School of Computing Science

Provenance, Propagation and Quality of Biological Annotation

Michael J. Bell

Submitted for the degree of Doctor of Philosophy in the School of Computing Science, Newcastle University

September 2014

Abstract

Biological databases have become an integral part of the life sciences, being used to store, organise and share ever-increasing quantities and types of data. Biological databases are typically centred around raw data, with individual entries being assigned to a single piece of biological data, such as a DNA sequence. Although essential, a reader can obtain little information from the raw data alone. Therefore, many databases aim to supplement their entries with *annotation*, allowing the current knowledge about the underlying data to be conveyed to a reader. Although annotations come in many different forms, most databases provide some form of free text annotation.

Given that annotations can form the foundations of future work, it is important that a user is able to evaluate the quality and correctness of an annotation. However, this is rarely straightforward. The amount of annotation, and the way in which it is curated, varies between databases. For example, the production of an annotation in some databases is entirely automated, without any manual intervention. Further, sections of annotations may be reused, being propagated between entries and, potentially, external databases. This provenance and curation information is not always apparent to a user.

The work described within this thesis explores issues relating to biological annotation quality. While the most valuable annotation is often contained within free text, its lack of structure makes it hard to assess. Initially, this work describes a generic approach that allows textual annotations to be quantitatively measured. This approach is based upon the application of Zipf's Law to words within textual annotation, resulting in a single value, α. The relationship between the α value and Zipf's principle of least effort provides an indication as to the annotations quality, whilst also allowing annotations to be quantitatively compared.

Secondly, the thesis focuses on determining annotation provenance and tracking any subsequent propagation. This is achieved through the development of a visualisation

framework, which exploits the reuse of sentences within annotations. Utilising this framework a number of propagation patterns were identified, which on analysis appear to indicate low quality and erroneous annotation.

Together, these approaches increase our understanding in the textual characteristics of biological annotation, and suggests that this understanding can be used to increase the overall quality of these resources.

Declaration

I declare that this thesis is my own work unless otherwise stated. No part of this thesis has previously been submitted for a degree or any other qualification at Newcastle University or any other institution.

Michael Bell

September 2014

Publications

Portions of the work within this thesis have been documented in the following publications:

Michael J. Bell, Colin S. Gillespie, Daniel Swan and Phillip Lord. An approach to describing and analysing bulk biological annotation quality: A case study using UniProtKB. *Bioinformatics*, 2012, 28, i562-i568.

Michael J. Bell, Matthew Collison, and Phillip Lord. Can Inferred Provenance and Its Visualisation Be Used to Detect Erroneous Annotation? A Case Study Using UniProtKB. *PLoS ONE*, 2013, 8(10): e75541.

Acknowledgements

First and foremost I would like to thank my primary supervisor, Dr. Phillip Lord, for giving me the chance to pursue a Ph.D. at Newcastle University. Your advice, guidance and continued support throughout the project has been truly invaluable.

I would also like to thank my secondary supervisor Dr. Daniel Swan for the regular discussions and support during the early stages of my project. This thanks is extended to Dr. Paolo Missier for agreeing to become my secondary supervisor when Daniel left Newcastle. Taking over mid project could not have been easy, but you made the transaction seamless with numerous ideas and suggestions arising from our discussions.

In addition to my supervisors, I am also indebted to Dr. Colin Gillespie from the school of Mathematics & Statistics for his work regarding the power-law framework (as described in Chapter 3). I greatly benefited from your assistance and insightful comments that arose from this collaboration.

Another person deserving specific mention is Matthew Collison. Your assistance with analysing large quantities of biological data that I had extracted was indispensable. Your input and enthusiasm allowed significantly more data to be analysed than I could have done individually. I am also grateful to Dr. Allyson Lister, for insightful and helpful discussions regarding the UniProt Knowledgebase, and to Jennifer Warrender for many helpful discussions and offers to proof read my work.

I was fortunate to complete my Ph.D. within a department consisting of many wonderful people, including Dr. Keith Flanagan, Dr. Katherine James, Beth Lawry, Dr. Goksel Misirli, Joseph Mullen and Sungshic Park. It has been a privilege to share many discussions with you all.

I am especially grateful to my friends and family for their support over the last four years. A very special thanks must go to my parents and Sheela. Your unfaltering support and encouragement helped me reach the finish line. Thank you.

Finally, I am grateful to the School of Computing Science and the European and

Physical Sciences Research Council (EPSRC) for funding this project.

Contents

List of Figures

List of Tables

Acronyms

ASCII American Standard Code for Information Interchange

BANE Biological ANnotation Extraction framework

BIND Biomolecular Interaction Network Database

BLAST Basic Local Alignment Search Tool

BMRB Biological Magnetic Resonance Bank

CDF Cumulative Distribution Function

CDS Coding Sequence

CSV Comma Separated Values

DDBJ DNA Data Bank of Japan

DPI Dots Per Inch

EBI European Bioinformatics Institute

EC Enzyme Commission

EMBL The European Molecular Biology Laboratory

ENA European Nucleotide Archive

FTP File Transfer Protocol

GAQ GO Annotation Quality

GDB Genome DataBase

GO The Gene Ontology

GOA Gene Ontology Annotation

HAMAP High-quality Automated and Manual Annotation of Proteins

HMMs Hidden Markov Models

HTML HyperText Markup Language

IDF Inverse Document Frequency

INSDC International Nucleotide Sequence Database Collaboration

IPI International Protein Index

JBC Journal of Biological Chemistry

JSON JavaScript Object Notation

KEGG Kyoto Encyclopedia of Genes and Genomes

NAR Nucleic Acids Research

NCBI National Center for Biotechnology Information

NHGRI National Human Genome Research Institute

NIH National Institute of Health

NLP Natural Language Processing

ORF Open Reading Frame

PDB Protein Data Bank

PIR Protein Information Resource

PMF Probability Mass Function

QUALM QUALity Metric

RDF Resource Description Framework

SAAS Statistical Automatic Annotation System

SIB Swiss Institute of Bioinformatics

SMOG Simple Measure of Gobbledygook

SRA Sequence Read Archive

TF-IDF Term Frequency-Inverse Document Frequency

UML Unified Modeling Language

UniMES The UniProt Metagenomic and Environmental Sequences

UniParc The UniProt Archive

UniProt The Universal Protein Resource

UniProtKB The UniProt Knowledgebase

UniRef The UniProt Reference Clusters

UniRule Unified Rule

VIPeR Visualising annotatIon PRopagation

wwPDB worldwide Protein Data Bank

WWW World Wide Web

XML eXtensible Markup Language

1

Introduction

Contents

Introduction

A major scientific landmark of the last decade was the complete sequencing of the human genome. Initiated in 1990, the human genome project took a total of thirteen years to complete at a cost of just under \$3 billion [4]. Since the completion of the project, the costs associated with sequencing have reduced drastically; sequencing a genome in 2013 could be done within one to two days and for under $\$5,000$ [5].

The sequence data resulting from the human genome project was made publicly accessible in biological databases, such as GenBank and the European Nucleotide Archive (ENA). These biological databases provide a solution for the storing, organising and dissemination of biological knowledge. With the improvements in sequencing technology, the data being added to biological databases continues to grow exponentially. For example, over eight million entries were added to the GenBank database in 2013, bringing the total sequences in the database to over 169 million [6]. The analysis of this data presents a distinct set of challenges.

Whilst databases are centred around raw biological data, such as DNA sequences, many also provide annotation as a mechanism to record and convey the knowledge known about the underlying biology. The types of annotation added to an entry can include, for example, information on how the data was obtained, possible roles within disease or links to external databases. Historically this information was identified, analysed and added manually, which is a significant bottleneck that is exacerbated given the exponential increase in data being deposited.

In an attempt to reduce the deficit between annotated and unannotated data, many databases also provide annotation that is generated computationally. This can result in annotation attached to one sequence essentially being copied to another sequence based on similarities shared between the two sequences.

Biological annotation can be pivotal to users who rely upon the information to form their understanding of the underlying biology and could potentially influence their future research and work. Given the importance that can be placed on an annotation it would be expected that a user could easily evaluate the quality and correctness should they wish. However, due to the textual nature of annotation, there is a distinct

lack of methods and tooling that allow annotation quality to be analysed.

One possible approach to assess annotation quality is to consider the method used to create the annotation, with manually produced annotations being considered of higher quality than those created computationally. However, this is a rather crude approach with a number of caveats, such as: not all manual annotations will be correct or of high quality; various techniques exist for producing automatic annotation that will produce annotations of differing quality; manually produced annotation will also vary in quality depending upon the author of the annotation and the resources available to them; and the quality of the annotation will likely differ depending upon the knowledge known about the underlying biology (e.g. is the underlying sequence from a model organism?).

The issue of annotation quality is further complicated if we consider that an annotation may have been copied between entries intra or inter-database. For example, if an automated annotation is produced by copying the annotation from another entry, how was this original annotation created? Was it produced manually or computationally? This complication raises a further question: if an annotation is found to be incorrect and is updated, then are the entries it has been copied to also updated?

However, the source of an annotation is not always easy to determine; many databases do not document the source of an annotation, or do so inadequately. While the identification of the source of an annotation can allow a user to gain confidence in its correctness, it can also provide an insight into how annotations are reused. For example, is a certain type of annotation more likely to be reused? Does reuse appear to be having a detrimental effect on the overall annotation quality in a database?

Within this thesis, we provide new mechanisms for analysing textual annotation within biological databases. Specifically, we wish to explore possible methods and approaches to allow users to analyse and evaluate an annotation. We also wish to employ these methods to analyse the quality of existing annotations, and to explore how the quality and properties of annotation change over time.

1.1 Contributions of This Thesis

The overall focus of the research presented in this thesis is on textual annotation in biological databases. The main outcome of this work was the development of two tools: QUALity Metric (QUALM) and Visualising annotatIon PRopagation (VIPeR).

The first of these tools, QUALM, is based on a power-law function being applied to word distributions. This approach means that QUALM only requires a list of how frequently each word occurs within a textual annotation, allowing any textual resource to be analysed. The outcome from QUALM is a single value, α, which is linked to Zipf's principle of least effort to provide an indication of quality.

To test the suitability of QUALM, we perform an in-depth analysis of The UniProt Knowledgebase (UniProtKB). This analysis explores the change in various subsets of annotation over time, including taxonomic groups and annotations of different ages. Linking these results to our *a priori* judgements suggests that QUALM provides an indication of annotation quality. To perform these analyses, we developed the Biological ANnotation Extraction framework (BANE) which allows textual annotation to be extracted from a database in the necessary form.

The second tool developed, VIPeR, is a visualisation approach that allows the provenance and propagation of sentences to be identified. More specifically, VIPeR shows the entries that a given sentence occurs in and for which database versions, visualising how it is distributed through a database over time. This tool also makes use of BANE.

The application of VIPeR to sentences in UniProtKB led to the identification of various propagation patterns. These patterns include the missing origin pattern, which identifies sentences that are removed from the first entry they appear in but subsequently remain in the database after this point. A number of sentences following the missing origin pattern were analysed and found to be erroneous, suggesting that these patterns can be used to indicate low quality and erroneous annotations.

Due to the lack of a gold standard dataset, UniProtKB was chosen for the initial analysis of QUALM and VIPeR. The documentation, structure and number of archived versions available made UniProtKB an ideal resource to test these tools. We later

extended this analysis to include annotation from the InterPro, PROSITE, PRINTS, TIGRFAMs and neXtProt databases. This extended analysis provided further confidence that the tools hold value for the analysis and exploration of textual annotation.

1.2 Thesis Structure

This thesis is divided into the following chapters:

- Chapter 2 introduces the different types of biological annotation and the role of biological databases, exploring their history and characteristics. Following this, we review existing work in the area of annotation quality and correctness before exploring the features and properties of UniProtKB.

- Chapter 3 describes the development of our quality metric (QUALM) that aims to allow textual annotation to be quantitatively assessed and compared. Applying QUALM to textual annotation produces a single value – α – that is related to Zipf's principle of least effort and used to evaluate textual annotation. QUALM is generic, allowing it to be applied to any textual resource.

- Chapter 4 applies QUALM to the textual annotation in UniProtKB. This analysis involves developing a framework (BANE) to extract data in the necessary format from UniProtKB, comparing manual and automated annotation and investigating how annotation in UniProtKB is changing over time.

- Chapter 5 discusses a range of techniques that could allow annotation provenance and propagation to be inferred. No ideal visualisation technique was identified through this analysis, so the development of a unique visualisation is described (VIPeR). VIPeR provides dynamic web-based graphs that allow the occurrences of individual sentences to be visualised over time.

- Chapter 6 applies VIPeR to sentences extracted from UniProtKB. This visualisation allows the provenance and subsequent propagation of a sentence to be inferred. Analysing these graphs identified a number of interesting propagation patterns. The sentences adhering to each propagation pattern were extracted, with the analysis suggesting that certain propagation patterns could be used as indicators of low quality and erroneous annotations.

- Chapter 7 presents results obtained from the application of QUALM and VIPeR to the neXtProt, InterPro, PRINTS, TIGRFAMs and PROSITE databases.

- Chapter 8 reviews the limitations of QUALM and VIPeR and discusses potential refinements and future work. The chapter concludes with a general discussion of the annotation landscape, and possible ways in which it could be improved.

2

Background Research

Contents

Introduction

This background chapter initially starts with a discussion of biological databases and analyses the main types of database and their history (Section 2.1). This discussion then focuses on the growth of biological databases and the sustainability issues that many databases face and the impact this can have on both users and other databases.

The sustainability of databases is important for our work; our analyses depend upon the availability of historical data. More specifically, our work focuses on annotation, which is used by many databases to record information about the underlying biology. Annotation is an overloaded term and can refer to various different types and forms of information. Therefore, we identify and discuss the various types of biological annotation and explore the different methods that are used to produce annotations (Section 2.2).

As there are various approaches and types of annotation, it is inevitable that their quality and correctness will vary. Therefore, we survey the broad area of annotation quality, discussing what is meant by "quality" and exploring existing techniques that are used to measure and assess annotation quality (Section 2.3). This survey highlighted how an error in one annotation can propagate to other annotations and impact the accuracy of new annotations. Identifying error propagation is difficult, partly due to provenance information not always being adequately documented. Therefore, we consider a number of methods that could be used to reconstruct missing provenance.

Although there are many different forms of annotation, our work is focused on textual annotation. Surveying the existing methods for analysing the quality and correctness of annotations identified a distinct lack of methods that can be suitably applied to textual annotations. This is potentially due to a lack of gold standard datasets for textual annotation. To alleviate this issue we explore The UniProt Knowledgebase (UniProtKB) in significant detail (Section 2.4). UniProtKB was chosen as it is a well established and popular database consisting of both manual and automated annotation. This analysis covers the history of the database, its structure and curation processes and protocols, which is relied upon for our work later in this thesis.

2.1 Biological Databases

Biological databases are a cornerstone for many biology-related fields. These databases provide a mechanism to store, organise and disseminate ever-increasing quantities of biological data [7, 8], providing the foundations required for modern-day biomedical research [3, 9]. Consulting biological databases has become routine for bioinformaticians wishing to obtain existing data for initial, or further, analysis [10, 11].

"Biological database" is an encompassing term used to capture a wide range of resources [12]. For example, although distinctly different, both GenBank and MEDLINE are regarded as biological databases; the former acts as a repository for DNA sequences, whilst the latter indexes bibliographic information. Broadly, a biological database can be defined as a "library of life sciences information, collected from scientific experiments, published literature, high-throughput experiment technology, and computational analyses" [13], which "aim to serve as a source of information to support experimental research scientists, and as a basis for computational analysis" [14].

Biological databases can be generally categorised into either *primary* or *secondary* databases [15–17]. Primary databases, such as DNA Data Bank of Japan (DDBJ), are those which collect and store original sequence data, often with minimal supporting documentation. Secondary databases, such as the Protein Information Resource (PIR), derive their contents from data stored in primary databases and are supplemented with curated annotation. However, this distinction between primary and secondary databases is becoming less meaningful [18]. For example, the Swiss-Prot database provides information that categorises it as both a primary and a secondary database, whilst resources such as The Gene Ontology (GO) and the bibliographic database MEDLINE do not fall into either category.

The distinction between primary and secondary databases is of little interest for the majority of users; the main interest is focused on the types of data and features provided by a given database [19]. Therefore, this section primarily focuses on the information and features made available by databases, rather than being concerned with specific categories or low-level technical aspects, such as the underlying database management system.

The advent of biological databases was initiated by the pioneering work of Dr Margaret Dayhoff. It was during her work of developing computational methods to allow protein sequences to be compared[1] that she identified the importance of recording identified protein sequences, as she explained in a letter to a colleague [20]:

> “There is a tremendous amount of information regarding evolutionary history and biochemical function implicit in each sequence and the number of known sequences is growing explosively. We feel it is important to collect this significant information, correlate it into a unified whole and interpret it.
>
> *Dr Margaret Dayhoff* ”

This vision led to the production of the “Atlas of Protein Sequence and Structure” [21], or more simply Atlas, the first biological database. The first edition of Atlas was published in 1965 and contained 65 protein sequences [22].

Early versions of the Atlas database were severely hindered by the necessity to distribute the database in book format. These restrictions eased with the uptake of the Internet and the introduction of more portable and high-density storage media; it was in the 1980s that the Atlas database started to be distributed electronically [20]. The change in the way the database was distributed and stored coincided with the database being renamed to the Protein Information Resource (PIR) [9]. In addition to the PIR, a number of other biological databases began to appear in the 1980s. Many of these databases, such as Swiss-Prot, The European Molecular Biology Laboratory (EMBL) and GenBank, remain active today. Analysing the history of these databases is aided by the availability of archival versions, allowing database releases from different decades to be downloaded and analysed. This historical data allows, for example, the growth of data over time to be quantified and analysed.

[1] This work was undertaken in an era of computing where punch-cards were used for data storage. Given the storage restrictions of punch-cards, Dr. Dayhoff stored amino acids with a single-letter code, rather than their typical 3-letter abbreviation. This approach also aided the readability of sequence alignment. For example, the amino acid Lysine can be identified by the one-letter symbol “K”. The one-letter codes developed by Dr. Dayhoff gained significant popularity, and still remain in common usage.

In addition to the growth of individual databases, there has also been an exponential increase in the number of database papers published since 1980 [23]. Bolser *et al.* analysed this increase by extracting the number of papers published each year with "database" in their title from PubMed. These results, which have been extended to show results from additional years, are illustrated in Figure 2.1. This figure shows that in 2012 a total of 1,440 papers were published compared to only two in 1975.

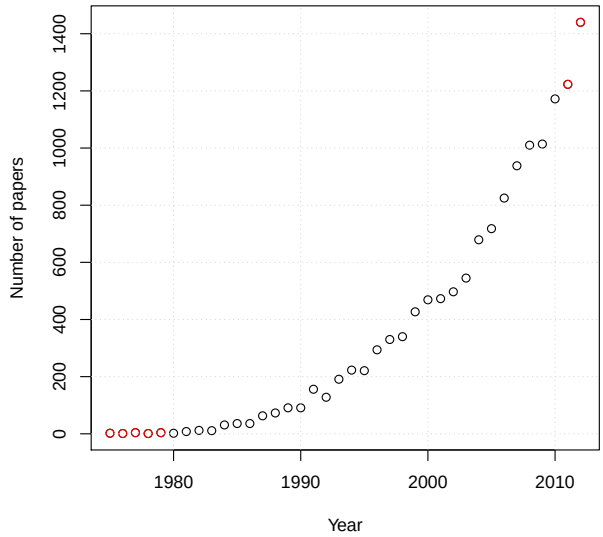


Figure 2.1: Figure illustrating the number of papers published each year with the word "database" in the title. This search was restricted to PubMed, which only indexes biomedical literature. Original data taken from the analysis by Bolser *et al.* and extended to include results from additional years (highlighted in red) [23].

The growth in this literature suggests that biological databases started to gain popularity within the 1980s; a trend that continues to grow within the biomedical domain. Due to the increased popularity, Oxford University Press introduced "Database: The Journal of Biological Databases and Curation" in 2009, which is a journal dedicated to biological databases [24, 25]. However, whilst the study by Bolser *et al.* provides an indication into the growth of biological databases, the results are flawed; not all papers will introduce or describe a new biological database. For example, a database may be created without publishing an associated paper (such as MEDLINE), the corresponding paper may not mention "database" in the title (as is true with OrthoDB [26]) or results may become redundant with databases that publish yearly or bi-yearly updates

(such as BIOGRID [27–30]). A more accurate estimate of the number of databases in existence can be drawn from collections that aim to collate biological databases.

These collections of biological databases, more colloquially referred to as "databases of databases", include MetaBase [23], the Bioinformatics links directory [31] DBcat [32] and the Wikipedia article "List of biological databases" [33]. However, the most recognised collection is the molecular biology database collection, which is produced by Nucleic Acids Research (NAR). The NAR collection is produced annually, with the 2013 release listing a total of $1,512$ online databases [34].

2013 marked the 20th annual database issue of NAR. The first issue, in 1993, consisted of 24 articles[2] compared to the 176 articles listed in 2013. However, the published articles do not relate to the total number of databases collected by NAR; this is recorded, since 1999, in a corresponding summary paper. These summary papers show a yearly increase in the number of databases listed by NAR, as illustrated in Figure 2.2. The growth in the NAR list has also been met with an increase in the number of downloads and citations that each issue receives [35–37].

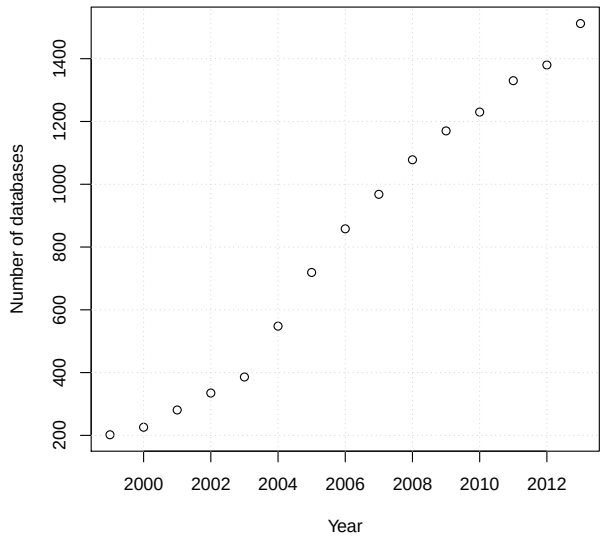


Figure 2.2: The number of databases listed in the annual NAR database collection between 1999 and 2013. Data taken from the yearly summary papers [34–48].

Although its growth continues, the NAR list does not capture all available biological

[2]There were two issues prior to 1993, however they were not formally labelled as database issues. These issues were published in 1991 and 1992 and contained 18 and 19 articles, respectively.

databases; only databases judged to be "of high value to the biological community" are eligible for inclusion [46, 49]. Eligibility is based upon a number of factors, including: the databases usability; perceived sustainability; scope and; data and curation quality [37, 48]. Therefore, there are likely a number of other databases in existence. For example, in 2000 the DBcat collection listed 513 databases compared to NAR which only listed 226.

A more recent comparison of NAR and DBcat is not possible; the last update to the DBcat catalogue was in 2000, with the catalogue subsequently being shut down[3]. Comparisons to alternative catalogues, such as MetaBase, show incompleteness (i.e. they contain fewer databases) or essentially mirror the NAR list.

The DBcat catalogue became inaccessible when the French INFOBIOGEN centre, who was responsible for hosting and maintaining DBcat, was forced to close due to severe budgetary difficulties [36, 51]. Obtaining sustainable funding is a challenge faced by the majority of biological databases [44, 52], which has resulted in a number of major databases coming under threat of closure. For example, the Swiss-Prot, PROSITE and ENZYME databases encountered a significant funding crisis in 1996 [53, 54], whilst the Sequence Read Archive (SRA) [55] faced closure after the National Center for Biotechnology Information (NCBI) discontinued funding in 2011 [56, 57]. Further resources, such as EcoCyc, face funding difficulties at the time of writing [58].

Databases under the threat of closure may be successful in securing funding from alternative sources. For example, the SRA obtained public funding from the National Institute of Health (NIH), with continued funding from the DDBJ and the European Bioinformatics Institute (EBI), allowing it to continue [48, 55]. Public funding provides significant support to many databases, however other funding approaches may be employed if a database is unsuccessful in obtaining public funding [59]. For example, since 2011 access to the Kyoto Encyclopedia of Genes and Genomes (KEGG) [60] File Transfer Protocol (FTP) site requires a subscription [61]. Whilst imposing subscription fees or becoming commercialised may allow the database to survive, it can be detrimental to the databases success. For example, such actions may result in its removal from the NAR list and can discourage users from participation, whilst the

[3]Historical versions of the DBcat catalogue were accessed via the Internet Archive [50] in 2006.

interests of a database may become dictated by commercial priorities [62].

Although a number of resources have proved to be sustainable, it is inevitable that some databases will be forced into closure. For example, the Genome DataBase (GDB) and Peptidome [63] were forced to shut down in 2008 and 2011, respectively [46, 48, 64]. Databases are also liable to close due to other factors. For example, the International Protein Index (IPI) closed in 2011 after it became superseded by the Ensembl and UniProtKB databases [65, 66].

Database closures are reflected in the NAR list with obsolete, unreachable or unsuitable databases being removed [45]. Overall, the databases listed within NAR have shown reasonable resilience [36]; only 44 databases were removed in 2013 [34], whilst 20 were removed in 2012 [48]. In total only 161 out of $1,673$ databases have been removed from the NAR list.

The NAR list requires ongoing maintenance to ensure the resources listed are both suitable and reachable. For example, a database URL may move, with previous URLS becoming outdated (i.e. suffer from URL decay). High levels of URL decay ($\sim$35%) have been shown in an analysis of MEDLINE abstracts [67, 68], whilst in 2011 URLs for 30 databases listed in NAR had to be updated [48].

Whilst factors such as database commercialisation and URL decay can be troublesome for users, the closure of a database may significantly affect a user's research. For example, the potential closure of Biological Magnetic Resonance Bank (BMRB) and EcoCyc has resulted in numerous scientists writing letters of support for these resources opposing their funding cuts [58, 62, 69]. Although ensuring a database remains sustainable involves dealing with numerous issues, these letters of support from users emphasises their value.

However, the closure of a particular database does not just affect individual users; there may also be repercussions for other biological databases. For example, many databases often contain information derived from other databases [35]. This includes the neXtProt database [70] which derives its raw data from UniProtKB, whilst UniProtKB derives a large portion of its data from International Nucleotide Sequence Database Collaboration (INSDC). Databases deriving raw data from external databases are

becoming increasingly common as many new databases are developed to focus exclusively on model organisms [34, 47]. For example, neXtProt, which was created in 2011, focuses exclusively on *Homo sapiens*, whilst EcoCyc [71] is dedicated to the bacterium *Escherichia coli*.

Additionally, many databases also provide cross-references to other databases [52]. The number of cross-references may be vast; for example, UniProtKB cross-references 140 other databases [72, 73], whilst the majority of databases listed by NAR provide GO annotations [37]. Therefore, the potential reference map can become far-reaching, as illustrated in Figure 2.3.

The attachment of cross-references to database entries in UniProtKB is an example of *metadata*. Metadata can be defined as "data about data" [75] and can cover many different types of data. For example, a database entry can include descriptive metadata (e.g. name of the entry) structural metadata (e.g. links to related database entries) and administrative metadata (e.g. entry version number) [76]. The attachment of metadata can aid the discovery of information, organisation of data and support the archiving of data.

In order to facilitate these features, metadata is often presented in a standardised format (or schema), examples of which include Dublin Core [77] and Darwin Core [78]. Metadata schemes are generally recorded in a machine readable language such as Resource Description Framework (RDF) or eXtensible Markup Language (XML).

Many biological databases provide some form of metadata. For example, UniProtKB provides an RDF format for entries which includes provenance and evidence metadata [79], whilst the neXtProt database provides metadata regarding the criteria used to grade the quality of an annotation [70].

Accessing and querying the metadata within a database is often done through web services. For example, metadata in the UniProtKB and European Nucleotide Archive (ENA) databases can be accessed in a RESTful manner [80]. This allows programmatic access to data, allowing end users to obtain specific data sets for further analysis and for other databases to integrate or cross-reference data. Such examples, that are provided UniProtKB, include identifying database entries that: are integrated within a given

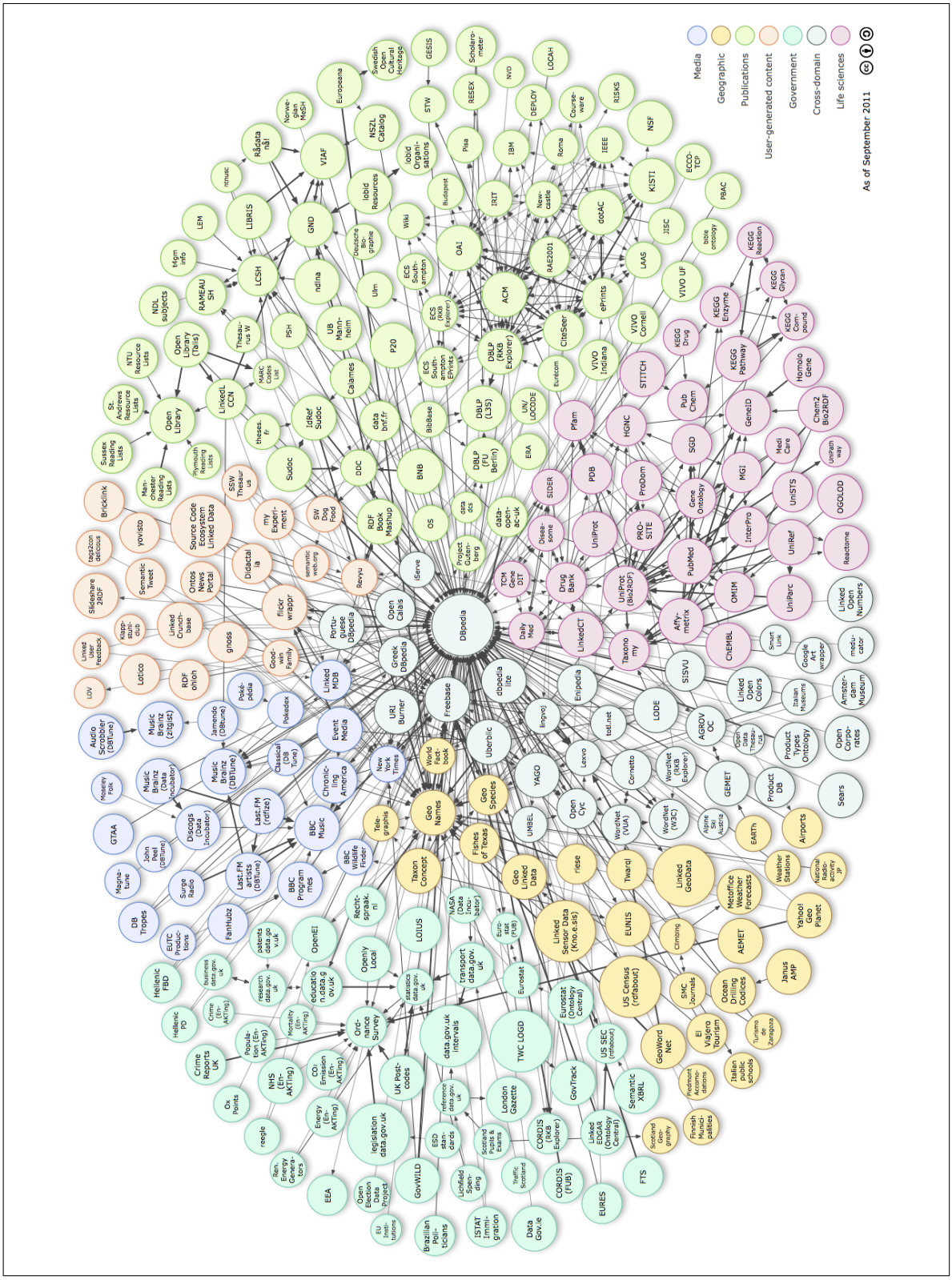


Figure 2.3: Cloud diagram illustrating the relationship between various resources. Each resource is represented as a circle, the size of which indicates the resources size, whilst the colour indicates the resources type. Arrows are shown between resources sharing over 50 links, with arrow thickness illustrating a higher number of links. The Figure is taken from [74], which is based upon resources listed on the Data Hub (`http://datahub.io`).

date range; have a cross-reference to a certain external database; and that are related to given ENA accession [81].

Within biological databases the types and amount of metadata that are available, like their resources and focus, will vary. However, a common feature of many databases is the provision of *annotation*. Annotation is used as a mechanism to document and convey known knowledge about the underlying biological data to a reader. In the following section (Section 2.2) we discuss the different types of biological annotation, and explore various curation methods.

2.2 Biological Annotation

There are many different types of biological data. The information that the data contains, and the way it is best represented, can be diverse [82]. For example, DNA sequences can be represented as a string of characters, whilst results of a microarray experiment are often stored as scanned graphical images [83]. Data in this form can be referred to as the *raw data*; that is the uninterpreted data relating to the biological organism [17]. In practice, the actual definition of raw data varies depending upon its usage and context. For example, output from DNA sequencers, such as the Ion Torrent [84], contain supporting trace data in addition to the nucleotide sequence, providing information such as quality scores. Whilst trace data are beneficial in certain applications, such supplementary information is not required to represent the underlying biology or to perform many common tasks such as a Basic Local Alignment Search Tool (BLAST) search [85]. Moreover, because the size of the trace data is so much larger than sequence data, often only the latter is kept, and a DNA sequence is often referred to as the raw data, giving the *ad hoc* definition of uninterpreted data or as close to it as is practical.

In itself, a piece of raw biological data often provides little information to a reader [15, 86]. For example, whilst an amino acid sequence is required for performing a BLAST search, a user cannot identify the function of the protein by looking at the raw amino acid sequence. Therefore, biological data stored in databases is often supplemented with additional information. For example, DNA sequences within the EMBL database [87] are stored with appropriate literature references and keywords, whilst ArrayExpress stores data from microarray experiments, including detailed information relating to the organism sampled [88]. This addition of information to raw data is known as *annotation*.

Within biology, annotation has a somewhat different meaning than in general usage: EMBL-EBI define it as "the process of attaching additional information to biological entities" [89]; compared to the Collins English Dictionary definition — "a note added in explanation, etc, esp of some literary work" [90]. In the context of biology, annotation may be one of many different types. For example, an annotation in Swiss-Prot could

refer to a number of items: a protein function; associated diseases; sequence conflicts; post-translation modification(s) or similarities to other proteins, amongst others [91]. Here we describe the main types of biological annotation.

Broadly, annotations can be categorised into two groups [89, 92]: *structural*, or "low-level", annotation and *functional*, or "high-level", annotation:

Structural Annotation

Structural annotation is the initial annotation performed on raw biological data. Broadly, this annotation involves identifying key elements and features of the sequence [89]. For example, the structural annotation of a nucleotide sequence involves the identification of Open Reading Frames (ORFs), which paves the way for identifying genes [93, 94].

The identification of ORFs can be considered part of nucleotide-level annotation; a prerequisite stage for protein-level annotation. Protein-level annotation involves classifying the gene into a protein family and determining the proteins' nomenclature [95].

Structural annotations are typically generated and produced in a form that allows for computational interpretation [92].

Functional Annotation

Functional annotation refers to the knowledge about elements and features identified by structural annotation [89]. Functional annotation encompasses a number of different types of information. For example, it can refer to a list of publications related to the sequence, cross-references to related entries in a database or the author responsible for identifying the sequence [16].

Perhaps the most valuable component of functional annotation is the attachment of current knowledge relating to the underlying biology. These annotations are used to convey information, such as a protein's known function and involvement in disease, to the reader [96, 97]. Therefore, functional annotations are generally composed of free text (English), and thus intended for human, rather than

machine, comprehension [92]. For clarity, the free text segment of functional annotation is referred to as *textual annotation.*

Collectively, the process of creating both functional annotation and structural annotation is often referred to as *genome annotation* [98]. Genome annotation consists of various stages and is often generated by annotation pipelines [99]. Annotation pipelines package various tools and techniques into a series of steps that can provide both structural annotation and functional annotation. An example of a general pipeline, showing the various stages required to perform genome annotation, is shown in Figure 2.4.

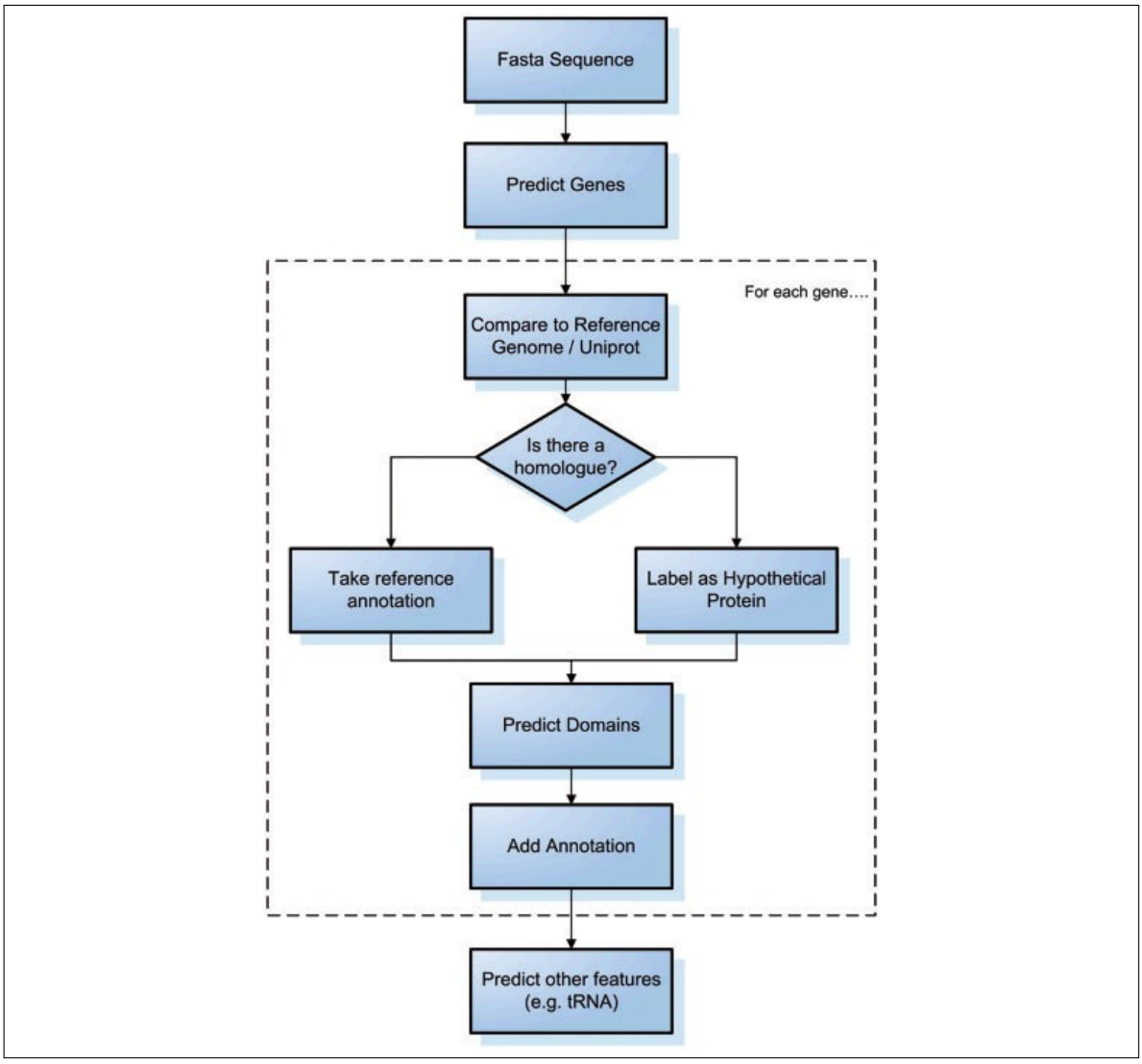


Figure 2.4: Illustrating the various stages and steps required to provide genome annotation for a sequenced genome. Figure taken from [1].

