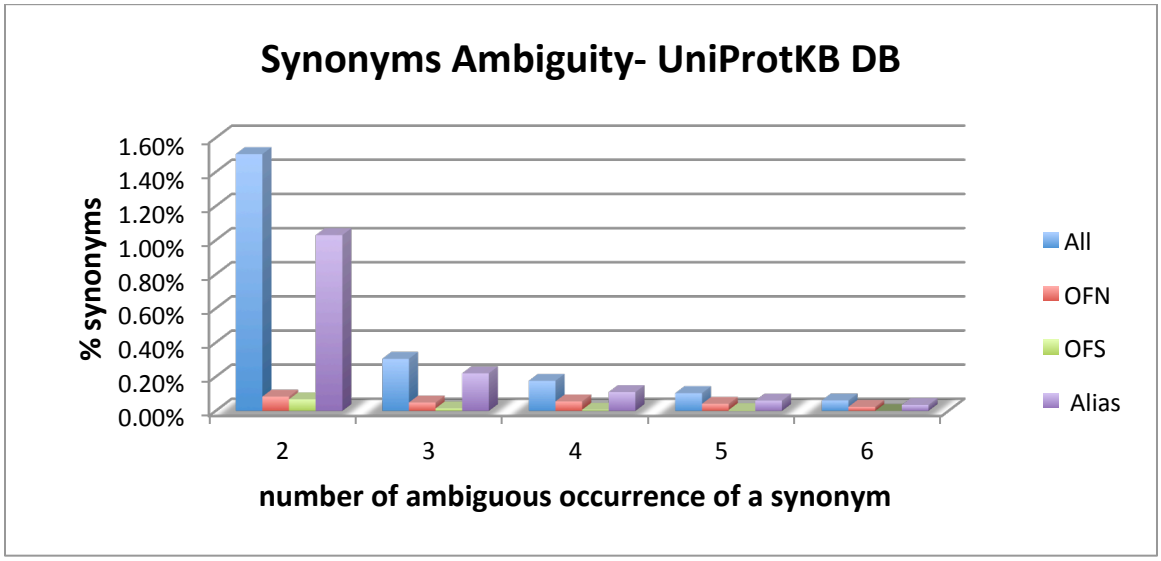

Figure 5. Synonyms ambiguity in UniProtKB for *Homo Sapiens*.

IV.3.2 Ambiguity Between Databases

Table 12 below shows that the ambiguity with reference to UniProtKB considering all synonyms is about five times the ambiguity with reference to Entrez Gene. The large number of aliases and their ambiguities, in both data sources, influences these percentages.

Table 12. Ambiguity between databases for *Homo Sapiens*.

Database	All Synonyms	OFN	OFS	Aliases
Entrez Gene	3.83%	7.16%	1.06%	2.32%
UniProtKB	18.29%	11.75%	6.25%	16.63%

IV.3.3 Cross-species Ambiguity

The cross-species ambiguities between *Homo Sapiens* synonyms and all the other species in Entrez Gene and UniProtKB are shown in Table 13 and Table 14 below, respectively. The tables show the species, the cross-species degree of ambiguity (CDA) and the strong degree of ambiguity SDA with considering the case insensitive and case sensitive matching of the

synonyms.. The notation we use in the SDA column is: A/B, where A represents the SDA (the proportion of strongly ambiguous synonyms out of the cross-species matched synonyms) and B represents the proportion of the strongly ambiguous synonyms, with respect to the set of all unique synonyms. The highest degree of ambiguity (CDA) is with *Mus Musculus*, as reported in the literature [26]. However, all other species show higher strong degrees of ambiguity (SDA).

Table 13. Cross-species ambiguity in Entrez Gene for *Homo Sapiens*.

***Homo Sapiens* Cross-species – Entrez Gene**				
Species	**Case insensitive matching**		**Case sensitive matching**	
	CDA%	**SDA%**	**CDA%**	**SDA%**
Mus musculus	16.30	40.19 / 6.55	7.73	8.26 / 0.64
Arabidopsis Thaliana	0.84	49.55 / 0.42	0.78	50.63 / 0.39
Oryza Sativa	0.03	93.00 / 0.03	0.02	98.08 / 0.02
Pseudomonas Fluorescens	0.12	49.87 / 0.06	0.0019	16.67 / 3E-4
Bacillus subtilis	0.12	44.47 / 0.05	0.00098	0

Table 14. Cross-species ambiguity in UniProtKB for *Homo Sapiens*.

***Homo Sapiens* Cross-species – UniProtKB**				
Species	**Case insensitive matching**		**Case sensitive matching**	
	CDA%	**SDA%**	**CDA%**	**SDA%**
Mus musculus	64.50	3.46 / 0.23	64.35	3.36 / 2.16
Arabidopsis Thaliana	3.47	21.90 / 0.76	3.43	22.05 / 0.76
Oryza Sativa	1.59	36.27 / 0.58	1.58	36.45 / 0.58
Pseudomonas Fluorescens	0.97	47.64 / 0.46	0.96	47.55 / 0.46
Bacillus subtilis	1.31	45.92 / 0.60	1.30	45.71 / 0.59

IV.4 Mus Musculus

IV.4.1 Intra-species Ambiguity

Figures 6, 7 show the ambiguity distributions for levels from 2 to 6 in Entrez Gene and UniProtKB databases, respectively.

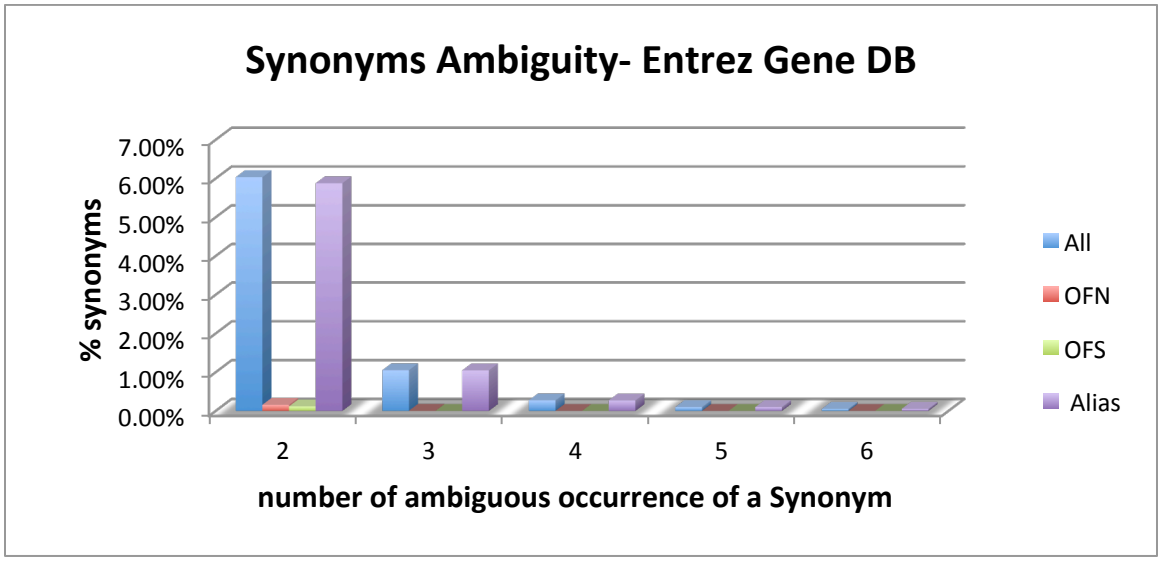

Figure 6. Synonyms ambiguity in Entrez Gene for *Mus Musculus*.

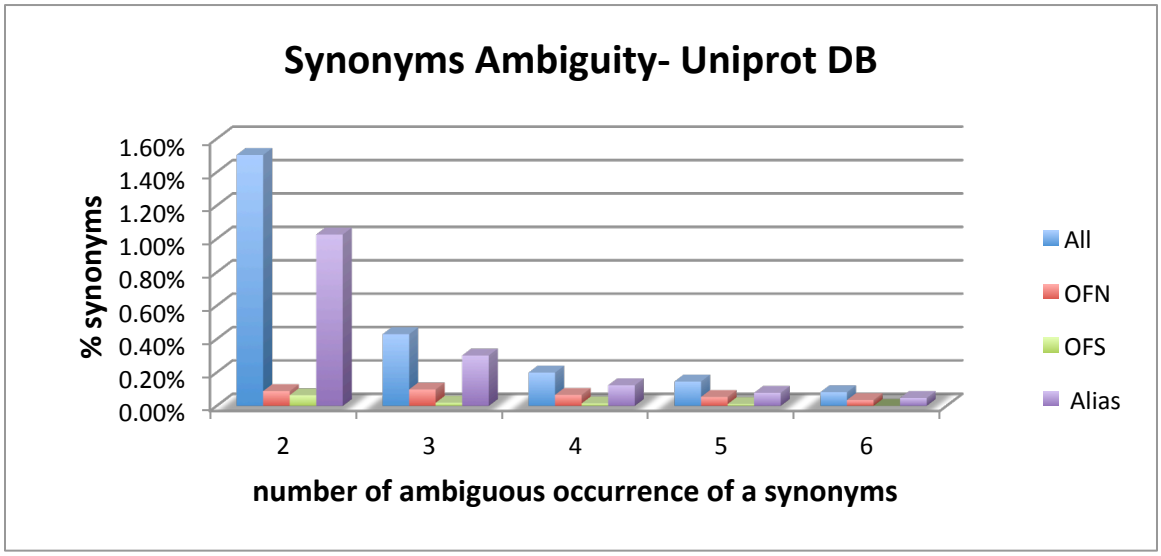

Figure 7. Synonyms ambiguity in UniProtKB for *Mus Musculus*.

IV.4.2 Ambiguity Between Databases

Table 15 below shows that the ambiguity with reference to UniProtKB considering all synonyms is about five times the ambiguity with reference to Entrez Gene.

Table 15. Ambiguity between databases for *Mus Musculus*.

Database	All Synonyms	OFN	OFS	Aliases
Entrez Gene	3.27%	3.69%	0.66%	1.85%
UniProtKB	15.99%	12.17%	7.06%	12.56%

IV.4.3 Cross-species Ambiguity

The cross-species ambiguities between *Mus Musculus* synonyms and all the other species in Entrez Gene and UniProtKB are shown in Tables 16 and 17 below. The highest degree of ambiguity (CDA) happens with the *Homo Sapiens*. With other species, however, the CDAs are relatively low.

Table 16. Cross-species ambiguity in Entrez Gene for *Mus Musculus*.

Mus Musculus Cross-species – Entrez Gene				
Species	Case insensitive matching		Case sensitive matching	
	CDA%	SDA%	CDA%	SDA%
Homo Sapiens	20.78	39.51 / 8.21	9.86	7.09 / 0.70
Arabidopsis Thaliana	0.96	52.26 / 0.50	0.24	11.61 / 0.03
Oryza Sativa	0.04	43.75 / 0.02	0.02	34.69 / 0.01
Pseudomonas Fluorescens	0.12	49.32 / 0.06	0.01	12.50 / 8E-4
Bacillus subtilis	0.13	48.00 / 0.06	0.01	27.27 / 2E-3

Table 17. Cross-species ambiguity in UniProtKB for *Mus Musculus*.