For consistency and clarity, we use the following annotation definitions throughout this thesis:

Structural annotation: Identification and classification of features from raw biological sequences.

Functional annotation: The attachment of high-level knowledge to structural annotation.

Textual annotation: The free text component of functional annotation.

Genome annotation: The combination of both functional annotation and structural annotation.

The process of creating annotations is known as the *curation process*, which is performed by *curators*; individuals who are trained to provide accurate and complete annotations [100]. Becoming a curator requires significant training – curators typically possess a Ph.D. [101] and can take many months to train. For example, curators within the FlyBase database [102] undergo six months training [103], whilst annotator training in the Protein Data Bank (PDB) [104] takes between two and six months [105]. Curators dealing exclusively with biological data are often termed biocurators, but the terms "annotators", "curators" and "biocurators" are effectively synonymous, often used interchangeably [106].

Biological data and their associated annotations are typically deposited into *biological databases*. As discussed in Section 2.1, biological databases vary in their approaches to storing, managing and curating data. For example, curation in FlyBase is performed on an article basis [103], whilst curation in UniProtKB is done on an entry (i.e. a protein) basis [2]. Database resources dedicated to curation also vary. For example, the worldwide Protein Data Bank (wwPDB) [107] reports having twenty annotators [105], whilst The Universal Protein Resource (UniProt) list 43 dedicated biocurators on their staff pages [108] and the *Candida* Genome Database [109] list only three dedicated curators [110].

Whilst the curation process and strategies inevitably vary between databases, a distinction between two main categories of curation can be made: *computational* and *manual* [111].

Manually curated annotations may be produced either exclusively, or by a combination of, the original authors of a sequence, dedicated curators or the wider scientific community [16, 112].

A condition imposed by many biological journals is that authors make available the data associated with their publication. In the case of biological data, this requirement mostly involves the requirement for raw biological data to be submitted to a relevant database[4]. For example, many journals require the submission of DNA sequences to one of the databases in INSDC; i.e. ENA, GenBank or the DDBJ [114]. Indeed, this is a requirement for submission to the Journal of Biological Chemistry (JBC) [115], Nature [116], and NAR [117].

Databases, such as GenBank and DDBJ, that accept raw biological data vary in the levels of annotation that are required with a submission. For example, data is submitted to UniProtKB via SPIN [118], a submission system that requests information from authors regarding the proteins' name and properties [119]. However, whilst annotations can be submitted by authors to UniProtKB, approximately 98% of entries in UniProtKB are made up of sequences derived from translations of the coding regions in the INSDC [120], resulting in most entries being created in their entirety by UniProtKB curators (the manual curation process employed by UniProtKB is summarised in Figure 2.5). In some cases there are databases, such as the PDB, where authors are entirely responsible for the production of annotations. The role of PDB annotators is to ensure quality and consistency by performing integrity checks on the submitted data and providing assistance to authors [105].

Finally, some databases generate or supplement their annotations through involvement with the wider scientific community. For example, in April 2011 [121], the protein family database Pfam [122] began to replace curator-produced annotations with Wikipedia articles [123]. There have also been annotation efforts, termed *annotation jamborees*, that involve collecting a group of scientists together to perform genome annotation [124]. Examples of species that have been involved in annotation jamborees include *Drosophila melanogaster* (common fruit fly) [125, 126], *Ciona in-*

[4]However, papers have been published where the authors failed to submit such data to a relevant database [113].

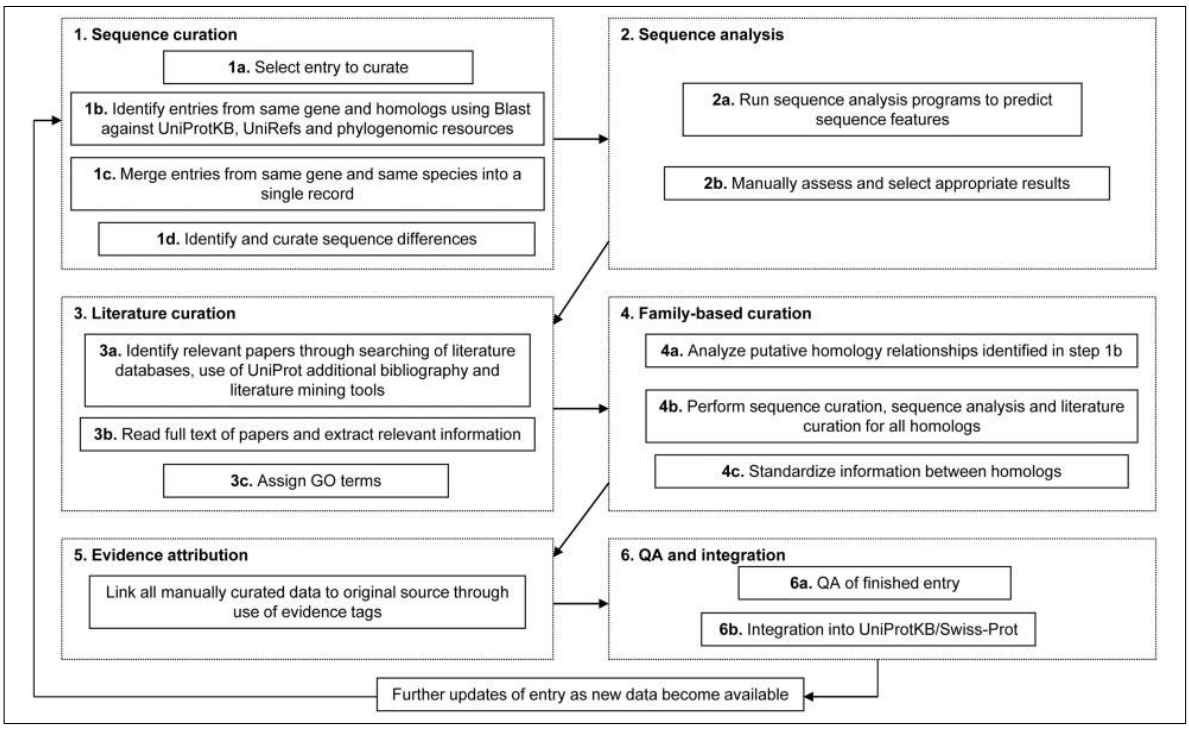


Figure 2.5: An overview of the manual curation process used by UniProtKB curators. Image taken from [2].

testinalis (vase tunicate) [127], *Escherichia coli* strain K-12 (bacteria) [128] and *Oryza sativa* ssp. *japonica* cultivar Nipponbar (Japanese rice) [129, 130].

Annotations that are manually curated are generally held as the 'gold standard' [131, 132]. However, manual curation is a significant bottleneck. For example, in the FlyBase database it can take between two and four months for an article to be manually curated [103]. With ever-increasing amounts of raw data, this bottleneck means that manual curation is insufficient; Baumgartner *et al.* estimated in 2007 that it will take over three decades to provide structural annotation for mice, with this estimate including the assumption that no new data will be added [133]. Therefore, many databases have introduced computational methods to assist, complement or replace manual curation.

One database that incorporates automated methods in a number of forms is UniProtKB. UniProtKB consists of two sections: Swiss-Prot which is manually curated and reviewed; and TrEMBL, which is automated and unreviewed [134]. Whilst this provides a clear distinction between automated and manually curated entries, UniProtKB do incorporate a number of automated methods to assist their manual curation pro-

cess [2]. For example: curators are presented with an interface showing potential sequence features, which are predicted by a variety of tools (Figure 2.5, step 2a); text mining tools are used to help identify relevant publications for literature curation (Figure 2.5, step 3a); and curated entries undergo a number of automated quality checks prior to release (Figure 2.5, step 6a).

These automated methods are used in a complementary manner, rather than as a manual replacement. For example, the automated quality checks (Figure 2.5, step 6a) are performed prior to a manually performed review. These automated checks ensure an entry conforms to a number of syntactic and biological rules, allowing manual review to be more focused on the entries semantics. This can be considered analogous with performing a spell-check on a document before proof reading; it aids the user in identifying spelling (i.e. syntactic) errors, whilst the user can focus more on the content trying to be conveyed (i.e. semantic).

In UniProtKB the usage of Natural Language Processing (NLP) and text mining tools[5] are also used in an assistive capacity, aiding the curator in identifying relevant literature (Figure 2.5, step 3a). This usage of text mining and NLP tools has gained popularity as the mechanism for reporting results and advancements in biological knowledge continues to be predominantly literature based. Coinciding with biological data, biological literature is also growing rapidly, reportedly at a double-exponential rate [136]. For example, PubMed [137] indexes bibliographic information for over 22 million biomedical publications [138], $1,047,008$ of which were published in 2012.

With over 1 million articles a year being published, it is clear that the uptake of text mining tools is inevitable; manually scouring for relevant articles requires significant time and effort. Automated methods for the identification of relevant literature can significantly reduce the time and effort required by curators. For example, the usage of the PreBind literature-mining system within the Biomolecular Interaction Network Database (BIND) database [139] saved a total of 176 days of curator time over a calendar year [140].

With a current average of over $87,000$ new articles being indexed in PubMed per

[5]Broadly, NLP and text mining are concerned with the ability to computationally analyse natural language, allowing relevant information within the text to be identified and extracted [18, 135]. NLP and text mining are areas of active research.

month, it is likely that a number of these will be of interest to a biological database, and relevant for curation. It appears that in many cases a significant number of articles are identified for curation. For example, BIND, which extracted biomolecular interaction and pathway information from the literature, identified that approximately 1,950 relevant interactions were published in the literature on a monthly basis[6] [141]. Similarly, high levels of relevant articles appear apparent in UniProtKB. Version 2013_03 of UniProtKB reported a total of 829,697 journal references in Swiss-Prot [142] and 15,365,925 in TrEMBL [143], being amassed over 25 years. However, this figure is redundant, highlighting that references are reused within UniProtKB. A more accurate figure, published in 2010, reported that UniProtKB contained around 228,000 distinct (i.e. non-redundant) PubMed citations, 67% of which were manually curated (i.e. around 150,000 citations in Swiss-Prot), with a further 350,000 citations contributed from external databases, that had not yet undergone annotation [144].

Although many databases have been able to reduce the time and effort required to identify relevant literature, the extraction of information from the literature is still required. Within UniProtKB, for example, each identified publication is read in full by a curator who then subsequently extracts relevant data from the paper into various forms of annotation [2]; this extraction process is an overwhelming bottleneck. Like UniProtKB, many other databases utilise NLP and text mining tools to automate, or semi-automate, the extraction of information during annotation curation [145, 146]. For example, the BIND database utilised PreBIND to extract protein-protein interaction information [140]; the LSAT system [147], which extracts information regarding transcript diversity from MEDLINE abstracts, was utilised by the alternative splicing database [136, 148]; whilst UniProtKB has recently extended its usage of text mining techniques to extract post-translational modifications information [149].

The usage of automated tools varies between databases. For example, in Swiss-Prot, the curation process is mostly manual, and is entirely manually reviewed, whilst in Xenbase the literature curation process involves 17 steps, ten of which are automated [150]. However, there are a number of databases, such as TrEMBL, whose curation process

[6]This survey was performed over three months in early 2004, covering 110 journals. The interactions extracted were taken from papers published in 79 journals.

is entirely automated.

Automated annotations have become an essential feature for biological databases wishing to avoid unannotated entries; it has simply become impossible for manual curation to keep pace with the number of sequenced genomes [151]. The approaches and tools used to computationally generate annotations vary between databases [98]. For example, automated annotations in GO [152] are generated from sequence similarity and keyword mapping techniques [132, 153], whilst annotations in TrEMBL are generated from rule based systems, such as Unified Rule (UniRule) and Statistical Automatic Annotation System (SAAS) [2, 154].

Given the various techniques that are used for curating annotations, it is to be expected that the quality will vary both within a database and between databases. For example, UniProtKB is often regarded as providing high quality annotation, whilst manually curated annotations are generally perceived as being of better quality than those produced computationally. However, how is this quality quantified and what methods are available to a user for assessing annotation quality? These questions form the basis for the following section (Section 2.3), which explores existing methods for analysing annotation quality and correctness.

2.3 Annotation Quality and Correctness

Quality is a commonly used term. For example, over 700,000 articles in PubMed contain the word quality, whilst it is in the top 550 most frequently used words within Wikipedia[7]. However, it is common for quality to be mentioned or discussed without any quantification; what makes an annotation be of high quality? How is one database of better quality than another? How can this quality be measured?

The Collins English Dictionary provides nine definitions for quality, which include "a distinguishing characteristic, property, or attribute" and "degree or standard of excellence, esp a high standard" [155]. These varying definitions suggest that quality is dependent upon a given application or end user. For example, a user requiring a specific database feature will likely rate a database providing the feature as better than an alternative database which does not; this behaviour would fit with the definition "a distinguishing characteristic, property, or attribute".

These definitions of quality are understandably broad; a definition of quality has to cover a range of areas and applications. In certain applications the usage of quality can be specifically defined or measured. For example, the quality of a sequenced DNA region can be based on the correctness of each nucleotide base. This approach avoids ambiguity as each base has either been sequenced correctly or incorrectly. This is the basis of the Phred quality score [156, 157], which is a widely used approach to quantitatively describe DNA sequence quality [158].

Whilst a DNA sequence can be evaluated based on the number of bases correctly sequenced, many situations have a multiple number of variables that can impact quality. For example, the quality of a printed image can be measured in Dots Per Inch (DPI), but other features, such as the type of ink and paper brightness, will also impact the image quality. In other situations, there may be competing metrics. For example, the quality of an academic journal can be evaluated based on its impact factor score, but this approach has come under criticism as a meaningful metric, resulting in alternative metrics, such as Eigenfactor (ES), being developed [159]. These cases highlight how a quality measure may not be entirely conclusive, or how there may be no definitive or

[7]This figure was obtained from an analysis of Wikipedia performed in Section 3.1

agreed measure for certain applications. These issues are also evident in higher levels of biological data, such as functional annotation and databases.

Additionally, for higher levels of biological data it is also important to acknowledge that there are different aspects that can be assessed. For example, when discussing the quality of an annotation that is attached to a database entry we may consider the correctness of the annotation, the breadth of knowledge that the annotation covers or the richness of the information contained within the annotation. This distinction is important as just because an annotation may be very detailed does not mean that is necessarily correct.

For annotation quality we consider two main components: correctness and richness. Although related, these components can be assessed individually; studies often evaluate the richness of an annotation without considering its correctness or vice versa. Therefore we first explore quality approaches that primarily consider the richness of annotation and then investigate those studies which base their definition of quality on annotation correctness.

As previously discussed, a database can contain various forms of functional annotation. One form of functional annotation that is commonly provided by protein databases is GO annotation. For example the UniProtKB curation process involves attaching relevant GO terms to a UniProtKB entry as part of the Gene Ontology Annotation (GOA) program [160].

GO terms provide a standardised vocabulary and allow the molecular function, biological process and cellular component of a gene to be described. GO annotation involves the attachment of a GO term to a database entry along with the terms source and an evidence code relating to the evidence of the source [161]. There are a large number of publications regarding GO, with a subset of these papers focused on analysing the quality of GO terms and annotation.

One of the most cited of these studies developed a metric, named GO Annotation Quality (GAQ), that provides a quantitative measure of GO annotation for a set of gene products [162]. This GAQ score is calculated based on the number of available GO annotations, the level of detail of the annotation and the associated evidence codes.

The GAQ score has been utilised in a number of situations, including: being used to score GO annotation in AgBase [163]; being used to evaluate the improvement of annotation quality following the re-annotation of germinal vesicle oocyte and cumulus proteins [164]; and being used to merge redundant GO annotations [165].

The GAQ score calculates the level of detail by combining the breadth (coverage of gene product) and the depth (level of detail) for the terms in GO. However, while deeper nodes within an ontology are generally more specialised, these measures are problematic; first GO has three root domains and second an ontology, such as GO, is a graph not a tree, therefore depth is not necessarily meaningful. GAQ also utilises annotation evidence codes to score an annotation, however the GO annotation manual explicitly states that evidence codes should *not* be used in this way [166], describing rather the type of evidence not its strength. Although evidence codes should not be used as indicators of strength, a number of approaches use evidence codes as a basis for assessing an annotations reliability [167–169].

Evidence codes have been used in this manner as electronically inferred annotations are generally perceived to be of dubious quality. However, a recent study suggests that these annotations are actually of similar quality to annotations curated manually from non-experimental annotations [170]. This study analysed the changes in electronic annotations between different versions of GOA, determining if an electronic annotation remains the same in a subsequent release or is changed; if it is changed it may be replaced by another electronic annotation, an experimental annotation or removed entirely. The authors used this information to determine the reliability, coverage and specificity of electronic annotations, concluding that the quality of electronic annotations are improving over time.

A similar study is presented by Gross *et al.* who explored the evolution of GO annotations [171]. This study analysed how the stability of GO annotations change over time, suggesting that older and more stable annotations have more reliability than those that undergo more regular changes. However, concerns about this approach are raised by Clarke *et al.*, as the quality of the ontology structure is not taken into account [172]. Instead, Clarke *et al.* evaluate the quality of GO annotations by determining if a given GO term is correctly attached to a set of expected genes, based on experimental

knowledge from the literature.

One drawback of the approach presented by Clarke *et al.* is that users need to manually identify and extract information related to a GO term from the literature. A more straightforward approach that requires minimal user intervention is presented by Kalankesh *et al.* [173]. This approach is based on the application of Zipf's Law to the frequency of individual GO terms within sets of annotation from GOA. The application of Zipf's law returns a single value, α, that is used to evaluate a set of annotations.

Although GO annotations provide numerous benefits, not all databases contain GO annotation. As previously discussed, the types and amount of functional annotation can vary between databases, although most databases will carry some form of unstructured free text (i.e. textual annotation). However, unlike GO annotation, the amount of research exploring textual annotation is very limited. Indeed, one study which analysed annotation in Swiss-Prot explicitly omitted textual annotation from their analysis as it "is not easily machine-parseable" [174].

To evaluate the quality of textual annotation we could consider adapting the approaches used for assessing GO annotations. However, most of these approaches rely upon additional information, which would make such an analysis problematic. For example, not all resources use evidence codes and these codes are not comparable between resources [175]; likewise, it is not generally possible to use methods based on an ontological hierarchy for non-ontological resources. However, the Zipf's Law approach, presented by Kalankesh *et al* [173], is an exception as it only requires information regarding the occurrences of GO terms and is not dependent upon any specific database features. Given the simplicity of Zipf's Law, it could be applied to the frequency of words within textual annotation.

Although Zipf's Law is reasonably simplistic, this is not necessarily a limitation or disadvantage. For example, there are a number of studies that have devised different methods for analysing the quality of Wikipedia articles (see, for example, [176–178]). These approaches exploit specific wiki features and are of varying complexity, but their results are no more effective than an approach that is based simply on counting the number of words within an article [179].

Devising a quality metric to analyse Wikipedia involves overcoming certain difficulties [180], namely: handling a vast quantity of exponentially increasing articles; regularly changing content; a diverse number of subject areas; contributors of varying backgrounds; and abuse, such as vandalism. Many of these properties are also shared by textual annotation, suggesting that a quality approach that is relatively straightforward would be most suitable.

A further commonality shared between textual annotation and Wikipedia is that, in their simplest form, they are composed of English text[8]. This allows the quality approaches based on word count and Zipf's law to be applied to both resources. Additionally, other quality approaches that are based on, or only require, English text can be considered.

However, although the word count metric appears to indicate quality within Wikipedia articles, we are unsure how meaningful it would be if applied to textual annotation; an article in Wikipedia can contain thousands of words, whilst entries in, for instance, UniProtKB generally contain textual annotation consisting of less than 100 words. Indeed, it is possible that the analysis of Wikipedia was based on a gold standard dataset that contained more words simply by coincidence.

Although assessing a textual annotation based on its size may be flawed, counting how frequently each individual word occurs can be used to indicate *information content*; words which occur rarely generally contain more relevant and specific information than those occurring frequently. This statistic, first described in 1972 [181], is commonly known as Inverse Document Frequency (IDF) and when combined with term frequency can be used to identify the most relevant words within a document [182, 183]. Applying Term Frequency-Inverse Document Frequency (TF-IDF) to textual annotation in a given database entry would highly rank an annotation if it consists of words that occur rarely within the overall database, but frequently within the database entry. For example the term 'aspirin' occurs very infrequently within UniProtKB as whole, yet it occurs relatively frequently within a handful of entries; TF-IDF would determine that 'aspirin' has a high value within these entries.

TF-IDF could be used to determine a quality score for a textual annotation based on

[8]References to Wikipedia are to the English-language edition of Wikipedia.

its information content in relation to the rest of the database. However, the implementation of this would not be straightforward, as annotations are not static; each time an annotation is changed the TF-IDF scores for the whole database corpus would need to be recalculated. Although this recalculation could be automated, the constant growth of textual annotation means it would be computationally demanding. Further, obtained scores would be database-specific, meaning that results from different resources could not be compared directly.

Alternative metrics for evaluating a text include the Simple Measure of Gobbledygook (SMOG) [184, 185] and Flesch-Kincaid [186, 187]. To calculate SMOG for a given text, the formula

$$1.043\sqrt{30 \times \frac{\text{number of polysyllables}}{\text{number of sentences}}} + 3.1291$$

is used, whilst the formula

$$0.39 \times \left(\frac{\text{total words}}{\text{total sentences}}\right) + 11.8 \times \left(\frac{\text{total syllables}}{\text{total words}}\right) - 15.59$$

is used to calculate Flesch-Kincaid. These formulae return single values that estimate how many years of education a reader requires in order to understand a text. Broadly, the higher the value the more complicated and difficult the text is to understand [188]. However, these metrics are based around readability, or reading-age; that is the literary quality of the text, rather than the quality of the subject matter. Additionally, readability metrics, including SMOG and Flesch-Kincaid, have come under criticism [189]. For example, one criticism of these metrics is that the readability score for a collection of words will always be the same regardless of its ordering; the syntactic and semantic complexity of a text is not evaluated. Further, computationally identifying syllables is not a trivial matter, whilst the SMOG formula is dependent upon texts consisting of 30 or more sentences.

Whilst these approaches have limitations, they both remain popular choices for assessing readability [188]. For example, Flesch-Kincaid is used within Microsoft Word to grade a documents readability [190]. Although the two metrics were both introduced

over forty years ago, no subsequent metric has been developed to render them redundant. This is likely caused by text evaluation being difficult and highly subjective; many factors can impact a texts quality, including: the subject matter of the text; its formatting and presentation; the inclusion of supporting illustrations; and its suitability for its intended purpose [191]. Although related research fields, such as NLP, remain active there is still no definitive measure for evaluating text.

All of the surveyed quality approaches highlight how there is no agreed approach for assessing annotation quality, with the definition of quality varying between these approaches. Out of the surveyed options, the Zipf's Law approach appears most plausible for assessing textual annotation. This approach uses the principle of least effort to define quality; an annotation that places the least effort onto the reader is deemed high quality. We explore this principle and Zipf's law in more detail in the Chapter 3.

Whilst the principle of least effort provides one definition of quality, it possible that an annotation may contain errors, omissions or outdated information yet still be classified as being of high quality. Such errors can impact whether a text is deemed suitable for its intended purpose (i.e. its *fitness for use*) [191]. Conversely, however, it is possible to have an annotation that contains no errors or omissions but is incomprehensible, making it difficult for a user to fully understand the information that the author is trying to convey. Therefore, we can consider annotation quality as having two parts: the effectiveness of the annotation in conveying the information stored about the underlying biology; and the correctness and accuracy of this underlying biology. Therefore, we also need to consider the latter of these parts.

As also encountered with the analysis of existing quality metrics, there is a distinct lack of studies that have explicitly explored the accuracy and correctness of textual annotation. However, there are studies that explore the correctness of other forms of annotation. For example, one study, by Artamonova *et al.*, developed an association rule mining approach to identify errors in the organelle, organism, feature table, database cross-reference and keyword annotations in Swiss-Prot [174]. Their approach generates rules based on this annotation to identify an outcome that corresponds to a set of features. For example, the rule `Nuclear localization` ⇒ `Origin: eu-`

`karyota` states that entries which are annotated as being located within the nucleus would also be annotated as having a eukaryotic origin. As there are often exceptions to a rule, each generated rule has an associated strength rating to indicate how many entries adhere to the rule. The authors extracted annotations that were exceptions to high value rules and manually analysed their correctness, with a number of these being found erroneous; the authors estimate that the error rate of these Swiss-Prot annotations is between 33% and 43%. A similar error rate (28% to 30%) was also estimated for curated GO sequence annotations in GOSeqLite [169].

Another approach, by Keseler *et al.*, analysed manually curated assertions in the EcoCyc and CGD databases [192]. This study involved manually cross-checking information within these databases against the source publication that it was attributed to, with information that was not found in its attributed publication being deemed erroneous. In total 633 assertions were manually analysed, with 10 (i.e. 1.58%) being found erroneous. Another study, which analysed partial Enzyme Commission (EC) codes in KEGG, found an error rate of 6.8% within the annotations of *Escherichia coli* genes [193].

With these studies identifying errors in various forms of annotation, it is likely that errors will also exist in textual annotation. However, although these studies have used different approaches, none can be suitably applied to textual annotation. For example, the study by Artamonova *et al.* relies upon annotations which take fixed values, whilst the studies of GO and KEGG both rely upon an ontological hierarchy. However, whilst the approach by Keseler *et al.* could be adapted to analyse textual annotation, it is heavily dependent upon domain experts manually analysing annotation.

These approaches have resulted in different error rates being estimated, with a more recent study suggesting that the actual error rate in many databases is higher than these previous estimates [194]. This study also suggests that error rates are increasing over time and identified the issue of *error propagation*.

These two findings are not unrelated; new annotations are commonly based upon preexisting annotations [18] and can lead to 'chains of misannotation' [195]. Such chains occur when, for example, a new annotation is produced based on an erroneous annotation, which was also created from an incorrect annotation. Whilst such errors

can percolate within a database it is also possible for errors to propagate to external databases. For example, an annotation with the incorrect spelling of 'synthase' in RefSeq was found to have propagated to entries in UniProtKB, KEGG and xBASE [1], whilst the incorrect classification of at least 18 proteins in PIR propagated to entries in UniProtKB, RefSeq and GenBank [196].

In attempt to measure error propagation one study developed a probabilistic framework to theoretically model the percolation of structural annotation errors [195, 197]. Although this model determined that the accuracy of annotations decline as the size of the database increases, it does not provide information about how individual annotations propagate.

Whilst databases copy data and annotation as a matter of protocol, it can be hard, even impossible, to determine the source from which an error has propagated [162, 195], meaning that propagated errors may never be corrected [169]. This is due, in part, to the source, or *provenance*, of an annotation not always being documented, or being done so inadequately [198]. The provision of provenance information is often regarded as being essential for validating the correctness of data [199–201].

The issue of annotations being propagated without formal provenance has been acknowledged by studies exploring potential approaches for allowing the recording of provenance information [198, 202]. However, these studies do not suggest methods for identifying the provenance or propagation of existing annotations. Whilst certain annotations, such as GO annotations, are stored along with their source, many textual annotations do not have any provenance evidence attributed to them. This has resulted in many existing textual annotations being stored with no formal provenance information; as the identification of an annotations provenance can help determine its correctness, is there a way to reconstruct the provenance of these existing textual annotations?

Incomplete or missing provenance is not just restricted to textual annotation, with studies from an increasing number of fields exploring methods to reconstruct missing provenance [203]. For example, there have been attempts at identifying the provenance and subsequent flow and reuse of information between newspaper articles [204, 205] and between websites [206] (see [203] for further examples). However, unlike many ap-

proaches that rely upon additional information for reconstructing missing provenance, the studies exploring information flow in newspapers and websites were achieved using only textual data. For example, the approaches used techniques such as TF-IDF to identify information reuse between different documents at the sentence level.

Another approach that could be used to help identify the provenance of a text is plagiarism detection software. Plagiarism detection software, such as Turnitin [207] and Dupli Checker [208], are commonly used for checking the originality of a piece of text by identifying any sections of the text which appear to have been copied from another source. The textual resources which are scanned by plagiarism detection software vary, but academic journals and Web pages are commonly included.

Given the importance of identifying plagiarism within areas such as academia, there have been continued developments in the tooling and support that such software provides. For example, Turnitin can highlight areas of a text that have been duplicated and provide a link to the original source along with summary statistics. Providing this additional information helps a user to gain confidence in the predicted provenance.

However, whilst there are advantages to using plagiarism detection software for identifying annotation provenance, utilising the software would be troublesome as many are commercial projects with proprietary detection algorithms. This means we would be restricted to the existing features implemented in the software and unable to make any refinements or extensions to adequately analyse textual annotation.

As textual annotations can be propagated between database entries as a matter of protocol, these approaches provide confidence that the provenance of a textual annotation within a database can also be reconstructed. Specifically, it appears reasonable that the provenance and propagation of an annotation can be inferred by tracking the occurrence of individual sentences over time. We discuss possible methods to allow the provenance and propagation of a sentence to be explored in Chapter 5.

With our work involving the analysis of textual annotation, it is of benefit to analyse how annotations are curated and managed within each database. However, as discussed in Section 2.1, there are over $1,500$ active databases making it infeasible to analyse each database in-depth. Further, given the broad range of specialisations and features,

it is not meaningful to attempt an all-encompassing generic analysis. Therefore, an extended analysis and focus is given to the UniProtKB database in the following section (Section 2.4).

UniProtKB is an ideal resource to analyse. The database is well established, maintained, at the core of many other databases and has extensive support and documentation. Crucially, UniProtKB contains both manual and automated forms of textual annotation which is the foundations required for the work within this thesis.

2.4 The Universal Protein Resource

The Universal Protein Resource (UniProt) is a resource created and maintained by the UniProt Consortium [134]. The UniProt consortium, formed in 2002, is composed of three members: the European Bioinformatics Institute (EBI); the Protein Information Resource (PIR) and the Swiss Institute of Bioinformatics (SIB). Similar to the INSDC, which provides the key databases responsible for storing and disseminating DNA and RNA sequences [209], the UniProt consortium aims to be the authoritative resource for amino acid sequences, with their main goals stated within their mission statement:

> “The mission of UniProt is to support biological research by providing a freely accessible, stable, comprehensive, fully classified, richly and accurately annotated protein sequence knowledgebase, with extensive cross-references and querying interfaces.
>
> *The UniProt Consortium* [210] ”

To achieve these goals, UniProt contains four key components:

The UniProt Knowledgebase (UniProtKB)

UniProtKB is a comprehensive database of protein sequences which consists of the manually curated Swiss-Prot and its automated counterpart TrEMBL [134].

The UniProt Archive (UniParc)

UniParc is a repository of non-redundant protein sequences. Sequences are mostly taken from public resources and are stored in UniParc with a unique identifier and their origin [211].

The UniProt Reference Clusters (UniRef)

UniRef provides clusters of sequences from UniProtKB based on their similarity. These clusters reduce sequence redundancy, aiding sequence similarity searches, and are available at resolutions of 100%, 90% and 50% identity [212].

The UniProt Metagenomic and Environmental Sequences (UniMES)

UniMES, introduced in 2007, is a database providing metagenomic and environmental data [213, 214].

As illustrated in Figure 2.6, sequence data enters the UniProt databases through UniParc, with UniMES and UniRef also depending upon data contained within UniProtKB. UniProtKB is often regarded as the core database of UniProt [215], due to the additional information that it provides; most UniProtKB entries include functional annotation and cross-references to other databases. Additionally, UniProt aim to attach as much information as possible to each protein entry in UniProtKB [216]. It is for these reasons that UniProtKB is often regarded and used as the gold standard for protein information and annotation [95, 169, 217, 218].

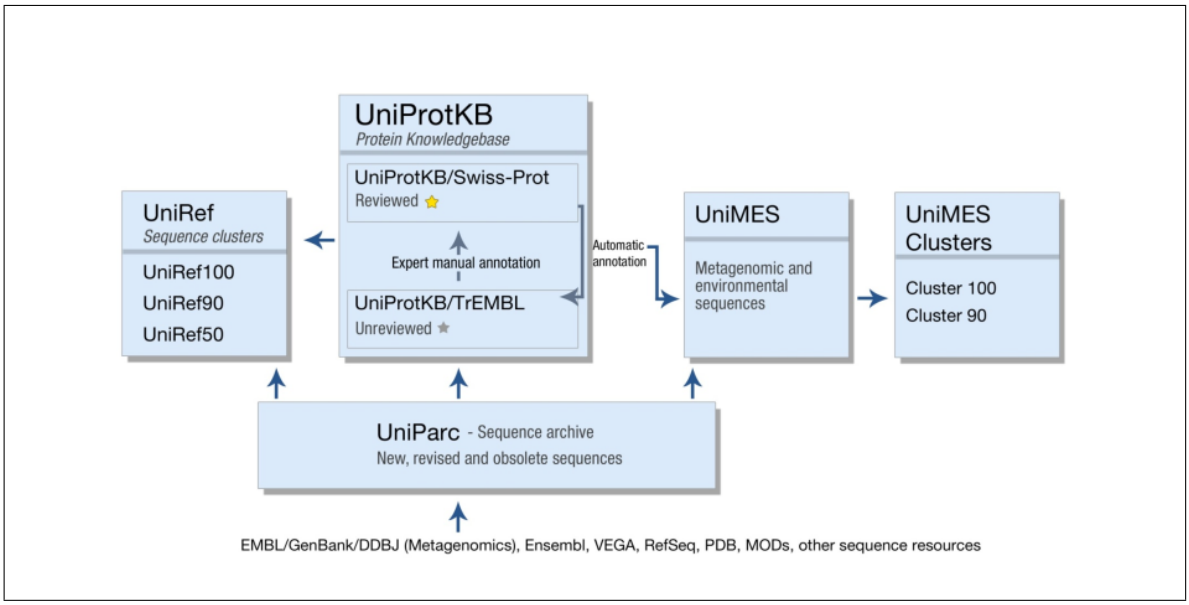


Figure 2.6: Relationship between the four UniProt databases: UniProtKB; UniParc; UniRef & UniMES. Image taken from [219].

Of the UniProt databases, UniProtKB is the only database that contains unique annotation: UniParc only stores sequence and provenance information; UniRef takes its annotation from UniProtKB; whilst UniMES is only available in FASTA format. Therefore, as the work within this thesis is focused on textual annotation, UniProtKB needs to be explored in greater detail.

UniProtKB consists of two sections: Swiss-Prot, which is reviewed and manually curated, and TrEMBL which is unreviewed and automatically annotated. These sec-

tions are technically referred to as UniProtKB/Swiss-Prot and UniProtKB/TrEMBL, however this has not always been the case. The first release of UniProtKB came in December 2003 following the formation of the UniProt consortium [220], yet the initial version of Swiss-Prot was created in 1986 by Amos Bairoch [221].

The birth of Swiss-Prot came about due to a number of issues encountered by Bairoch when using and distributing PIR. These involved difficulties with parsing the PIR file format and a lack of supporting cross-references and textual annotation. Bairoch raised these issues with PIR, but after they went unaddressed he opted to create a new database, which he named Swiss-Prot [222]. The Swiss-Prot file format was developed to closely resemble EMBL files, allowing the lack of annotation and parsing issues to be overcome. These files consist purely of American Standard Code for Information Interchange (ASCII) text and are referred to as the *flat file* format. The flat file format still remains in common usage today, with the flat file for Swiss-Prot entry P0C9E9 shown in Figure 2.7.

Swiss-Prot flat files are line-oriented with each line encoding a specific piece of information, which is indicated by the two letters at the start of the line. The order, occurrence and format of these lines follow a strict structure, as detailed in the UniProtKB user manual [223]. For example, each entry must start with an ID line, that can only occur once, and consists of the entries name, entries status (i.e. reviewed or unreviewed) and the sequence length. This structure allows entries to be machine parsable, whilst also remaining human readable. There are total of 26 possible line types, as summarised in Table 2.1.

Although early versions of Swiss-Prot were only produced as flat files, more recent releases of the database are stored in a relational database and are distributed in multiple formats, including XML and FASTA. It is also possible to download and view individual entries on the UniProtKB website. For example, Figure 2.8 shows the Web view for Swiss-Prot entry P0C9E9, which corresponds to the flat file shown previously in Figure 2.7.