Mus Musculus Cross-species – UniprotKB				
Species	Case insensitive matching		Case sensitive matching	
	CDA%	SDA%	CDA%	SDA%
Homo Sapiens	84.24	3.14 / 2.65	84.04	3.04 / 2.56
Arabidopsis Thaliana	4.36	21.69 / 0.94	4.30	21.89 / 0.94
Oryza Sativa	2.02	34.83 / 0.7	2.00	34.97 / 0.7
Pseudomonas Fluorescens	1.22	46.88 / 0.57	1.21	46.95 / 0.57
Bacillus subtilis	1.67	45.44 / 0.76	1.66	45.41 / 0.75

IV.5 Arabidopsis Thaliana

IV.5.1 Intra-species Ambiguity

Figures 8, 9 show the ambiguity distributions for levels from 2 to 6 in Entrez Gene and UniProtKB databases, respectively.

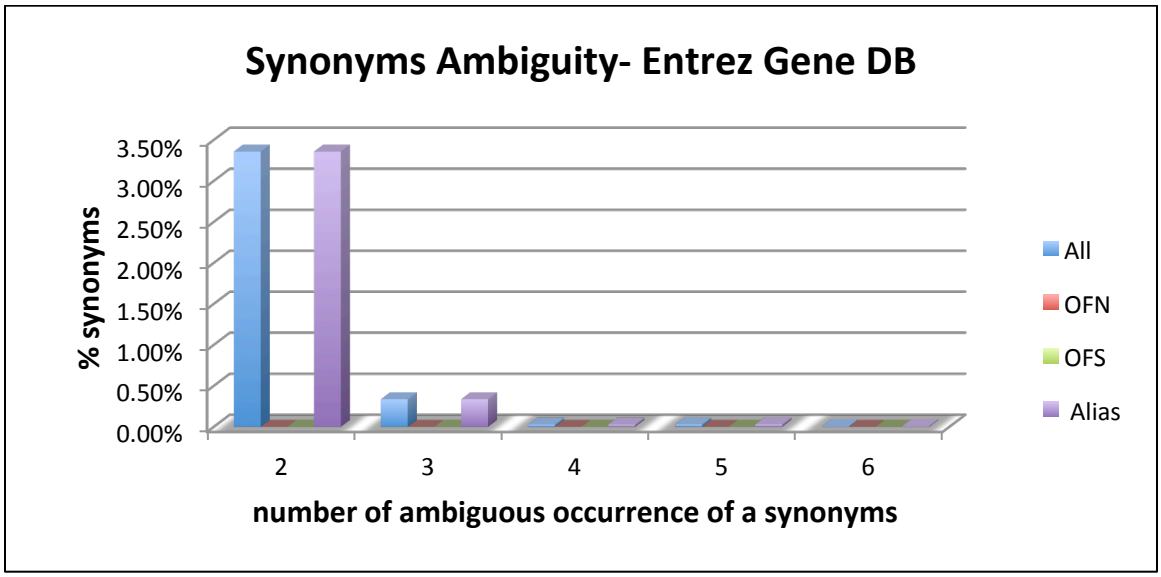

Figure 8. Synonyms ambiguity in Entrez Gene for *Arabidopsis Thaliana*.

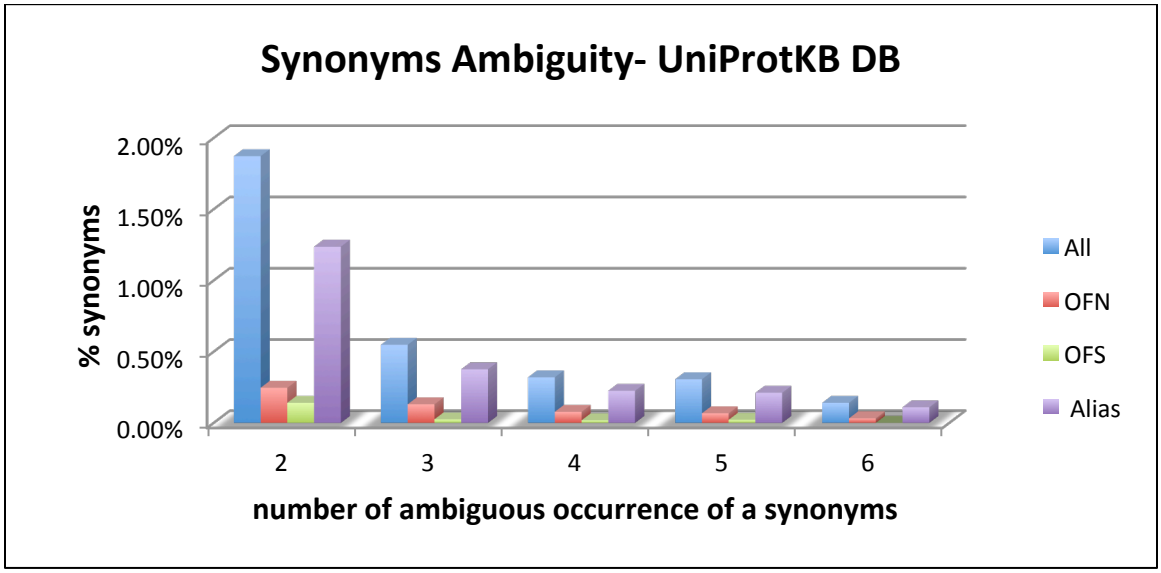

Figure 9. Synonyms ambiguity in UniProtKB DB for *Arabidopsis Thaliana*.

IV.5.2 Ambiguity Between Databases

Table 18 below shows that the ambiguity with reference to UniProtKB considering all synonyms is about three times the ambiguity with reference to Entrez Gene. Since there are no official names and symbols in Entrez Gene, their degree of ambiguity is zero. A note that Aliases ambiguity is different from All synonyms ambiguity. This is because in the aliases fields we only compare those alias synonyms from Entrez Gene against aliases in UniProtKB, while for all synonyms, we compare all the synonyms, regardless of the synonym type, and hence aliases might be ambiguous with official symbols, or with official full names.

Table 18. Ambiguity between databases for *Arabidopsis Thaliana*.

Database	All Synonyms	OFN	OFS	Aliases
Entrez Gene	4.69%	0%	0%	2.17%
UniProtKB	12.76%	0%	0%	13.09%

IV.5.3 Cross-species Ambiguity

The cross species ambiguities between *Arabidopsis Thaliana* synonyms and all the other species in Entrez Gene and UniProtKB are shown in Tables 19 and 20 below. The highest degrees of ambiguity (CDA) in Entrez Gene for *Arabidopsis Thaliana* occur with *Mus Musculus* and *Homo Sapiens*. These also cause the highest strong degree of ambiguity. In UniProtKB, however, rice causes the highest degree of ambiguity, but the least strong degree of ambiguity.

Table 19. Cross-species ambiguity in Entrez Gene for *Arabidopsis Thaliana*.

***Arabidopsis Thaliana* Cross-species – Entrez Gene**				
Species	**Case insensitive matching**		**Case sensitive matching**	
	CDA%	**SDA%**	**CDA%**	**SDA%**
Mus Musculus	2.94	58.14 / 1.71	0.75	46.97 / 0.35
Homo Sapiens	3.29	58.96 / 1.94	3.06	58.78 / 1.8
Oryza Sativa	0.22	35.71 / 0.08	0.21	32.72 / 0.07
Pseudomonas Fluorescens	0.17	46.27 / 0.08	0.04	25.00 / 0.01
Bacillus subtilis	0.24	39.25 / 0.09	0.07	5.17 / 3.8E-3

Table 20. Cross-apecies ambiguity in UniProtKB for *Arabidopsis Thaliana.*

***Arabidopsis Thaliana* Cross-species – UniProtKB**				
Species	**Case insensitive matching**		**Case sensitive matching**	
	CDA%	**SDA%**	**CDA%**	**SDA%**
Mus Musculus	7.45	24.28 / 1.81	7.35	24.43 / 1.79
Homo Sapiens	7.76	24.03 / 1.86	7.66	24.21 / 1.85
Oryza Sativa	11.64	16.63 / 1.94	11.58	16.71 / 1.94
Pseudomonas Fluorescens	2.04	57.76 / 1.18	2.03	57.96 / 1.18
Bacillus subtilis	2.59	55.63 / 1.44	2.57	55.93 / 1.44

IV.6 Oryza Sativa

IV.6.1 Intra-species Ambiguity

Figures 10, 11 show the ambiguity distributions for levels from 2 to 6 in Entrez Gene and UniProtKB databases, respectively.

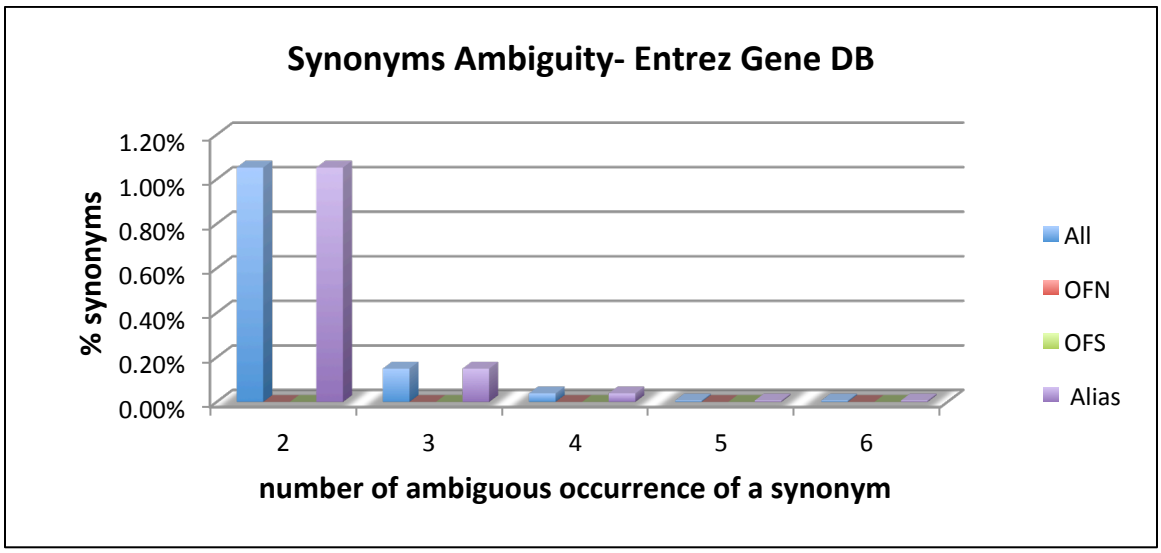

Figure 10. Synonyms ambiguity in Entrez Gene for *Oryza Sativa*.

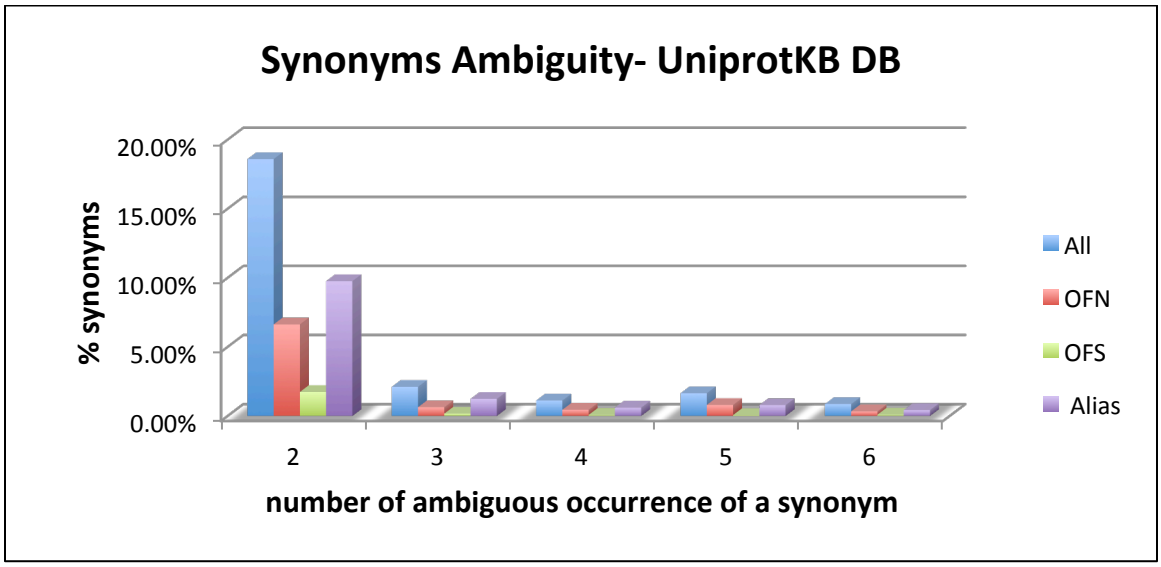

Figure 11. Synonyms ambiguity in UniProtKB for *Oryza Sativa*.

IV.6.2 Ambiguity Between Databases

Table 21 shows that the in-between databases ambiguity with reference to UniProtKB considering all synonyms is about four times the ambiguity with reference to Entrez Gene. However, this degree is negligible and is too low to be considered.

Table 21. Ambiguity in between databases for *Oryza Sativa*.

Database	All Synonyms	OFN	OFS	Aliases
Entrez Gene	0.013%	0%	0%	0.013%
UniProtKB	0.051%	0%	0%	0.12%

IV.6.3 Cross-species Ambiguity

The cross-species ambiguities between rice and all the other species in Entrez Gene and UniProtKB are shown in Tables 22 and 23 below. The highest degree of ambiguity (CDA) happens with the *Arabidopsis Thaliana* in both databases.

Table 22. Cross-species ambiguity in Entrez Gene for *Oryza Sativa*.

***Oryza Sativa* Cross-species – Entrez Gene**				
Species	**Case insensitive matching**		**Case sensitive matching**	
	CDA%	**SDA%**	**CDA%**	**SDA%**
Mus Musculus	0.31	68.75 / 0.21	0.16	71.43 / 0.11
Homo Sapiens	0.32	69.00 / 0.22	0.17	69.23 / 0.11
Arabidopsis Thaliana	0.53	73.81 / 0.39	0.52	72.84 / 0.38
Pseudomonas Fluorescens	0.10	50.00 / 0.05	0.10	50.0 / 0.05
Bacillus subtilis	0.19	74.58 / 0.14	0.19	74.58 / 0.14

Table 23. Cross-species ambiguity in UniProtKB for *Oryza Sativa*.

***Oryza Sativa* Cross-species – UniProtKB**				
Species	**Case insensitive matching**		**Case sensitive matching**	
	CDA%	**SDA%**	**CDA%**	**SDA%**
Mus Musculus	12.61	53.77 / 6.78	12.52	54.05 / 6.77
Homo Sapiens	12.95	53.82 / 6.97	12.89	54.08 / 6.97
Arabidopsis Thaliana	42.54	35.94 / 15.29	42.34	35.81 / 15.16
Pseudomonas Fluorescens	5.29	81.31 / 4.30	5.26	81.22 / 4.27
Bacillus subtilis	6.37	79.03 / 5.03	6.33	78.90 / 4.99

IV.7 Pseudomonas Fluorescens

IV.7.1 Intra-species Ambiguity

Figures 12, 13 show the ambiguity distributions for levels from 2 to 6 in Entrez Gene and UniProtKB databases, respectively.

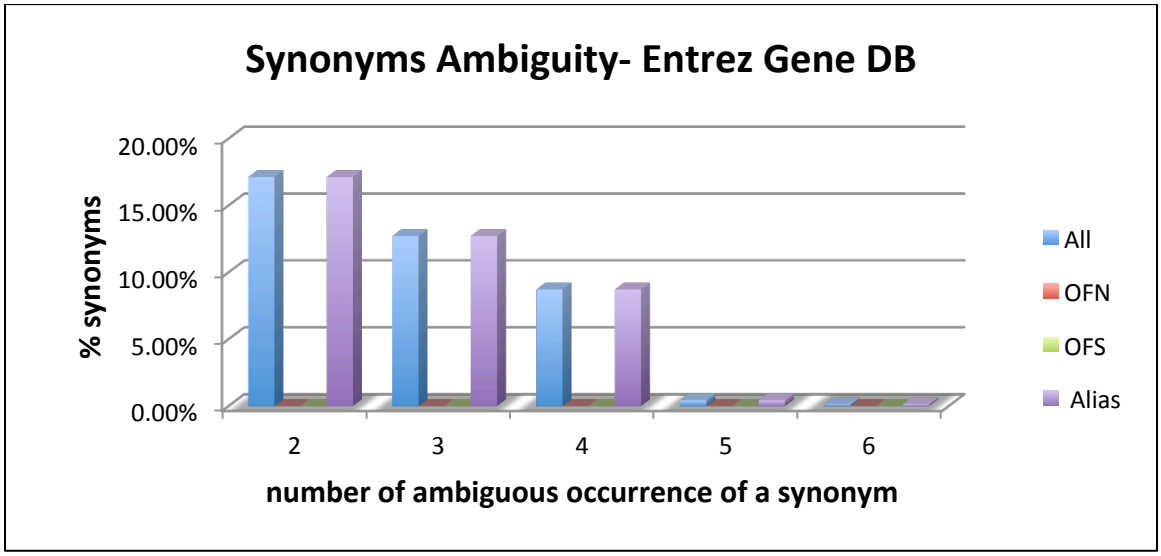

Figure 12. Synonyms ambiguity in Entrez Gene for *Pseudomonas Fluorescens*.

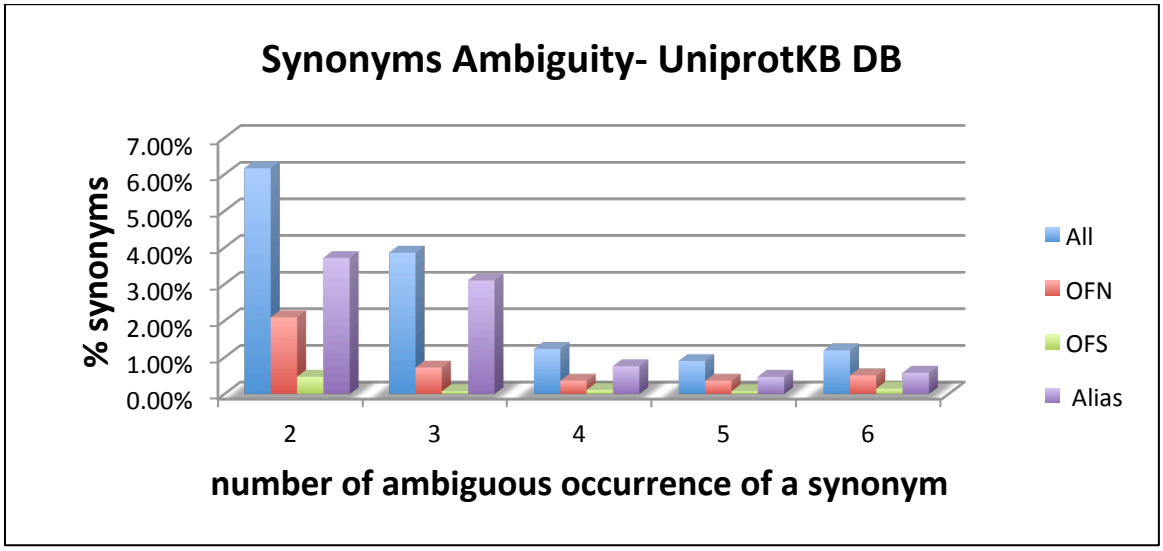

Figure 13. Synonyms ambiguity in UniprotKB for *Pseudomonas Fluorescens*.

IV.7.2 Ambiguity Between Databases

Table 24 below shows that the ambiguity with reference to UniProtKB considering all synonyms is almost the same the ambiguity with reference to Entrez Gene.

Table 24. Ambiguity between databases for *Pseudomonas Fluorescens*.

Database	All Synonyms	OFN	OFS	Aliases
Entrez Gene	0.49%	0%	0%	0.09%
UniProtKB	0.58%	0%	0%	0.22%

IV.7.3 Cross-species Ambiguity

The cross-species ambiguities between *Pseudomonas Fluorescens* synonyms and all the other species in Entrez Gene and UniProtKB are shown in Tables 25 and 26 below. The highest cross-species degree of ambiguity (CDA) is with the other bacteria, *Bacillus Subtilis,* in both databases.

Table 25. Cross-species ambiguity in Entrez Gene for *Pseudomonas Fluorescens*.

Pseudomonas Fluorescens Cross Species – Entrez Gene				
Species	**Case insensitive matching**		**Case sensitive matching**	
	CDA%	**SDA%**	**CDA%**	**SDA%**
Mus Musculus	8.96	53.74 / 4.81	0.49	68.75 / 0.34
Homo Sapiens	11.43	53.33 / 6.09	0.18	33.33 / 0.06
Arabidopsis Thaliana	4.08	61.19 / 2.5	0.98	34.38 / 0.34
Oryza Sativa	0.98	40.63 / 0.40	0.98	40.63 / 0.40
Bacillus subtilis	30.56	70.99 / 21.69	30.50	70.93 / 21.63

Table 26. Cross-species ambiguity in UniProtKB for *Pseudomonas Fluorescens*.

Pseudomonas Fluorescens Cross Species – UniProt				
Species	**Case insensitive matching**		**Case sensitive matching**	
	CDA%	**SDA%**	**CDA%**	**SDA%**
Mus Musculus	21.38	92.24 / 19.72	21.27	92.20 / 19.61
Homo Sapiens	22.17	91.54 / 20.30	22.06	91.50 / 20.19
Arabidopsis Thaliana	20.91	96.55 / 20.19	20.84	96.54 / 20.12
Oryza Sativa	14.85	96.60 / 14.35	14.78	96.59 / 14.28
Bacillus subtilis	51.55	94.62 / 48.77	51.44	94.60 / 48.67

IV.8 Bacillus Subtilis Ambiguity

IV.8.1 Intra-species Ambiguity

Figures 14, 15 show the ambiguity distributions for levels from 2 to 6 in Entrez Gene and UniProtKB databases, respectively.