Since its introduction Swiss-Prot has shown significant growth, as illustrated in Figure 2.9a. Recent releases of Swiss-Prot have now surpassed half a million entries; a significant increase from 25 years ago when Swiss-Prot contained just under $9,000$

```
ID   1001R_ASFWA             Reviewed;         124 AA.
AC   P0C9E9;
DT   05-MAY-2009, integrated into UniProtKB/Swiss-Prot.
DT   05-MAY-2009, sequence version 1.
DT   11-JAN-2011, entry version 2.
DE   RecName: Full=Protein MGF 100-1R;
GN   OrderedLocusNames=War-018;
OS   African swine fever virus (isolate Warthog/Namibia/Wart80/1980)
OS   (ASFV).
OC   Viruses; dsDNA viruses, no RNA stage; Asfarviridae; Asfivirus.
OX   NCBI_TaxID=561444;
OH   NCBI_TaxID=6937; Ornithodoros (relapsing fever ticks).
OH   NCBI_TaxID=85517; Phacochoerus aethiopicus (Warthog).
OH   NCBI_TaxID=41426; Phacochoerus africanus (Warthog).
OH   NCBI_TaxID=273792; Potamochoerus larvatus (Bushpig).
OH   NCBI_TaxID=9823; Sus scrofa (Pig).
RN   [1]
RP   NUCLEOTIDE SEQUENCE [LARGE SCALE GENOMIC DNA].
RA   Kutish G.F., Rock D.L.;
RT   "African swine fever virus genomes.";
RL   Submitted (MAR-2003) to the EMBL/GenBank/DDBJ databases.
CC   -!- FUNCTION: Plays a role in virus cell tropism, and may be required
CC       for efficient virus replication in macrophages (By similarity).
CC   -!- SIMILARITY: Belongs to the asfivirus MGF 100 family.
CC   ----------------------------------------------------------------------
CC   Copyrighted by the UniProt Consortium, see http://www.uniprot.org/terms
CC   Distributed under the Creative Commons Attribution-NoDerivs License
CC   ----------------------------------------------------------------------
DR   EMBL; AY261366; -; NOT_ANNOTATED_CDS; Genomic_DNA.
PE   3: Inferred from homology;
KW   Complete proteome.
FT   CHAIN         1    124       Protein MGF 100-1R.
FT                               /FTId=PRO_0000373169.
SQ   SEQUENCE   124 AA;  15327 MW;  C1EC5CC5B6D3E2BB CRC64;
     MVRLFRNPIK CIFYRRSRKI QEKKLRKSLK KLNFYHPPED CCQIYRLLEN VPGGTYFITE
     NMTNDLIMVV KDSVDKKIKS IKLYLHGSYI KIHQHYYINI YMYLMRYTQI YKYPLICFNK
     YYNI
//
```

Figure 2.7: An example entry from UniProtKB shown in flat file format.

Line code	Content	Occurrence in an entry
ID	Identification	Once; starts the entry
AC	Accession number(s)	Once or more
DT	Date	Three times
DE	Description	Once or more
GN	Gene name(s)	Optional
OS	Organism species	Once or more
OG	Organelle	Optional
OC	Organism classification	Once or more
OX	Taxonomy cross-reference	Once
OH	Organism host	Optional
RN	Reference number	Once or more
RP	Reference position	Once or more
RC	Reference comment(s)	Optional
RX	Reference cross-reference(s)	Optional
RG	Reference group	Once or more (Optional if RA line)
RA	Reference authors	Once or more (Optional if RG line)
RT	Reference title	Optional
RL	Reference location	Once or more
CC	Comments or notes	Optional
DR	Database cross-references	Optional
PE	Protein existence	Once
KW	Keywords	Optional
FT	Feature table data	Once or more in Swiss-Prot, optional in TrEMBL
SQ	Sequence header	Once
	Sequence data	Once or more
//	Termination line	Once; ends the entry

Table 2.1: Each of the possible line types that can appear in a UniProtKB flat file, presented in the order in which they must appear in an entry. Data taken from the UniProtKB user manual [223].

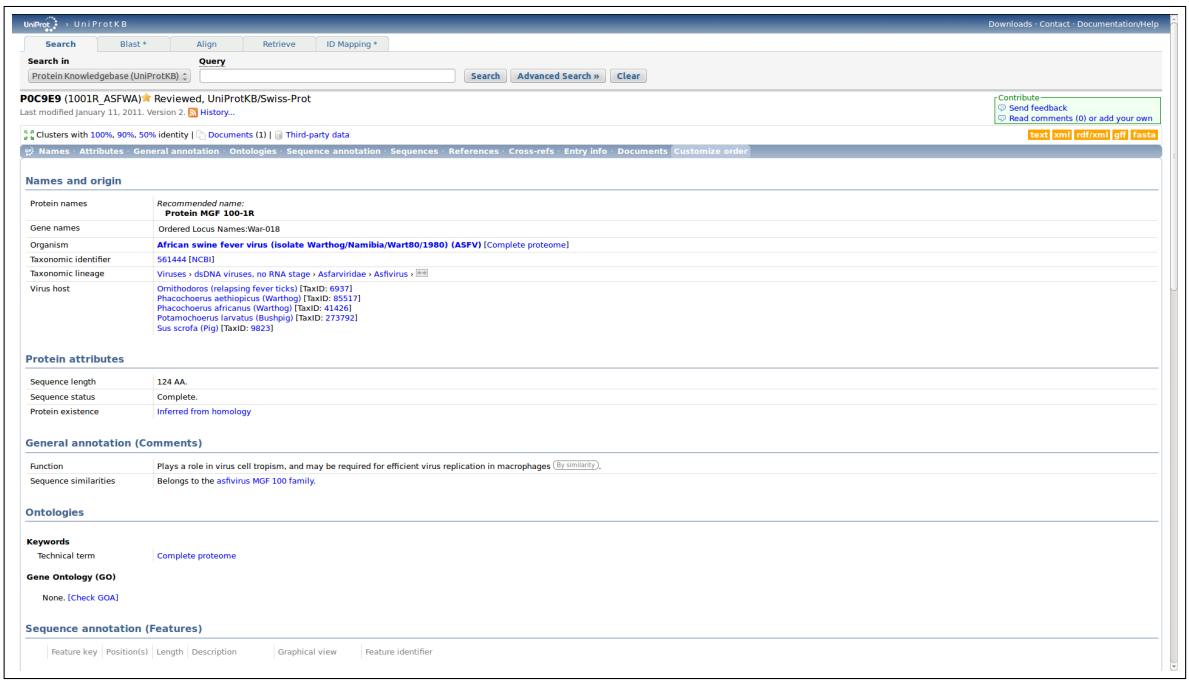
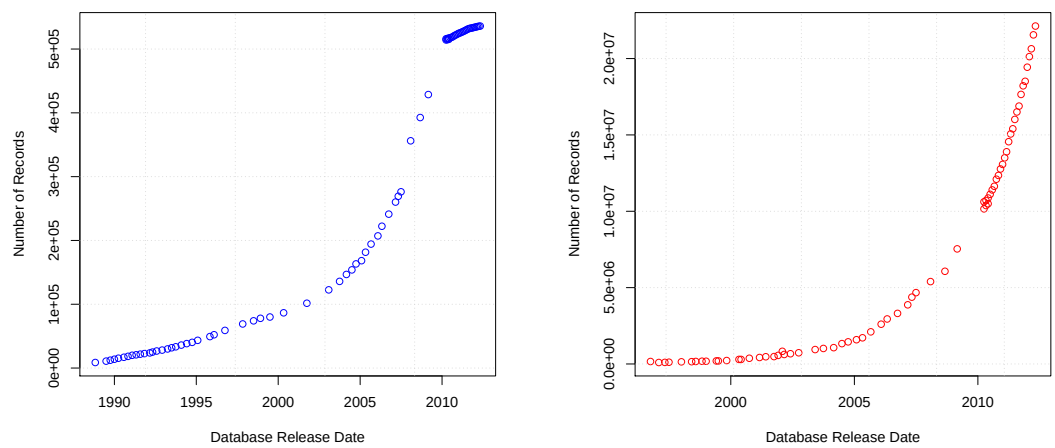


Figure 2.8: Screenshot of entry P0C9E9 shown on the UniProtKB website. This entry corresponds to Web view for the flat file shown in Figure 2.7.

entries. It is likely that Bairoch never imagined how large Swiss-Prot would become as he had intended to step away from the production of Swiss-Prot in late 1986, but opted to remain temporarily involved due to the volume of data being submitted and added to the database at that time [224]. This temporary involvement has spanned the entire life of Swiss-Prot, with Bairoch currently acting as a senior scientific adviser to Swiss-Prot having stepped down as director in 2009 [225].

The growth of data added to Swiss-Prot is mainly a result of sequences from genome projects, such as the human genome project, being deposited into the database [226]. Developments and improvements in sequencing technology has resulted in more genome projects being initiated, leading to a constantly increasing volume of data entering Swiss-Prot [227]. This is likely to continue with the cost of sequencing a genome becoming more accessible; the National Human Genome Research Institute (NHGRI) have reported that sequencing costs have dropped from $\sim \$100M$ to $\sim \$5,000$ over the last 10 years [228]. However, whilst sequencing speeds and costs are constantly reducing, manual curation cannot be realistically sped up without a compromise to curation quality, or an influx of curators to match the databases' growth.

This increase of sequence data left Swiss-Prot at a crossroads. The provision of man-



(a) Swiss-Prot (b) TrEMBL

Figure 2.9: Total number of records (entries) in (a) Swiss-Prot and (b) TrEMBL over time.

ually curated entries ensures Swiss-Prot continues to provide high quality data, yet manual curation is both time consuming and costly. Therefore, Swiss-Prot could either delay the release of sequence data until they were manually curated or incorporate the data into Swiss-Prot with minimal analysis. Neither of these solutions were deemed acceptable. Therefore TrEMBL, a database providing computer generated annotations, was introduced in November 1996 [91]. The TrEMBL database allows sequence data to be made rapidly available without impacting the quality of Swiss-Prot.

The TrEMBL database was built upon a series of tools developed by Thure Etzold at EMBL that translated Coding Sequences (CDSs) from EMBL into flat file entries. The TrEMBL database is named after these tools, which Thure named TREMBL ("TRanslation from EMBL") [91, 229]. Sequence data in TrEMBL is still obtained from the automatic translation of CDSs from the INSDC databases.

The TREMBL tools were extended so that TrEMBL and Swiss-Prot shared a number of similarities. For example, both databases follow the same flat file structure and sequences already incorporated within Swiss-Prot are excluded from the TrEMBL database. This allows interoperability between the databases from both a data management and usability point of view. For example, searching the UniProtKB website for "pax6" returns results that integrate both Swiss-Prot and TrEMBL entries in the

same interface, as illustrated in Figure 2.10, whilst a standardised format allows entries to become merged.

Results Customize

› Show only reviewed (62) (UniProtKB/Swiss-Prot) or unreviewed (470) (UniProtKB/TrEMBL) entries

› Restrict term "pax6" to gene name (195), protein name (278), web resource (1)

Entry	Entry name	Status	Protein names	Gene names
P63015	PAX6_MOUSE	★	**Paired box protein Pax-6**	**Pax6** Pax-6 Sey
P26367	PAX6_HUMAN	★	**Paired box protein Pax-6**	**PAX6** AN2
P47237	PAX6_CHICK	★	**Paired box protein Pax-6**	**PAX6**
P63016	PAX6_RAT	★	**Paired box protein Pax-6**	**Pax6** Pax-6 Sey
O73917	PAX6_ORYLA	★	**Paired box protein Pax-6**	**pax6**
O18381	PAX6_DROME	★	**Paired box protein Pax-6**	**ey** pax6 CG1464
P55864	PAX6_XENLA	★	**Paired box protein Pax-6**	**pax6**
P47238	PAX6_COTJA	★	**Paired box protein Pax-6**	**PAX6**
Q1LZF1	PAX6_BOVIN	★	**Paired box protein Pax-6**	**PAX6**
G3V3Q9	G3V3Q9_HUMAN	☆	**Paired box protein Pax-6**	**PAX6**
F7BI93	F7BI93_MOUSE	☆	**Paired box protein Pax-6**	**Pax6**
B2FDB2	B2FDB2_MOUSE	☆	**Paired box protein Pax-6**	**Pax6**
B7ZBX1	B7ZBX1_MOUSE	☆	**Paired box protein Pax-6**	**Pax6**

Figure 2.10: Screenshot of the UniProtKB website showing the search results for the term "pax6". The status column distinguishes between Swiss-Prot (gold star) and TrEMBL (silver star) entries. The search can also be easily refined to only show reviewed (Swiss-Prot) or unreviewed (TrEMBL) entries.

The merging of entries in Swiss-Prot and TrEMBL is done to avoid database redundancy. Within Swiss-Prot, entries representing the same gene product in a species are merged, whilst entries with identical sequences from the same species are merged in TrEMBL [230]. When entries become merged the first accession number, which is decided based on alphanumerical order, becomes the primary accession with all additional accession numbers becoming secondary accessions. Entries are also subject to deletion and becoming demerged (i.e. split into two or more entries), although this is relatively rare. The deletion of a UniProtKB entry can occur if the original nucleotide sequence is removed from the source INSDC database or if a protein was found to have been incorrectly predicted [231]. Certain entries in Swiss-Prot began to be demerged in 2010 after the merging policy was updated; originally entries in Swiss-Prot from the same species were merged if they shared identical sequences [232, 233].

The introduction of TrEMBL, combined with Swiss-Prot, meant that all publicly available protein sequences could be covered by the two databases. This changed the way in which sequence data was added to Swiss-Prot, with entries now being moved from TrEMBL into Swiss-Prot and manually annotated. However, the number of entries in TrEMBL quickly outgrew Swiss-Prot, with TrEMBL continuing to grow exponentially as shown in Figure 2.9b. For example, UniProtKB/TrEMBL Version 2012_05

contains over 22 million entries compared to just over half a million entries contained in UniProtKB/Swiss-Prot Version 2012_05. Therefore, curators have to prioritise the order in which entries in TrEMBL are to undergo manual review. Entries are more likely to be chosen for curation if they are involved in current research, are from a model organism or requested by a user [234, 235].

Choosing an entry to undergo manual curation is the first step of the Swiss-Prot curation pipeline [2, 235, 236], as outlined in Figure 2.5. This manual curation process involves six key stages:

Sequence curation

As previously discussed, entries in Swiss-Prot become merged when sequences from the same gene and organism are identified; this is the main output from sequence curation. To identify these entries, a BLAST search is performed against other UniProtKB entries. Before merging two entries any discrepancies between the sequences, such as alternative splicing, are identified and documented within the merged entries annotation.

A BLAST search is also performed to identify entries containing homologous sequences. These identified entries are compared and analysed to identify any sequence errors.

Additionally, a curator takes ownership of the sequence they are going to curate by "locking" the associated entry. This avoids concurrent access from other curators, eliminating any conflicts that could arise.

Sequence analysis

This curation stage is performed through an analysis platform that provides various tools and features to the curator. The features of a sequence including its topology, domain and post-translational modifications, are predicted by executing approximately 25 analysis tools. These prediction tools can be executed automatically by the analysis platform.

Results from these predictions are assessed by a curator with features deemed suitable and relevant to the entry being added as annotation. This annotation

is generated automatically by the analysis platform through a series annotation rules. For example, Figure 2.11 shows the annotations that would be generated for a predicted domain, with an example of the annotation generated from a series of predictions shown in Figure 2.12.

General rule information [?]

Accession	PRU00494
Dates	16-NOV-2005 (Created) 9-FEB-2009 (Last updated, Version 3)
Data class	Domain
Predictors	PROSITE; PS51150; AGOUTI_2
Name	Agouti domain
Function	The agouti domain is a Cys-rich C-terminal module, which is responsible for melanocortin receptor binding activity in vitro.

Propagated annotation [?]

Comments [?]

Similarity	Contains # agouti domain.

Cross-references [?]

PROSITE	PS60024; AGOUTI_1; 1;

Keywords [?]

case <FTTag:disulf>
Disulfide bond
end case

Features [?]

From: PS51150

Key	From	To	Description	Tag	Condition	FTGroup
DOMAIN	from	to	Agouti #			
DISULFID	1	16	By similarity	disulf	C-x*-C	
DISULFID	8	22	By similarity	disulf	C-x*-C	
DISULFID	15	33	By similarity	disulf	C-x*-C	
DISULFID	19	40	By similarity	disulf	C-x*-C	
DISULFID	24	31	By similarity	disulf	C-x*-C	

Figure 2.11: An example of an annotation rule used within the curators analysis platform. This allows annotation to be automatically generated for a predicted domain that matches the rule. Image taken from [235].

Literature curation

Annotations predicted from the previous stage are supplemented with experimental data obtained from the literature. This involves the curator identifying relevant journal articles, which are read in full, with the experimental data then being extracted.

As previously discussed, text mining tools are utilised by Swiss-Prot curators to identify relevant publications. Curators also find relevant articles through the UniProt Additional Bibliography, which includes references imported from external databases, and by manually searching bibliographic databases such as MEDLINE.

The majority of the information extracted from the literature is added to the general annotation section (textual annotation) of the entry. This section includes

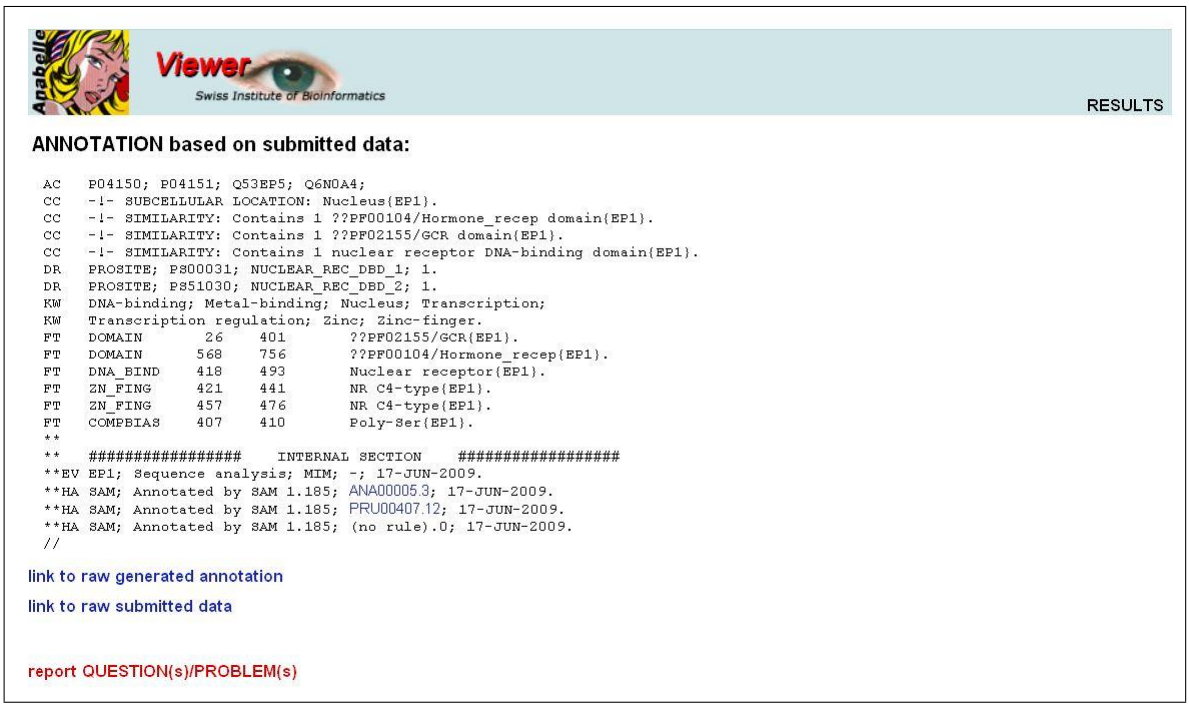


Figure 2.12: An example of an annotation generated from annotation rules for a series of predicted features. Image taken from [235].

information about the proteins function, subcellular location and involvement in disease. The extracted data is also used to attach relevant GO terms to the entry.

Each journal article used within this stage is added to the reference list of the entry.

Family-based curation

Family-based curation involves standardising the annotation between homologous proteins to provide consistency between database entries. This involves propagating annotation between the newly curated entry and the homologous entries identified during the sequence curation stage. This also includes propagating any relevant GO terms.

Evidence attribution

The annotation attached to an entry can originate from various sources and methods. Therefore, curators include various evidence codes within the entry to allow data to be attributed to its original source and record the annotation method used. These evidence codes are included within the XML format of an

entry but are not available within the flat file format and only partially available through the entry view on the UniProtKB website.

Quality assurance and integration

Following the attachment of evidence, each entry undergoes a number of automated and manual checks prior to its integration into Swiss-Prot. Automated syntax checks are performed to verify the structure and formatting of the entry prior to a manual review by a senior curator. The senior curator manually checks that all relevant sequence features and literature have been included and that annotation has been attached correctly.

Upon completion, the entry is "unlocked" by the curator so it can be accessed and altered by other curators.

Manual curation of an entry is an ongoing process, with these six stages reapplied to an entry when new data becomes available. Ongoing annotation is also true for entries in TrEMBL, however the curation process differs from Swiss-Prot as it is automated. Entries in TrEMBL are annotated using two complementary systems [2, 134, 237]:

The Unified Rule System (UniRule)

UniRule is rule-based system that is composed of three (originally independent) systems: High-quality Automated and Manual Annotation of Proteins (HAMAP) [238–240]; the PIR rule systems [241, 242]; and RuleBase [154, 243]. Rules in these systems are manually curated and are described in the UniRule format, which follows a similar structure to the UniProtKB flat file format. Each individual rule specifies a set of conditions and corresponding annotations. Annotations are suitable for propagation to entries that meet the conditions of a rule.

The amount of annotation that is eligible for propagation varies between each rule but can include: keywords; GO annotation; protein and gene names; and textual annotation. Similarly the number of conditions contained within a rule can also vary, but may include checks for a particular sequence feature or family membership.

Before each UniProtKB release, annotations predicted by the UniRule system are evaluated against Swiss-Prot annotations. Inconsistencies between the annotations results in the corresponding annotation rule being flagged for manual review.

The Statistical Automatic Annotation System (SAAS)

Like UniRule, SAAS, previously known as Spearmint [244], is also a rule based system. However the generation of rules, and subsequent annotation, is fully automated by applying the C4.5 data mining algorithm to annotation in Swiss-Prot.

The C4.5 algorithm is an algorithm that generates a decision tree, based on a training dataset, and is used to classify an unknown piece of data; essentially a classification is based upon the information contained within the training dataset.

In SAAS the training dataset contains information about proteins in Swiss-Prot, such as their taxonomy and sequence length. A decision tree is built using this information and is used to determine which entries contain annotation that should be propagated. SAAS is used to predict a variety of annotation in TrEMBL, with the exception of sequence features and protein names. Annotations in UniProtKB that are produced from either UniRule or SAAS have the associated annotation rule attached as evidence, as illustrated in Figure 2.13. This evidence is only available in the XML and web-view of an entry.

The decision trees produced by SAAS are also generated as annotation rules. These automated rules can potentially be moved in the UniRule system, or be used as a basis for manually curated rules.

The automated nature of TrEMBL means that annotations are reproduced for each database release, ensuring that annotations are based on the latest rules and that entries contain the latest knowledge. UniProtKB is currently released on a monthly basis, with its version number corresponding to its release date (version numbers take the form YYYY_MM). This release cycle and numbering scheme was introduced in 2010, starting from UniProtKB Version 2010_04 [245]. Prior to this, UniProtKB made

Figure 2.13: Web view of TrEMBL entry G0HS88. Sections of the entry, such as the subcellular location and protein names, were automatically generated from annotation rules. The accession number of the annotation rule is added after the predicted annotation and is hyperlinked to the annotation rule.

a distinction between major and minor versions with version numbering taking the form XX.YY, where X and Y are positive integers (XX.0 indicated a major release).

The first major release of UniProtKB was in 2004. Prior to this UniProtKB release, Swiss-Prot and TrEMBL versions were not synchronised and released independently. The last independent releases of Swiss-Prot and TrEMBL were Versions 43 and 26, respectively, with UniProtKB Version two[9] containing Swiss-Prot Version 44 and TrEMBL Version 27. The release history of these databases, along with the number of entries contained in each entry, is shown in Table 2.2.

The history of Swiss-Prot, TrEMBL and UniProtKB releases can make discussing and referencing the databases problematic. Therefore, for clarity, the following naming approach is used when referencing database versions:

- **Swiss-Prot** – Refers to Swiss-Prot entries prior to the formation of the UniProt Consortium.
- **TrEMBL** – Refers to TrEMBL entries prior to the formation of the UniProt Consortium.
- **UniProtKB** – Refers to the combination of both Swiss-Prot and TrEMBL datasets.

[9]UniProtKB Version 2 was the first major release of UniProtKB.

When necessary the fully qualified names, UniProtKB/Swiss-Prot or UniProtKB/TrEMBL, will be made explicit to avoid ambiguity. This naming scheme also allows post-UniProtKB versions of UniProtKB/Swiss-Prot and UniProtKB/TrEMBL to be referred to with the same version number.

Number of Entries	Swiss-Prot Version	Release Date	TrEMBL Version	Number of Entries
8702	Swiss-Prot Version 9	November 1988	-	-
10856	Swiss-Prot Version 11	July 1989	-	-
12305	Swiss-Prot Version 12	October 1989	-	-
13837	Swiss-Prot Version 13	January 1990	-	-
15409	Swiss-Prot Version 14	April 1990	-	-
16941	Swiss-Prot Version 15	August 1990	-	-
18364	Swiss-Prot Version 16	November 1990	-	-
20024	Swiss-Prot Version 17	February 1991	-	-
20772	Swiss-Prot Version 18	May 1991	-	-
21795	Swiss-Prot Version 19	August 1991	-	-
22654	Swiss-Prot Version 20	November 1991	-	-
23742	Swiss-Prot Version 21	March 1992	-	-
25044	Swiss-Prot Version 22	May 1992	-	-
26706	Swiss-Prot Version 23	August 1992	-	-
28154	Swiss-Prot Version 24	December 1992	-	-
29955	Swiss-Prot Version 25	April 1993	-	-
31808	Swiss-Prot Version 26	July 1993	-	-
33329	Swiss-Prot Version 27	October 1993	-	-
36000	Swiss-Prot Version 28	February 1994	-	-
38303	Swiss-Prot Version 29	June 1994	-	-
40292	Swiss-Prot Version 30	October 1994	-	-
43470	Swiss-Prot Version 31	February 1995	-	-
49340	Swiss-Prot Version 32	November 1995	-	-
52205	Swiss-Prot Version 33	February 1996	-	-
59021	Swiss-Prot Version 34	October 1996	TrEMBL Version 1	104955
-	-	February 1997	TrEMBL Version 2	116379
-	-	May 1997	TrEMBL Version 3	126995
-	-	July 1997	TrEMBL Version 4	139208
69113	Swiss-Prot Version 35	November 1997	-	-
-	-	January 1998	TrEMBL Version 5	166361
-	-	June 1998	TrEMBL Version 6	177757
74019	Swiss-Prot Version 36	July 1998	-	-
-	-	August 1998	TrEMBL Version 7	193860
-	-	November 1998	TrEMBL Version 8	224543

Continued on next page

Number of Entries	Swiss-Prot Version	Release Date	TrEMBL Version	Number of Entries
77977	Swiss-Prot Version 37	December 1998	-	-
-	-	January 1999	TrEMBL Version 9	221422
-	-	June 1999	TrEMBL Version 10	244862
80000	Swiss-Prot Version 38	July 1999	TrEMBL Version 11	245761
-	-	November 1999	TrEMBL Version 12	276472
86593	Swiss-Prot Version 39	May 2000	TrEMBL Version 13	353156
-	-	June 2000	TrEMBL Version 14	351834
-	-	October 2000	TrEMBL Version 15	431424
-	-	March 2001	TrEMBL Version 16	489620
-	-	June 2001	TrEMBL Version 17	540195
101602	Swiss-Prot Version 40	October 2001	TrEMBL Version 18	558150
-	-	December 2001	TrEMBL Version 19	636825
-	-	March 2002	TrEMBL Version 20	700753
-	-	June 2002	TrEMBL Version 21	751148
-	-	October 2002	TrEMBL Version 22	821014
122564	Swiss-Prot Version 41	February 2003	TrEMBL Version 23	921952
-	-	June 2003	TrEMBL Version 24	1043240
135850	Swiss-Prot Version 42	October 2003	TrEMBL Version 25	1117376
146720	Swiss-Prot Version 43	March 2004	TrEMBL Version 26	1069649
153871	UniProtKB/Swiss-Prot Version 2	July 2004	UniProtKB/TrEMBL Version 2	1333917
163235	UniProtKB/Swiss-Prot Version 3	October 2004	UniProtKB/TrEMBL Version 3	1449374
168297	UniProtKB/Swiss-Prot Version 4	February 2005	UniProtKB/TrEMBL Version 4	1589670
181577	UniProtKB/Swiss-Prot Version 5	May 2005	UniProtKB/TrEMBL Version 5	1714475
194317	UniProtKB/Swiss-Prot Version 6	September 2005	UniProtKB/TrEMBL Version 6	2105517
207132	UniProtKB/Swiss-Prot Version 7	February 2006	UniProtKB/TrEMBL Version 7	2605584
222289	UniProtKB/Swiss-Prot Version 8	May 2006	UniProtKB/TrEMBL Version 8	2948323
241242	UniProtKB/Swiss-Prot Version 9	October 2006	UniProtKB/TrEMBL Version 9	3313264
261513	UniProtKB/Swiss-Prot Version 10	March 2007	UniProtKB/TrEMBL Version 10	3874166
269293	UniProtKB/Swiss-Prot Version 11	May 2007	UniProtKB/TrEMBL Version 11	4377315
276256	UniProtKB/Swiss-Prot Version 12	July 2007	UniProtKB/TrEMBL Version 12	4672908
356194	UniProtKB/Swiss-Prot Version 13	February 2008	UniProtKB/TrEMBL Version 13	5395414
392667	UniProtKB/Swiss-Prot Version 14	September 2008	UniProtKB/TrEMBL Version 14	6070084
428650	UniProtKB/Swiss-Prot Version 15	March 2009	UniProtKB/TrEMBL Version 15	7537442
516081	UniProtKB/Swiss-Prot Version 2010_04	April 2010	UniProtKB/TrEMBL Version 2010_04	10618387
516603	UniProtKB/Swiss-Prot Version 2010_05	May 2010	UniProtKB/TrEMBL Version 2010_05	10706472

Continued on next page

Number of Entries	Swiss-Prot Version	Release Date	TrEMBL Version	Number of Entries
517100	UniProtKB/Swiss-Prot Version 2010_06	June 2010	UniProtKB/TrEMBL Version 2010_06	10862351
517802	UniProtKB/Swiss-Prot Version 2010_07	July 2010	UniProtKB/TrEMBL Version 2010_07	11109684
518415	UniProtKB/Swiss-Prot Version 2010_08	August 2010	UniProtKB/TrEMBL Version 2010_08	11397958
519348	UniProtKB/Swiss-Prot Version 2010_09	September 2010	UniProtKB/TrEMBL Version 2010_09	11636205
521016	UniProtKB/Swiss-Prot Version 2010_10	October 2010	UniProtKB/TrEMBL Version 2010_10	12098541
522019	UniProtKB/Swiss-Prot Version 2010_11	November 2010	UniProtKB/TrEMBL Version 2010_11	12347303
523151	UniProtKB/Swiss-Prot Version 2010_12	December 2010	UniProtKB/TrEMBL Version 2010_12	12769092
524420	UniProtKB/Swiss-Prot Version 2011_01	January 2011	UniProtKB/TrEMBL Version 2011_01	13069501
525207	UniProtKB/Swiss-Prot Version 2011_02	February 2011	UniProtKB/TrEMBL Version 2011_02	13499622
525997	UniProtKB/Swiss-Prot Version 2011_03	March 2011	UniProtKB/TrEMBL Version 2011_03	13897064
526969	UniProtKB/Swiss-Prot Version 2011_04	April 2011	UniProtKB/TrEMBL Version 2011_04	14555721
528048	UniProtKB/Swiss-Prot Version 2011_05	May 2011	UniProtKB/TrEMBL Version 2011_05	15062837
529056	UniProtKB/Swiss-Prot Version 2011_06	June 2011	UniProtKB/TrEMBL Version 2011_06	15400876
530264	UniProtKB/Swiss-Prot Version 2011_07	July 2011	UniProtKB/TrEMBL Version 2011_07	16014672
531473	UniProtKB/Swiss-Prot Version 2011_08	August 2011	UniProtKB/TrEMBL Version 2011_08	16504022
532146	UniProtKB/Swiss-Prot Version 2011_09	September 2011	UniProtKB/TrEMBL Version 2011_09	16886838
532792	UniProtKB/Swiss-Prot Version 2011_10	October 2011	UniProtKB/TrEMBL Version 2011_10	17651716
533049	UniProtKB/Swiss-Prot Version 2011_11	November 2011	UniProtKB/TrEMBL Version 2011_11	18215214
533657	UniProtKB/Swiss-Prot Version 2011_12	December 2011	UniProtKB/TrEMBL Version 2011_12	18510272
534242	UniProtKB/Swiss-Prot Version 2012_01	January 2012	UniProtKB/TrEMBL Version 2012_01	19434245
534695	UniProtKB/Swiss-Prot Version 2012_02	February 2012	UniProtKB/TrEMBL Version 2012_02	20127441
535248	UniProtKB/Swiss-Prot Version 2012_03	March 2012	UniProtKB/TrEMBL Version 2012_03	20639311
535698	UniProtKB/Swiss-Prot Version 2012_04	April 2012	UniProtKB/TrEMBL Version 2012_04	21552793
536029	UniProtKB/Swiss-Prot Version 2012_05	May 2012	UniProtKB/TrEMBL Version 2012_05	22128511

Table 2.2: Showing the release date for each Swiss-Prot, TrEMBL and UniProtKB release. The number of entries within each database Version is also listed.

3

A Quality Metric for Bulk Biological Annotation Quality

Contents

Introduction

Textual annotations are an integral part of biological databases, being utilised and depended upon by a range of users. The information contained within these annotations can be used in numerous ways, including as a basis for future research. Users, however, cannot easily assess the quality and correctness of a textual annotation partly due to the lack of quality metrics. Within this chapter we propose QUALity Metric (QUALM), a generic approach which aims to allow textual annotation to be quantitatively assessed and compared.

QUALM is based upon Zipf's principle of least effort which states that, when achieving a goal, humans will naturally take the path of least resistance. For example, when producing a textual annotation, curators can put the least effort onto readers by ensuring that annotations are detailed and unambiguous. Alternatively, a curator can produce more generic and less detailed annotations, placing the least effort onto themselves. Using this definition, textual annotations where the least effort is placed onto the reader, rather than the curator, are deemed to be of high quality. Relating this principle to textual annotation is achieved through the application of Zipf's Law, which relates word occurrences to their relative ranks. Shown graphically, Zipfian data broadly follows a straight line with the exponent of the relationship, α, defining the slope of a fitted regression line. This obtained α value is used to characterise the text by relating it to the principle of least effort (Section 3.1).

The application of Zipf's law has become pervasive, resulting in numerous cases of Zipfian distributions being reported. Many of these reports have been met with doubts regarding the suitability and validity of Zipf's Law due to the way they were produced and analysed; many claims are based solely on visual inspection. This is subjective, and is especially troublesome with Zipfian graphs, as they often exhibit irregularities caused by numerous words occurring with the same frequency. To alleviate these irregularities we explore Pareto's Law, which removes duplicate points (Section 3.2).

Pareto's Law has many similarities to Zipf's law; both laws are essentially different ways of looking at the same thing. However, values of α still have to be extracted through visual inspection, which is both subjective and error prone. To extract the

value of α more rigorously, we introduce a statistical framework (Section 3.3). This framework provides methods for estimating values of α and deriving associated p-values to provide confidence in estimated α values.

We conclude this chapter with a summary and discussion of QUALM (Section 3.4).

3.1 Zipf's Principle of Least Effort and Zipf's Law

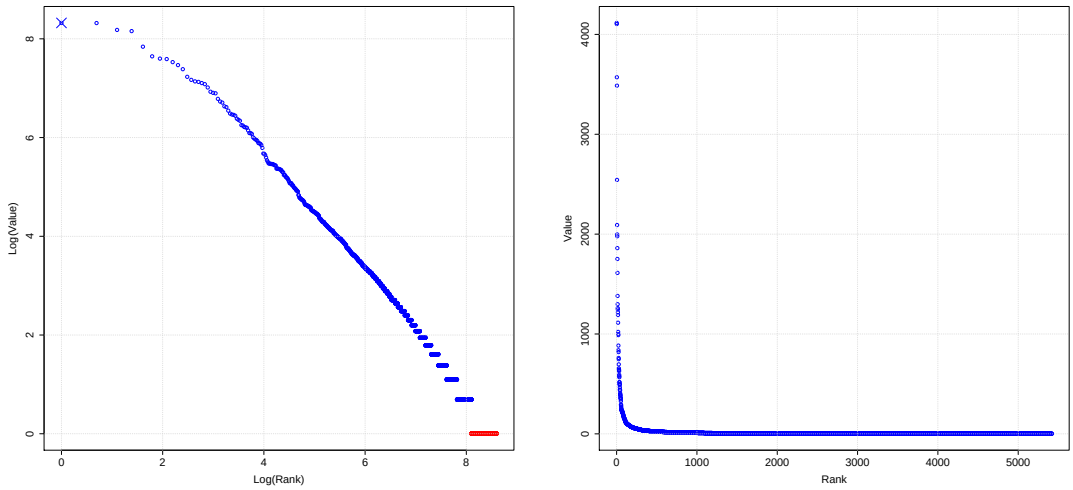
The principle of least effort was first formally proposed by the American linguist George Kingsley Zipf in 1949 [246], following on from his earlier work of analysing the behaviour of language [247]. Zipf published over 35 related articles, however these books comprise Zipf's preeminent work, resulting in him being attributed as the founder of quantitative linguistics [248]. Although originally published over sixty years ago, these works remain of both importance and interest. His 1949 book "Human Behaviour and the Principle of Least Effort" has undergone a recent (2012) reprint [249] and an issue of the Glottometrics journal was dedicated to Zipf, in the year that would have been his 100'th birthday [248].

The purpose of Zipf's work was to, in part, establish the principle of least effort that governs the behaviour of natural language. This led Zipf to analyse the distribution of word usage, hypothesising that word distributions change depending on the effectiveness of how information is conveyed within a text. This analysis identified two interesting and key patterns. Firstly, it is rare for a word to occur very frequently in a text; most words occur very infrequently. Secondly, the frequency of a word is inversely proportional to its rank (i.e. the most commonly occurring word has a rank one, the second most commonly occurring word has rank two, and so on). The latter of these two patterns is the definition of Zipf's eponymous Zipf's law, the work that Zipf is most known and recognised for. More formally, Zipf's Law relates the size (that is its frequency, rather than length) of a word (x) to its rank (r), which can be represented as:

$$x \sim r^{-\alpha} \tag{3.1}$$

The simplest way to investigate a Zipf's law dataset is to draw a graph. Zipfian graphs, like the one shown in Figure 3.1a, plot the rank of each word against its corresponding occurrence with each data point representing how commonly occurring a given ranked word is. In Figure 3.1a the most commonly occurring word, with rank number one, is the upper leftmost point (highlighted with a blue cross), whilst the bottom rightmost points represents all of the words occurring a single time (these points are coloured

red). If the overall data is mostly comprised of relatively few words that occur with high frequency, the data has a positive-skew. When represented graphically, data that is skewed shows little symmetry. Positively-skewed distributions have the majority of the data in the leftmost side of the graph, whilst the rightmost side of the graph has a number of smaller points; more specifically the right tail of the distribution will be longer. Therefore, if these graphs were to be shown on a linear scale plot, they would exhibit an almost perfect "L" shape, as illustrated in Figure 3.1b. Given the large distribution of values, logarithmic scales are used to gain a more even spacing. Graphs that are said to follow Zipf's Law have data points that broadly correspond to a straight line. When this feature is evident, then the data is said to exhibit a *Zipfian distribution* [250]. Zipfian graphs often have a regression line fitted to the curve, the slope of which corresponds to the exponent of Zipf's Law (i.e. α).