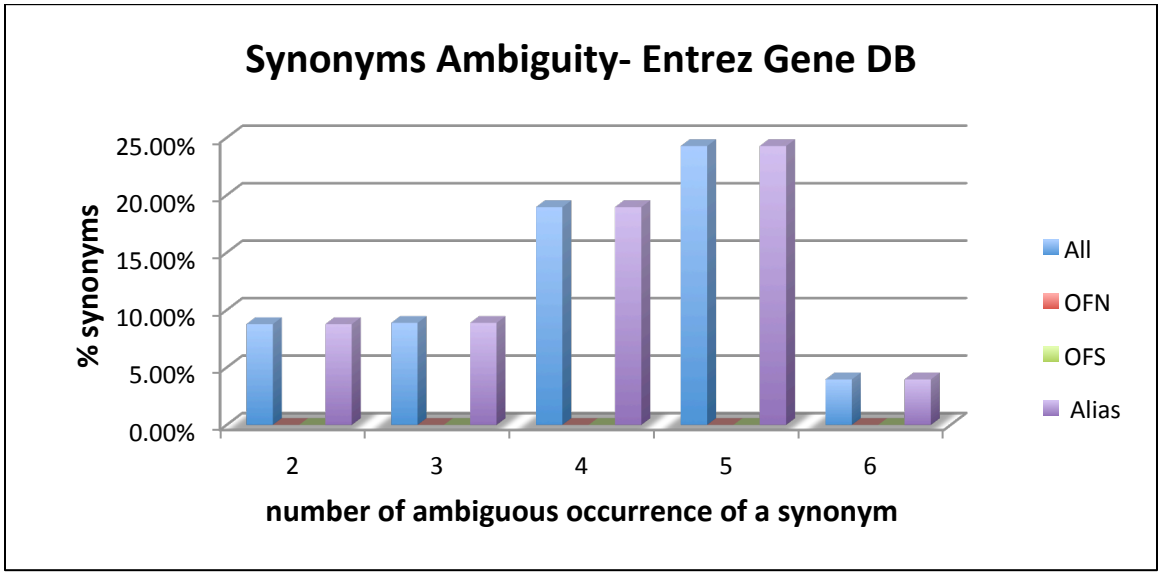

Figure 14. Synonyms ambiguity in Entrez Gene for *Bacillus Subtilis*.

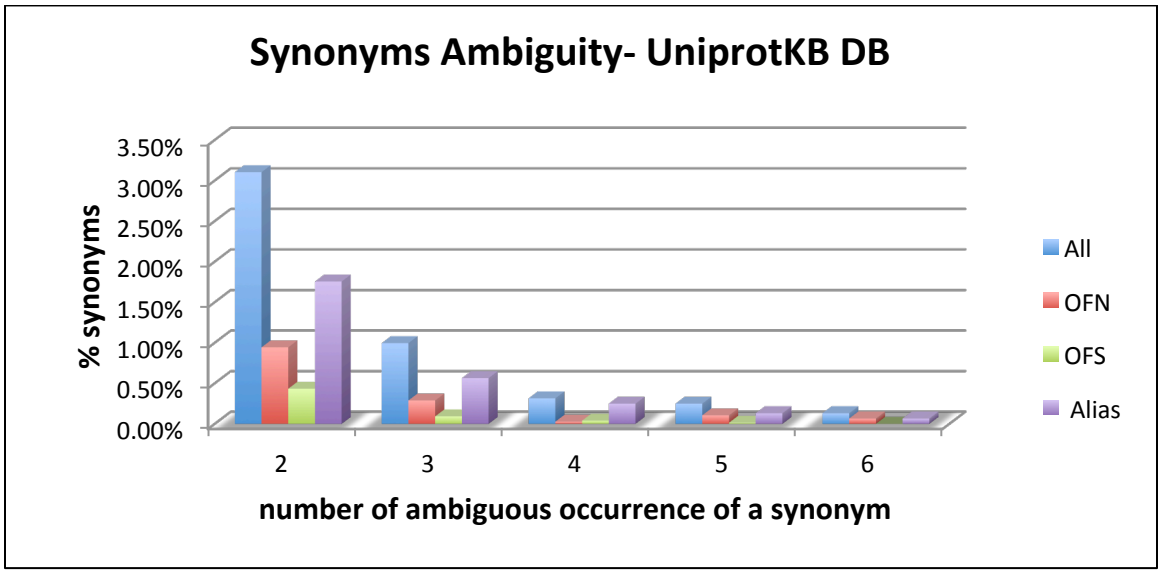

Figure 15. Synonyms ambiguity in UniProtKB for *Bacillus Subtilis*.

IV.8.2 Ambiguity Between Databases

Table 27 below shows that the ambiguity with reference to UniProtKB considering all synonyms is close to the ambiguity with reference to Entrez Gene.

Table 27. Ambiguity between databases for *Bacillus Subtilis*.

Database	All Synonyms	OFN	OFS	Aliases
Entrez Gene	0.43%	0%	0%	0.13%
UniProt	0.31%	0%	0%	0.27%

IV.8.3 Cross-species Ambiguity

The cross species ambiguities between *Bacillus Subtilis* synonyms and all the other species in Entrez Gene and UniProtKB are shown in Tables 28 and 29 below. The highest degrees of ambiguity and strong ambiguity happen with the *Pseudomonas Fluorescens* species. This is the case in both databases.

Table 28. Cross-species ambiguity in Entrez Gene for *Bacillus Subtilis*.

Bacillus Subtilis Cross Species – Entrez Gene				
Species	Case insensitive matching		Case sensitive matching	
	CDA%	SDA%	CDA%	SDA%
Mus Musculus	5.37	73.00 / 3.92	0.39	45.45 / 0.18
Homo Sapiens	6.64	74.12 / 4.92	0.05	66.67 / 0.04
Arabidopsis Thaliana	3.33	53.23 / 1.77	1.04	24.14 / 0.25
Oryza Sativa	1.06	25.42 / 0.27	1.06	25.42 / 0.27
Pseudomonas Fluorescens	17.95	84.15 / 15.11	17.92	84.12 / 15.07

Table 29. Cross-species ambiguity in UniProtKB for *Bacillus Subtilis*.

Bacillus Subtilis Cross Species – UniProt				
Species	Case insensitive matching		Case sensitive matching	
	CDA%	SDA%	CDA%	SDA%
Mus Musculus	10.63	63.92 / 6.79	10.55	64.02 / 6.75
Homo Sapiens	10.92	64.03 / 6.99	10.82	64.21 / 6.95
Arabidopsis Thaliana	9.65	75.03 / 7.24	9.59	75.17 / 7.21
Oryza Sativa	6.49	81.85 / 5.31	6.45	81.95 / 5.29
Pseudomonas Fluorescens	18.72	88.39 / 16.54	18.68	88.37 / 16.51

IV.9 Discussion

Recently, various studies have been conducted to explore the problems related to gene and protein names ambiguities in text. These studies aim to understand and address naming problems, and to propose potential solutions to them. In fact, the degrees of ambiguities in gene/protein names might vary for different species, leading to problems at different levels of importance and impacts. For example, some species might experience high degrees of ambiguities intra-species, between data sources, and cross-species. In such cases the standards and naming conventions used for these species need to be examined and reviewed, for example, as we found for *Bacillus Subtilis*. Naming in other species, however, might suffer from ambiguities, but at levels that are too low to represent a concern or worth spending times and efforts to address them. An example to that is *Oryza Sativa* in Entrez Gene database, as we found in this study. Hence, the process to analyse gene and protein nomenclatures is important, particularly in text mining, because it helps understanding the ambiguities problems, the sizes and the importance of these problems, and may guide solutions to reduce them.

By analysing the distributions of records and synonyms, we find that *Homo Sapiens* and *Mus Musculus* together represent a large percentage, about 12%, of the overall synonyms in Entrez Gene database. However, their data records only represent about 4% of the overall Entrez Gene records, which apparently indicates large synonyms to record ratio. This is also true for *Arabidopsis Thaliana*. The remaining three species have synonyms to record ratio close to 0.5. In UniProtKB, on the other hand, all species have synonyms to record ratio close to 1.

When we analysed the intra-species ambiguities in Entrez Gene database, we found that all species, except *Bacillus Subtilis*, have the highest degree of ambiguity at level 2. That is, most of the ambiguous synonyms are shared only between two Entrez Gene records. *Bacillus Subtilis* has its highest ambiguity degree at level 5. For *Homo Sapiens,* about 23.29% of the overall synonyms are ambiguous, and 8% of the synonyms are ambiguous at level 2. Rice (*Oryza Sativa*), however, has the least ambiguity degree in Entrez Gene; its ambiguous synonyms represent about 5.47% in total, and those ambiguous synonyms at level 2 represent only 1% of the total rice synonyms in Entrez Gene. In UniProtKB, on the other hand, the highest degree of ambiguity for all synonyms is at level 2, where for *Homo Sapiens,* 17.88% of the synonyms are ambiguous, and 1.4% of the synonyms are ambiguous at level 2. For *Oryza Sativa*, ambiguous synonyms represent about 49% of its total synonyms; and ambiguous synonyms at level 2 represent about 18%. *Mus Musculus* has the least ambiguity in UniProtKB at about 11%.

Beside intra-species ambiguities in the aliases, official full names and official symbols could be ambiguous as well. Some records in Entrez Gene refer to genes reported on different sequences or chromosomes, and have exactly the same official names and may have the same or different official symbols. For example, Entrez Gene records (Entrez Gene IDs: 112026, 492996, 100126086) in Table 30, agree on the official full names, but not on the official symbols.

Table 30. *Mus Musculus* Entrez Gene records with ambiguous official full names.

Entrez Gene ID	Official Full Name	Official Symbol
112026	liver weight QTL 1	Lwq1
492996	liver weight QTL 1	Lvrq1
100126086	liver weight QTL 1	Lvwtq1

Our analysis on the overlapping ambiguities between data sources show that *Homo Sapiens*, *Mus Musculus* and *Arabidopsis Thaliana* have much higher overlapping ambiguities over the other species. Another observation is that for all species except *Bacteria*, the matched synonyms between these databases represent percentages in UniProtKB that are about 2.7 to 6 times higher that the percentages they represent in Entrez Gene. *Bacteria*, however, has relatively close percentages of ambiguities, with respect to both databases. We report that rice (*Oryza Sativa*) has the least ambiguity between databases; about 0.013% in Entrez Gene and 0.051% in UniProtKB, which are too low to be considered as a concern.

Considering the cross-species ambiguity in Entrez Gene, we expected that the highest cross-species degree of ambiguity (CDA) would be between the species that belong to the same pair. However, we noticed that this is not always the case, as *Arabidopsis Thaliana* has higher CDAs with *Homo Sapiens* and *Mus Musculus* over the CDA with its pair, *Oryza Sativa.* We also found that ambiguities between *Homo Sapiens* and *Mus Musculus*, and between *Pseudomonas Fluorescens and Bacillus Subtilis* are much higher than the ambiguity between *Arabidopsis Thaliana* and *Oryza Sativa*. In fact, we observed that rice causes the least cross-species ambiguity with all the other species, except with Arabidopsis. Another interesting observation is that many cross-species ambiguous synonyms are already ambiguous within the species (have strong degree of ambiguity). Overall, the SDAs range from 40% to 90%.

When we studied the effects of case sensitive matching of names, we found that case sensitivity had minimal impacts on ambiguities between species that belong to the same pair, and higher effects on ambiguities across-species from different pairs. *Homo Sapiens* and *Mus*

Musculus make an exception to this, as their CDAs are also reduced by considering case sensitivity. In addition, we found that rice's cross-species ambiguity is affected the least by case sensitive matching; the CDAs are almost the same across all other species, except with human and mouse. On the other hand, Bacteria CDAs are the most affected and reduced by case sensitive matching, as their CDAs with the other species become negligible.

The cross-species analysis in UniProtKB shows that species in the same pair have always the highest CDAs when compared to others. In general, cross-species ambiguities in UniProtKB are noticably higher than those in Entrez Gene. We also note that Bacteria have the highest strong degrees of ambiguity. Additionally, it is interesting to report that case sensitivity has negligible effect on the CDA or SDA between any two species in UniProtKB.

Based on our observations, cross-species ambiguity is highly influenced by the intra-species ambiguity; in most cases, at least 40% of the cross-species ambiguous synonyms are already ambiguous within the species. Therefore, we interpret that resolving the intra-species ambiguity, would largely help in reducing that across species. One solution we propose is to maintain a database of the ambiguous synonyms per species, and to encourage researchers to avoid using these synonyms, whenever possible. Another important observation is that *Oryza Sativa* has the least ambiguity in Entrez Gene database, very low overlapping ambiguity between databases, and causes the least degrees of ambiguity across species. Therefore, it might not be worth the efforts to study the source of these ambiguities, as they might have very little impacts. In addition, we can use *Oryza Sativa* nomenclature as a guideline model for naming other species genes.

Chapter V

Conclusion and Future Work

In this study, we developed an automated process to analyse gene and protein nomenclatures of different species, based on the records available in public biological databases. The process is composed of different components that analyse the synonyms types, their frequencies, intra-species, between data sources and cross-species ambiguities, with considering the case insensitive and case sensitive matching of the synonyms.

We extracted the data available in Entrez Gene and UniProtKB databases, and we used our process to separate and analyse records related to six different species, grouped in pairs, where each pair belongs to a biological Kingdom. These are: *Homo Sapiens and Mus Musculus* from the *Animalia* kingdom*, Arabidopsis Thaliana and Oryza Sativa* from the *Plantae* kingdom, and *Pseudomonas Fluorescens and Bacillus Subtilis* from the Bacteria kingdom. In addition, we studied the relationship between intra-species and cross-species ambiguities. Overall, we found that at least 40% of the cross-species ambiguity is caused by intra-species ambiguous synonyms. We also found that among the six species, rice has the best naming model in Entrez Gene database, and it has low overlapping ambiguities between data sources, and across-species. These results are useful for text mining and the analysis of high throughput experiments where lists of genes/proteins are analysed. One of our suggested solutions is to construct a database of ambiguous synonyms for different species, and to build a system that evaluate the biological texts and use this database to find any ambiguous synonym in the text.

Based on that, the system can recommend alternative gene synonyms that are not ambiguous, to be used instead.

This work can be extended in different directions to support other general or species-specific biological databases. In addition, we intend to support databases for other biological concepts such as chemicals and diseases, and to analyze the ambiguities in the names of these entities, as well as the overlapping ambiguities between different entities.

REFERENCES

[1] M. Weeber, B. J. Schijvenaars, E. M. Van Mulligen, B. Mons, R. Jelier, C. C. Van Der Eijk*, et al.*, "Ambiguity of human gene symbols in LocusLink and MEDLINE: creating an inventory and a disambiguation test collection," *AMIA Annu Symp Proc,* pp. 704-708, 2003.

[2] O. Tuason, L. Chen, H. Liu, J. A. Blake, and C. Friedman, "Biological nomenclatures: a source of lexical knowledge and ambiguity," *Pac Symp Biocomput,* pp. 238-249, 2004.

[3] S. o. t. C. o. B. Diversity, *Guide to the Global Taxonomy Initiative*: Secretariat of the Convention on Biological Diversity, 2008.

[4] G. Edwards, *Biology: The Easy Way*: Barron's Educational Series, Incorporated, 2000.

[5] L. R. Berg, *Introductory Botany: Plants, People, and the Environment [With Online Access]*: Thomson Brooks/Cole, 2007.

[6] B. Alberts, *Molecular Biology of the Cell: Reference Edition*: Garland Publishing, Incorporated, 2008.

[7] I. C. o. G. Symbols and Nomenclature, *Report of the International Committee on Genetic Symbols and Nomenclature, August 1957*: U.I.S.B., 1957.

[8] L. Hirschman, A. A. Morgan, and A. S. Yeh, "Rutabaga by any other name: extracting biological names," *J Biomed Inform,* vol. 35, pp. 247-259, Aug 2002.

[9] D. Maglott, J. Ostell, K. D. Pruitt, and T. Tatusova, "Entrez Gene: gene-centered information at NCBI," *Nucleic Acids Res.,* vol. 39, pp. D52-57, Jan 2011.

[10] G. O. Consortium, "Creating the gene ontology resource: design and implementation," *Genome Res.,* vol. 11, pp. 1425-1433, Aug 2001.

[11] M. A. Harris, J. Clark, A. Ireland, J. Lomax, M. Ashburner, R. Foulger*, et al.*, "The Gene Ontology (GO) database and informatics resource," *Nucleic Acids Res.,* vol. 32, pp. D258-261, Jan 2004.

[12] D. Swarbreck, C. Wilks, P. Lamesch, T. Z. Berardini, M. Garcia-Hernandez, H. Foerster*, et al.*, "The Arabidopsis Information Resource (TAIR): gene structure and function annotation," *Nucleic Acids Res.,* vol. 36, pp. D1009-1014, Jan 2008.

[13] R. A. Drysdale and M. A. Crosby, "FlyBase: genes and gene models," *Nucleic Acids Res.,* vol. 33, pp. D390-395, Jan 2005.

[14] K. Fundel and R. Zimmer, "Gene and protein nomenclature in public databases," *BMC Bioinformatics,* vol. 7, p. 372, 2006.

[15] U. Consortium, "The Universal Protein Resource (UniProt)," *Nucleic Acids Res.,* vol. 35, pp. D193-197, Jan 2007.

[16] H. Liu, A. R. Aronson, and C. Friedman, "A study of abbreviations in MEDLINE abstracts," *Proc AMIA Symp,* pp. 464-468, 2002.

[17] H. Yu, G. Hripcsak, and C. Friedman, "Mapping abbreviations to full forms in biomedical articles," *J Am Med Inform Assoc,* vol. 9, pp. 262-272, 2002.

[18] M. Yoshida, K. Fukuda, and T. Takagi, "PNAD-CSS: a workbench for constructing a protein name abbreviation dictionary," *Bioinformatics,* vol. 16, pp. 169-175, Feb 2000.

[19] E. Adar, "SaRAD: a Simple and Robust Abbreviation Dictionary," *Bioinformatics,* vol. 20, pp. 527-533, Mar 2004.

[20] W. Zhou, V. I. Torvik, and N. R. Smalheiser, "ADAM: another database of abbreviations in MEDLINE," *Bioinformatics,* vol. 22, pp. 2813-2818, Nov 2006.

[21] H. Xu, P. D. Stetson, and C. Friedman, "A study of abbreviations in clinical notes," *AMIA Annu Symp Proc,* pp. 821-825, 2007.

[22] R. Malik, L. Franke, and A. Siebes, "Combination of text-mining algorithms increases the performance," *Bioinformatics,* vol. 22, pp. 2151-2157, Sep 2006.

[23] M. H. Coletti and H. L. Bleich, "Medical subject headings used to search the biomedical literature," *J Am Med Inform Assoc,* vol. 8, pp. 317-323, 2001.

[24] L. Tanabe and W. J. Wilbur, "Tagging gene and protein names in biomedical text," *Bioinformatics,* vol. 18, pp. 1124-1132, Aug 2002.

[25] K. Fukuda, A. Tamura, T. Tsunoda, and T. Takagi, "Toward information extraction: identifying protein names from biological papers," *Pac Symp Biocomput,* pp. 707-718, 1998.

[26] K. Fundel, D. Guttler, R. Zimmer, and J. Apostolakis, "A simple approach for protein name identification: prospects and limits," *BMC Bioinformatics,* vol. 6 Suppl 1, p. S15, 2005.

[27] B. Settles, "ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text," *Bioinformatics,* vol. 21, pp. 3191-3192, Jul 2005.

[28] L. Hirschman, A. Yeh, C. Blaschke, and A. Valencia, "Overview of BioCreAtIvE: critical assessment of information extraction for biology," *BMC Bioinformatics,* vol. 6

Suppl 1, p. S1, 2005.

[29] M. Krallinger, A. Morgan, L. Smith, F. Leitner, L. Tanabe, J. Wilbur*, et al.*, "Evaluation of text-mining systems for biology: overview of the Second BioCreative community challenge," *Genome Biol.,* vol. 9 Suppl 2, p. S1, 2008.

[30] D. Hanisch, J. Fluck, H. T. Mevissen, and R. Zimmer, "Playing biology's name game: identifying protein names in scientific text," *Pac Symp Biocomput,* pp. 403-414, 2003.

[31] M. T. Heinz JF, Dach H Oster M Hofmann-Apitius M, "ProMiner: Recognition of Human Gene and Protein Names using regularly updated Dictionaries," ed, 2007.

[32] A. A. Morgan, L. Hirschman, M. Colosimo, A. S. Yeh, and J. B. Colombe, "Gene name identification and normalization using a model organism database," *J Biomed Inform,* vol. 37, pp. 396-410, Dec 2004.

[33] T. Ono, H. Hishigaki, A. Tanigami, and T. Takagi, "Automated extraction of information on protein-protein interactions from the biological literature," *Bioinformatics,* vol. 17, pp. 155-161, Feb 2001.

[34] Y. Tsuruoka and J. Tsujii, "Improving the performance of dictionary-based approaches in protein name recognition," *J Biomed Inform,* vol. 37, pp. 461-470, Dec 2004.

[35] D. W. Nebert and H. M. Wain, "Update on human genome completion and annotations: gene nomenclature," *Hum. Genomics,* vol. 1, pp. 66-71, Nov 2003.

[36] M. T. Davisson, "Rules and guidelines for nomenclature of mouse genes. International Committee on Standardized Genetic Nomenclature for Mice," *Gene,* vol. 147, pp. 157-160, Sep 1994.

[37] M. Demerec, E. A. Adelberg, A. J. Clark, and P. E. Hartman, "A proposal for a uniform nomenclature in bacterial genetics," *J. Gen. Microbiol.,* vol. 50, pp. 1-14, Jan 1968.

[38] S. R. McCouch, "Gene Nomenclature System for Rice," *Rice,* vol. 1, pp. 72-84, 2008.

[39] E. A. Bruford, "Highlights of the 'gene nomenclature across species' meeting," *Hum. Genomics,* vol. 4, pp. 213-217, Feb 2010.

[40] L. Chen, H. Liu, and C. Friedman, "Gene name ambiguity of eukaryotic nomenclatures," *Bioinformatics,* vol. 21, pp. 248-256, Jan 2005.

[41] M. J. Schuemie, M. Weeber, B. J. Schijvenaars, E. M. van Mulligen, C. C. van der Eijk, R. Jelier*, et al.*, "Distribution of information in biomedical abstracts and full-text publications," *Bioinformatics,* vol. 20, pp. 2597-2604, Nov 2004.