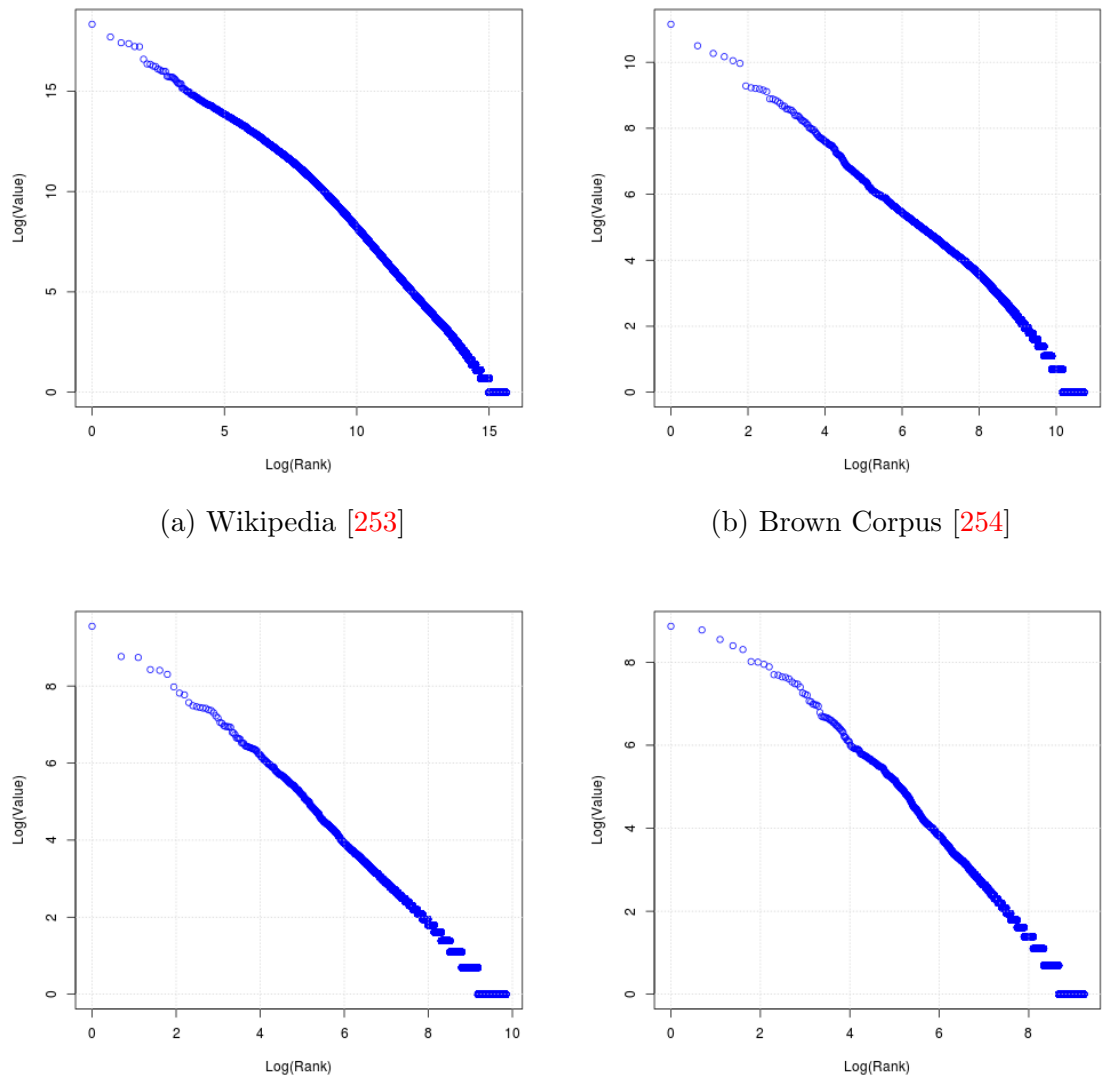


(a) With logarithmic scales (b) Without logarithmic scales

Figure 3.1: An example of a Zipfian graph with and without logarithmic scales. In (a), the blue cross indicates the most commonly occurring word (rank 1), whilst the red points represent words occurring only a single time. These graphs were produced using the poweRlaw package [251] (discussed in Section 3.3) and are based on the occurrences of words in Jane Austen's novel "Sense and Sensibility" [252].

Calculating Zipf's Law is algorithmically straightforward but was historically computationally intensive. Modern processing speeds have alleviated this bottleneck, resulting in a rise of studies applying Zipf's Law to a wide array of texts. A selection of graphs depicting Zipf's Law applied to a set of textual corpora are shown in Figure 3.2. The data points on these graphs all follow straight lines and can therefore be said to exhibit

Zipfian distributions.



(a) Wikipedia [253]

(b) Brown Corpus [254]

(c) *Moby Dick* by Hermann Melville [255]

(d) *Great Expectations* by Charles Dickens [256]

Figure 3.2: Graphical representations of Zipf's Law being applied to four textual corpora. Graphs are produced using the poweRlaw package [251] (discussed in Section 3.3), with references indicating where the data was obtained from.

Although the graphs in Figure 3.2 exhibit Zipfian distributions, the distributions vary between the different text corpora. Visually these differences can be subtle. For example, although not immediately obvious, the head and the tail of the distribution within Wikipedia (Figure 3.2a) differ; the tail (the rightmost part of the graph) is steeper than the head. This pattern is also noticeable in other large corpora, such as the Wall Street Journal [257].

Unlike Wikipedia, the graphs representing the text from Moby Dick (Figure 3.2c) and the Brown Corpus (Figure 3.2b) do not have a head and tail that differ in steepness. However, all of the graphs in Figure 3.2 have noisy tails, which is particularly noticeable in Figures 3.2c and 3.2d. This noise is a result of the large number of words that occur with the same frequency (i.e. many words occur only once, slightly less occur twice, and so on). This is identifiable by a number of horizontally aligned points with gaps between the vertically adjacent points. This noisy tail is less evident in Wikipedia as the dataset is several magnitudes larger.

This visual inspection of graphs shows that a number of interesting traits about the underlying data can be identified from their distributions. For example, the steepness of a distribution gives an indication as to the levels of word reuse. Given that each graph and corresponding α value are based on the underlying text, it is plausible that the α value could provide a quantitative measure to give an indication as to the underlying textual quality. Clearly, this hypothesis fits with Zipf's work on the principle of least effort and analysis of textual quality.

This hypothesis has previously been explored by Ferrer-i-Cancho [258]. Ferrer-i-Cancho has made use of the textual studies that have applied Zipf's law to a given text and published the extracted α from their analysis. The results from this paper, as summarised in Table 3.1, show a correlation between the α value and the corresponding domain that the analysed text was extracted from. These results suggest that the value of α is related to Zipf's principle of least effort and can be used to give an indication of the quality of the text. The results shown in Table 3.1 have been further supplemented by additional studies extracted from the literature. These studies, although not included in the original paper by Ferrer-i-Cancho [258], continue to suggest a correlation between α and textual quality.

Therefore it appears that by applying Zipf's law to a list of word occurrences, and their associated rank, an indication of textual quality can be obtained by relating the extracted α value to the categories described in Table 3.1. The only prerequisite for this approach is the ability to provide a list of word occurrences; if necessary, the rank can easily be derived from this data. Therefore, it appears plausible that this approach can be applied to biological annotation.

α value	Examples in literature	Least effort for
$\alpha < 1.6$	Advanced schizophrenia [246, 259], young children [259, 260]	-
$1.6 \leq \alpha < 2$	Military combat texts [259], Wikipedia [261], Web pages listed on the open directory project [261]	Annotator
$\alpha = 2$	Single author texts [262]	Equal effort levels
$2 < \alpha \leq 2.4$	Multi author texts [263], Moby Dick [255]	Audience
$\alpha > 2.4$	Fragmented discourse schizophrenia [259]	-

Table 3.1: The relationship between $\boldsymbol{\alpha}$ and Zipf's principle of least effort. For α values less than 1.6 or greater than 2.4, there is no corresponding effort level as the text is treated as incomprehensible.

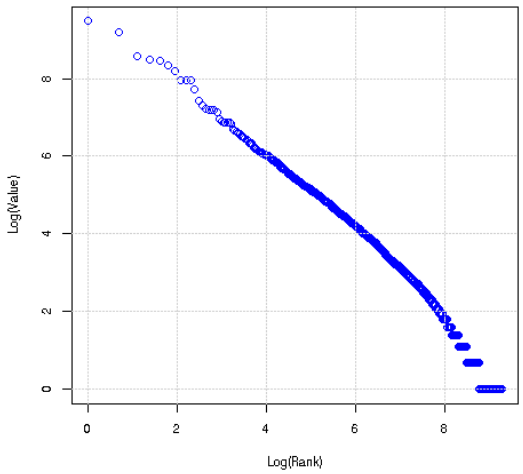
As discussed in Section 2.2, the structure of annotation varies between databases. Some databases cluster annotation into topic sections of related information, such as a proteins function, whilst others have little or no structure. Whilst this structure can aid the presentation of data, often the richest biological knowledge is contained within the natural language; that is, the textual annotation.

Whilst textual annotation is appropriate for human comprehension, computationally analysing and interpreting natural language is a notoriously difficult task. This is due, in part, to the ambiguity of natural language. For example, when trying to computationally determine the semantics of a single sentence, lexical problems, such as *polysemy* — words with multiple meanings — need to be overcome.

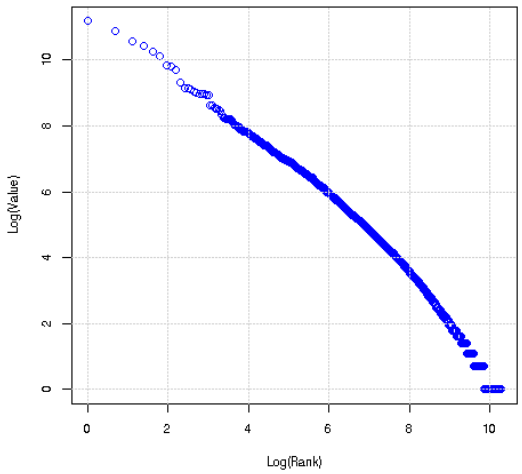
Given the issues posed by natural language processing, many studies avoid the analysis of free text annotation. However, the application of Zipf's Law to free text annotation requires minimal language processing; only the correct extraction of words and their frequency is required. This is reasonably straightforward. For example, Figure 3.3 shows the resulting graphs for a number of Swiss-Prot versions, which suggests that annotation in Swiss-Prot exhibits a Zipfian distribution[1].

Such is the generality of the approach that there are a number of studies that have applied Zipf's law to non-text resources and data and claimed to have found Zipfian distributions. These diverse studies include deaths from terrorist attacks [264], rates of paper citations [265], Web page links on the World Wide Web (WWW) [266] and earthquake intensities [255]. The resulting graphs for these four studies are shown in

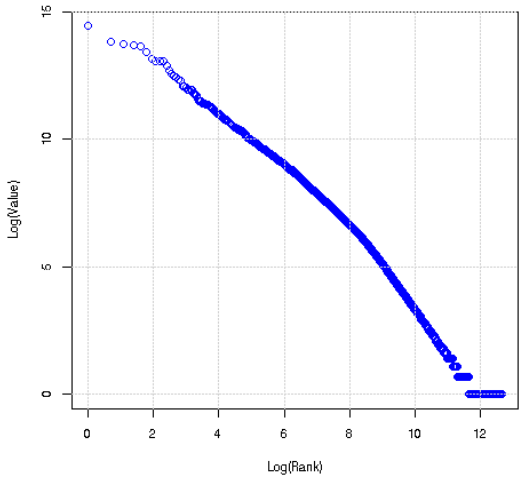
[1]The analysis of Swiss-Prot annotation is undertaken in Chapter 4.



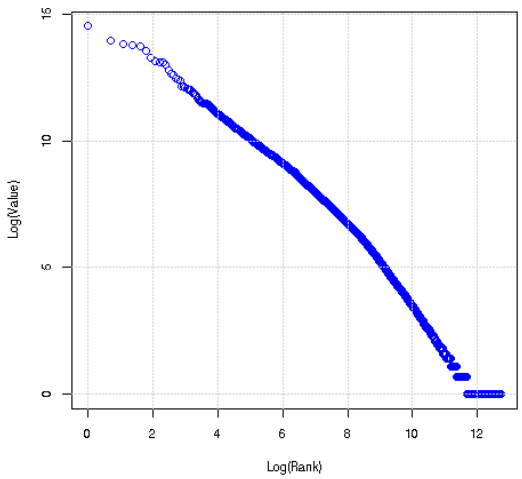
(a) Swiss-Prot Version 9



(b) Swiss-Prot Version 30



(c) UniProtKB/Swiss-Prot Version 2010_10



(d) UniProtKB/Swiss-Prot Version 2012_05

Figure 3.3: Application of Zipf's Law to four versions of Swiss-Prot.

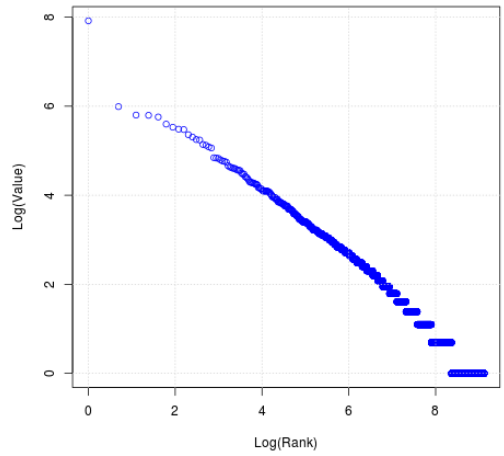
Figure 3.4.

The majority of the graphs shown in Figure 3.4 show less consistency than those produced from textual resources, as shown in Figure 3.2. For example, Figures 3.4c and 3.4d only very broadly follow a straight line. Further, Figure 3.4a shows a substantial gap between the first two points. This gap reflects the fact that the number of deaths in the largest terrorist attack (i.e. the September 11 terrorist attack on America in 2001) was substantially larger than the next largest terrorist attack. However, these inconsistencies are not reflective of all non-textual data. For example, Figure 3.4b is based on data from a Web crawl and appears to exhibit a Zipfian distribution.

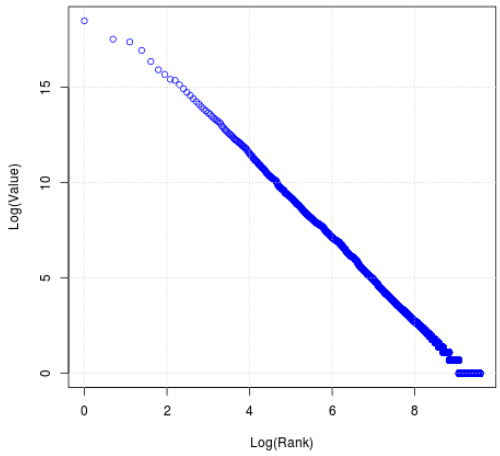
Extending the application of Zipf's law beyond textual data was also done by Zipf himself, who applied his approach and theory to the distribution of city sizes [267]. Given this apparent ubiquity of Zipf's law in natural and man-made phenomena, a number of studies have come under scrutiny. For example, a study into non-coding regions of DNA suggests that they show greater linguistic features than DNA coding regions [268]. Following the publication of this paper, a number of letters and comments were published raising concerns as to the validity of the claims within the paper [269–272]. These included: the inability to recreate the differences between intron and exon data when using annotated data from the GenBank database [272]; DNA being very different to natural language – DNA "words" are composed from a very small alphabet (i.e. A, C, G and T) [270]; and that a control study accounting for noise was missing [271].

The latter issue raised is one that is a commonality between the papers. Indeed, a similarity between a human DNA sequence and that of a random sequence (with identical length and nucleotide frequencies) is presented by one of these letters [270]. Due to the limited alphabet and tuple size for DNA, this is not surprising. A similar idea is also explored by an earlier paper, which appears to show that randomly generated texts (i.e. words consisting of letters from the English alphabet) exhibit Zipfian distributions [273]. However, this paper also comes under scrutiny due to the way random texts were generated and that no visual comparison between real and random texts was made [274].

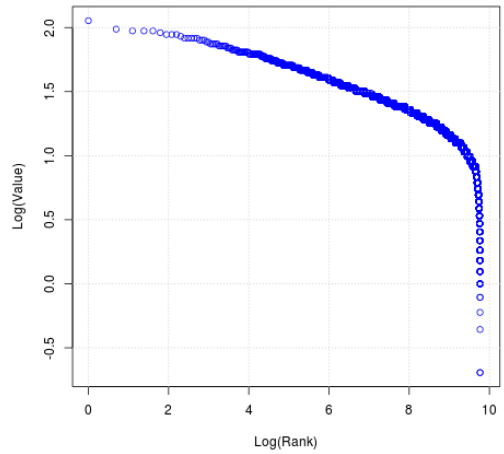
The ubiquity of studies claiming to have observed Zipfian distributions means any in-



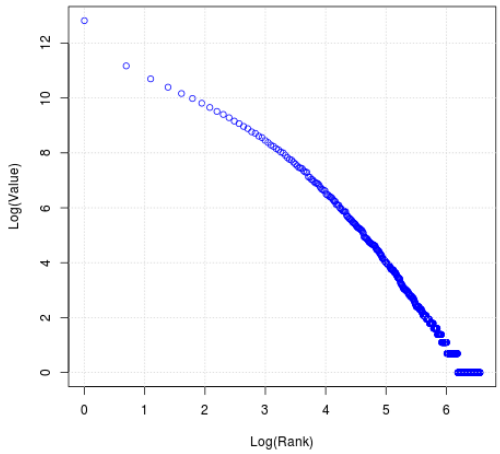
(a) Deaths resulting from terrorist attacks between 1968 and 2006 [264].



(b) Number of links to websites found in a Web crawl of 200 million pages [266].



(c) Earthquake intensities occurring in California between 1910 and 1992 [255]



(d) Citation distribution of papers catalogued by the Institute of Scientific Information between 1981 and 1997 [265]

Figure 3.4: Graphical representations of Zipf's Law being applied to a range of natural and man-made phenomena. Graphs are produced using the poweRlaw package [251] (discussed in Section 3.3), with references indicating where the data was obtained from.

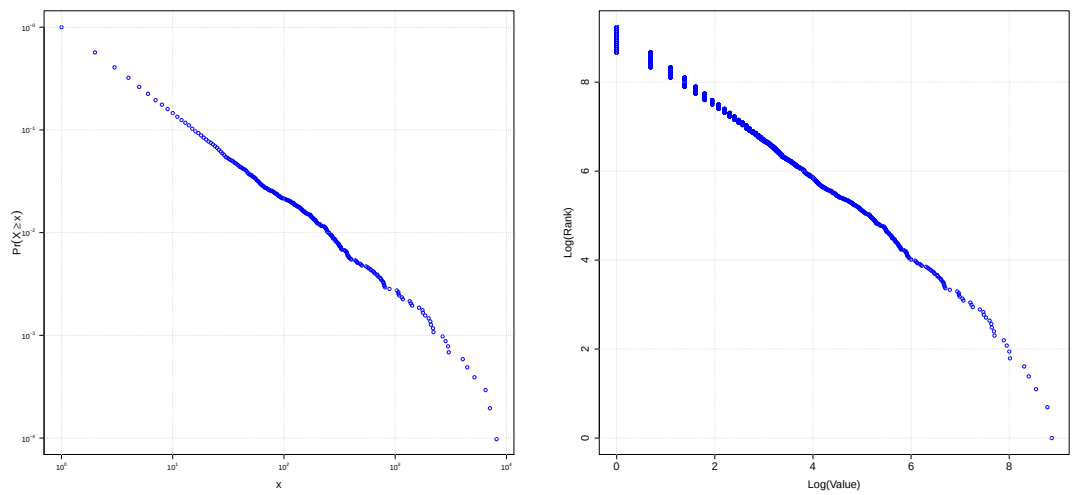
terpretations should be made with caution. If Zipf's law is omnipresent, as facetiously stated by one author [275], then any interpretations are quite probably meaningless. However, visual inspection of Zipfian graphs, like those in Figure 3.2, highlight characteristics of the underlying text, suggesting that graphs can provide insights into the underlying data.

Many of the previous studies rely solely upon visual inspection as a basis for their conclusions. With the majority of Zipfian graphs having noisy tails, which impairs the visual inspection of graphs, this approach is often inconclusive and always subjective, leading to studies coming under scrutiny. If there is any value to be extracted from Zipfian graphs, then it is first necessary that the approach used to plot graphs and extract α values is more formal and rigorous.

3.2 Pareto's Law

In many cases, determining if a given dataset exhibits a Zipfian distribution is based upon visual inspection. However, given the way that Zipfian data is represented, visual inspections are often difficult; Zipfian graphs often suffer from noisy tails. The impact of this noise is dependent upon the type and magnitude of the data. As we have seen earlier, the noisy tail in Figure 3.4a is more pronounced than the noisy tail in Figure 3.4b. As well as making visual inspection more difficult, this noise can also impact the corresponding regression line, and thus the value of α obtained.

One approach to alleviate this noise is to only plot a single point for each frequency. This can be achieved by calculating the Cumulative Distribution Function (CDF) for the data. For example, Figure 3.5a shows the CDF graph for the Great Expectations dataset.



(a) Pareto's Law (b) Zipf's Law with inverted axes

Figure 3.5: Graphical comparison of Zipf's Law and Pareto's Law, when applied to the Great Expectations dataset.

In Figure 3.5a multiple points for a single frequency have been removed, and the graph is subsequently much smoother. In these graphs, the Y-axis represents the probability of a word occurring x or more times (i.e. $Pr[X \geq x]$), with the size (or frequency) shown along the X-axis. Therefore, the data in the top left of the figure represents the probability of a word occurring one or more times, whilst the bottom right point represents the most frequently occurring word. Data presented in this form is referred

to as *Pareto's Law* [276]. Pareto's Law is attributed to the Italian economist Vilfredo Pareto whose original work focused on the distribution of wealth [277] and is also commonly referred to as the 80/20 rule [278]. Pareto's law has a number of similarities to Zipf's Law: they are both empirical laws used to describe phenomena where large events are infrequent and small events are common (i.e. processes that do not follow a normal distribution) [255]. However, unlike Zipfian data which is concerned with how many times the xth ranked word occurs, data represented by Pareto's law is concerned with how many words occur x or more times:

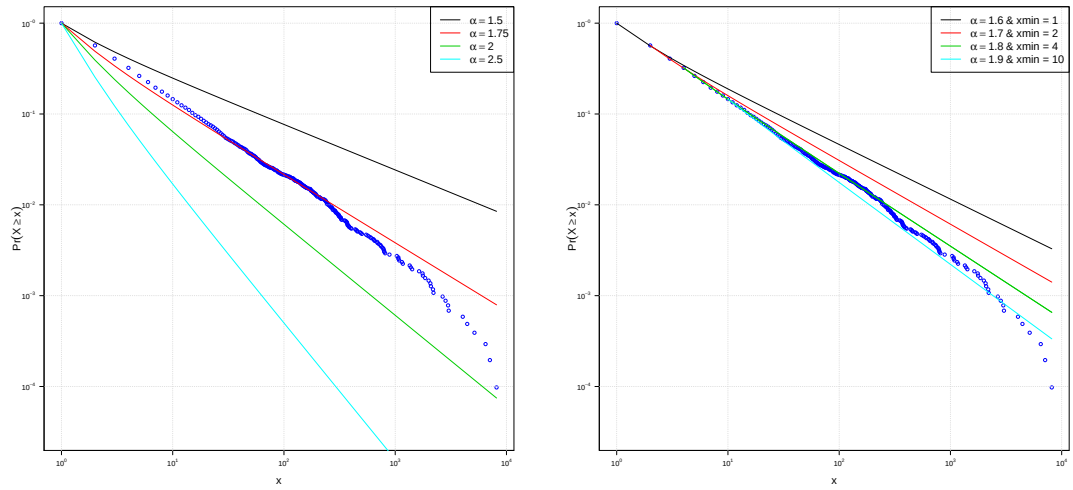
$$P[X \geq x] \sim x^{-\alpha} \tag{3.2}$$

Although there are differences between the two distributions, they are essentially different ways of looking at the same thing [255, 279, 280]. Therefore, in this thesis, graphs following a straight line on a CDF plot are also said to follow a Zipfian distribution. Figure 3.5b illustrates the similarities between Pareto and Zipfian representations by showing the corresponding Zipf's Law graph for the Great Expectations dataset with the axes inverted (the original Zipf's representation is shown in Figure 3.2d). Although these two distributions are visually similar, the individual points represent different values. For example, whilst the bottom right point in both graphs represents the most commonly occurring word in Great Expectations (the word "the"), within a Zipfian distribution (i.e. Figure 3.5b) this point illustrates that this is the first ranked word and occurs $8,145$ times ($\sim log9$), whilst within the Pareto graph (i.e. Figure 3.5a), this same point states that the probability of a chosen word occurring $8,145$ or more times is approximately 0.0001 (10^{-4}).

Given that points on a Pareto graph represent probability, the upper leftmost point on a Pareto graph represents the probability of a word occurring one or more times, which is always one, as only words within the corpus are considered (a word picked at random from the corpus will always occur at least a single time).

By using Pareto's Law as opposed to Zipf's Law for visualising the data, the issues of the noisy tail can be overcome. This not only aids visual inspection, but also the fitting of a regression line. For example, the application of various regression lines to the

Great Expectations dataset, as shown in Figure 3.6a, becomes less obstructed without the noisy tail. Within Figure 3.6a, it would appear that an α value of around 1.75 would provide the most suitable fit for the dataset. Although this gives a reasonable approximation, and overlaps with a number of the data points, the regression line with an α value of two appears to run parallel to the majority of the data. However, as all lines are plotted from the first point, any outliers will impact the fitting of the entire regression line. If a number of these points were discarded when fitting the regression line, then a line more accurately reflecting the majority of the data could be obtained.



(a) Various α values (b) Various $x_{\min}$ and α values

Figure 3.6: Manually fitting various regression lines to the Great Expectations dataset.

It is common for many datasets to exhibit Zipfian distribution only in part; the distribution is often only exhibited in the tail of a graph [255]. Whilst this is highlighted in the Great Expectations dataset, it is more apparent in other datasets, such as the deaths from terrorist attacks as shown in Figure 3.4a. Specifically, this approach considers only those values that are greater than a minimum value, termed $x_{\min}$, when fitting a regression line. This value of $x_{\min}$ corresponds to the starting point of the regression line.

By varying the values of $x_{\min}$, in addition to the value of α, a more accurate regression line can be obtained for the Great Expectations dataset, as shown in Figure 3.6b. This figure suggests that an α value between 1.8 and 1.9 is a more accurate representative than the previous approximation of 1.75.

Attempting to decide which of these two α values gives the best approximation through visual inspection is highly subjective, whilst the calculation of $x_{\min}$ requires careful consideration as data is discarded. Determining these values manually through visual inspection is both labour-intensive and error-prone. Therefore, the calculation of these values needs to be performed in a manner that provides reproducibility and confidence in the estimated values.

3.3 Power-Law Distributions

Whilst visual inspection provides an indication as to whether a dataset exhibits a Zipfian distribution, extracting an α value based purely on visual inspection alone is insufficient. A method to estimate α, and provide an associated confidence score, is required.

One such approach is presented by Clauset *et al.* [281]. In this paper, a statistical framework is presented that allows values of $x_{\min}$ and α to be estimated for a given dataset. Further, this framework presents an approach to calculate the plausibility that a given dataset can be accurately represented by a *power-law distribution*. A power-law distribution is similar to Pareto and Zipfian distributions; they are all a type of power-law. However, whilst Pareto's law is concerned with how many words occur x or more times and Zipf's Law is concerned with the rank of a word x, power-law distributions are concerned with how many words occur exactly x times. As previously discussed, Pareto and Zipfian distributions are essentially different ways of looking at the same thing; this is also true for power-law distributions. The terms "Zipfian distribution" and "power-law distribution" are used interchangeably within this thesis.

The discrete power-law distribution presented by Clauset *et al.* has the form

$$p(x) = \Pr(X = x) = Cx^{-\alpha} \tag{3.3}$$

where C is a normalising constant. Although the authors present complementary approaches for continuous data, only the discrete forms are considered within this thesis, as word frequencies can only take the form of positive integers. The power-law distribution in Equation 3.3 diverges when $x = 0$, meaning that an $x_{\min}$ value that is greater than zero must exist. By calculating the normalising constant C, then the Probability Mass Function (PMF) is given as

$$p(x) = \Pr(X = x) = \frac{x^{-\alpha}}{\zeta(\alpha, x_{\min})} \tag{3.4}$$

where

$$\zeta(\alpha, x_{\min}) = \sum_{n=0}^{\infty} (n + x_{\min})^{-\alpha} \tag{3.5}$$

is the generalised zeta function. As discussed, the power-law distribution, and therefore the PMF, is concerned with the probability of individual values. However, datasets are visualised as CDFs. Therefore, to apply the α value extracted from the power-law distribution to these graphs, the cumulative probabilities are required. The required CDF is calculated using Equation 3.6.

$$P(X \leq x) = \frac{\zeta(\alpha, x)}{\zeta(\alpha, x_{\min})} \tag{3.6}$$

The definition of the power-law distribution is clearly dependent upon the values of $x_{\min}$ and α, which need to be estimated. To estimate α, Clauset *et al* use the maximum likelihood estimator

$$\hat{\alpha} \simeq 1 + n \left[\sum_{i=1}^{n} \ln \frac{x_i}{x_{\min} - \frac{1}{2}} \right]^{-1} \tag{3.7}$$

where x_i are the observed data values and $x_i \geq x_{\min}$. To estimate the value of $x_{\min}$, the authors use the Kolmogorov-Smirnov approach

$$D = \max_{x \geq x_{\min}} |S(x) - P(x)| \tag{3.8}$$

where $S(x)$ represents the CDF for the observational data and $P(x)$ represents the CDF of the theoretical model (for $x \geq x_{\min}$ in both cases).

Having provided mechanisms to obtain estimates for $x_{\min}$ and α, Clauset *et al.* provide a goodness-of-fit test, which produces an associated p-value, to assess the plausibility that the given dataset follows a Zipfian distribution. This goodness-of-fit test is also based on the Kolmogorov-Smirnov statistic. A dataset can be deemed to exhibit a Zipfian distribution when a p-value greater than 0.1 is obtained (i.e. $p \leq 0.1$ rules out the plausibility that a dataset follows a Zipfian distribution).

Using these methods, the authors performed an analysis of 24 datasets that have been published and claimed to exhibit Zipfian distributions. Of these 24 datasets, 17 had

sufficient evidence to support the claim that a Zipfian distribution was evident. The earthquake intensities dataset, as shown in Figure 3.4c, was one of the seven datasets deemed to not exhibit a Zipfian distribution, whilst the Moby Dick dataset, as shown in Figure 3.2c, exhibited the most convincing fit.

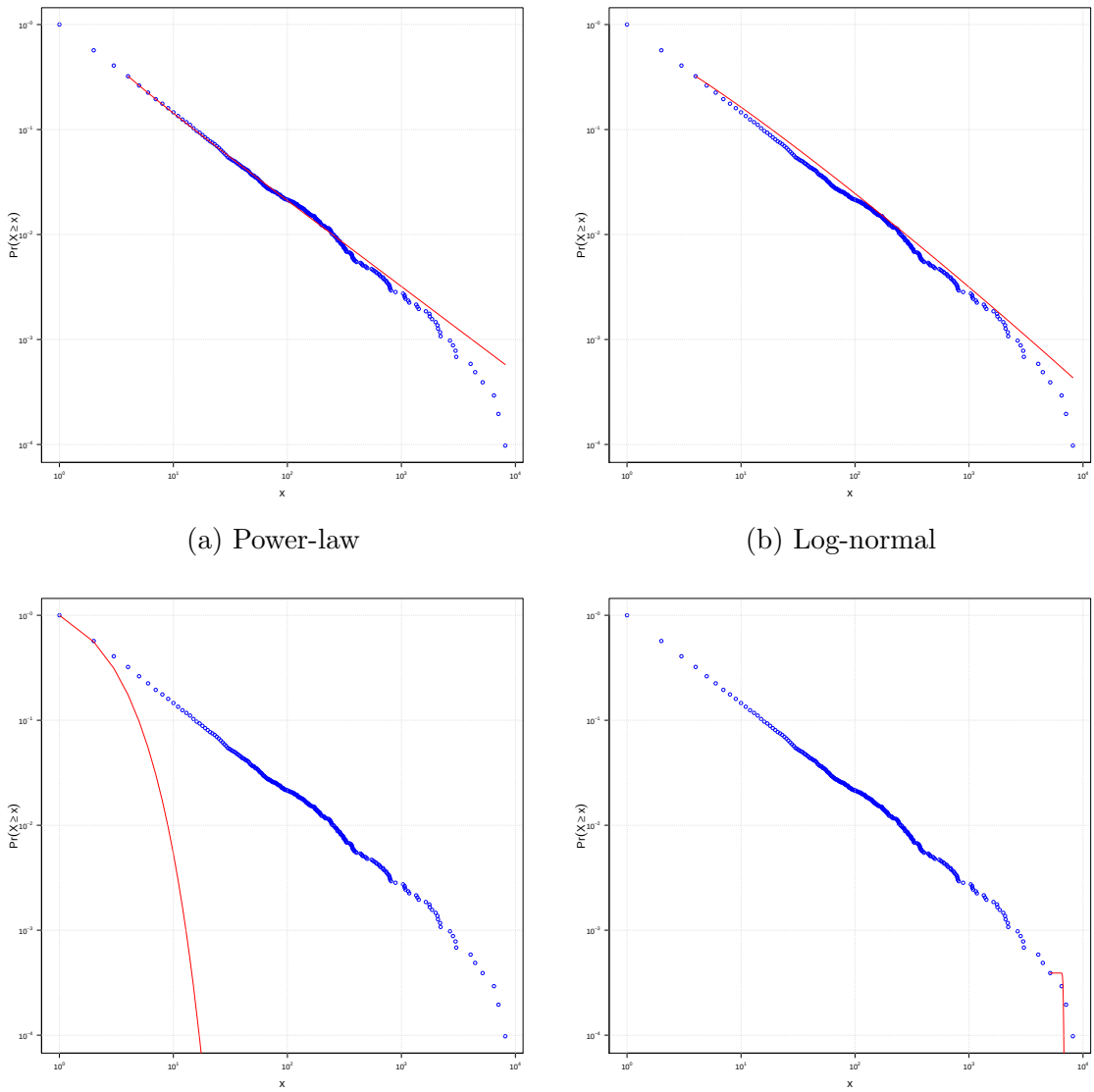
These approaches have been implemented as frameworks in a number of languages, including MatLab [281] Python [282] and R [251]. Of these implementations, the R framework "poweRlaw", which was developed in collaboration with Colin Gillespie, implements a number of additional features not available in the other frameworks, such as multithreading support.

To illustrate the application of the poweRlaw package, and the methods described here, the framework can be applied to the Great Expectations dataset. The poweRlaw framework estimates that the values of α and $x_{\min}$ for the Great Expectations dataset are 1.82 and 5, respectively. This appears highly plausible given the manual fitting performed previously, as shown in Figure 3.6, which estimated an α of between 1.8 and 1.9 and an $x_{\min}$ of between 4 and 10. Calculating the p-value for this dataset returns a value of 0.55, which provides sufficient confidence that the Great Expectations dataset does indeed exhibit a Zipfian distribution.

However, although the extracted p-value concludes that the dataset can be suitably characterised by a power-law model, it is possible that an alternative model may provide a more accurate fit. Therefore, the poweRlaw package provides methods for fitting alternative distributions (namely the exponential, log-normal & Poisson distributions) to a dataset and for comparing the suitability of these models. The resulting graphs for each of these models applied to the Great Expectations dataset is shown in Figure 3.7.

From Figure 3.7, it is clear that the Poisson (Figure 3.7d) and exponential (Figure 3.7c) distributions show a poor fit. However, the log-normal distribution appears to provide a reasonable fit. Comparing these distributions within the poweRlaw package returns a p-value of 0.226, which means that the dataset is more accurately represented by a power-law than a log-normal distribution.

Whilst the poweRlaw package provides p-values to gain confidence in the suitability of the power-law distribution, it also provides mechanisms to analyse the accuracy



(a) Power-law

(b) Log-normal

(c) Exponential

(d) Poisson

Figure 3.7: Figures representing a variety of distributions applied to the Great Expectations dataset.

of the obtained $x_{\min}$ and α values. To measure the accuracy of the obtained $x_{\min}$ and α values a bootstrapping procedure is used [283]. The bootstrapping procedure involves estimating the $x_{\min}$ and α values a large number of times from samples of the dataset, whilst covering all possible values of $x_{\min}$. By applying a bootstrap procedure to the Great Expectations dataset, using $5,000$ bootstrap samples, the output shown in Figure 3.8 is obtained. This Figure shows the estimates for $x_{\min}$ and α over $5,000$ iterations, with a 95% confidence level. These estimates can also be illustrated as histograms, as shown in Figure 3.9.