APPENDICES

A Frequency Diagrams

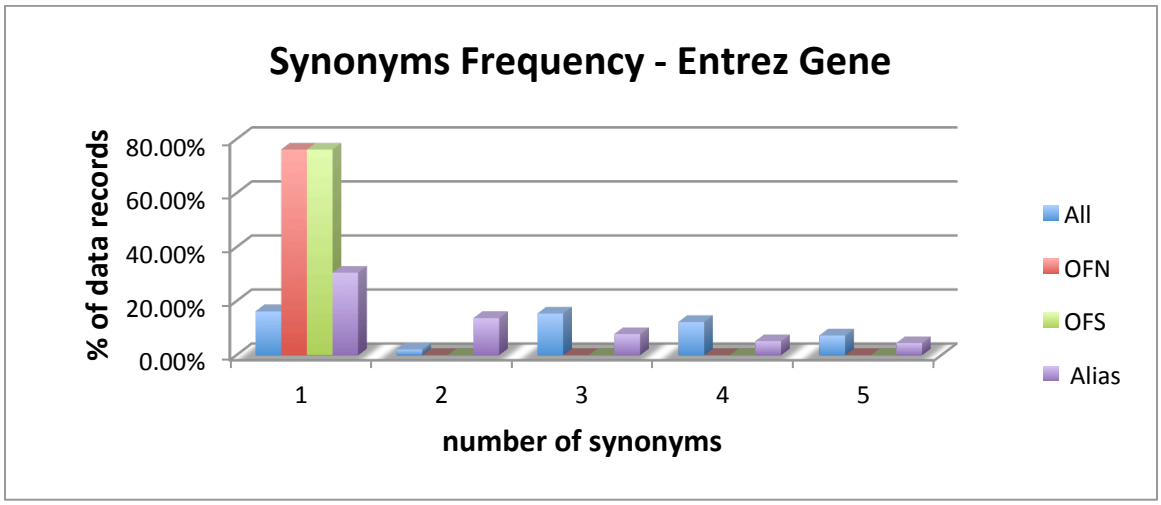

Figure 16. Synonyms frequency in Entrez Gene for *Homo Sapiens*.

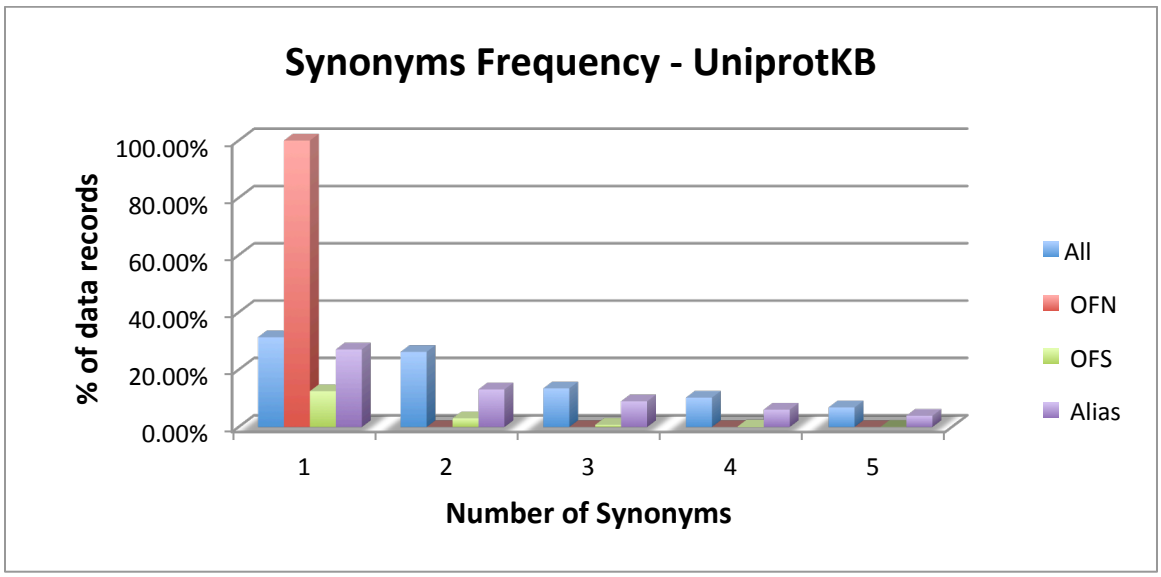

Figure 17. Synonyms frequency in UniProtKB for *Homo Sapiens*.

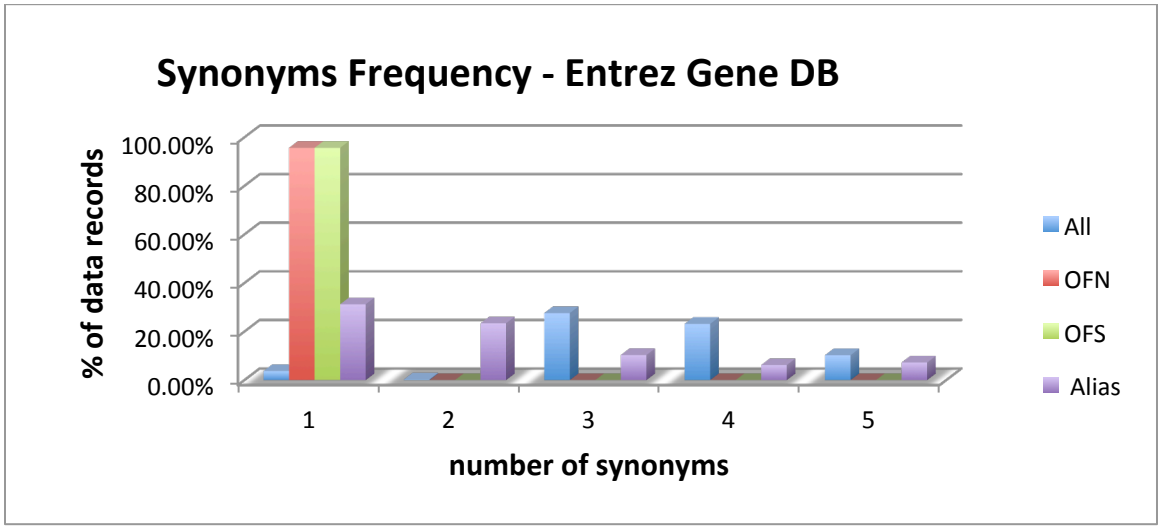

Figure 18. Synonyms frequency in Entrez Gene for *Mus Musculus*.

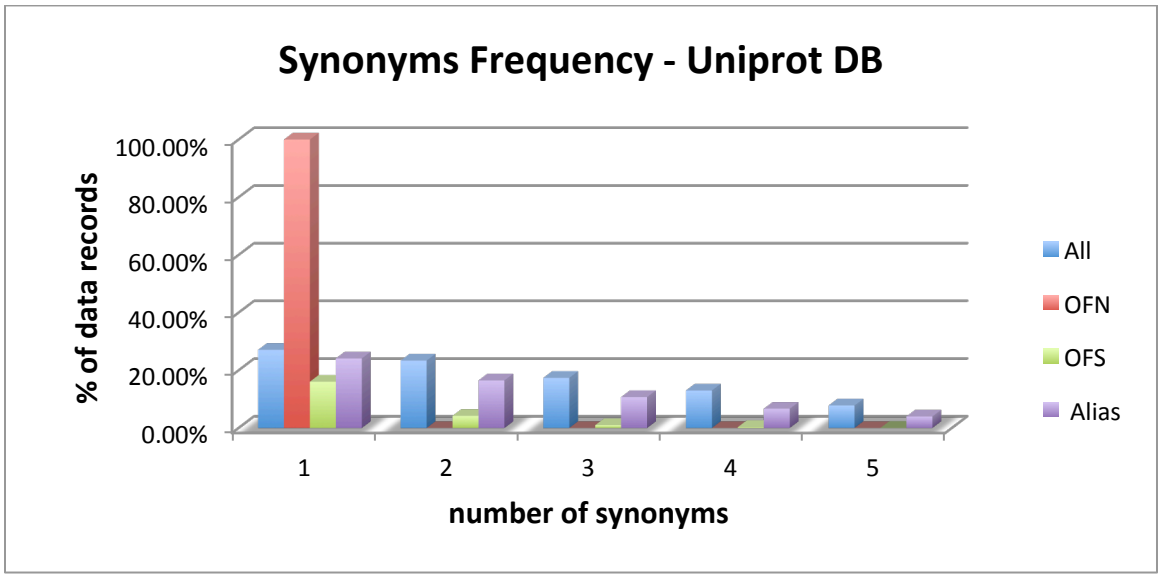

Figure 19. Synonyms frequency in UniProtKB for *Mus Musculus*.

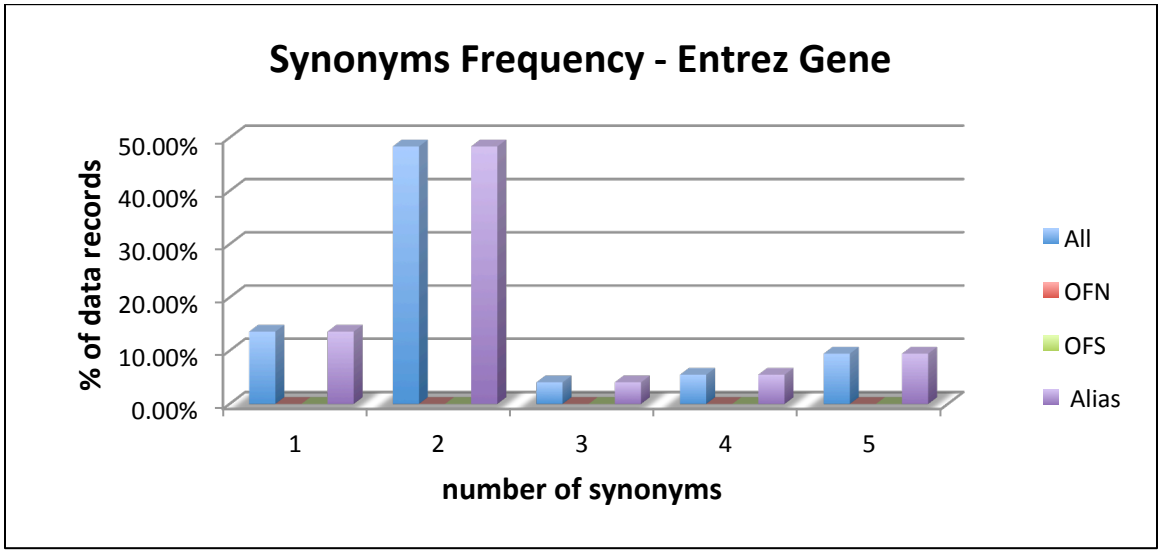

Figure 20. Synonyms frequency in Entrez Gene for *Arabidopsis Thaliana*.

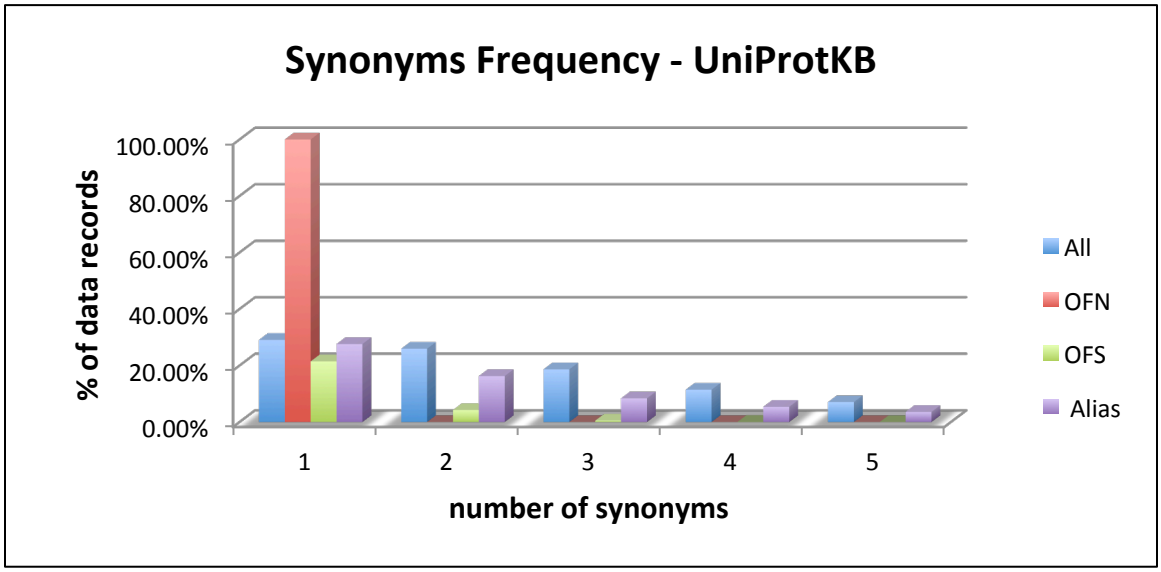

Figure 21. Synonyms frequency in UniProtKB for *Arabidopsis Thaliana*.