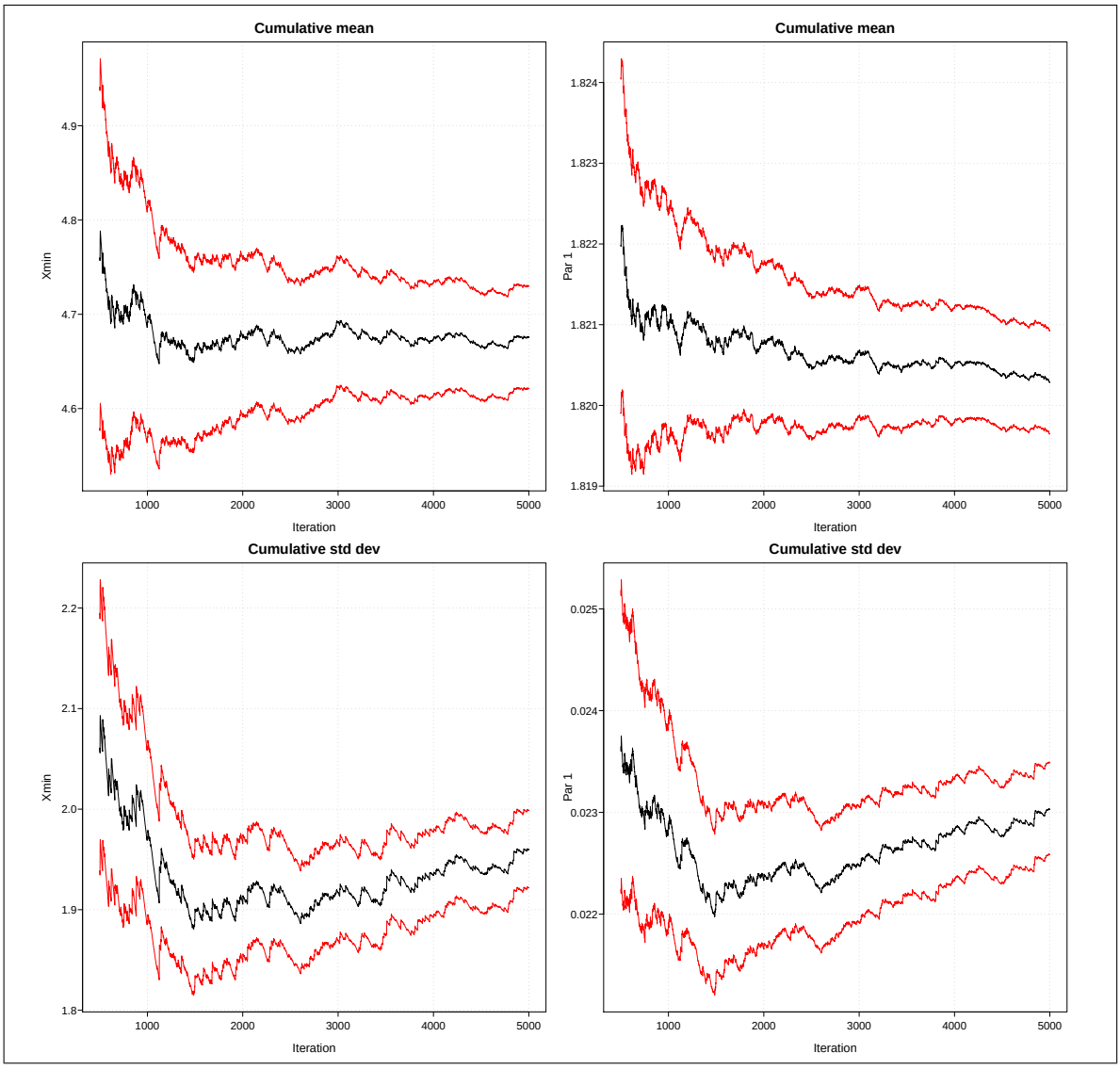
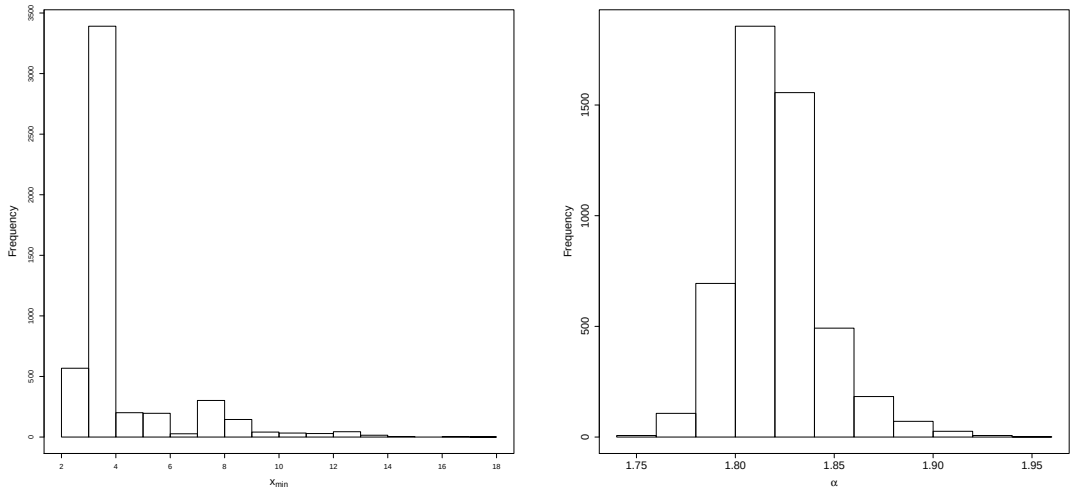


Figure 3.8: Results from $5,000$ iterations of the bootstrapping procedure. The top row shows the mean estimate for $x_{\min}$ and α, whilst the bottom row shows the estimated standard deviation of $x_{\min}$ and α.

These figures show that the estimated values of α and $x_{\min}$ are converging; the change



(a) $x_{\min}$ (b) α

Figure 3.9: Histograms representing the results from the bootstrapping procedure after estimating $x_{\min}$ (standard deviation 1) and α (standard deviation 0.03).

in the average value over time is minimal. Additionally, the estimated parameters both have reasonably low standard deviations, which provides further confidence in the extracted parameters.

3.4 Discussion

Current methods for assessing the quality of textual annotation are limited. Clearly, a method that allows any form of textual annotation to be quantitatively assessed and compared would be of benefit to both curators and end users alike. One potential approach, as presented within this chapter, is QUALM which is based upon α values obtained from texts that follow a power-law distribution.

Power-laws are observed in a number of natural and man-made phenomena, identifiable by a few large events and many small events. For example, in Wikipedia, approximately 6% of the corpus is composed of the most commonly occurring word ("the"), whilst the second most occurring word ("of") accounts for approximately 3% of the corpus. Conversely, there are almost 3 million words that occur only a single time within the entire Wikipedia website.

Such surprising patterns of word reuse and distributions have sparked a substantial interest in power-law distributions. However, many studies use flawed approaches to both fit and conclude power-law distributions; many are simply based on visual inspection. Therefore, it was necessary, and indeed inevitable, that methods to statistically assess and estimate such distributions would emerge. The methods presented by Clauset *et al.* have gained substantial usage and essentially become the *de facto* standard for fitting power-law distributions to data. Originally published in early 2009 the paper has, at the time of writing, amassed over $1,800$ citations – an average of seven new citations per day.

The statistical framework presented by Clauset *et al.* and implemented by the poweR-law package provides straightforward and consistent methods for deriving both α and $x_{\min}$ values. Crucially, the confidence in an observed power-law distribution and its estimated parameters can be evaluated by the p-value returned from the framework. Using this approach, an α value of 1.82 was estimated for the Great Expectations dataset. Relating this result to the categories presented in Table 3.1 suggests that, between the audience and the author, the least effort was placed with the author Charles Dickens. A similar comparison can be made to Moby Dick, written by Herman Melville, which has a reported α value of 2.20 [255]. This result suggests that a

reader will have to exert more effort to fully understand the information being conveyed within the Great Expectations novel compared to that of the Moby Dick novel. Using the principle of least effort as the definition of quality, this comparison would conclude that Moby Dick is of greater quality than Great Expectations.

The α value extracted for the Great Expectations dataset was based upon a power-law distribution, with the Poisson, log-normal and exponential distributions ruled out. However, whilst these distributions are commonly suggested as potential alternatives to power-law distributions, it is possible that another distribution could still provide a more suitable fit. Alternative approaches, such as least-squares fitting [284] or logarithmic binning [285], could be considered for estimating the α and $x_{\min}$ parameters. Whilst consideration is given to a number of alternative approaches, performing an exhaustive search of all potential distributions and approaches for estimating parameters would unlikely prove beneficial; both visual inspection and extracted p-values provide suitable confidence in the plausibility of the distribution and its extracted parameters.

The power-law model achieves a good balance between model parsimony and fit. These two features make the approach appealing as a potential generic quality metric for biological annotation. Specifically, this approach is straightforward and provides consistent and comparable results for any textual annotation. However, as discussed in Section 3.1, Zipf's Law has come under scrutiny. This scrutiny is partly due to the approaches employed; Clauset *et al.* [281] showed this was warranted as they found insufficient evidence to support many studies claiming to have identified a Zipfian distribution. Whilst using the developed power-law model eradicates dubiousness regarding the fitting and estimation of parameters, the level of value and meaning that can be extracted from a dataset exhibiting a Zipfian distribution remains unclear. This is explored in the following chapter by applying QUALM to textual annotation in The UniProt Knowledgebase (UniProtKB).

4

Analysing Annotation Quality in the UniProt Knowledgebase

Contents

Introduction

In the previous chapter a potential quality metric, QUALM, was proposed. QUALM is based on a power-law model and is applied to the occurrences of words extracted from textual annotation. Zipf's principle of least effort suggests that the exponent of this power-law model, α, offers an indication of quality; that is annotations which puts the least effort onto the reader, rather than the curator, are deemed to be of high quality.

One way of assessing the performance of QUALM would be to use an explicit gold standard dataset; unfortunately, there is no obvious gold standard to use in this case. Therefore, in this chapter QUALM is applied to various sets of annotation, with the results related to our *a priori* judgements and knowledge to assess its suitability as a measure of quality. For example, it is generally held that manually curated annotations are of higher quality than those generated computationally. If this metric is a measure of quality, then we would expect it to determine that manual annotations are of higher quality than automated annotations.

This analysis is performed on annotation taken from UniProtKB. UniProtKB provides an ideal resource to test the suitability of the metric for a number of reasons: it is a comprehensive resource composed of both manual and automated textual annotations; it is well established, with an abundance of historical versions; and it allows proteins from individual species, or entire taxonomic divisions, to be extracted. These features are supported by numerous publications and a helpful and responsive help desk. This support aids the application and analysis in a number of ways, such as the ability to query UniProt curators regarding specific details that are not otherwise available.

To perform this analysis, an extraction framework – Biological ANnotation Extraction framework (BANE) – was developed (Section 4.1). BANE extracts lists of word occurrences from annotation, allowing the power-law to be applied to UniProtKB. Initially, QUALM is applied to Swiss-Prot annotation to determine if annotation obeys a power-law (Sections 4.2 and 4.3). UniProtKB provides over twenty years worth of data, for both manual and automatic annotations. This history allows us to analyse annotation both over time and at a specific point in time. It also allows for a compari-

son between manual and automated annotation (Section 4.4) as well as analysing how annotations change over time for a subset of mature entries (Section 4.5). Following this, the differences in annotation between various species and taxonomic ranges is investigated (Section 4.6). Finally, the chapter concludes with a discussion of QUALM and its suitability as a measure of quality (Section 4.7).

4.1 Data Extraction

In order to apply QUALM to textual annotation in the UniProtKB database, a list of all words and their occurrences are required. Annotation within UniProtKB is spread across numerous entries, some of which may contain no textual annotation. Given these characteristics, a bulk retrieval of entries from UniProtKB is required. The retrieval of entries from UniProtKB is available via multiple methods, including programmatic access via RESTful URLs. However, given the database size and the number of historical versions, the most efficient approach is to obtain the database dumps from the UniProtKB File Transfer Protocol (FTP) server[1].

The UniProtKB FTP server contains compressed downloads of complete UniProtKB versions in eXtensible Markup Language (XML), FASTA and flat file formats. For the extraction of annotation, the FASTA format is unsuitable as the free text annotation is removed. Additionally, while XML files provide a convenient syntax, not all historical versions of UniProtKB are available in XML format. Therefore all historical versions of UniProtKB were obtained in flat file format, with the exception of Swiss-Prot Versions 1-8 and 10, which were never archived[2].

As previously discussed in Section 2.4, the flat file format, as shown in Figure 2.7, follows a strict structure. Each line starts with two upper-case letters, used to identify the content type contained within the line. An entry is identified by a line beginning with "ID", whilst the end of an entry is indicated by two forward slashes ("//"). Various lines can appear within an entry, with the textual annotation being contained within the comment (i.e. general annotation) lines. Comment lines begin with the characters "CC", allowing textual annotation to be easily identified.

Given that the flat file format is strictly defined and well established, a number of programming tools and libraries have been developed that provide a framework to handle and manipulate UniProtKB entries. Such projects, including BioJava [286], BioPerl [287] and Swissknife [288], provide varying degrees of UniProtKB support.

[1] `ftp://ftp.uniprot.org/`

[2] Additionally, early versions of TrEMBL (i.e. those prior to the formation of UniProtKB) are not available on the FTP server. However, these were kindly made available by The Universal Protein Resource (UniProt) upon request.

However, the extraction of words from textual annotation is a niche requirement and, unsurprisingly, not provided by these tools. Therefore a custom parsing and extraction program, which we name *Biological ANnotation Extraction framework (BANE)*, is required.

Developing BANE to read flat files and identify the type of each line is relatively straightforward; the difficulty lies with the correct identification and extraction of words from the comment lines. Historically, the comment lines were composed almost entirely of free text. However, textual annotations have evolved over time, with annotations in later versions of UniProtKB becoming more structured. Most notable is the usage of topic blocks, used to group related annotations into topics. Topic blocks are identified by the characters `-!-` followed by a topic block name. At the time of writing, there are a total of 29 topic blocks, as summarised in Table 4.1.

The frequency of topic blocks within an entry is variable; a topic may occur zero or more times. Additionally, within early versions of Swiss-Prot the topic block identifier (`-!-`) was often used without an attached topic block name. In both cases, comments contained within a topic block have at least one line, but may span multiple lines.

The structure of each topic block is also variable. Whilst the majority contain only free text, certain topics contain a number of subtopics. Specifically, the topic "alternative products" has eight possible subtopics, whilst the "biophysicochemical properties" topic block can have five possible subtopics. These subtopics follow a strict syntax, as summarised in Tables 4.2 and 4.3. Both topics and subtopics may also contain properties, which consist of the property name followed by the corresponding value, for example "Note=No experimental confirmation available".

In order to correctly extract words contained within textual annotation, such structural and formatting information has to be removed, otherwise the words used in this structure would have unduly high rates of occurrence (see Section 4.2 for an example of this). This process includes the removal of identifiers and headings used in each topic, subtopic and property as well as the "CC" identifier. Additionally, English punctuation and formatting is also removed in certain instances. For example, a comment that contains "(by similarity)", will extract two words, "by" and "similarity", with the parentheses being removed. However, not all punctuation is removed. For example,

Topic	Description
Allergen	Information relevant to allergenic proteins
Alternative products	Description of the existence of related protein sequence(s) produced by alternative splicing of the same gene, alternative promoter usage, ribosomal frameshifting or by the use of alternative initiation codons
Biophysicochemical properties	Description of the information relevant to biophysical and physicochemical data and information on pH dependence, temperature dependence, kinetic parameters, redox potentials, and maximal absorption
Biotechnology	Description of the use of a specific protein in a biotechnological process
Catalytic activity	Description of the reaction(s) catalyzed by an enzyme
Caution	Warning about possible errors and/or grounds for confusion
Cofactor	Description of any non-protein substance required by an enzyme for its catalytic activity
Developmental stage	Description of the developmentally-specific expression of mRNA or protein
Disease	Description of the disease(s) associated with a deficiency of a protein
Disruption phenotype	Description of the effects caused by the disruption of the gene coding for the protein
Domain	Description of the domain structure of a protein
Enzyme regulation	Description of an enzyme regulatory mechanism
Function	General description of the function(s) of a protein
Induction	Description of the compound(s) or condition(s) that regulate gene expression
Interaction	Conveys information relevant to binary protein-protein interaction
Mass spectrometry	Reports the exact molecular weight of a protein or part of a protein as determined by mass spectrometric methods
Miscellaneous	Any comment which does not belong to any of the other defined topics
Pathway	Description of the metabolic pathway(s) with which a protein is associated
Pharmaceutical	Description of the use of a protein as a pharmaceutical drug
Polymorphism	Description of polymorphism(s)
PTM	Description of any chemical alternation of a polypeptide (proteolytic cleavage, amino acid modifications including crosslinks). This topic complements information given in the feature table or indicates polypeptide modifications for which position-specific data is not available.
RNA editing	Description of any type of RNA editing that leads to one or more amino acid changes
Sequence caution	Description of protein sequence reports that differ from the sequence that is shown in UniProtKB due to conflicts that are not described in feature table "conflict" lines, such as frameshifts, erroneous gene model predictions, etc.
Similarity	Description of the similaritie(s) (sequence or structural) of a protein with other proteins
Subcellular location	Description of the subcellular location of the chain/peptide/isoform
Subunit	Description of the quaternary structure of a protein and any kind of interactions with other proteins or protein complexes; except for receptor-ligand interactions, which are described in the topic "function"
Tissue specificity	Description of the tissue-specific expression of mRNA or protein
Toxic dose	Description of the lethal dose, paralytic dose or effective dose of a protein
Web resource	Description of a cross-reference to a network database/resource for a specific protein

Table 4.1: A list of all the possible topic blocks that can occur within the comment lines of a UniProtKB entry. Taken from the UniProtKB user manual [223].

Topic	Description
Event	Biological process that results in the production of the alternative forms. It lists one or a combination of the following values (Alternative promoter usage, Alternative splicing, Alternative initiation, Ribosomal frameshifting). Format: Event=controlled vocabulary; Example: Event=Alternative splicing;
Named isoforms	Number of isoforms listed in the topics 'Name' currently only for 'Event=Alternative splicing'. Format: Named isoforms=number; Example: Named isoforms=6;
Comment	Any comments concerning one or more isoforms; optional; Format: Comment=free text; Example: Comment=Experimental confirmation may be lacking for some isoforms;
Name	A common name for an isoform used in the literature or assigned by Swiss-Prot; currently only available for spliced isoforms. Format: Name=common name; Example: Name=Alpha;
Synonyms	Synonyms for an isoform as used in the literature; optional; currently only available for spliced isoforms. Format: Synonyms=Synonym_1[,Synonym_n]; Example: Synonyms=B, KL5;
IsoId	Unique identifier for an isoform, consisting of the Swiss-Prot accession number, followed by a dash and a number. Format: IsoId=acc#-isoform_number[, acc#-isoform_number]; Example: IsoId=P05067-1;
Sequence	Format: Sequence=VSP_#[, VSP_#]|Displayed|External|Not described; Example: Sequence=Displayed; Example: Sequence=VSP_000013, VSP_000014;
Note	Lists isoform-specific information; optional. It may specify the event(s), if there are several. Format: Note=Free text; Example: Note=No experimental confirmation available;

Table 4.2: List of the subtopic blocks that can occur under the "alternative products" topic block, with the syntax of each subtopic also shown. Taken from the UniProtKB user manual [223].

Property	Description
Absorption	Indicates the wavelength at which photoreactive proteins such as opsins and DNA photolyases show maximal absorption
Kinetic parameters	Mentions the Michaelis-Menten constant (KM) and maximal velocity (Vmax) of enzymes
pH dependence	Describes the optimum pH for enzyme activity and/or the variation of enzyme activity with pH variation
Redox potential	Reports the value of the standard (midpoint) oxido-reduction potential(s) for electron transport proteins
Temperature dependence	Indicates the optimum temperature for enzyme activity and/or the variation of enzyme activity with temperature variation; the thermostability/thermolability of the enzyme is also mentioned when it is known

Table 4.3: List of the subtopic blocks that can occur under the "biophysicochemical properties" topic block. Taken from the UniProtKB user manual [223].

"DNA-binding" would be extracted as a single word, without the removal of the hyphen. Aside from the removal of structural and formatting information, there is no further manipulation of textual annotations. For example, abbreviations and shorthands are extracted verbatim, rather than expanding them into their full form. This light-weight extraction ensures that the process is computationally undemanding and was implemented in BANE.

The flat file of a complete database version contains all entries within a single file. BANE progresses through the flat file sequentially, first identifying comment lines and then the individual words contained within these lines. Each identified word is recorded along with a corresponding occurrence. When the end of the file is reached, a list of all the words encountered with their associated occurrences are output; the light-weight nature of this process allows in-memory processing and makes check-pointing unnecessary. The overall extraction process can be summarised in four key stages, as shown in Figure 4.1.

The correct extraction of words and their corresponding occurrences is pivotal; incorrect data can significantly impact the application of the power-law. Therefore, a number of checks and safeguards were implemented to gain confidence that the results from the parsing process are indeed correct. Given the evolution of the flat files and annotation, checks were performed for all database versions.

The first implemented check was to ensure that an equal number of ID and end-of-entry (i.e. "//") lines were encountered by BANE. Given the format of flat files, and

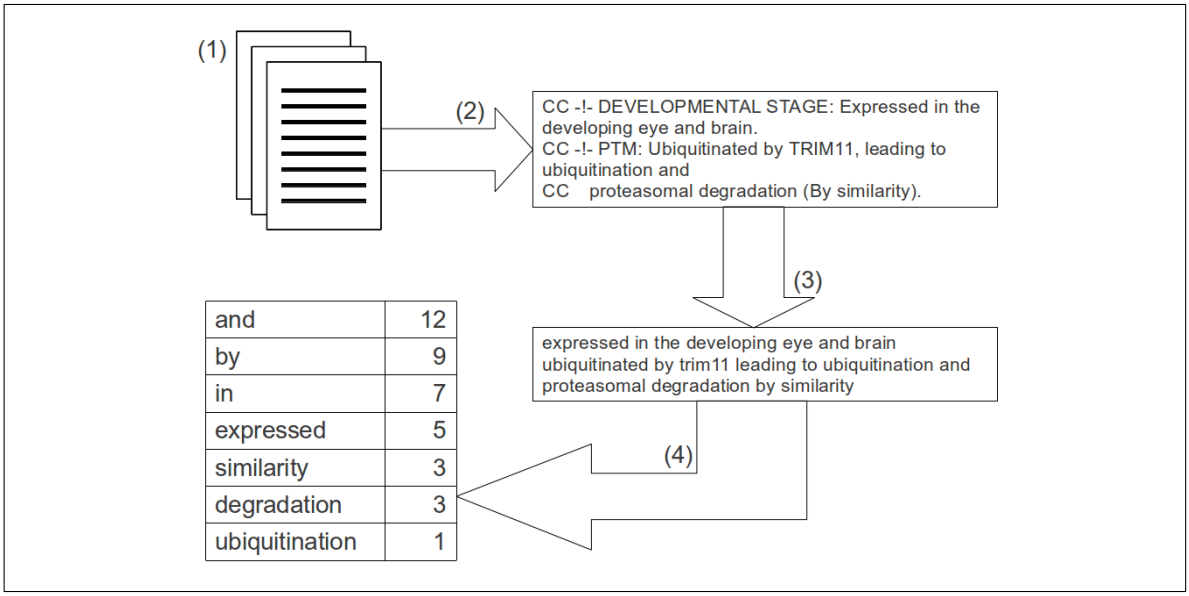


Figure 4.1: Outline view of the data extraction process. (1) Initially a complete dataset for a given database version was downloaded in flat file format. (2) Comment lines were then extracted (lines beginning with 'CC', the comment indicator). (3) Punctuation, 'CC', brackets, comment blocks and properties (as defined in the UniProtKB manual [223]) were removed, whilst making words lower case, so as to treat them as case insensitive. (4) Finally, the total occurrences of each word identified was updated.

the strong dependence on line identifiers, an incorrectly formatted line could cause all downstream lines to be incorrectly parsed. Such an error could be especially problematic for later versions of UniProtKB which contain numerous lines. For example, the flat file for UniProtKB/TrEMBL Version 2012_05 is over 55GB in size and contains over 1.1 billion lines. The number of ID and end-of-entry lines encountered by BANE was also checked against the expected number of entries, as reported by the UniProtKB release notes[3]. These values correctly matched for all database versions, with the exception of UniProtKB/Swiss-Prot Version 5 and UniProtKB/Swiss-Prot Version 10. In these cases the number of entries parsed did not match the number reported by UniProt (a mismatch of 6 and $1,338$ entries, respectively). After contacting the UniProt help desk this inconsistency was determined to be an error within the UniProtKB release notes, rather than with the actual flat files or BANE. These release notes have since been corrected by UniProt [289].

Having confidence that BANE is correctly identifying individual entries within a file means focus can be given to testing that words are correctly identified and extracted

[3]Available at `http://www.uniprot.org/statistics/`

from the comment lines. The initial checks performed were to ensure that structural formatting was correctly removed. This involved noting the list of headings (comment blocks and properties) removed, along with their frequency and checking these matched the headings expressed in the UniProtKB manual. This list contained a number of topic headings not listed in the UniProtKB manual, including a number of incorrectly spelt topic blocks such as "similartity" (P15321, Swiss-Prot Version 14), "functon" (P21127, Swiss-Prot Version 17) and "tissue specificifity" (Q04735, Swiss-Prot Versions 29, 30 & 31). Further, within earlier UniProtKB versions there are a number of headings identified, such as "enzymatic regulation" and "alternative splicing" that are not stated within the current user manual, but were likely to have been previously defined as topic headings (for example, "enzymatic regulation" was likely changed to the current topic block "enzyme regulation", whilst "alternative splicing" was later added as a subtopic of "alternative products"). Although not listed in the current UniProtKB manual, these headings were added to the removal list.

To supplement these checks, 100 records were manually analysed. These records were selected randomly to ensure a broad range of entries, covering different types of topic-blocks and varying levels of annotation, were analysed. These checks and safeguards provide confidence in the extraction of words from UniProtKB comment lines allowing the power-law model to be applied to UniProtKB annotation, as discussed in the following section.

4.2 Does UniProtKB Obey a Power-Law?

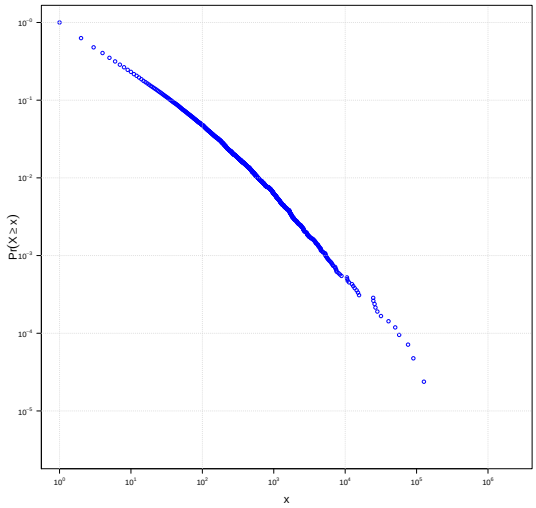
The result from BANE is a list of all words and their occurrences for each version of UniProtKB, which allows QUALM, as described in Chapter 3, to be applied to this annotation. The aim of QUALM is to allow annotations to be quantitatively assessed and scored.

QUALM can only be used effectively if annotation in Swiss-Prot actually obeys a power-law. Swiss-Prot was chosen for the initial analysis as it is commonly held that manual annotation is of higher quality than automated annotation. If this is true, and a power-law distribution is a measure of quality, then it would be expected that a power-law distribution is more likely to occur in human curated annotation rather than annotations produced automatically.

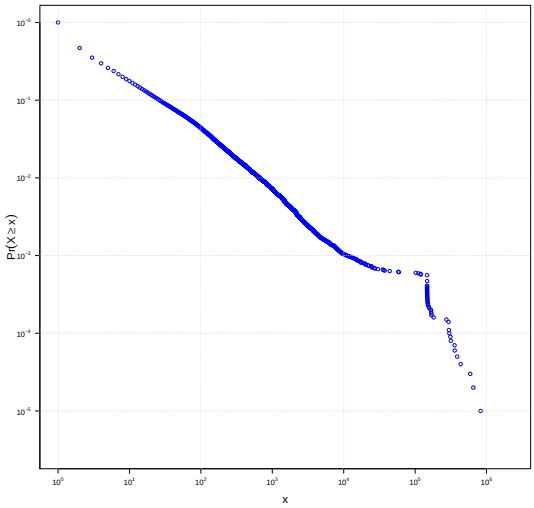
QUALM was applied to all historic versions of Swiss-Prot, four of which are shown in Figure 4.2. This figure shows that the annotation in Swiss-Prot does broadly obey a power-law, however there is a noticeable structure, or "kink", in Figures 4.2b, 4.2c and 4.2d. This kink is visible in the tail of the power-law, in the bottom right portion of these graphs. Specifically, the structure is visible at approximately 10^5 in Figures 4.2b and 4.2c and between 10^5 and 10^6 in Figure 4.2d. Although this structure is initially very distinct, it becomes less apparent over time.

An inspection of the words in this region, as shown in Table 4.4, identified that this structure is artifactual, resulting not from annotation *per se* but from copyright and license information. This information is identifiable by a series of high-rank words occurring with the same frequency. For example, there are eight words occurring exactly $72,307$ times. This copyright information was initially introduced into Swiss-Prot at Version 37, as a result of the funding crisis faced by Swiss-Prot in 1996.

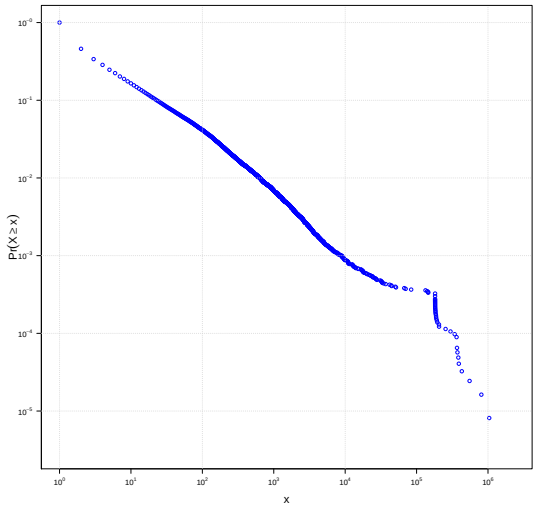
The copyright statements are placed within the comment lines of all UniProtKB entries (i.e. within the "CC" lines). These statements are placed after all other textual annotation, as well as being enclosed by comment lines composed entirely of hyphens, as shown in Figure 4.3. Figure 4.3 shows the initial copyright statement that was present between Swiss-Prot Versions 37 and 45. This initial copyright statement was subsequently replaced by a shorter copyright statement in UniProtKB Version 4, as



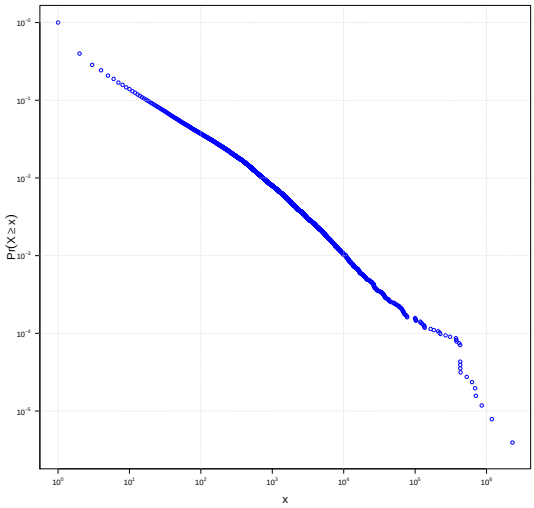
(a) Swiss-Prot Version 35



(b) Swiss-Prot Version 44



(c) UniProtKB/Swiss-Prot Version 5



(d) UniProtKB/Swiss-Prot Version 15

Figure 4.2: Figure showing the power-law model applied to four versions of Swiss-Prot. Figures (b), (c) and (d) show a distinct structure in the tail of the power-law, caused by the introduction of copyright and licence information. Specifically, this structure is visible close to 10^5 in Figures (b) and (c) and between 10^5 and 10^6 in Figure (d).

Rank	Word	Occurrences
1	the	372,301
2	is	320,122
3	and	280,013
4	a	188,938
5	by	176,442
6	of	173,556
7	to	162,853
8	as	156,783
9	this	156,380
10	its	147,955
11	no	145,050
12	———————————	144,614
13	institute	144,614
14	bioinformatics	144,614
..	...	...
42	agreement	72,308
43	commercial	72,308
44	swiss	72,308
45	restrictions	72,307
46	swiss-prot	72,307
47	institutions	72,307
48	entities	72,307
49	non-profit	72,307
50	send	72,307
51	copyright	72,307
52	outstation	72,307

Table 4.4: List of the most commonly occurring words in Swiss-Prot Version 37.

shown in Figure 4.4. A further revision was applied to reduce the copyright statement to four lines, as shown in Figure 4.5. This revised copyright statement was introduced in UniProtKB Version 7, and has remained in all of the future UniProtKB releases.

```
CC   ---------------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between  the Swiss Institute of Bioinformatics  and the  EMBL outstation -
CC   the European Bioinformatics Institute.  There are no  restrictions on  its
CC   use  by  non-profit  institutions as long  as its content  is  in  no  way
CC   modified and this statement is not removed.  Usage  by  and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------------
```

Figure 4.3: Copyright statement added to Swiss-Prot annotation between Versions 37 and 45. This version of the copyright statement contains a total of 9 lines.

```
CC   ---------------------------------------------------------------------------
CC   This Swiss-Prot entry is copyright. It is produced through a collaboration
CC   between  the Swiss Institute of Bioinformatics  and the  EMBL outstation -
CC   the European Bioinformatics Institute.  There are no  restrictions on  its
CC   use as long as its content is in no way modified and this statement is not
CC   removed.
CC   ---------------------------------------------------------------------------
```

Figure 4.4: Copyright statement added to UniProtKB annotation between Versions 4 and 6. This version of the copyright statement contains a total of 7 lines.

```
CC   ----------------------------------------------------------------------
CC   Copyrighted by the UniProt Consortium, see http://www.uniprot.org/terms
CC   Distributed under the Creative Commons Attribution-NoDerivs License
CC   ----------------------------------------------------------------------
```

Figure 4.5: Copyright statement added to the annotation in UniProtKB Version 7. This version of the copyright statement contains a total of 4 lines and remains in the latest version of UniProtKB.

The refinement of the copyright statement has reduced the footprint required within each UniProtKB entry, with a total of five lines being removed from the initial version. These revisions were likely done to reduce the size of the flat files. For example, if the original copyright statement remained in UniProtKB/TrEMBL Version 2012_05, then the flat file would be over 110 million lines longer and approximately 8GB larger. The

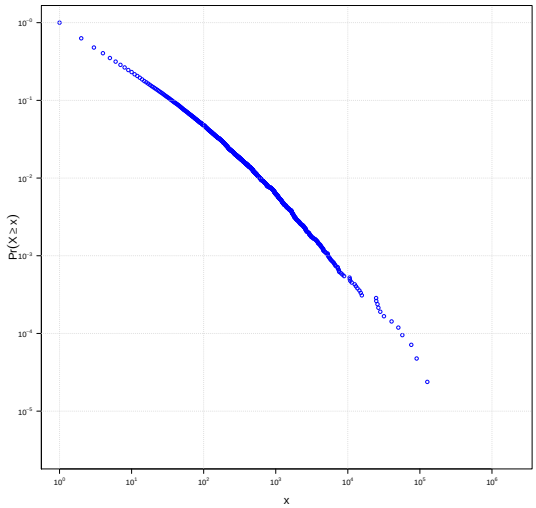
refinement of the copyright statement explains why the kink in these figures become less distinctive over time.

This analysis shows that the introduction of a large amount of material into the annotation with no biological significance can be detected. Additionally, revisions to this material can also be detected. Therefore, this analysis demonstrates that the power-law model can be used as a partial measure of quality, albeit for detecting artefacts.

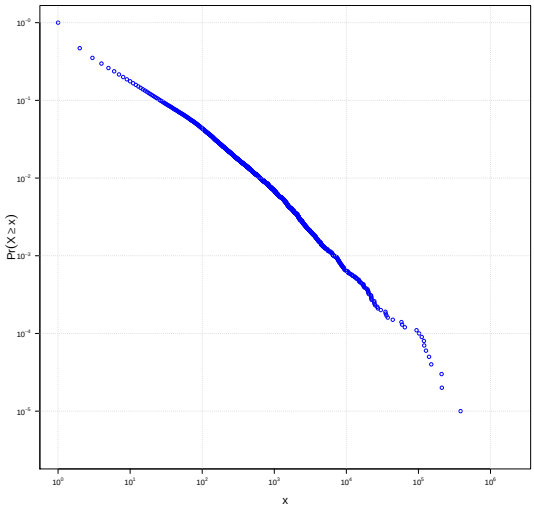
With the identification of the copyright and licence information, BANE was extended to allow copyright information to be excluded from the parsed output. The removal of copyright can be achieved by simply identifying comment lines consisting entirely of hyphens, and excluding those lines, along with the comment lines contained between them. Following the extension of BANE, the series of tests previously discussed were re-run to ensure data was still parsed correctly.

Updated graphs, with copyright statements removed, are shown in Figure 4.6. These graphs clearly show the impact that the removal of the copyright has on the fitting of the power-law; the tail of the graphs are subsequently much smoother. However, the power-law model applied to Swiss-Prot Version 35 is identical in both Figure 4.2a and Figure 4.6a, as entries within Swiss-Prot Version 35 contain no copyright. This provides further confidence that the attempted removal of copyright in those database versions without copyright statements are unaffected.

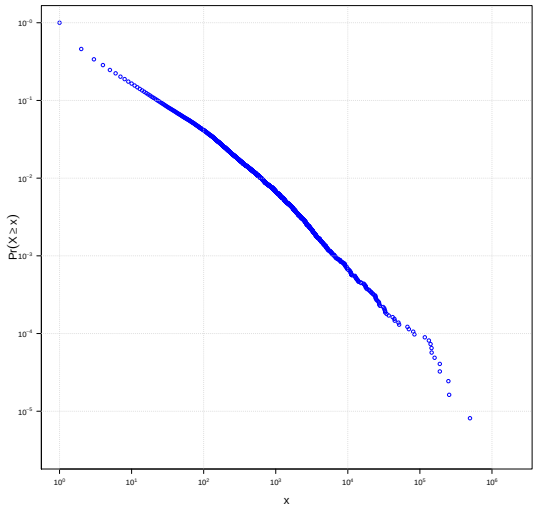
Visual inspection of these graphs show that the power-law changes over time. For example, the graph contains more data points, as the amount of annotation increases, in addition to the head and tail having different gradients. This is a marked two-slope behaviour which is commonly seen for mature resources, such as large complex natural languages [257, 290]. This change over time is analysed in the following section.



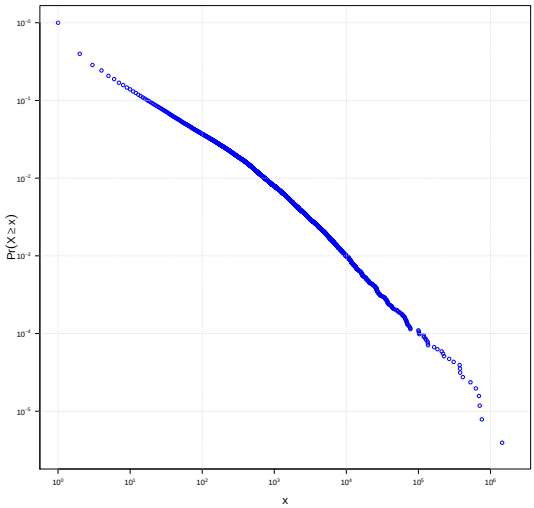
(a) Swiss-Prot Version 35



(b) Swiss-Prot Version 44



(c) UniProtKB/Swiss-Prot Version 5



(d) UniProtKB/Swiss-Prot Version 15

Figure 4.6: Figure showing the application of the power-law model to four versions of Swiss-Prot. These versions are shown with copyright and licence statements removed; the same versions without the removal of copyright are shown in Figure 4.2. The tail of each power-law is significantly smoother following the removal of the kink caused by the copyright statements. However, Swiss-Prot versions prior to the introduction of copyright are not impacted by this removal, as demonstrated by Figure (a) being identical to Figure 4.2a.

4.3 Analysing Swiss-Prot Annotation Over Time

Visual inspection of the power-law graphs are clearly beneficial, although further analysis is somewhat troublesome. There are over 70 available Swiss-Prot versions, each of which has a corresponding power-law graph. Attempting to display each of these graphs in a paper-based format is problematic given the obvious space restrictions. Further, the changes over time are best viewed continuously; a feature not easily implemented given the static nature of paper-based publication.

Although these issues can be alleviated by various techniques, such as providing a subset of graphs or producing a flip book, the largest restriction is that the change in behaviour over time cannot be easily quantified or measured from visual analysis alone. However, as discussed in Section 3.1, a dataset can be characterised by relating the α value obtained from QUALM to Zipf's principle of least effort. Therefore, by extracting the value of α from each version of Swiss-Prot the change in behaviour over time can be analysed, as shown in Figure 4.7.

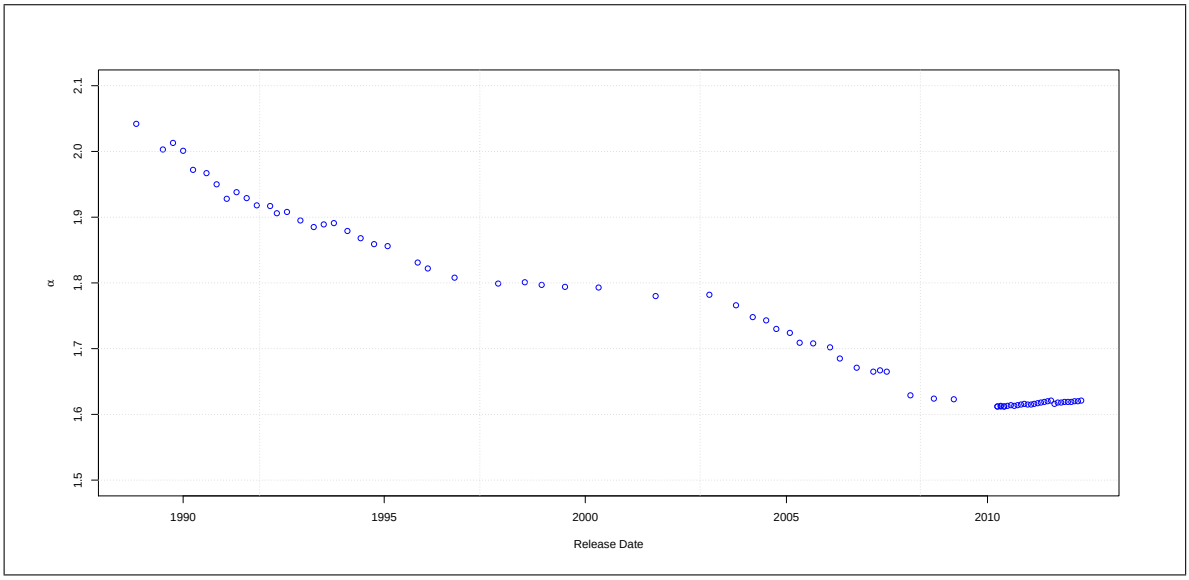


Figure 4.7: α values over time for each version of Swiss-Prot.

Figure 4.7 shows a decline in α values over time; a measure that could not be drawn from visual inspection alone. This decline in α values shows that the annotation in Swiss-Prot is changing in its nature over time. This is further supported by the development of two slopes in the power-law graphs, as visually identifiable in Figure 4.6.

Based on this relationship, the decline in α value appears to suggest that Swiss-Prot is becoming more optimised for the annotator, rather than the reader, over time. Specifically, the α values for the latest versions of Swiss-Prot are just above 1.6, which deems the annotation to have been least effort for the annotator. Conversely, the α value for initial versions of Swiss-Prot was above 2, which categorises the annotation as least effort for the reader.

This optimisation is also reflected by the change in the fifty most commonly occurring words between Swiss-Prot Version 9 and UniProtKB/Swiss-Prot Version 2012_05, as shown in Table 4.5. Between these versions there has been a reduction in words which commonly occur in general English language. For example, the words "be", "other" and "has" are no longer in the top 50 most popular words, being replaced by words such as "subunit", "biosynthesis" and "ribosomal". The increase in these biological terms appears to be caused by a standardisation in annotation [291].

This conclusion also fits with previous research from Baumgartner *et al.* [133], which suggests that the enormous increase in the number of proteins requiring annotation is outstripping the ability to provide this annotation. Indeed, this issue has been acknowledged by UniProt, with their introduction of automated annotation. This automated annotation in UniProtKB (i.e. TrEMBL) is investigated in the following section.

Word	Occurrences
the	13,056
of	9,764
and	5,241
is	4,918
in	4,771
a	4,166
to	3,607
this	2,880
+	2,868
protein	2,865
are	2,276
by	1,676
for	1,501
with	1,382
proteins	1,329
=	1,313
chains	1,311
from	1,269
that	1,045
two	1,000
it	970
which	968
as	960
an	956
chain	928
one	810
2	773
other	753
&	732
sequence	725
be	666
at	666
enzyme	663
complex	644
identical	613
or	604
gene	571
membrane	566
alpha	560
dna	530
c	508
binding	493
beta	482
cell	478
component	468
i	457
binds	453
has	453
three	442
cells	439

(a) Swiss-Prot Version 9

Word	Occurrences
the	2,017,869
by	1,126,283
of	1,000,621
to	934,857
similarity	919,294
and	763,469
in	592,064
+	509,051
family	484,643
belongs	482,979
a	444,911
protein	350,003
is	313,421
membrane	290,081
with	259,327
=	245,879
1	234,422
for	187,929
from	186,402
biosynthesis	180,928
domain	170,772
cytoplasm	166,211
subunit	164,160
contains	162,625
cell	144,803
complex	142,870
binds	139,425
2	134,047
as	117,890
atp	109,410
that	103,136
step	102,658
ribosomal	102,293
at	101,636
proteins	94,871
involved	94,266
an	94,175
phosphate	94,141
catalyzes	93,775
it	90,740
or	90,165
dna	86,000
may	85,903
interacts	82,115
which	80,106
subfamily	79,978
potential	75,581
are	72,931
this	72,430
activity	69,242

(b) UniProtKB/Swiss-Prot Version 2012_05

Table 4.5: Change in the top 50 words, and their occurrences, between Swiss-Prot Version 9 and UniProtKB/Swiss-Prot Version 2012_05.

4.4 Swiss-Prot Vs. TrEMBL

Within UniProtKB, proteins are initially annotated automatically and placed into TrEMBL. Eventually they are manually annotated and placed into Swiss-Prot. Therefore, TrEMBL and Swiss-Prot are ideal resources by which to compare equivalent human and automated annotations. However, a comparison between the early versions of the two resources is not straightforward.

As previously described in Section 2.4, the first version of TrEMBL was introduced in 1996, ten years after the first version of Swiss-Prot. Until the formation of the UniProt consortium, TrEMBL and Swiss-Prot releases were not synchronized, with TrEMBL being released more frequently than Swiss-Prot. Therefore, to allow versions prior to UniProtKB version two[4] to be compared, the version of TrEMBL released most closely in time to each version of Swiss-Prot is used. This mapping is shown in Table 4.6 and allows the behaviour between Swiss-Prot and TrEMBL to be investigated at equivalent points in time.

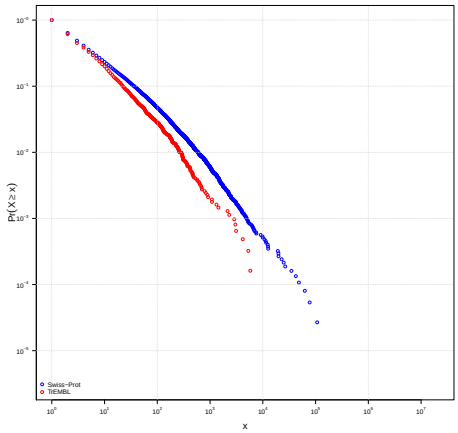
Date	Swiss-Prot version	Date	TrEMBL version
Oct-96	34	Nov-96	1
Nov-97	35	Jan-98	5
Jul-98	36	Aug-98	7
Dec-98	37	Jan-99	9
Jul-99	38	Aug-99	11
May-00	39	May-00	13
Oct-01	40	Oct-01	18
Feb-03	41	Mar-03	23
Oct-03	42	Oct-03	25
Mar-04	43	Mar-04	26

Table 4.6: Mapping between TrEMBL and Swiss-Prot release dates. Each version of Swiss-Prot is associated to the nearest version of TrEMBL based on release date.

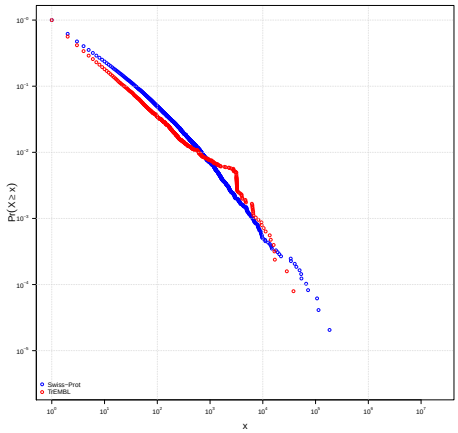
Using this mapping, an evenly spaced subset of the resulting graphs from the power-law model is shown in Figure 4.8. These graphs combine the equivalent Swiss-Prot and TrEMBL versions within a single graph, allowing the resulting power-law model for the databases to be more easily compared.

Inspection of the TrEMBL graphs shows that a kink appears in UniProtKB Version 15, as shown in Figure 4.8d, at approximately 10^6. This kink is similar to the one caused

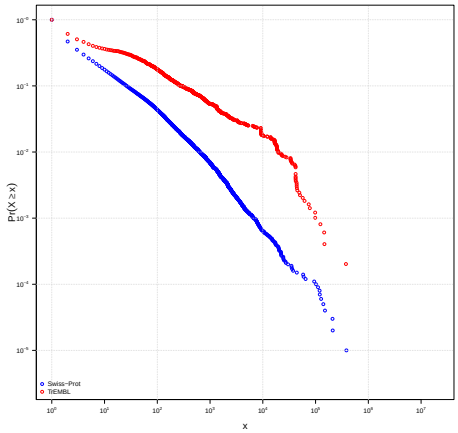
[4]Version 2 was the first major release of UniProtKB



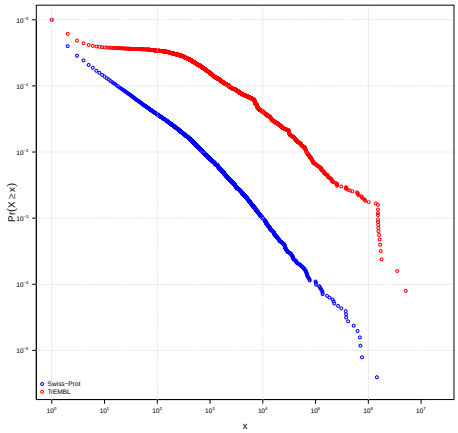
(a) TrEMBL Version 1 & Swiss-Prot Version 34



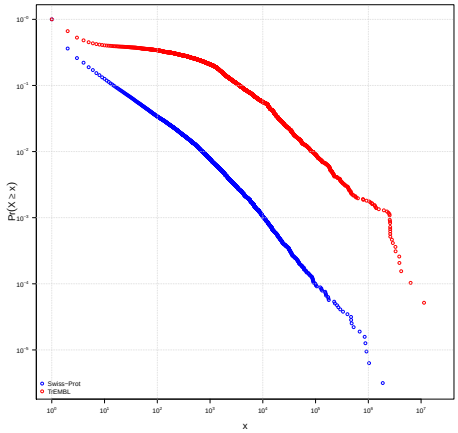
(b) TrEMBL Version 13 & Swiss-Prot Version 39



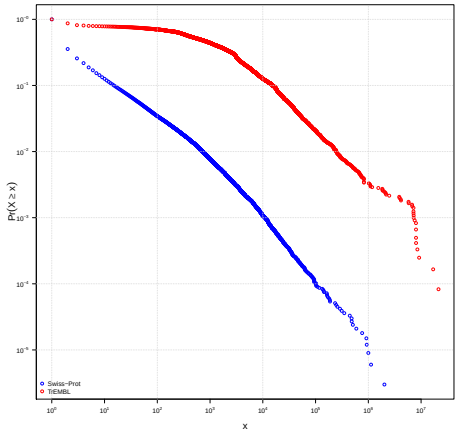
(c) UniProtKB Version 2



(d) UniProtKB Version 15



(e) UniProtKB Version 2010_12



(f) UniProtKB Version 2012_05

Figure 4.8: Figure showing the power-law model applied to six versions of Swiss-Prot and the nearest TrEMBL release.

by copyright statements in the Swiss-Prot graphs, although it remains significantly noticeable for all subsequent versions of TrEMBL. Inspecting the most commonly occurring words, as shown in Table 4.7b, identifies that the kink is caused by a number of sentences (such as "the sequence shown here is derived from an ensembl automatic analysis pipeline and should be considered as preliminary data." and "the sequence shown here is derived from an embl/genbank/ddbj whole genome shotgun (wgs) entry which is preliminary data.") appearing in a large number of entries. However, unlike the copyright statements which contain no biological information, these sentences represent biological knowledge about the underlying sequence. Therefore this information is not removed from the analysis and suggests that annotations are subject to high levels of reuse in TrEMBL.

It is also clear from Figure 4.8 that TrEMBL and Swiss-Prot diverge over time. This divergence is due to the head of the TrEMBL graphs becoming flatter over time and moving towards the top of the graph (i.e. towards 10^0 on the Y-axis). This flattening is a result of the probability being very similar for many values of X (i.e. the occurrence, or size, of a word). For example, within UniProtKB/TrEMBL Version 2012_05 (Figure 4.8f), the probability of a word occurring five or more times has almost the same probability as a word occurring 25 or more times. Overall, the behaviour shown in these graphs provides evidence that TrEMBL has much higher levels of re-use than Swiss-Prot, with the latter exhibiting more maturity.

The conclusions drawn from this visual analysis can be confirmed by analysing the underlying data. Examining the list of all words and their occurrences in UniProtKB Version 2012_05 shows that approximately 330 million words occur in TrEMBL, whilst approximately 32 million words occur in Swiss-Prot. To gauge how frequently each word is re-used, the corpus of words can be extracted; that is just the distinct (i.e. non-redundant) words. This shows that the Swiss-Prot corpus is composed of approximately $333,500$ words, compared to that of TrEMBL which is composed of approximately $12,000$ words. Further, of these words approximately $1,500$ occur only a single time within TrEMBL, whilst approximately $215,000$ occur a single time in Swiss-Prot. This means that all of the annotations in TrEMBL are composed from a total of $12,000$ words, with approximately $10,500$ of these words being subject to reuse. The

Word	Occurrences
+	5,794
the	5,307
of	4,166
and	3,117
to	3,024
=	2,874
a	2,308
in	2,150
is	1,442
protein	1,323
2	1,089
by	1,083
h2o	916
c	888
are	831
family	789
other	702
membrane	693
belongs	680
heme	646
cytochrome	631
an	615
with	576
which	566
co2	523
4	490
phosphate	483
for	468
subunit	466
ii	464
or	452
o2	437
this	422
atp	421
d-ribulose	421
proteins	413
as	412
binds	374
dna	373
cua	372
mitochondrial	366
n	358
similarity	358
form	343
cell	340
two	331
i	315
enzyme	308
nad+	307
may	306

(a) TrEMBL Version 1

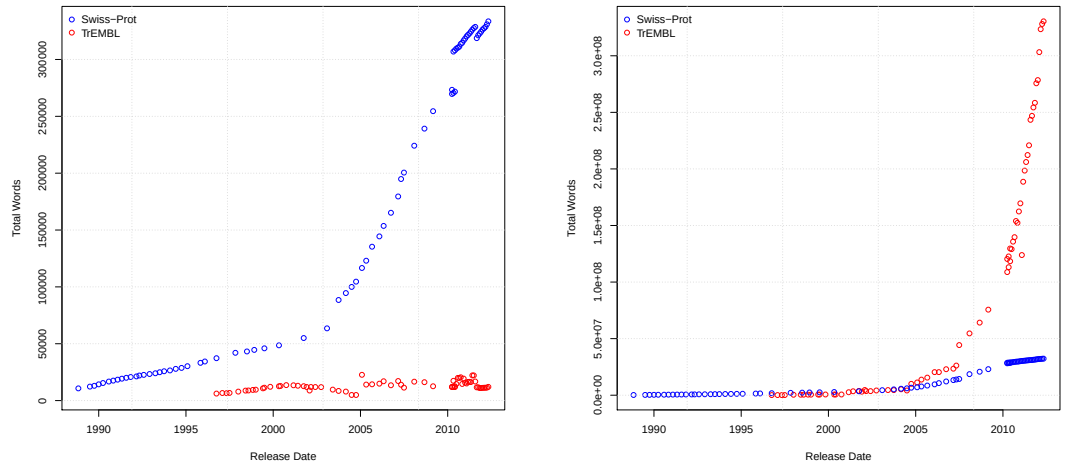
Word	Occurrences
the	21,392,505
is	16,943,254
from	9,182,969
an	8,506,108
sequence	8,030,743
derived	8,014,165
shown	8,010,762
here	8,010,762
data	8,003,183
preliminary	8,003,183
which	7,562,252
entry	7,248,388
genome	7,238,789
embl/genbank/ddbj	7,185,597
whole	7,178,041
shotgun	7,178,018
wgs	7,178,018
to	6,811,630
by	6,741,845
similarity	5,825,675
of	5,786,217
+	4,158,637
belongs	4,096,753
and	3,906,205
family	3,857,021
1	2,491,009
protein	2,181,028
in	2,138,830
a	2,100,318
membrane	1,992,650
=	1,869,842
contains	1,855,722
domain	1,842,929
subunit	1,547,867
as	1,182,001
2	1,097,217
c	1,088,339
be	1,075,645
complex	1,054,401
biosynthesis	1,023,416
should	826,209
considered	825,200
analysis	825,165
pipeline	825,165
ensembl	825,165
automatic	825,165
that	819,739
binds	807,398
cytoplasm	764,439
cell	745,682

(b) UniProtKB/TrEMBL Version 2012_05

Table 4.7: Change in the top 50 words, and their occurrences, between TrEMBL Version 1 and UniProtKB/TrEMBL Version 2012_05.

increase in word reuse is also illustrated in Table 4.7, which shows the change in the top 50 words between the initial version of TrEMBL and UniProtKB/TrEMBL Version 2012_05.

The number of redundant and non-redundant words in UniProtKB over time is shown in Figure 4.9. This figure shows that the corpus of words continues to expand within Swiss-Prot, whilst the size of the TrEMBL corpus fluctuates but typically remains below 20,000 words. These results confirm the conclusions drawn from visual inspection of the power-law graphs and illustrate that, compared to TrEMBL, Swiss-Prot appears to have a much richer use of vocabulary.



(a) Total Non-Redundant Words (b) Total Redundant Words

Figure 4.9: Total number of redundant and non-redundant words in UniProtKB annotation.

The richer use of vocabulary in Swiss-Prot is likely due to the differences in the way Swiss-Prot and TrEMBL entries are curated. However, it cannot be ruled out that the distinction between these two resources could be because they are annotating a different set of proteins. Unfortunately, it is not possible to check a protein's annotation in both Swiss-Prot and TrEMBL at the same point in time; once a record is migrated to Swiss-Prot, and manually annotated, it is removed from subsequent versions of TrEMBL. This is necessary as Swiss-Prot is used as a basis for the automated annotation of TrEMBL, so proteins not removed from TrEMBL could have their automated annotation based on their manual annotation in Swiss-Prot. However, the rapid

increase in size of both resources argues against this explanation; the set of proteins annotated by each resource also changes significantly over time.

Although visual inspection of the power-law graphs highlight that Swiss-Prot has a relatively regular progression, whilst TrEMBL does not, this view suffers the drawbacks discussed in Section 4.3; that is the inability to quantitatively measure and analyse change over time. Therefore, Figure 4.7, which shows the α values for each version of Swiss-Prot, is extended to include the α values from each version of TrEMBL. This updated graph is shown in Figure 4.10.

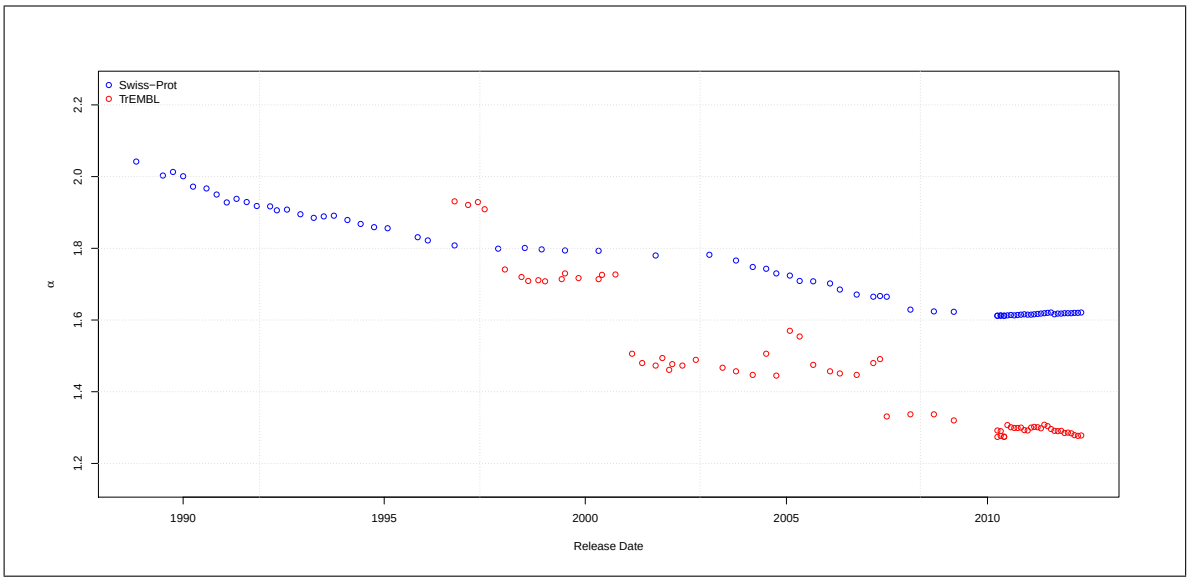


Figure 4.10: α values for all versions of Swiss-Prot and TrEMBL over time.

Figure 4.10 shows that, like Swiss-Prot, TrEMBL also has a decline in α values over time. However, the decline of α values is less regular in TrEMBL, with three significant disjuncts in the relationship, where large jumps occur between releases[5].

The first disjunct appears in 1998, between TrEMBL Versions 4 and 5. Based on historical events, it appears that this disjunct was the result of new procedures being introduced into the automated curation process [293]. These approaches include making use of the ENZYME database, specialised genomic databases and scanning for PROSITE patterns compatible with an entries taxonomic range. These PROSITE patterns are used to enhance the content of the comment lines by adding information such as protein function and subcellular location.

[5]The published account of this work (see [292]) only discusses the first two disjuncts as the third disjunct only became apparent after the addition of more recent UniProtKB versions.

The second disjunct appears in 2001, between TrEMBL Versions 15 and 16. Similar to the first disjunct, it appears that amendments to the curation process are also responsible for this disjunct. Specifically, in 2000 TrEMBL began the planning, introduction and development of annotation rules to supplement the existing curation process [226]. These annotation rules were developed with the aim of further increasing the coverage of automated annotation, by exploiting the existing annotation from within Swiss-Prot.

The third disjunct appears in 2007, between UniProtKB/TrEMBL Versions 11 and 12. Unlike the previous two disjuncts, the literature does not appear to document any fundamental changes that would significantly impact the textual annotation. However, around this time, PIRSF[6] site rules appear to have been introduced, which are used to propagate structural annotation [294]. Although the literature only refers to site rules (PIRSR), it is possible that name rules (PIRNR) were also introduced at this time, which would involve the propagation of textual annotation.

If these disjuncts are evidence of the curation process being changed to enhance the coverage of annotation, then it would be expected that new words and phrases would be introduced which, potentially, would affect the measures described here. Indeed, the total number of words between versions exhibiting the disjuncts shows a significant rise, compared to those for nearby releases, as shown in Table 4.8. Specifically a $\sim 200\%$ rise in total words is seen between TrEMBL Versions 4 and 5, with a $\sim 315\%$ rise is observed between TrEMBL Versions 15 and 16, whilst a rise of $\sim 70\%$ is seen between UniProtKB/TrEMBL Versions 11 and 12.

Whilst changes within the TrEMBL curation process appear to explain the first two disjuncts, it is possible that they are coincidental. Therefore the UniProt help desk was contacted to establish reasons that could be used to explain these disjuncts[7] [295]. Unfortunately, UniProt could not provide detailed explanations regarding early versions of TrEMBL, due to the age of the database. Features, such as a relational database, that UniProt currently rely upon for their detailed statistics and information were

[6]The Unified Rule (UniRule) system, as discussed in Section 2.4, incorporates the Protein Information Resource (PIR) rule systems PIRNR and PIRSR, which are based on PIRSF.

[7]This third disjunct was not discussed with UniProt due to it only becoming apparent after the main work in this thesis was complete.

TrEMBL version	Distinct words	Total number of words	% Change
3	6,527	123,548	−6.2%
4	6,757	135,757	9.9%
5	**7,907**	**406,480**	**199.4%**
6	8,785	437,785	7.7%
7	8,897	464,962	6.2%
...	...	...	...
14	12,846	624,471	0.6%
15	13,612	634,471	1.6%
16	**13,459**	**2,642,548**	**316.5%**
17	13,069	3,479,253	31.7%
18	12,671	3,793,128	9.0%
...	...	...	...
10 (UniProtKB)	17,216	23,450,288	2.1%
11 (UniProtKB)	13,956	26,190,723	11.7%
12 (UniProtKB)	**11,384**	**44,312,809**	**69.2%**
13 (UniProtKB)	16,647	54,626,812	23.3%
14 (UniProtKB)	16,192	64,166,377	17.5%

Table 4.8: The increase of both redundant and non-redundant words between certain versions of TrEMBL. We show the percentage change in redundant words between versions to emphasise the significant increase in total words between certain versions (highlighted in bold).

not available for these early versions. Further, many of personnel who worked on the database have since left, retired or would be unable to accurately recall specific details. UniProt could, however, confirm that extensive work was undertaken to improve the data in 2001, which does correspond to the given explanation for the second disjunct.

Whilst the literature suggests that the scanning of PROSITE was introduced after the initial few versions of TrEMBL, UniProt believe that it was actually in effect from the first version of TrEMBL. The scanning of PROSITE patterns, as previously discussed, was also introduced in the literature with a number of other procedures. It is possible, therefore, that this disjunct may be due to an alternative approach, such as the scanning of the ENZYME database. Given that changes to the annotation process appear responsible for the second disjunct, it is highly plausible, given this evidence, that such changes were also responsible for the first disjunct.

The increase in the total number of redundant words, as shown in Figure 4.9, correlates with the increase of entries into UniProtKB; the rate of data being added exhibits an exponential trend. This increase of data means that entries and annotations within UniProtKB are of mixed age. The current analysis only differentiates between au-

tomated and manual annotations on a bulk scale. Therefore, given the decline of α values in both Swiss-Prot and TrEMBL, it is beneficial to analyse and explore how subsets of entries based on age compare.

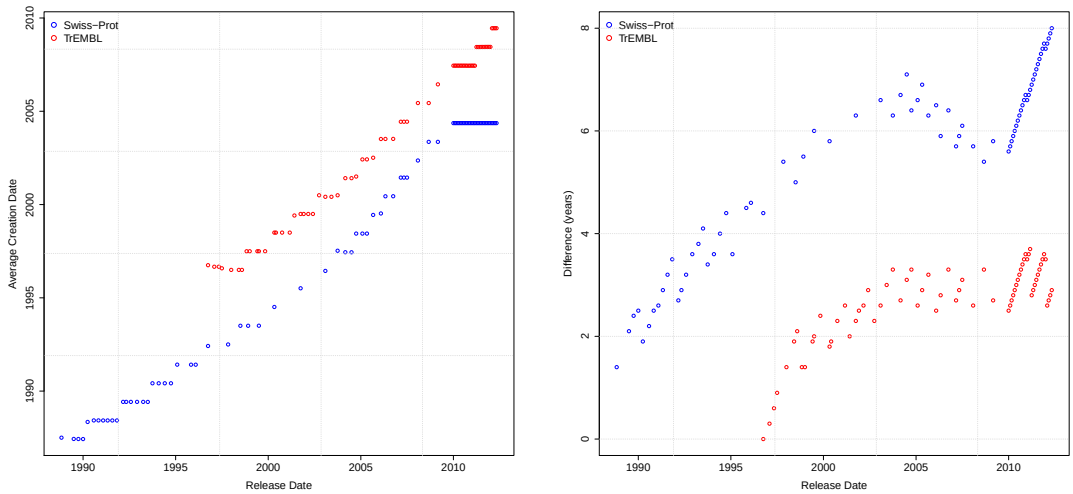
4.5 Analysing Maturity of Entries Over Time and the Impact of New Annotations

Previously, the analysis of UniProtKB has investigated annotation quality in bulk, without analysing how individual records are maturing. If quality is a function of maturity or age of a record, then it would be expected that individual entries should be improving over time, even if, due to the rapid increase in size of UniProtKB the data as a whole is not.

Each entry within UniProtKB contains three date stamps indicating: when the entry was first introduced into the database; the last modification date of the entry and the last modification date of the sequence. By extracting the creation date from each UniProtKB entry, the average record age can be calculated, as is illustrated in Figure 4.11a. Using this information it can be seen that the average age of a record has increased only slowly over the life span of UniProtKB as a whole. From this graph, it can be calculated that, although Swiss-Prot is currently around 25 years old, the average record age is actually around eight years old. This difference between the average age and release date for all versions of UniProtKB is illustrated in Figure 4.11b. For example, Figure 4.11b shows that Swiss-Prot Version 9 has an age difference of 1 year and 4 months, which is calculated based on the difference between the release date (November 1988) and the average entry release date (July 1987).

Figure 4.11 shows similar patterns for both Swiss-Prot and TrEMBL, accounting for the fact that Swiss-Prot is ten years older than TrEMBL. However, in Figure 4.11a, it is noticeable that Swiss-Prot, and to a lesser extent TrEMBL, maintain the same average age for a number of recent releases. This constant average age coincides with the introduction of more regular releases of UniProtKB, which has also seen a reduction in the number of Swiss-Prot entries being added, as shown in Figure 2.9a.

These figures emphasise the increasing size of UniProtKB and the corresponding effect on the average age of entries. Therefore, in order to assess whether individual records appear to be maturing, it is necessary to abstract away from the increasing size of the database. Such an analysis, however, is not straightforward; essentially, a set of records which relate to a defined set of proteins is needed.



(a) Average entry age (b) Age difference

Figure 4.11: The average entry age and the difference between release date and average age for each version of UniProtKB.

To achieve such an analysis, the annotations from entries common between Swiss-Prot Version 9 and other versions of Swiss-Prot were extracted. This extraction of annotation from entries common across all databases versions allows an equal comparison between a set of records over the history of the database. The resulting α values from this analysis are shown in Figure 4.12, in addition to the α value for the entries remaining in the database (i.e. those entries that are not in Swiss-Prot Version 9).

This result shows that the α value for the mature set of entries has decreased over time, correlating with the Swiss-Prot database as a whole. However, the overall decrease in α value is reasonably small compared to the α values for the remaining entries. Although the difference in α value between the subset of common entries and the remaining entries initially increased significantly, it has started to slowly reduce, with later versions showing only a minimal change in α value.

Given that the α value for mature entries has generally decreased over time, it is of interest to investigate the α values of entries that are new to each version of Swiss-Prot. To perform this analysis, the annotations from entries that appeared for the first time in a given database version were extracted. The results from this analysis are shown in Figure 4.13. It again would appear that the α value is decreasing over time, similar to that of other Swiss-Prot graphs, with later versions of Swiss-Prot starting to show

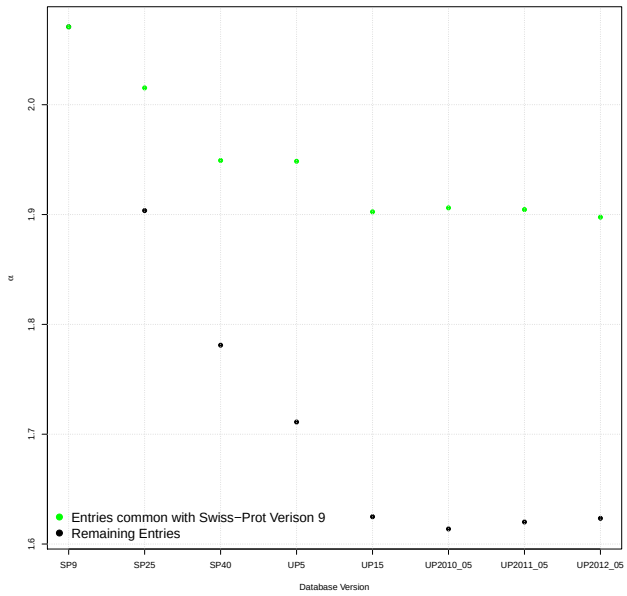


Figure 4.12: Figure showing the α value for all entries contained within Swiss-Prot Version 9 that are also in various other Swiss-Prot versions. In addition, the α value of the remaining entries for each Swiss-Prot version are shown (i.e. the annotation from all entries that weren't in Swiss-Prot Version 9).

improvement.

Since the new release cycle, the α values for Swiss-Prot annotations have steadily increased, with the age difference in UniProtKB/Swiss-Prot Version 2012_05 being at a high of eight years. It appears that changes to the release cycle and annotation procedure have started to slowly improve the quality of both new and existing annotations.

From these analyses, we conclude that there are differences between bulk annotation and individual sets of proteins, either as they mature over time, or as they first enter the database. However, the broad direction of change in the annotation is similar for these subsets as it is for the database as a whole. Therefore, we also conclude that the change in α value that we see in bulk is unlikely to result only from the increase in size of the database.

However, age is not the only factor that can have an impact on annotation quality. UniProtKB categorises proteins in relation to species and taxonomy; analysing these categories allows additional subsets of annotations to be analysed. Specifically, given that some species are model organisms, it would be expected that the quality and wealth of knowledge attached to these proteins would be of higher quality than those

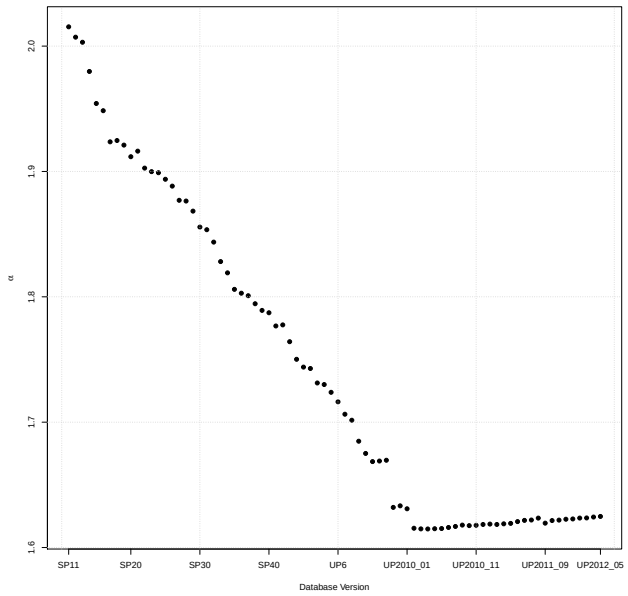


Figure 4.13: α value of annotations from entries new to each version of Swiss-Prot.

of less studied species. These taxonomic divisions are explored in the following section.

4.6 Taxonomic Divisions

In the previous section, UniProtKB annotations of subsets of proteins based on age were analysed. In this section, focus is given to subsets of proteins based on the taxonomy of the organism. While taxonomy describes the evolutionary relationship between two organisms, this analysis investigates the knowledge known about the organisms. By comparing two organisms with similar taxonomic groups, we can abstract away from the biology and investigate simply the distinction between the annotation and the knowledge about different organisms. It would be expected, for instance, that the level of annotation for an extremely well-studied model organism, such as *Saccharomyces cerevisiae* would be significantly more mature than that for any other yeast. UniProtKB provides a taxonomy hierarchy[8] that can be both navigated and searched, allowing the extraction of entire taxonomic groups as well searching for proteins from a specific species.

The taxonomy hierarchy of cellular organisms in UniProtKB is divided into three domains: Archaea; Bacteria; and Eukarya. These domains can be more broadly categorised into either prokaryotes (Archaea and Bacteria) or eukaryotes (Eukarya), based on the presence or absence of a nucleus. Applying the power-law model to annotation based on these two groupings, as shown in Figure 4.14, suggests that annotations from prokaryote entries exhibit more irregularity than those from eukaryote entries. This is also supported by eukaryote entries having a higher value of α (1.7) than those from prokaryotes (1.5). This result is understandable as the majority of model organisms are eukaryotes, as illustrated by over half of Swiss-Prot entries being eukaryotic, whilst the bulk of TrEMBL entries are prokaryotic.