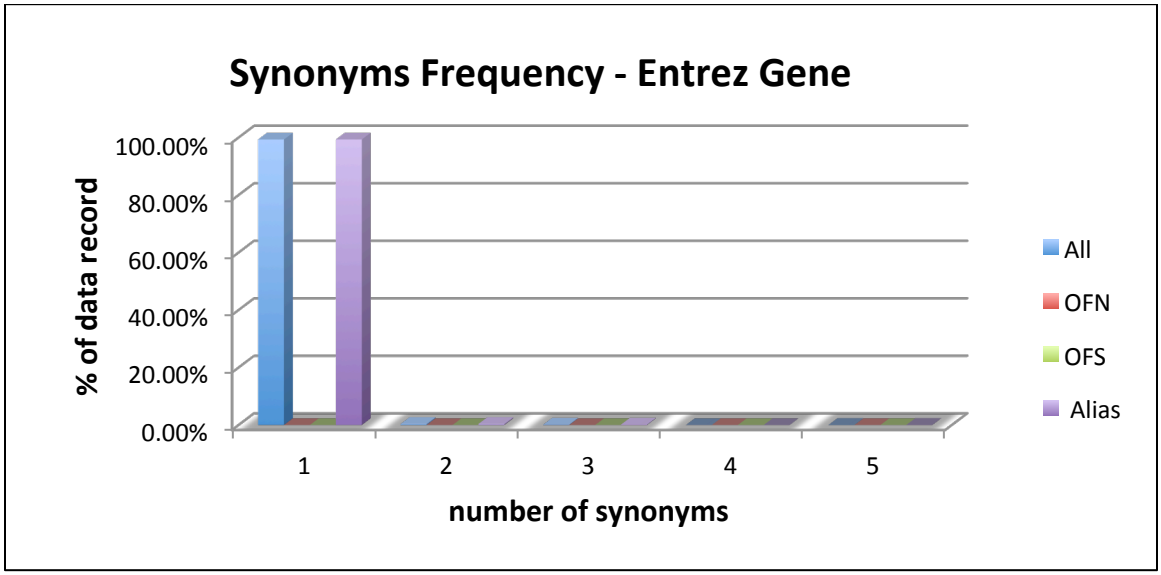

Figure 22. Synonyms frequency in Entrez Gene for *Oryza Sativa*.

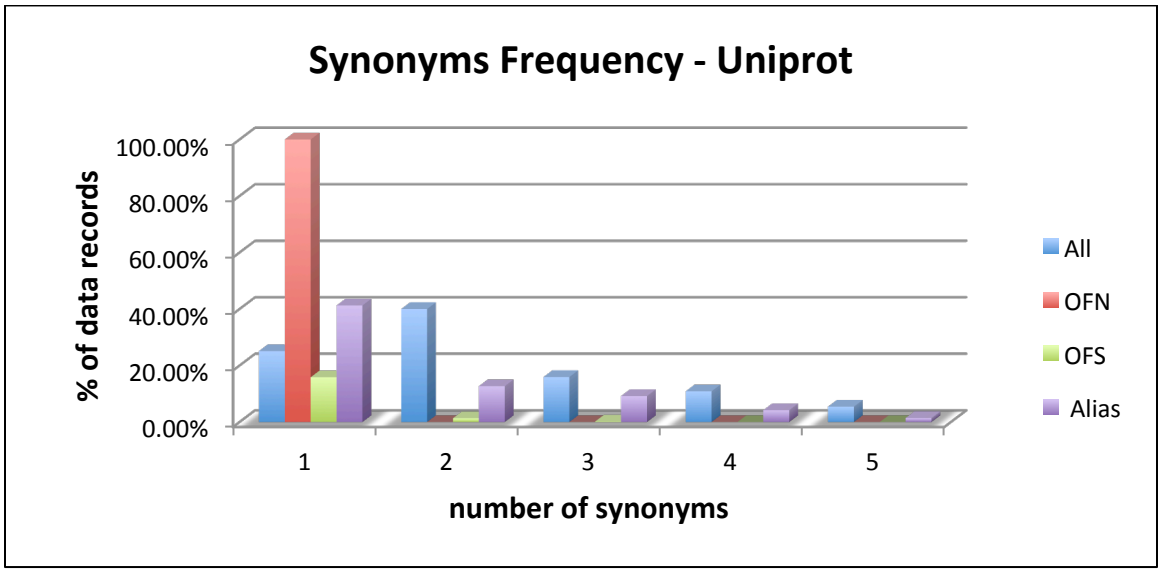

Figure 23. Synonyms frequency in UniProtKB for *Oryza Sativa*.

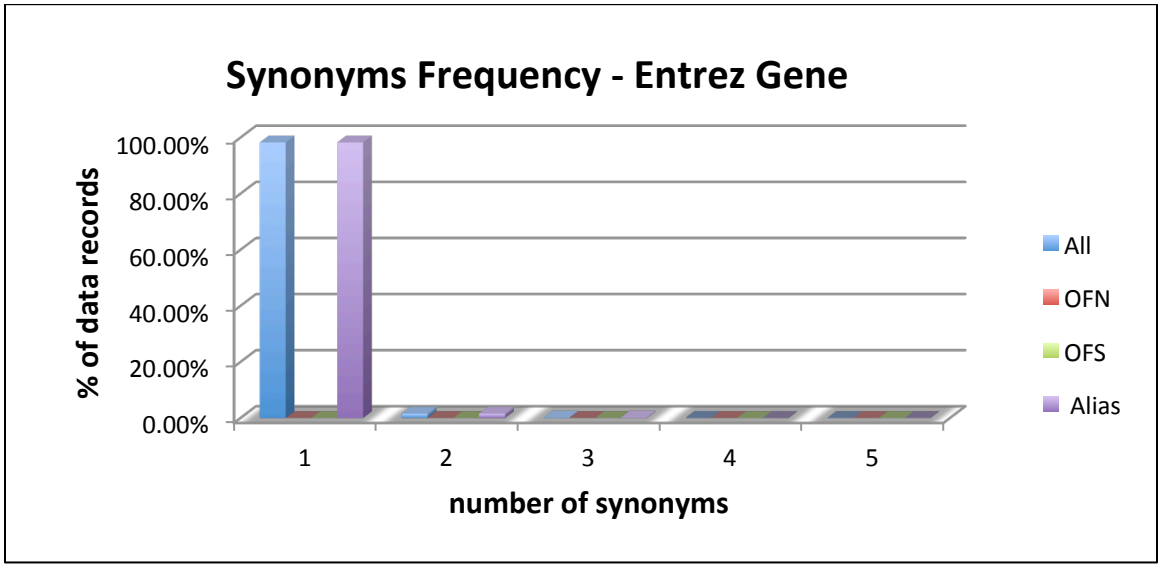

Figure 24. Synonyms frequency in Entrez Gene for *Pseudomonas Fluorescens*.

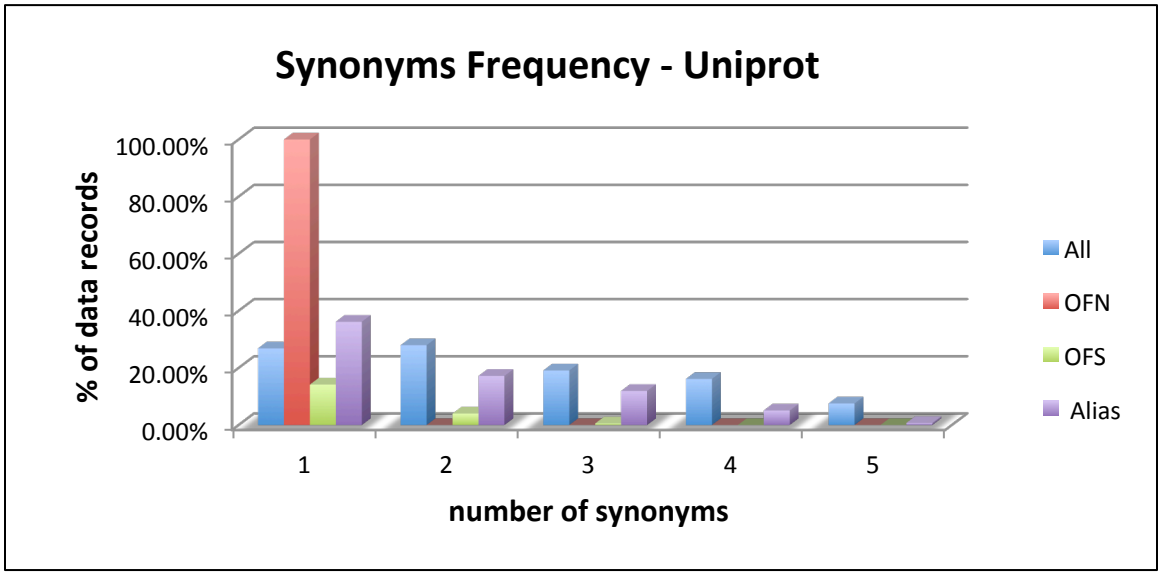

Figure 25. Synonyms frequency in UniProtKB for *Pseudomonas Fluorescens*.

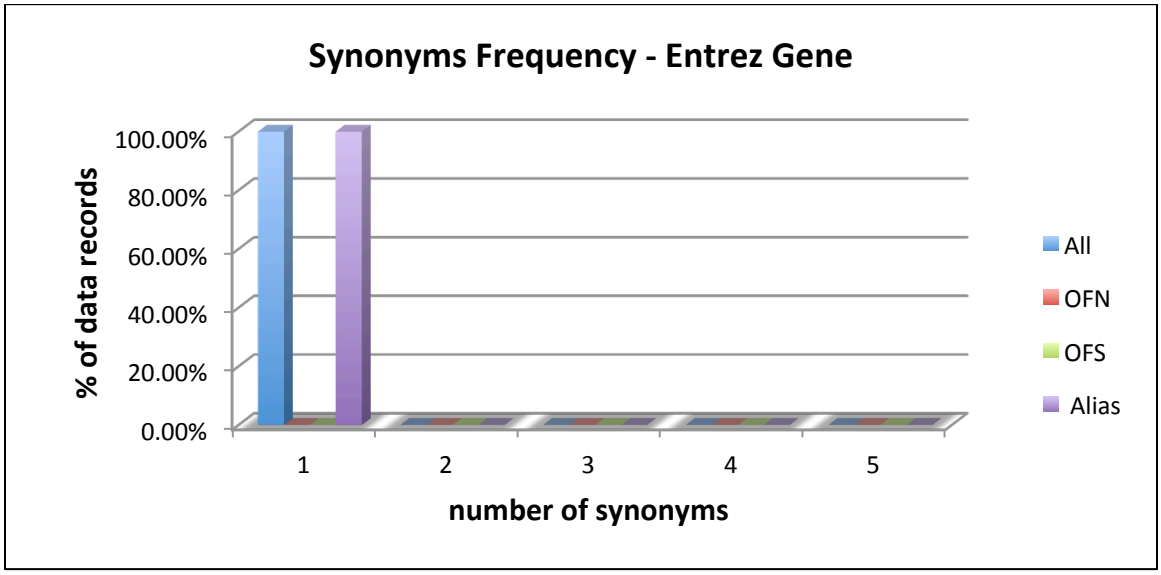

Figure 26. Synonyms frequency in Entrez Gene for *Bacillus Subtilis*.