Therefore, this analysis is extended to analyse more specific taxonomic subsets, whilst also making a distinction between Swiss-Prot and TrEMBL entries. UniProtKB make available ten taxonomic divisions (archaea, bacteria, fungi, human, invertebrates, mammals, plants, rodents, vertebrates and viruses) as flat files, for both Swiss-Prot and TrEMBL. Combined, these files encompass all of the entries for the entire database. However, although a number of entries could be present in multiple files (e.g. entries

[8] `http://www.uniprot.org/browse/uniprot/by/taxonomy/`

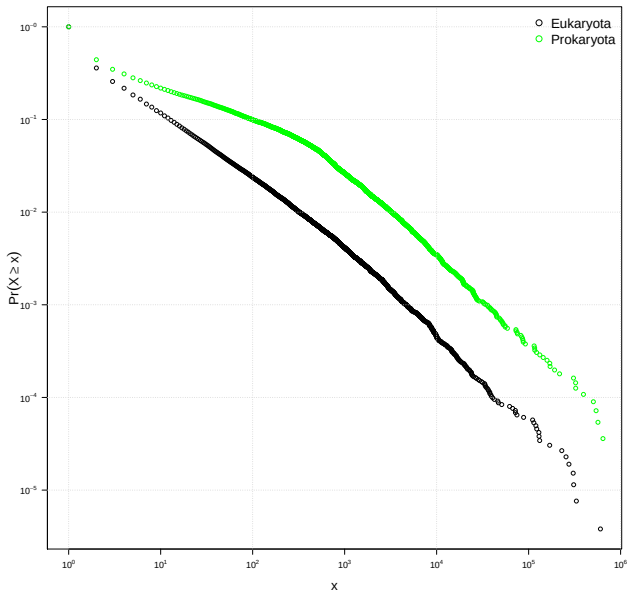


Figure 4.14: Power-law graph comparing eukaryotes and prokaryotes (which is the combination of Bacteria and Archaea). The α value for eukaryota is 1.7, whilst the α for prokaryota is 1.5.

in the human division could also be in the mammals and vertebrates files), each entry is only included a single time, in the most specific division. For example, the invertebrates division contains all eukaryotic entries with the exception of those contained in the vertebrates, fungi and plant files (whilst the vertebrates file does not contain entries contained within the mammals file).

The corresponding α values for each of these ten taxonomic groups, for Swiss-Prot and TrEMBL entries, is shown in Figure 4.15. Within Swiss-Prot the α value in the majority of taxonomic groups is above 1.75, although it is significantly lower for the archaea, bacteria and virus divisions, where it is just below 1.6. Within TrEMBL, the lowest α values obtained are also from the bacteria and viruses divisions, with all divisions having an α value less than their Swiss-Prot counterpart, with a gap of over 0.2 in most cases.

The highest α value observed in Figure 4.15 is from entries within the human division of Swiss-Prot, with the next highest α being obtained from the rodents dataset. These results are likely due to specificity of these divisions, compared to others such as Archaea, with the majority of entries being from well-studied and model organisms such as *Homo sapiens*, *Rattus norvegicus* and *Mus musculus*. The α value for these

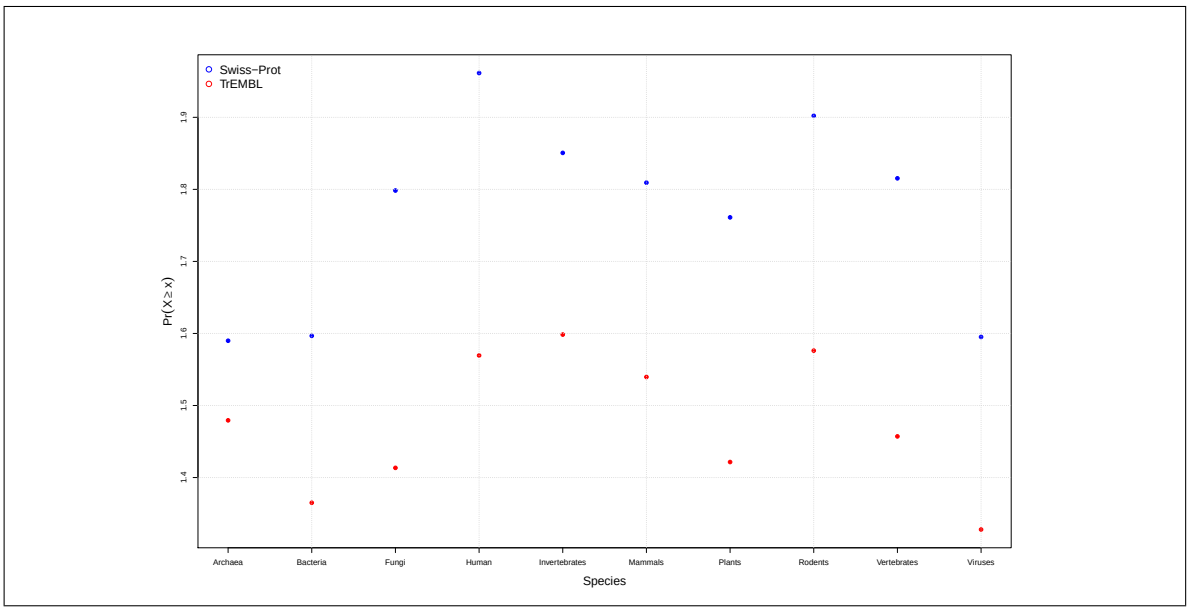


Figure 4.15: α values for Swiss-Prot and TrEMBL based on ten taxonomic groups.

three model organisms, along with a selection of other model and well-studied organisms[9], is shown in Figure 4.16. Although there is some variation in α between these organisms, their α values are higher than the overall database and taxonomic group. This is expected as manual curation efforts in UniProtKB are generally prioritised for model organisms.

To evaluate this further, each of the model organisms can be compared to a taxonomically similar, but less studied species. The resulting α values for a number of these species is shown in Figure 4.17. Unexpectedly, the α value for each of the less studied species is higher than the corresponding model organism. For example, the α value obtained for *Drosophila miranda* is over 2.8, which is significantly greater than the corresponding model organism *Drosophila melanogaster*. Relating this α to Zipf's principle of least effort (Table 3.1) suggests that the *Drosophila miranda* annotation is incomprehensible.

Inspection of the underlying *Drosophila miranda* and *Drosophila melanogaster* power-law graphs, as shown in Figure 4.18b, highlights that *Drosophila miranda* contains very few data points. This is also true for *Rattus rattus*, which is used as a comparison against *Rattus norvegicus*, as illustrated in Figure 4.18a. The lack of annotation

[9]These organisms were selected from UniProtKBs list of reference proteomes `http://www.uniprot.org/taxonomy/complete-proteomes`

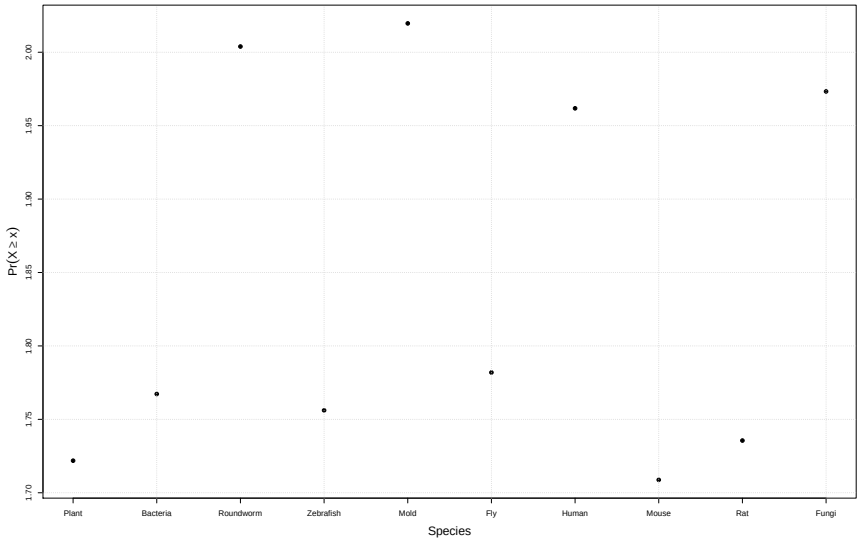


Figure 4.16: α values for a range of model organisms in Swiss-Prot. Plant = *Arabidopsis thaliana*, Bacteria = *Bacillus subtilis*, Roundworm = *Caenorhabditis elegans*, Zebrafish = *Danio rerio*, Mold = *Dictyostelium discoideum*, Fly = *Drosophila melanogaster*, Human = *Homo sapiens*, Mouse = *Mus musculus*, Rat = *Rattus norvegicus* and Fungi = *Saccharomyces cerevisiae*.

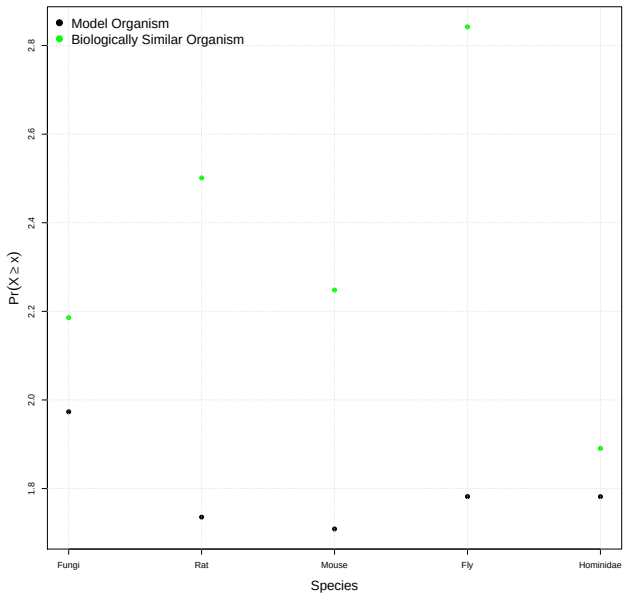
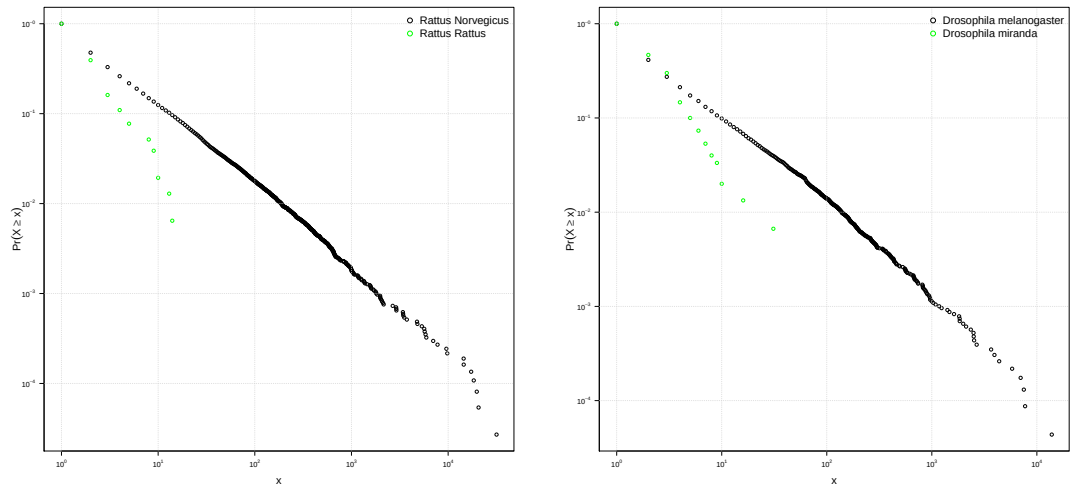


Figure 4.17: α values for a range of model organisms and biologically similar organisms in Swiss-Prot. Fungi = *Saccharomyces cerevisiae* and *Saccharomyces bayanus*; Rat = *Rattus norvegicus* and *Rattus rattus*; Mouse = *Mus musculus* and *Mus spretus*; Fly = *Drosophila melanogaster* and *Drosophila miranda*; Hominidae = *Homo sapiens* and *Pan Troglodytes*.

attached to less studied species is not uncommon, with some species, such as *Pan troglodytes troglodytes*[10], containing no manually annotated entries. This lack of data explains the unexpected α values, making it hard to draw any meaningful conclusions from the α values obtained.



(a) Rat (b) Fly

Figure 4.18: Power-law graphs comparing two species of Rat (*Rattus rattus* and *Rattus norvegicus*) and Fly (*Drosophila miranda* and *Drosophila melanogaster*).

[10] `http://www.uniprot.org/uniprot/?query=taxonomy:37011`

4.7 Discussion

Within this chapter QUALM, which may be a mechanism for assessing annotation quality (Chapter 3), was applied to textual annotation in UniProtKB. UniProtKB was chosen for the initial analysis as it provides various features and support which have allowed us to assess the suitability of QUALM. This evaluation approach was necessary due to a lack of an explicit gold standard dataset. The results from this application suggest that QUALM holds promise as a quality metric.

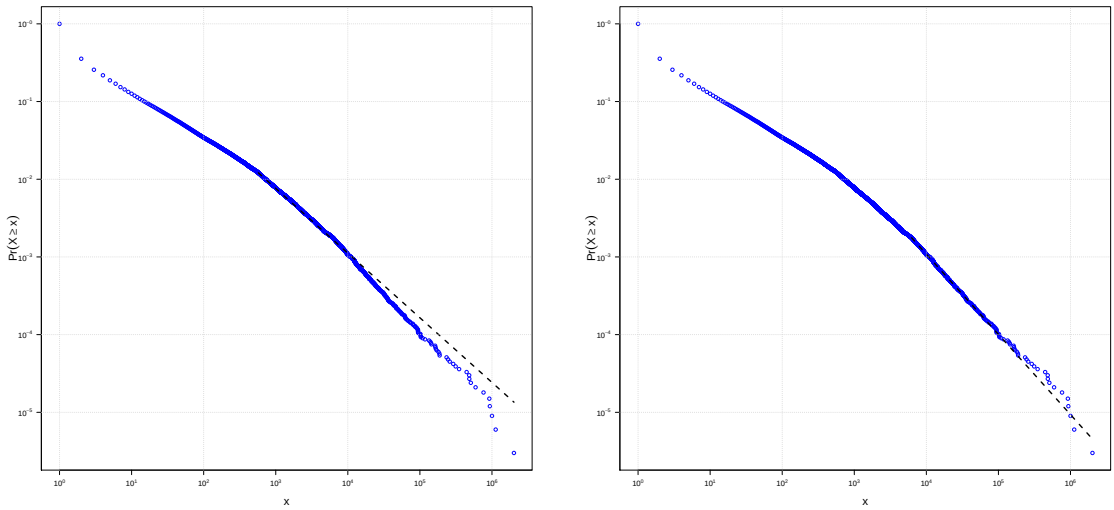
Specifically, the evaluation involved applying QUALM to various subsets of UniProtKB annotation and relating the results to our *a priori* knowledge. In the majority of cases, the results obtained matched our understanding of the annotation. For example, manual annotation was deemed to be of a higher quality than automated annotation (Figure 4.10), whilst model organisms were identified as being of high quality within their taxonomic divisions (Figure 4.15). However, in the latter analysis of investigating taxonomic divisions unexpected results were obtained; these results suggested that the annotations from the model organisms were of lesser quality than those from biologically similar but less studied species.

This result identified a drawback of QUALM, which is the requirement of a bulk corpus of annotation. Although QUALM can technically be applied to small subsets, as shown in Figure 4.18, the values obtained suggest that it provides no analytical value.

A further limitation of QUALM is its inability to handle graphs exhibiting two slopes. A marked two slope behaviour is a feature commonly exhibited in large corpora, such as Wikipedia (Figure 3.2a), and is seen in later versions of Swiss-Prot (e.g. Figure 4.6d). Within these graphs, the tail represents frequently occurring words, whilst more specialised and less frequent words are represented by the head. One possible explanation for this, as suggested by Ferrer-i-Cancho [290], is that all texts consist of a kernel corpus and an unlimited corpus which diverge as the text grows. Specifically, the kernel corpus consists of common and versatile words whilst the unlimited corpus is unbounded as it contains highly specific words which may be coined exclusively for a given text. For example, when curating an annotation for a UniProtKB entry, kernel words such as “the” and “by” are needed to help convey the message whilst words

such as gene names provide the specific details and are potentially unique to a single entry. As the database grows the core corpus will be continually utilised, whilst the specialist corpus will be added to and used infrequently, explaining why the head and tail increase at different rates.

The fitting of the power-law model to datasets exhibiting two slopes involves either discarding a significant portion of the graph (i.e. having a high value of $x_{\min}$) or focusing the fitting of the regression line on the head of the graph. For example, Figure 4.19a shows the power-law graph for UniProtKB/Swiss-Prot Version 2012_05, with the regression line fitted to just the tail of the graph, whilst Figure 4.19b shows the resulting graph when fitting the regression line from the head of the graph. α values of 1.55 and 1.83 are obtained for the head and tail of these graphs, respectively.



(a) $x_{\min} = 575$ (b) $x_{\min} = 3$

Figure 4.19: Power-law model applied to UniProtKB/Swiss-Prot Version 2012_05, illustrating the development of the marked two slope behaviour. The regression line is fitted to (a) the tail of the power-law and (b) the head of the power-law.

Performing a goodness-of-fit test[11] on these two α values results in p-values of less than 0.1 being obtained. Obtaining a p-value of less than 0.1 means it is implausible that the α value accurately characterises the dataset. Applying the goodness-of-fit test to all UniProtKB versions finds that confidence can only be gained in α values extracted from UniProtKB versions which do not exhibit two slopes. For example,

[11] Goodness-of-fit tests are provided by the poweRlaw package, as discussed in Section 3.3

Table 4.9 shows how the p-value for derived α values quickly deteriorates after Swiss-Prot Version 27, due to development of two slopes.

Version	α	p-value
Swiss-Prot Version 9	2.04	0.87
Swiss-Prot Version 11	2.00	0.55
Swiss-Prot Version 12	2.01	0.88
...	...	...
Swiss-Prot Version 27	1.90	0.67
Swiss-Prot Version 28	1.88	0.10
Swiss-Prot Version 29	1.87	0.08
Swiss-Prot Version 30	1.86	< 0.01
...	...	...
Swiss-Prot Version 39	1.79	< 0.001
Swiss-Prot Version 40	1.78	< 0.001

Table 4.9: Extracted α values for various Swiss-Prot versions and the p-value obtained from the corresponding goodness-of-fit test. A p-value above 0.1 provides confidence that the α value accurately represents the underlying dataset.

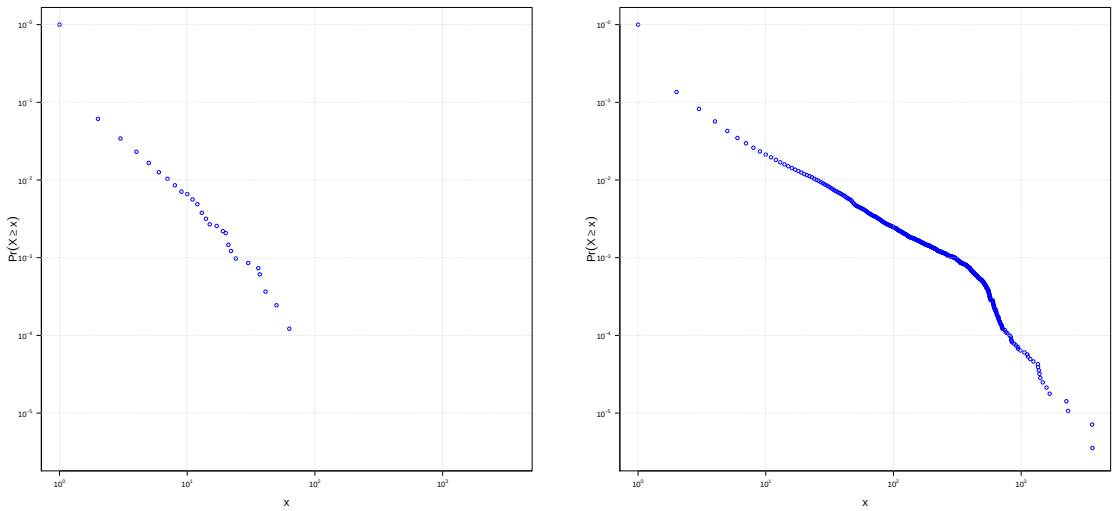
Similar results are also obtained for TrEMBL, which also develops a two slope behaviour, albeit it more pronounced due to high levels of reuse in the head of the graph (see, for example, Figure 4.8f). These results mean that the α values obtained for later versions of UniProtKB should be used cautiously. For example, a direct comparison to Zipf's principle of least effort determines that annotation in UniProtKB/TrEMBL Version 2012_05 is similar to writings produced by individuals with advanced schizophrenia.

Whilst confidence in an α value is vital for making specific quality claims, if used as an approximation it still remains beneficial. For example, analysing the α values for TrEMBL (Figure 4.10) identified disjuncts, which related to substantial revisions in the underlying curation process. Additionally, the analysis of the power-law graphs has also proven beneficial, with the introduction, and subsequent refinements, of copyright statements in Swiss-Prot being identified.

Prior to UniProtKB versions exhibiting two slopes, there is sufficient confidence to claim that early versions of Swiss-Prot and TrEMBL exhibit Zipfian distributions. There is also sufficient confidence for the other datasets analysed, such as mature entries (Figure 4.12). Mature entries were investigated to explore the hypothesis that by abstracting from the growth of UniProtKB, mature entries should improve with

age, as they gain more time from the curator.

However, as also seen for newly added annotations (Figure 4.13), a general decrease in α value is observed over time. One possible explanation for this decrease is the standardisation of annotations between homologous entries, which will involve reusing sections of annotation. To investigate this, the power-law model can be applied to sentences from UniProtKB annotation. Figure 4.20 shows a clear increase in sentence reuse over time, with two slopes also becoming evident.



(a) Swiss-Prot Version 11 (b) UniProtKB/Swiss-Prot Version 15

Figure 4.20: Power-law graphs applied to whole sentences in Swiss-Prot.

The reuse of annotations in UniProtKB has likely increased due to the exponential growth of entries being added. However, the source of an annotation is not always made explicit. If sentences are copied verbatim between entries, as Figure 4.20 suggests, then by analysing this reuse it may be possible to infer the original source of an annotation. The ability to analyse sentence propagation over time will supplement QUALM as an additional tool for users wishing to evaluate annotation quality and correctness. The following chapter explores methods for assessing this sentence reuse.

5

Methods for Visualising and Exploring Annotation Reuse

Contents

Introduction

A by-product of the annotation quality analysis was the identification of annotation reuse; annotations can be copied verbatim between entries, often as a matter of protocol. It also appears that the number of sentences being reused is increasing over time, likely due to the continued addition of raw data. If this is true, then we would expect a continued increase in the amount of annotation reuse over time to match the exponential amount of raw data being added to databases.

Reusing annotation can help reduce the number of unannotated entries and aid curators. However, the source, or *provenance*, of a specific annotation is not always known. Therefore, annotations are potentially based purely, or in part, on existing annotations of unknown origin. If an annotation is copied, or *propagated*, to other entries and is then found to be erroneous, are the entries it has propagated to also affected? If so, has the annotation been updated? We hypothesise that by being able to identify an annotations provenance, and tracking its subsequent propagation, confidence in an annotation's correctness can be obtained.

When propagating an annotation between entries, the amount of relevant annotation will vary. For example, for two particular entries, the annotation in the function topic block may be shared, while the subcellular location annotation may not be. As annotation is composed of free text, reuse can be analysed by splitting annotation into individual sentences. By calculating the various entries and database version(s) that each sentence occurs in, then it is possible that the provenance and propagation of an annotation can be identified.

A dataset of sentences, along with each entry and database versions they occur in, will quickly become vast. Large quantities of interconnected data can be problematic to analyse. Therefore in various disciplines, such as bioinformatics, visualisation techniques are employed to aid the analysis of large datasets. The usage of visualisation techniques provide a mechanism to view these datasets within a single combined image and can help identify patterns that would otherwise be difficult to identify. Within this chapter the suitability of various visualisation approaches that could be applied to sentence analysis are explored (Section 5.1). These approaches include commonly-

used visualisations, such as data flow diagrams (Section 5.1.2), and more specialised visualisations, such as IBM's History Flow tool (Section 5.1.4).

Although none of the analysed approaches provide an entirely suitable visualisation for sentence reuse, a number of key properties and features became evident. These identified features and properties were used to form a set of requirements that a visualisation must fulfil (Section 5.2). Based on these requirements a bespoke visualisation, named Visualising annotatIon PRopagation (VIPeR), was developed to allow the provenance and propagation of a sentence to be identified (Section 5.2.1). A discussion of the developed visualisations suitability concludes this chapter (Section 5.3).

5.1 Existing Visualisation Techniques

Analysing data is a key process within scientific research. However, in many disciplines, data analysis is complex and often compounded by the amount of data generated or collected. Therefore, visualisation is frequently used to ease these issues and aid data analysis. For example, visualising a dataset of protein-protein interactions can help identify clusters.

By visualising annotation reuse it is hypothesised that the provenance and subsequent propagation of a sentence can be identified. This visualisation needs to be based on a list of sentences which states the entry or entries it occurs in and for which database version(s). An example of a dataset that represents this information is shown in Table 5.1.

Sentence	Entry	Database Version
"key control step of glycolysis."	P12345	Swiss-Prot Version 9
"belongs to family 27 of glycosyl hydrolases."	P12345	Swiss-Prot Version 9
"key control step of glycolysis."	P12345	Swiss-Prot Version 11
"belongs to family 27 of glycosyl hydrolases."	P12345	Swiss-Prot Version 11
"belongs to family 27 of glycosyl hydrolases."	P12345	Swiss-Prot Version 12
"belongs to family 27 of glycosyl hydrolases."	P12345	Swiss-Prot Version 13
"belongs to family 27 of glycosyl hydrolases."	P12345	Swiss-Prot Version 14
"belongs to family 27 of glycosyl hydrolases."	Q54321	Swiss-Prot Version 13
"belongs to family 27 of glycosyl hydrolases."	Q54321	Swiss-Prot Version 14
"belongs to family 27 of glycosyl hydrolases."	Q54321	Swiss-Prot Version 15

Table 5.1: A hypothetical dataset representing sentences extracted from textual annotation. For each sentence, the entry (or entries) and database version(s) it occurs in is also recorded.

The dataset shown in Table 5.1 is the minimal amount of data that can be used to visualise sentence propagation, with a need for further data potentially limiting which databases could be analysed. This identifies two key requirements: the visualisation should only require a list of database entries and versions that each sentence appears in (RQ2); and the visualisation should allow any database that can extract the required data to be visualised (RQ3). In addition to these requirements, we also have the core requirement of the visualisation, which is to allow the provenance and propagation of a sentence to be identified (RQ1). Specifically, we can define three requirements:

RQ1 The visualisation should show sentence provenance and propagation over time.

RQ2 To visualise a sentence only the list of entries and database versions it occurs in should be required.

RQ3 The visualisation should be generic, allowing any database with textual annotation to be analysed.

There are numerous different techniques and approaches for visualising data. By having a set of requirements and knowing the type and structure of data to be visualised many unsuitable visualisations can be ruled out. For example, although traditional data visualisations, such as bar charts and pie charts, can be calculated from the dataset they do not allow the flow of a sentence through a database to be easily analysed, meaning they do not meet the core requirement.

However, there are a number of existing visualisations that we believe could provide a suitable visualisation. For each of these approaches we generate a visualisation from a sentence dataset and assess its suitability by determining if it meets our requirements.

5.1.1 Sankey diagrams

A Sankey diagram is a method for showing how an object, or objects, flows between various processes in a system. Sankey diagrams consist of one or more large arrows that are split into various sub-arrows to represent an objects distribution, with the size of an arrow representing the quantity of the object being distributed. For example, the Sankey diagram in Figure 5.1 shows the distribution of energy in a diesel engine. In this Figure there is a single input (fuel) which is distributed into six different outputs; this makes it clear that the majority of the inputted fuel is converted into power, whilst only a small percentage is lost as heat.

Figure 5.1 was produced using the R script "SankeyR" [297]. The Sankey diagram could also have been produced using the Excel macro "Sankey Helper" [298]. However, other implementations are limited as the majority are either not free software or have limited functionality. Additionally, manual creation of diagrams is not straightforward, as arrows need to be kept to scale.

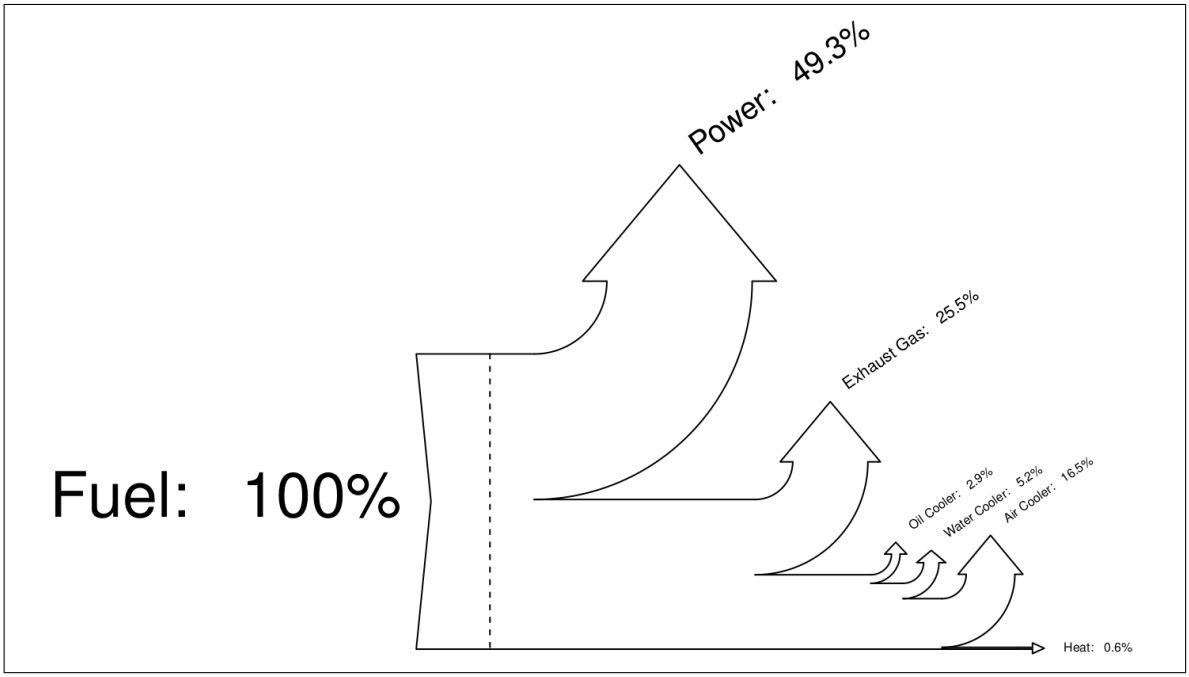


Figure 5.1: Sankey diagram representing the flow of energy in a marine diesel engine. Data taken from [296].

An attempt at creating a Sankey diagram to visualise sentence reuse, using SankeyR, is shown in Figure 5.2. This figure illustrates that a single sentence appears in four database versions, with the arrows size indicating the number of database versions each entry appears in. The exact database versions an entry occurs in is not clear from this view, meaning that the requirement for determining the propagation and provenance of a sentence is not fulfilled. However, this could be achieved by creating a new arrow for each database version and combining them sequentially, although implementing this extension would be difficult due to the lack of available tooling and would unlikely to be intuitive. These issues highlight further requirements that we require from the visualisation: the visualisation should not be limited by inadequate tooling and support (thus we refine RQ3); and the information presented in the visualisation should be intuitive to interpret (thus we refine RQ1).

Sankey diagrams would be of more benefit if the true propagation of a sentence was known. That is, for each sentence in an entry, the actual entry that it originated from was known. If this information was available, rather than being inferred, then a more beneficial visualisation could be produced.

As Sankey diagrams are used to visualise the flow of an objects through a system,

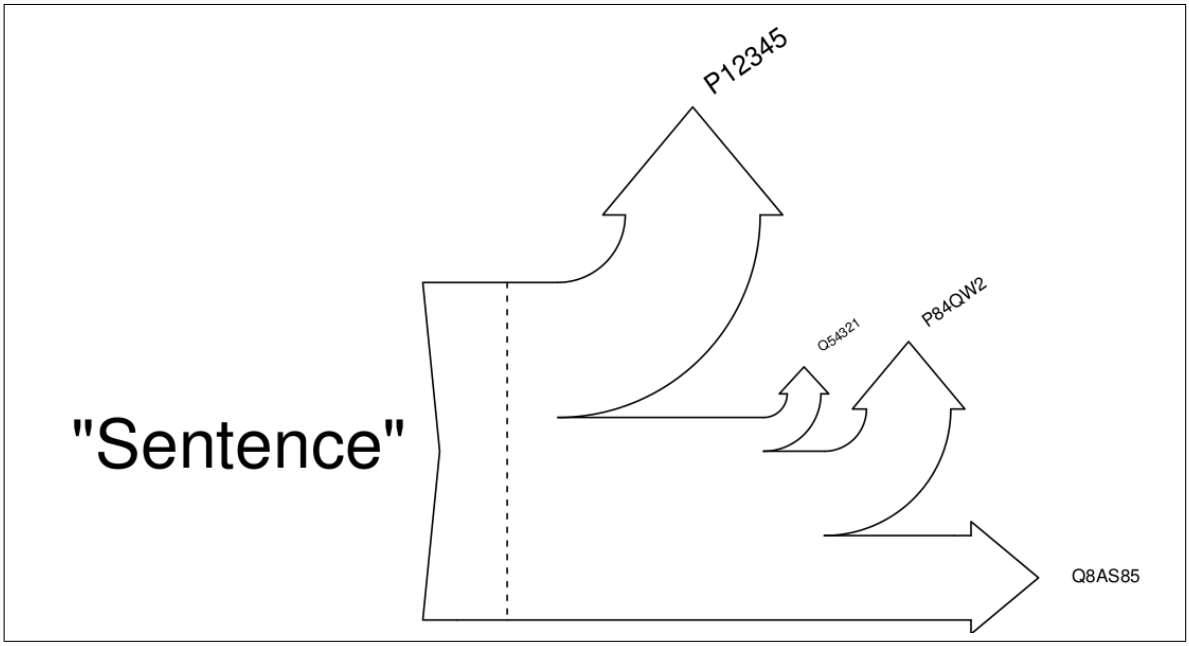


Figure 5.2: An example of a Sankey diagram which depicts the flow of a sentence through the database. The diagram has a single input ("sentence") which ends up in four different entries.

it appeared possible that they would provide an approach for visualising sentence propagation; in practice, however, the Sankey approach is not suitable and is better suited to systems involving numerical quantities, such as energy and currency.

5.1.2 Data flow diagrams

Data flow diagrams are a graphical approach used within computing to represent the flow of data within a system. Data flow diagrams share similarities with other graphical approaches for modelling computing systems, such as Unified Modeling Language (UML) diagrams and flowcharts; each diagram is built from a combination of lines and symbols.

Although these approaches share similarities, they each provide unique features making them more suited to different applications. For example, a flowchart can represent conditions, making it a suitable choice for illustrating stages of an algorithm, while UML diagrams are better suited for database modelling as they can distinguish between different table relationships. However, none of these approaches are ideally suited for visualising sentence reuse, which requires features from all three approaches.

Figure 5.3 provides an illustration of how the propagation of a sentence could be iden-

tified by applying features from each of these approaches. This diagram was produced using Dia [299], although various other programs, such as Microsoft Visio [300], could have been used.

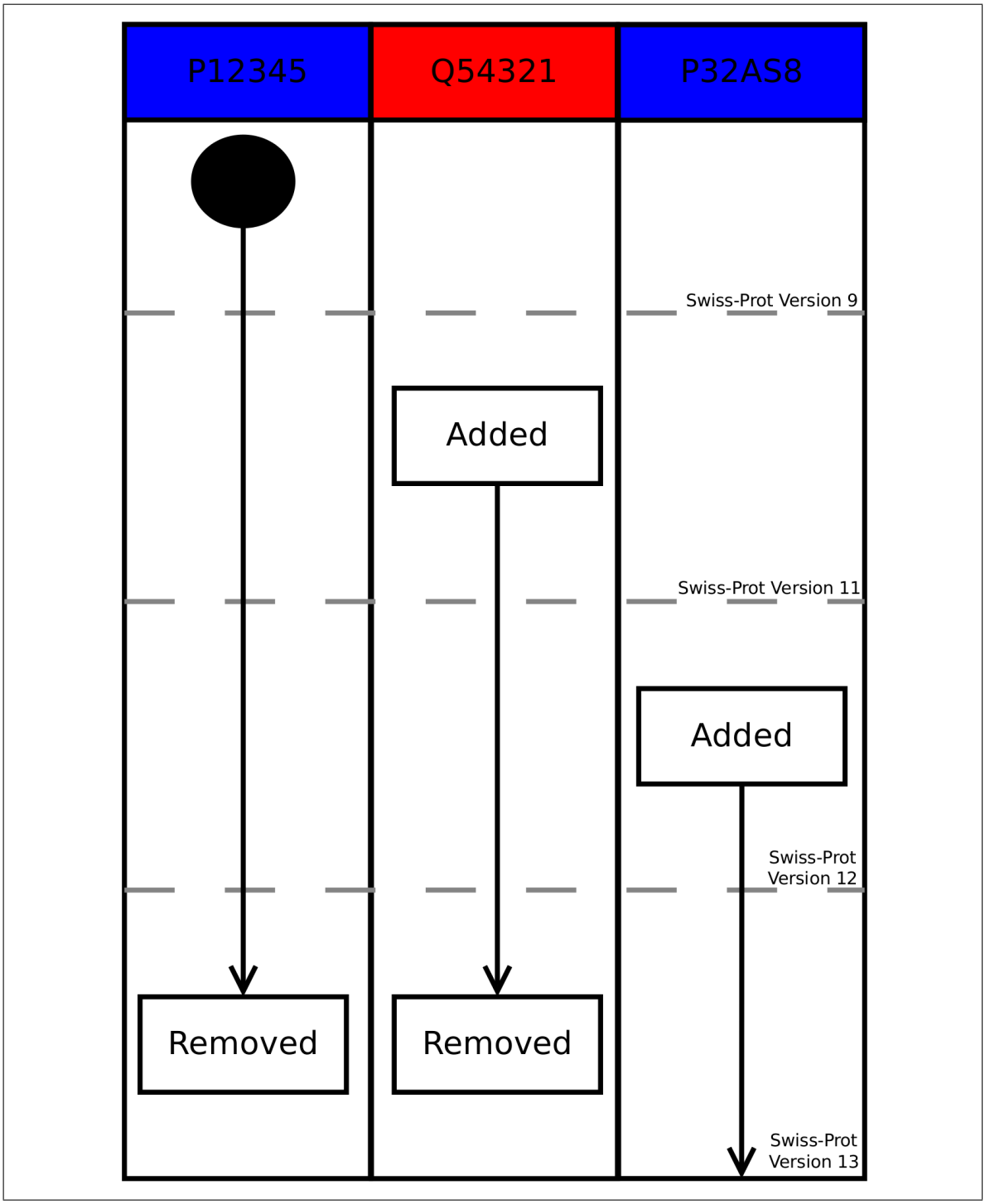


Figure 5.3: Showing the propagation of a sentence through UniProtKB based on features from UML, flowchart and data flow diagrams.