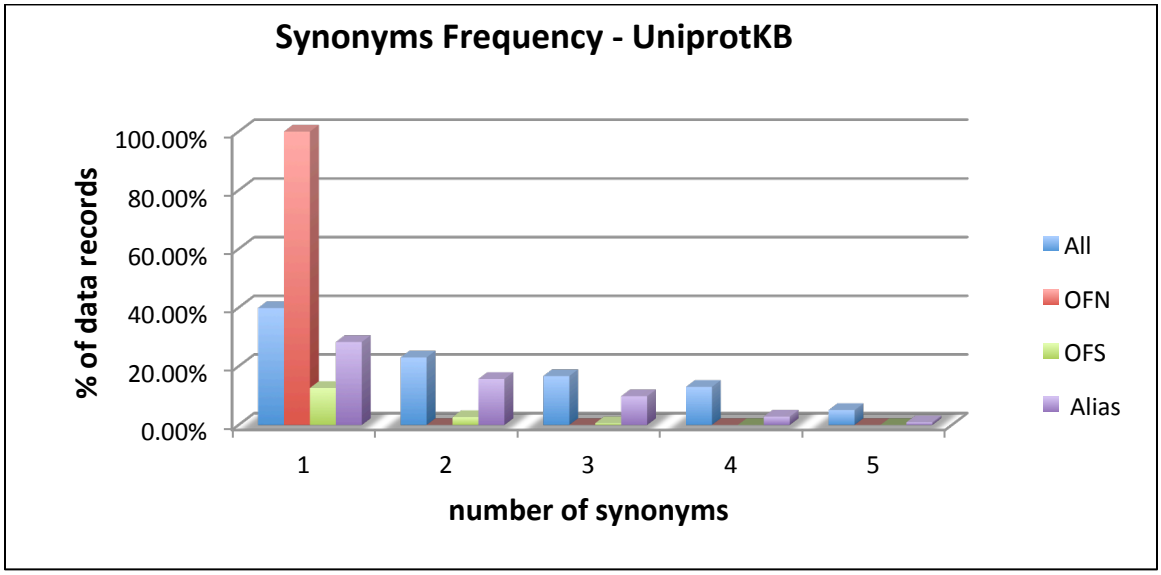

Figure 27. Synonyms frequency in UniProtKB for *Bacillus Subtilis*.

B Frequency and Ambiguity Tables

Table 31. *Homo Sapiens* frequency percentages.

	Entrez Gene					UniprotKB			
i	**All**	**OFN**	**OFS**	**Alias**		**All**	**OFN**	**OFS**	**Alias**
1	16.21%	76.16%	76.16%	30.54%		31.34%	100%	12.51%	27.07%
2	2.11%	0%	0%	13.67%		26.23%	0%	3.01%	13.11%
3	15.41%	0%	0%	7.78%		13.47%	0%	0.75%	8.96%
4	12.28%	0%	0%	5.17%		10.24%	0%	0.23%	6.05%
5	7.27%	0%	0%	4.43%		6.82%	0%	0.02%	3.94%

Table 32. *Mus Musculus* frequency percentages.

	Entrez Gene					UniprotKB			
i	**All**	**OFN**	**OFS**	**Alias**		**All**	**OFN**	**OFS**	**Alias**
1	3.73%	95.67%	95.67%	31.22%		27.07%	100%	16.06%	24.11%
2	0.27%	0%	0%	23.37%		23.38%	0%	4.22%	16.49%
3	27.60%	0%	0%	10.31%		17.37%	0%	0.97%	10.71%
4	23.15%	0%	0%	6.19%		13.00%	0%	0.30%	6.67%
5	10.24%	0%	0%	7.23%		7.79%	0%	0.03%	4.00%

Table 33. *Arabidopsis Thaliana* frequency percentages.

	Entrez Gene					UniprotKB			
i	**All**	**OFN**	**OFS**	**Alias**		**All**	**OFN**	**OFS**	**Alias**
1	13.57%	0%	0%	13.57%		29.06%	100.00%	21.46%	27.61%
2	48.34%	0%	0%	48.34%		25.95%	0%	4.25%	16.25%
3	4.05%	0%	0%	4.05%		18.64%	0%	0.48%	8.39%
4	5.49%	0%	0%	5.49%		11.43%	0%	0.04%	5.39%
5	9.44%	0%	0%	9.44%		7.05%	0%	0%	3.58%

Table 34. *Pseudomonas Fluorescens* frequency percentages.

	Entrez Gene					UniprotKB			
i	**All**	**OFN**	**OFS**	**Alias**		**All**	**OFN**	**OFS**	**Alias**
1	98.26%	0%	0%	98.26%		26.85%	100.00%	14.17%	36.16%
2	1.54%	0%	0%	1.54%		27.95%	0%	4.00%	17.18%
3	0.14%	0%	0%	0.14%		19.10%	0%	0.55%	11.93%
4	0.04%	0%	0%	0.04%		16.15%	0%	0%	5.09%
5	0.02%	0%	0%	0.02%		7.56%	0%	0%	0.73%

Table 35. *Oryza Sativa* frequency percentages.

	Entrez Gene					UniprotKB			
i	**All**	**OFN**	**OFS**	**Alias**		**All**	**OFN**	**OFS**	**Alias**
1	99.32%	0%	0%	99.32%		25.11%	100.00%	15.84%	41.25%
2	0.37%	0%	0%	0.37%		39.99%	0%	1.32%	12.67%
3	0.16%	0%	0%	0.16%		15.92%	0%	0.19%	9.19%
4	0.01%	0%	0%	0.01%		10.87%	0%	0%	4.20%
5	0.04%	0%	0%	0.04%		5.43%	0%	0%	1.40%

Table 36. *Bacillus Subtilis* frequency percentages.

	Entrez Gene					UniprotKB			
i	**All**	**OFN**	**OFS**	**Alias**		**All**	**OFN**	**OFS**	**Alias**
1	99.95%	0%	0%	99.95%		39.86%	100.00%	12.60%	28.27%
2	0.05%	0%	0%	0.05%		23.03%	0%	2.69%	15.67%
3	0%	0%	0%	0%		16.67%	0%	0.60%	9.77%
4	0%	0%	0%	0%		12.95%	0%	0%	2.83%
5	0%	0%	0%	0%		5.05%	0%	0%	0.90%

Table 37. Intra-species ambiguities.

Species	**Entrez Gene**	**UniProtKB**
Homo Sapiens	23.29%	17.88%
Mus Musculus	28.83%	10.99%
Arabidopsis Thaliana	17.54%	16.54%
Oryza Sativa	5.47%	49.15%
Pseudomonas Fluorescens	41.77%	86.93%
Bacillus Subtilis	74.70%	67.98%

Table 38 *Homo Sapiens* ambiguity percentages.

	Entrez Gene					UniprotKB			
i	**All**	**OFN**	**OFS**	**Alias**		**All**	**OFN**	**OFS**	**Alias**
2	7.92%	0%	0%	7.92%		1.50%	0.08%	0.07%	1.03%
3	1.34%	0%	0%	1.34%		0.31%	0.05%	0.01%	0.22%
4	0.34%	0%	0%	0.34%		0.18%	0.06%	0.01%	0.11%
5	0.10%	0%	0%	0.10%		0.10%	0.04%	0.00%	0.06%
6	0.04%	0%	0%	0.04%		0.06%	0.02%	0%	0.03%

Table 39 *Mus Musculus* ambiguity percentages.

	Entrez Gene					UniprotKB			
i	**All**	**OFN**	**OFS**	**Alias**		**All**	**OFN**	**OFS**	**Alias**
2	6.03%	0.16%	0.12%	5.88%		1.50%	0.09%	0.06%	1.03%
3	1.05%	0.0046%	0.0017%	1.05%		0.43%	0.10%	0.02%	0.30%
4	0.27%	0.0017%	0%	0.27%		0.20%	0.07%	0.01%	0.12%
5	0.10%	0%	0%	0.10%		0.15%	0.05%	0.01%	0.08%
6	0.05%	0%	0%	0.05%		0.08%	0.03%	0%	0.05%

Table 40 *Arabidopsis Thaliana* ambiguity percentages.

	Entrez Gene					UniprotKB			
i	**All**	**OFN**	**OFS**	**Alias**		**All**	**OFN**	**OFS**	**Alias**
2	3.36%	0%	0%	3.36%		1.87%	0.25%	0.14%	1.24%
3	0.34%	0%	0%	0.34%		0.55%	0.13%	0.02%	0.38%
4	0.03%	0%	0%	0.03%		0.32%	0.08%	0.02%	0.22%
5	0.03%	0%	0%	0.03%		0.31%	0.07%	0.02%	0.21%
6	0.01%	0%	0%	0.01%		0.14%	0.03%	0%	0.11%

Table 41 *Oryza Sativa* ambiguity percentages.

	Entrez Gene					UniprotKB			
i	**All**	**OFN**	**OFS**	**Alias**		**All**	**OFN**	**OFS**	**Alias**
2	1.05%	0%	0%	1.05%		18.59%	6.62%	1.73%	9.76%
3	0.15%	0%	0%	0.15%		2.11%	0.63%	0.17%	1.23%
4	0.04%	0%	0%	0.04%		1.12%	0.45%	0.08%	0.60%
5	0.01%	0%	0%	0.01%		1.64%	0.80%	0.06%	0.78%
6	0.01%	0%	0%	0.01%		0.86%	0.33%	0.09%	0.42%

Table 42 *Pseudomonas Fluorescens* ambiguity percentages.

	Entrez Gene					UniprotKB			
i	**All**	**OFN**	**OFS**	**Alias**		**All**	**OFN**	**OFS**	**Alias**
2	17.12%	0%	0%	17.12%		6.16%	2.09%	0.47%	3.71%
3	12.71%	0%	0%	12.71%		3.86%	0.72%	0.07%	3.10%
4	8.71%	0%	0%	8.71%		1.23%	0.36%	0.11%	0.76%
5	0.43%	0%	0%	0.43%		0.90%	0.36%	0.07%	0.47%
6	0.12%	0%	0%	0.12%		1.19%	0.50%	0.14%	0.58%

Table 43 *Bacillus Subtilis* ambiguity percentages.

	Entrez Gene					UniProtKB			
i	**All**	**OFN**	**OFS**	**Alias**		**All**	**OFN**	**OFS**	**Alias**
2	8.77%	0%	0%	8.77%		3.10%	0.94%	0.43%	1.75%
3	8.88%	0%	0%	8.88%		0.99%	0.29%	0.09%	0.56%
4	18.95%	0%	0%	18.95%		0.31%	0.03%	0.04%	0.25%
5	24.27%	0%	0%	24.27%		0.25%	0.10%	0.01%	0.13%
6	3.96%	0%	0%	3.96%		0.13%	0.07%	0%	0.07%