In this figure, each column represents an accession, with grey horizontal lines indicating

UniProtKB releases. The flow of a sentence is indicated by arrows, showing the entries and database versions a sentence appears. For example, the sentence originates in P12345 and remains in the entry for three database releases until it is removed in Swiss-Prot Version 13.

This visualisation allows the provenance and propagation of a sentence to be visualised. However, a major limitation of this approach is that the graph was produced manually. As this visualisation utilises features from three different approaches, no tools to automate the production of a diagram in this form were available. The ability for a user to produce a visualisation automatically with only minimal input is clearly a necessary requirement for the chosen visualisation. Therefore we derive a new requirement (RQ7).

5.1.3 Graph theory and visualisation

Graph theory involves the study of objects and the relationships between these objects. Within a graph, objects are represented as *nodes*, with the relationship between nodes stored as *edges*. When visualised, nodes are shown as circles with a line between two circles representing an edge. Lines, or arcs, can be either directed (indicated by an arrow representing the direction of a relationship) or undirected. A visualisation of a simple graph is shown in Figure 5.4. In this Figure there are three nodes ("A", "B" and "C") with undirected relationships between the nodes "A" and "B" and the nodes "B" and "C".

It is possible to visualise a graph using many standard drawing packages. However, with graph visualisations gaining popularity in fields such as bioinformatics and social network analysis, various tools and dedicated programs for generating graphs have been developed. For example, producing graphs in the R programming language can be done with the igraph package [301], whilst dedicated programs such as Cytoscape [302] and Ondex [303] provide additional features such as layout managers and filters. There are also other programs, such as Vizster [304], that have been developed for a specific application (Vizster was developed for the exploration of online social networks, such as Facebook).

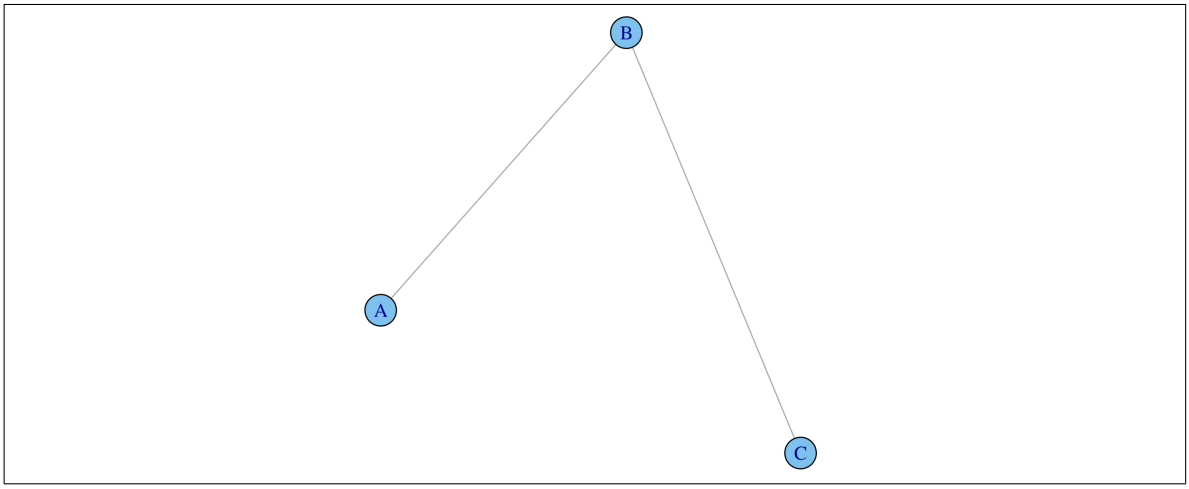
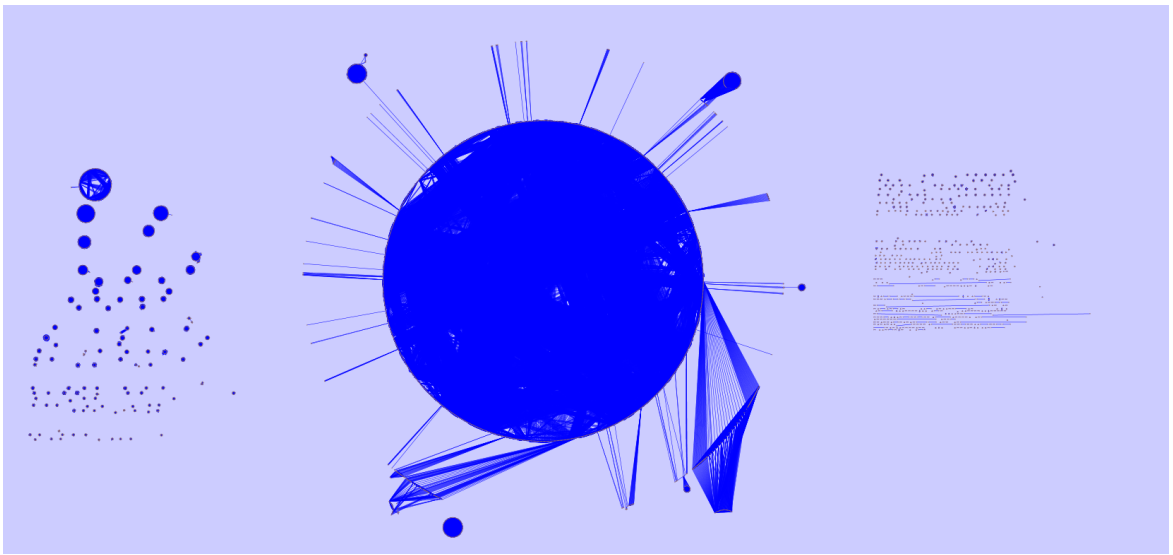


Figure 5.4: An example of a small graph. The graph is composed of three nodes ("A", "B" and "C") and two undirected relationships (between "A" and "B" and also between "B" and "C").

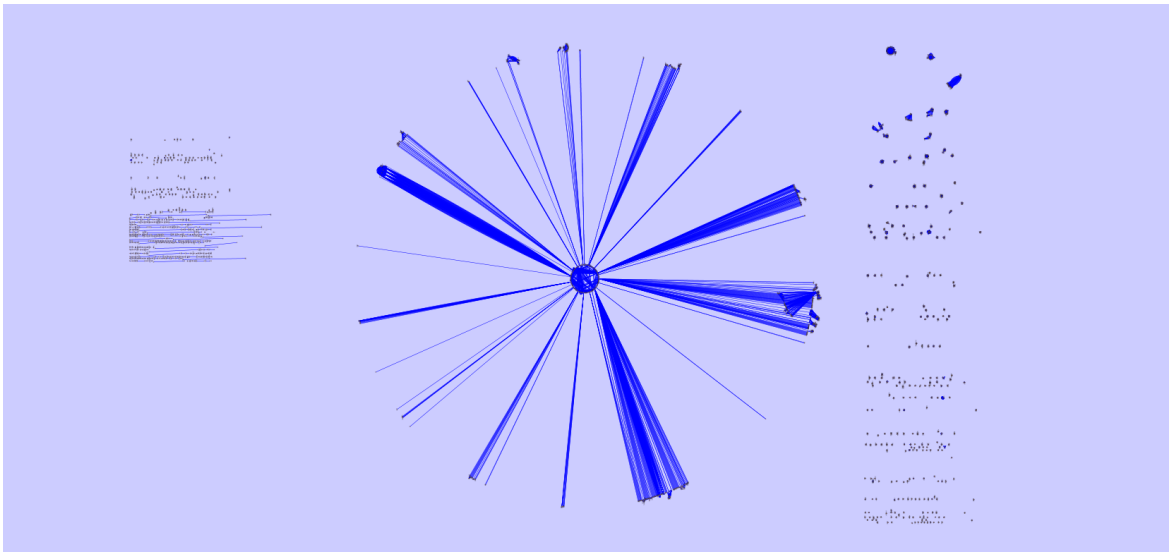
The sentence reuse dataset can be visualised as a graph by organising the data into the form [object – relationship – object]. For example, [P12345 – "key control step of glycolysis." – P54321] would represent two nodes (P12345 and P54321) and a relationship between these nodes (the sentence "key control step of glycolysis." occurs in both entries). This would provide a gene-centric view, illustrating the sentences shared between database entries. Alternatively, a sentence-centric view could be achieved using the form ["key control step of glycolysis." – P12345 – "belongs to family 27"]. In this case, nodes would represent individual sentences with edges representing entries. Examples of these views are shown in Figures 5.5a and 5.5b, which were produced in Cytoscape using the circular layout style.

These figures provide a unique perspective of sentence reuse in Swiss-Prot Version 9. In both views there are clusters of heavily connected nodes, which are often independent of other clusters. These independent clusters, which are more frequent in the gene-centric view, show that there are sets of sentences that are only propagated between a subset of entries. However, although uncommon, there are instances of sentences that are shared between two entries in separate clusters, as illustrated in Figure 5.6.

Although these views illustrate the relationship between entries and sentences within a database, the propagation of individual sentences is not clear. This is partly due to the way in which the data is organised within the visualisation. The current view displays the relationship between all sentences and entries in Swiss-Prot Version 9, but



(a) Gene-centric



(b) Sentence-centric

Figure 5.5: Visualisation of sentence reuse in Swiss-Prot Version 9 using Cytoscape. In (a) nodes represent Swiss-Prot entries, with sentences shared between two entries indicated by edges. Conversely, in (b) nodes represent sentences with edges representing entries. Graphs were organised using the Cytoscape circular layout style.

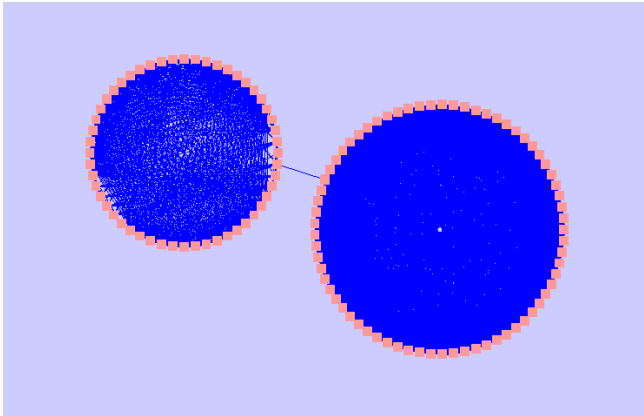


Figure 5.6: Swiss-Prot entries often cluster into groups and share the same annotation. However, this figure illustrates a single sentence ("the active-site selenocysteine is encoded by the opal codon, uga.") which is shared between two entries that are in separate clusters.

to visualise the propagation of a sentence all database versions need to be displayed simultaneously. For example, Figure 5.7 illustrates how the propagation of a sentence could be visualised using this approach. However, automatically producing duplicate nodes for different database versions and attaching the correct edges based on the database version would be problematic. Further, automatically organising nodes in a manner that allows the propagation and provenance of a sentence to be identified would be troublesome. Whilst Figure 5.7 demonstrates that a suitable visualisation is technically achievable using graph visualisation, these issues mean a significant amount of manual intervention would be required, which would quickly become cumbersome and error-prone as a dataset grows.

In addition to the main visualisation, Cytoscape provide features, such as zooming and link outs, that can be beneficial to a user. For example, the link out feature allows a user to visit a website based on a node from within Cytoscape. In the case of UniProtKB, where nodes represent database entries, a Web browser can be launched when the node is clicked to show the entries Web view on the UniProtKB website. These interactive features can aid an analysis, whilst features such as zooming allow dense graphs to be more easily explored. This would be especially beneficial in later database versions, where datasets can consist of many nodes and edges. Given that the exclusion of such interactive features could hinder many analyses, it is necessary to have zooming and link out features as requirements for the chosen visualisation (thus

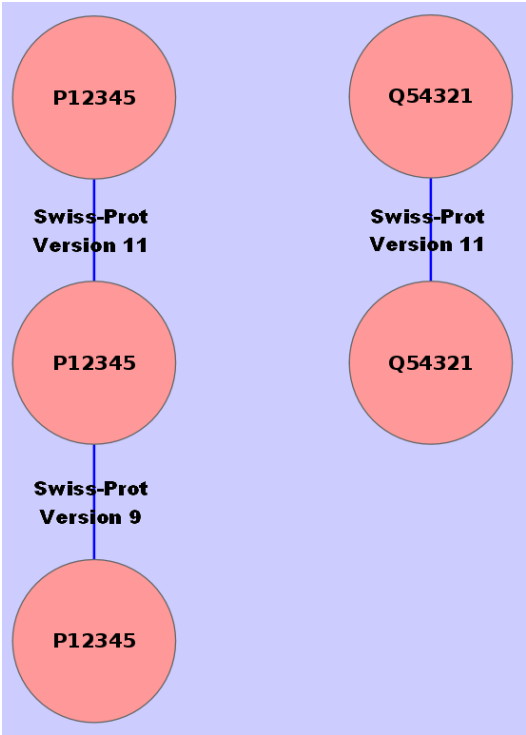


Figure 5.7: An example of how sentence propagation could be visualised in Cytoscape. The graph is based on a single sentence, with nodes duplicated to represent different database versions. The edge labels and specific ordering of nodes allows the propagation and provenance of a sentence to be visualised. For example, the sentence occurs in entry P12345 for Swiss-Prot Versions 9 and 11, whilst in Swiss-Prot Version 11 the sentence also appears in entry Q54321.

we derive RQ6).

Graph visualisation is beneficial in many applications, and is quite possibly of benefit in analysing annotation reuse. For example, Figure 5.6 identified that the sentence "the active-site selenocysteine is encoded by the opal codon, uga." is shared between two entries in separate clusters. Given this pattern occurs infrequently, it may hold some analytical value. However, the amount of manual intervention required means that this approach cannot be used to provide a suitable visualisation for analysing sentence propagation and provenance.

5.1.4 History flow

History Flow was a tool developed at IBM to visualise the relationships between multiple versions of a wiki [305]. Specifically, the tool generates an interactive visualisation showing changes to a text file over multiple revisions. A basic visualisation can be produced from a series of text files containing the different versions of the text. However, a more detailed visualisation can be produced if information about each revision is available, such as the date and time of each revision.

An example of a visualisation produced by the History Flow tool is shown in Figure 5.8. This Figure represents the various changes made to a small file containing a list of fruits. In total, the file undergoes five revisions (represented as columns) and is edited by two authors (identifiable by colour). The contents of the file after each revision is shown in Table 5.2.

Initial version	Revision 2	Revision 3	Revision 4	Revision 5	Revision 6
Apple	Apple	Apple	Apple	Apple	Apple
Kumquat	Banana	Banana	Banana	Banana	Banana
Watermelon	Kumquat	Kumquat	Kumquat	Kumquat	Kumquat
	Melon	Melon	Melon	Melon	Melon
	Watermelon	Pear	Pear	Pear	Pear
		Watermelon	Ugli	Watermelon	Ugli
			Watermelon		Watermelon

Table 5.2: The contents of the file visualised in Figure 5.8 after each revision.

Within this visualisation a coloured line indicates a line of text, with changes over time shown from left to right. For example, the top line represents "apple", which

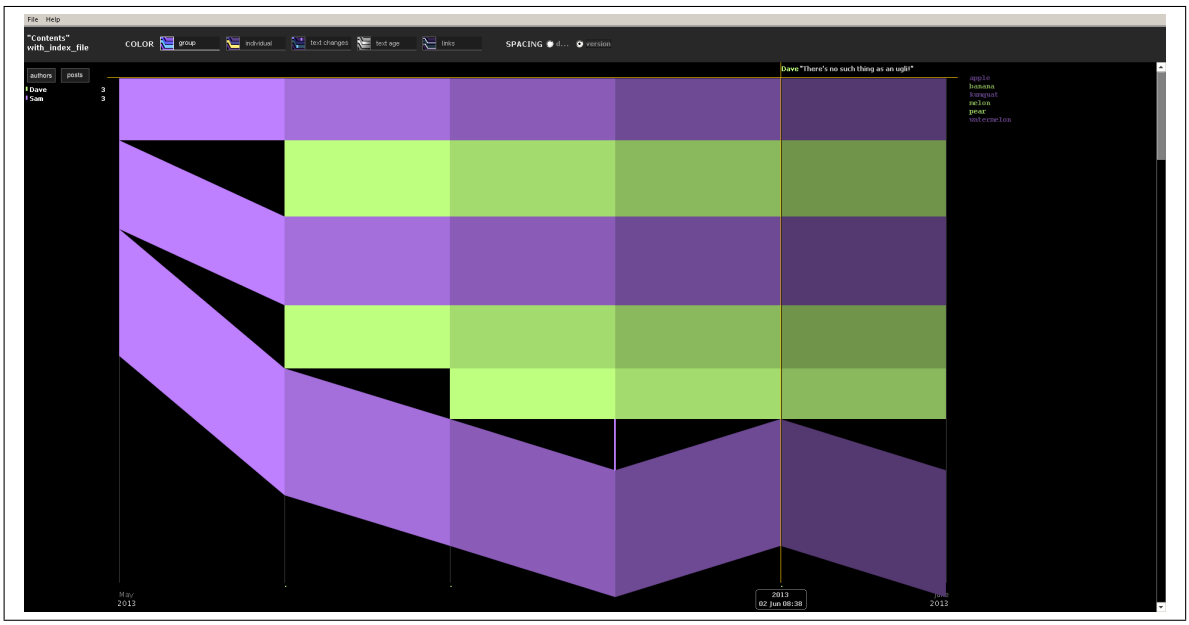


Figure 5.8: The visualisation produced by the History Flow tool to represent the revisions made to a file containing a list of fruit. The contents of the file after each revision is shown in Table 5.2.

was added by the author Sam, and remains in the file for all six revisions. This line remains at the top of the visualisation, as "apple" is constantly at the top of the file. However, the bottom line, which represents "watermelon", moves downwards to reflect new lines being added to the file. This line fluctuates between revisions three and six as the fruit "ugli" is added, removed and then re-added to the file.

Changes in the visualisation can be explored by clicking the intersection of a column and a row. When selected, the contents of the file at that point and, if available, an explanation for the revision with its time-stamp are shown. For example, in Figure 5.8 the top row and fifth column is selected (identifiable by crosshairs). At this point, the whole contents of the file is shown to the right of the visualisation, while an explanation for the revision is shown at the top of the column and the revision time-stamp shown at the bottom. This allows the reason for "ugli" being added and removed to be identified (the authors, Dave and Sam, did not agree that "ugli" was a fruit).

Applying the tool to UniProtKB can be done for either the accessions a sentence occurs in, or for the textual annotation within an entry. For example, Figure 5.9a shows the entries over time that the sentence "the active-site selenocysteine is encoded by the opal codon, uga." occurs in. However, this visualisation has a number of significant gaps

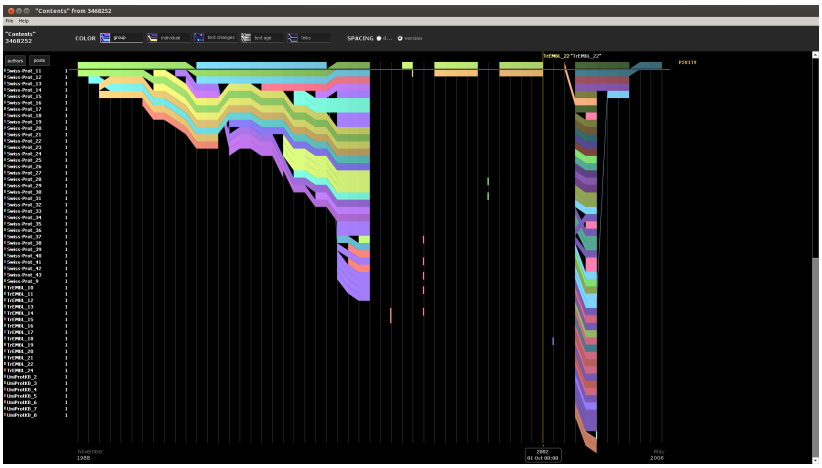
caused by the unsynchronised release dates of TrEMBL and Swiss-Prot. Removing the TrEMBL entries from this visualisation to just show Swiss-Prot entries, as shown in Figure 5.9b, results in a much more regular visualisation being obtained. From this figure it can be seen that the sentence first occurs in Swiss-Prot Version 9, with it being added to more Swiss-Prot entries overtime and then being removed from the majority of entries after Swiss-Prot Version 44.

Conversely, Figure 5.9c shows the change of textual annotation within UniProtKB entry P04395[1]. This view highlights that there has been a number of additions and deletions to the annotation over time. However, this visualisation becomes overly complicated and, therefore, difficult to interpret due to the reordering of the underlying text. Within textual annotation the order of sentences is relatively unimportant; in UniProtKB, for example, sentences can be added at any point, resulting in reordering with relatively little change in semantics. However, History Flow visualises these changes as they would be considered important in its original use case (a wiki). While this complexity can be reduced by the careful ordering of text within the file, they cannot always be eradicated. This is unfortunate as the ordering of lines for sentence reuse analysis is unnecessary.

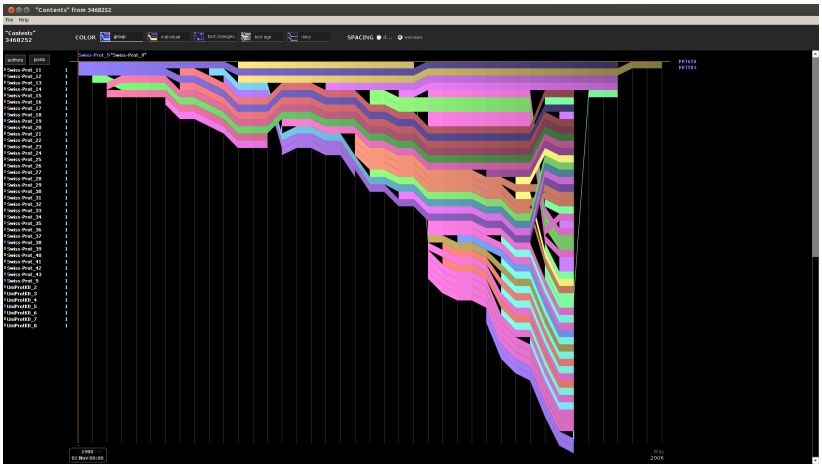
Although the visualisation can become complicated, Figure 5.9b shows that both the provenance and propagation of a sentence can be identified using the History Flow tool. The additional features provided by the tool can also aid sentence analysis. For example, the usage of colour allows entries adding the sentence within a particular database version to be easily identified and tracked over time. Unfortunately, however, the tool has a number of limitations. Crucially, the tool cannot handle the unsynchronised nature of early Swiss-Prot and TrEMBL releases, as illustrated in Figure 5.9a, meaning it cannot be used to visualise sentences appearing in both Swiss-Prot and TrEMBL. Being able to handle databases with differing release cycles is a necessary requirement of the visualisation (thus we derive RQ4).

An additional drawback of the tool is that it is a stand-alone program, requiring users to install the tool and then import data in the correct form for analysis. Although other approaches, such as Cytoscape, have a similar requirement, a visualisation that

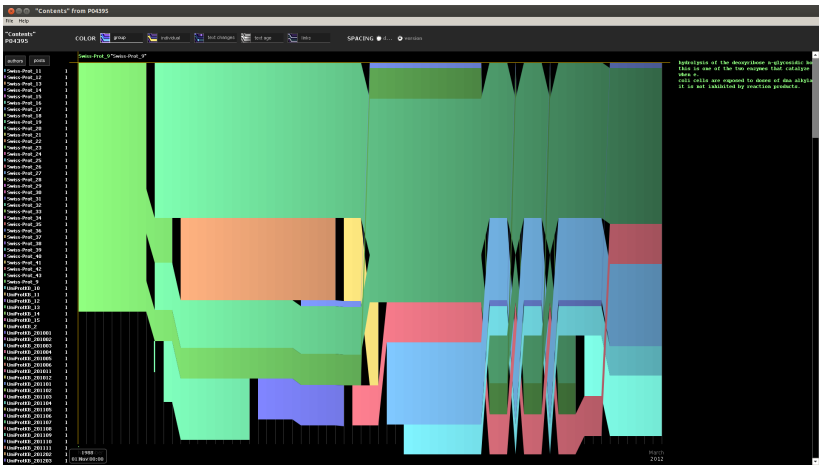
[1] `http://www.uniprot.org/uniprot/P04395`



(a) Propagation of the sentence through UniProtKB



(b) Sentence propagation in just Swiss-Prot



(c) Visualising sentences within an accession

Figure 5.9: Visualisation of sentence reuse using the History Flow tool. Figures (a) and (b) are based on the sentence "the active-site selenocysteine is encoded by the opal codon, uga.", whilst (c) is based on sentences from UniProtKB entry P04395.

could be generated from, or integrated into, a database entry view is more appealing. For example, the benefit of being able to generate graph visualisations online resulted in Cytoscape Web being developed, allowing networks to be embedded within websites [306]. The ability to view and produce graphs on the Web identifies a further requirement that is derived for the visualisation (RQ5).

Although the History Flow tool has drawbacks for analysing sentence reuse, the actual visualisation appears suitable. Unfortunately, the tool is no longer maintained or supported[2], meaning these drawbacks cannot be addressed.

5.1.5 Summary

To analyse sentence reuse, a visualisation allowing provenance and propagation to be inferred is required. However, as none of the existing approaches analysed provided a suitable solution, a bespoke visualisation is required, which we name VIPeR.

Although the explored visualisations were not deemed entirely suitable, there is much to learn from these approaches. Specifically we identified various features and properties that our visualisation should provide:

Provenance & Propagation

We identified that the provenance and propagation of a sentence could be depicted by connecting a series of Sankey diagrams together. However, this diagram would likely be unintuitive as it uses Sankey diagrams in an unconventional way. Conversely, the visualisation produced by the data flow diagram (Figure 5.3) was more conventional, and thus reasonably intuitive.

Therefore it is important that VIPeR depicts sentence reuse in a manner that intuitively shows all of the entries a particular sentence occurs in, and for which database versions. This should allow the first occurrence of an entry to be identified, whilst also enabling the subsequent flow of a sentence to be tracked through a database.

[2]The last release of the tool was in 2004, with the tools Web page redirecting to the IBM visualisation research groups homepage.

Automated Generation

The data flow diagram shown in Figure 5.3 and the Cytoscape graph shown in Figure 5.7 provide suitable visualisations that allow the provenance and propagation to be inferred. However, their production requires a significant amount of user skill and input. Depending upon user input increases the likelihood of incorrect visualisations being produced, whilst also meaning users are less likely to use and benefit from the visualisation.

Given this, we believe it is important that a user can generate a visualisation for a sentence with only minimal intervention. Further, generating a visualisation should only require minimal information, consisting of a list of entries and database versions that a sentence occurs, like shown in Table 5.1. The produced visualisation should also be reproducible, with a dataset always producing the same visualisation.

Web-based

The Cytoscape and History Flow tools both require a program to be downloaded and installed, with the user having to learn how to use each tools unique interface. Additionally, the History Flow tool is only available for the Windows operating system.

Therefore, we propose to develop VIPeR for the Web, as many users access biological databases and their contents through a Web browser. This provides a common interface for users and should not require specialist software to be downloaded. In practice, this is not a significant limitation, as web-based visualisation frameworks are now common with rich functionality.

Interactive Features

The analysis of the Cytoscape and History Flow tools identified the benefits of interactive features. Therefore, VIPeR should be augmented with interactive features to aid sentence analysis. Specifically, we should provide a zoom functionality to help alleviate difficulties when analysing dense graphs and provide hyperlinks to database entries to enable users to view data associated with a sentence.

Genericity

Although UniProtKB entries and sentences have been utilised in the previous analysis, the developed visualisation should allow sentence reuse in any textual resource to be visualised. VIPeR should also have the ability to show sentence reuse between multiple databases within a single visualisation. The importance of this was highlighted by the History Flow tool, which was unable to handle unsynchronised releases of Swiss-Prot and TrEMBL.

These identified features and properties will be used as a basis for the development of VIPeR.

5.2 Developing a Visualisation for Sentence Reuse

In the previous section a variety of different visualisations were explored. Although no single approach was deemed suitable, the analysis identified a number of beneficial visualisations and features that allowed sentence reuse to be analysed. In this section, VIPeR is developed to visualise sentence reuse and is based upon the beneficial approaches and features previously identified. Specifically, VIPeR should satisfy the following requirements:

RQ1 The visualisation should clearly depict the provenance and propagation of a sentence over time.

RQ2 To visualise a sentence only the list of entries and database versions it occurs in should be required.

RQ3 The visualisation should be generic, with adequate support and tooling, to allow any database with textual annotation to be analysed.

RQ4 The visualisation should be able to handle multiple databases with unsynchronised releases in a single graph.

RQ5 The visualisation should be web-based.

RQ6 The visualisation should be interactive, allowing a user to zoom into dense graphs and link through to relevant information.

RQ7 The visualisation should not require any user input other than the sentence that they wish to analyse.

5.2.1 Visualisation prototype

Prior to the actual implementation of VIPeR, it is first necessary to consider how the visualisation will be represented. One possible visualisation approach is illustrated in Figure 5.10, which shows a manually produced prototype for a hypothetical sentence.

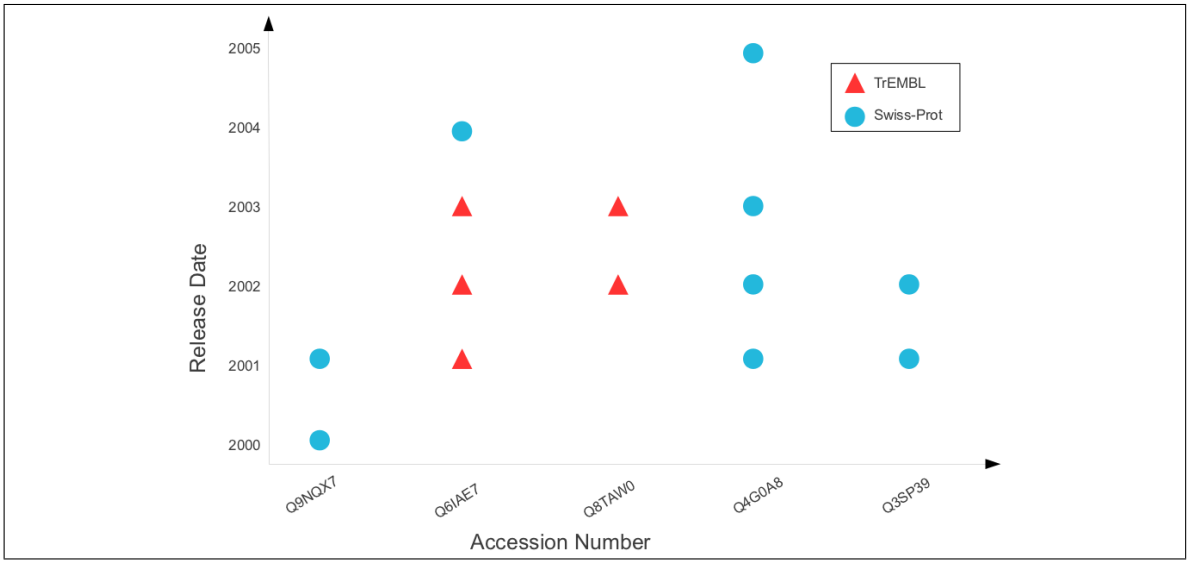


Figure 5.10: A manually produced prototype showing how the propagation of a single sentence can be visualised. Accession numbers are shown on the X-axis, with database release dates shown on the Y-axis. A point on the graph represents that the sentence occurs in an entry within a given database version. For example, the bottom left point shows that the sentence occurs in accession entry Q9NQX7 for Swiss-Prot in 2000 – this sentence remains in Q9NQX7 for one more version; it is removed in the following version (in 2002).

This prototype is based upon the benefits identified from previously analysed visualisations, increasing the likelihood that the visualisation will be suitable and intuitive to users.

The prototype shares similarities with a scatter plot, with data points representing a database entry (X-axis) and version (Y-axis) that the given sentence occurs in. This visualisation allows the first occurrence of the sentence to be identified (entry Q9NQX7 in 2000), with the propagation of the sentence also visible (the sentence occurs in a further four entries). The absence of a data point indicates that the sentence is not in that particular database entry and version, making it clear when a sentence is added and removed from an entry. For example, the sentence was initially added to entry Q4G0A8 in 2001 before being removed in 2004 and then re-added in 2005.

Although the prototype appears suitable for visualising sentence reuse, it has been manually produced based on a hypothetical dataset. To test if this visualisation can be applied to sentence reuse in UniProtKB, a real dataset is required. Table 5.3 shows a section of a dataset which represents how the sentence "the active-site selenocysteine

is encoded by the opal codon, uga." is reused in UniProtKB[3]. This dataset lists each of the database entries and corresponding versions that the sentence occurs in.

Entry	Database	Database Version
P07658	Swiss-Prot	9
P07203	Swiss-Prot	9
P07658	Swiss-Prot	11
P07203	Swiss-Prot	11
P04041	Swiss-Prot	11
...	...	...
P24183	UniProtKB/Swiss-Prot	7
P78261	UniProtKB/Swiss-Prot	7
P24183	UniProtKB/Swiss-Prot	8
P78261	UniProtKB/Swiss-Prot	8

Table 5.3: A section of the dataset, used to produce Figure 5.11, which describes how the sentence "the active-site selenocysteine is encoded by the opal codon, uga." is reused in UniProtKB. Each row of the dataset contains a database entry and corresponding version that the sentence occurs in.

To generate a visualisation of this data we used R, as it allows us to produce a prototype visualisation relatively quickly. This visualisation, as shown in Figure 5.11, plots the release date, rather than release version given in the dataset, on the Y-axis. This was achieved by mapping each database version to its release date[4], allowing Swiss-Prot and TrEMBL entries to be shown concurrently in the same visualisation. This satisfies the requirement of allowing multiple databases with unsynchronised releases to be visualised in a single figure (RQ4).

Crucially, Figure 5.11 also shows that a sentence's origin and subsequent propagation can be identified using this visualisation approach, which was generated based only on the data shown in Table 5.3. This satisfies a further two of our requirements (RQ1 and RQ2).

However, this prototype does not provide an interactive web-based visualisation, meaning two of our requirements are not satisfied (RQ5 and RQ6). The importance of this interactivity is emphasised by this prototype as the amount of data shown in the visualisation means it is not immediately obvious which database version or entry a particular point refers to. This difficulty is compounded by the release date, rather than the database version, being plotted on the Y-axis. By being able to, for example,

[3]The method used for extracting sentences from UniProtKB is described in Section 6.1.
[4]This mapping is shown in Table 2.2

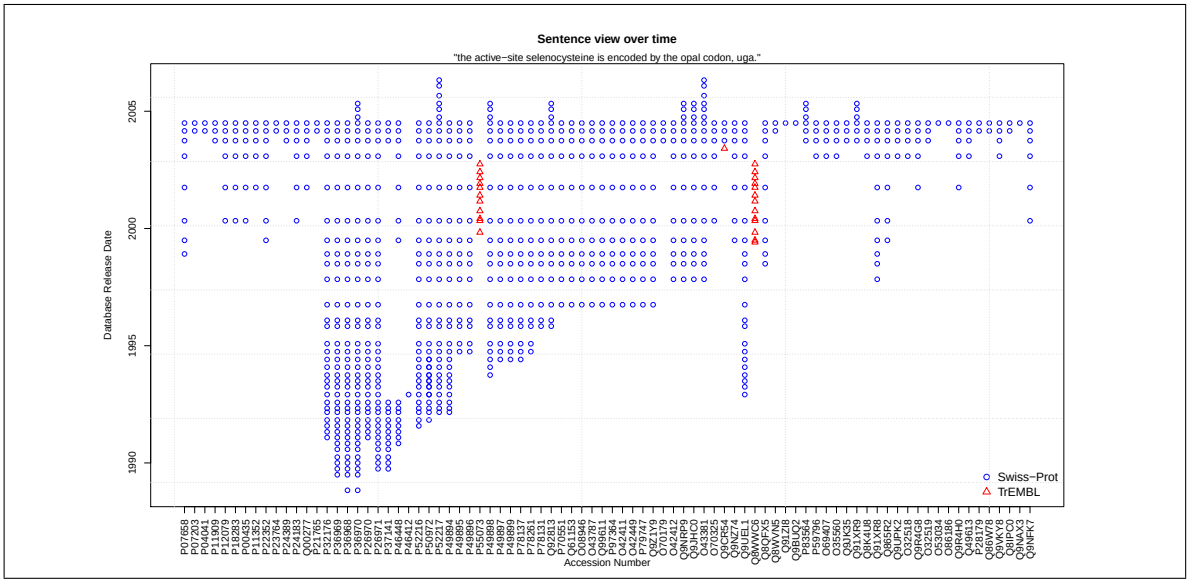


Figure 5.11: Visualising the propagation of the sentence "the active-site selenocysteine is encoded by the opal codon, uga." through UniProtKB. The graph is based upon the prototype illustrated in Figure 5.10 and was produced in R.

obtain the database entry and version by clicking on a data point, then our visualisation would be easier to interpret. Therefore, we need to identify a suitable tool or framework that allows us to generated Web based visualisations.

5.2.2 Developing a web-based visualisation with Highcharts

Although the visualisation shown in Figure 5.11 allows sentence reuse to be analysed, the implementation in R does not provide any interactivity. Therefore, an alternative approach for producing VIPeR, which also provides the features previously discussed in Section 5.2, is required.

Searching for web-based visualisation approaches returns a plethora of available tools and libraries, including Fusion Charts [307], pChart [308], PlotKit [309] and Google charts [310]. These tools provide varying degrees of functionality and supporting features. For example, some approaches only produce static graphs, offering no interactivity, whilst others have compatibility issues, with visualisations only being displayed correctly in a particular browser. Other potential issues or restrictions with using certain approaches include: licencing issues; lack of updates; poor documentation and support; or key features being deprecated with no alternatives being created.

Given this criteria, Highcharts appears to be appropriate software [311]. Highcharts is

a JavaScript library that provides support for producing a variety of web-based charts and graphs. Features of Highcharts include:

- A variety of interactive features, including zooming and tooltips, and an exporting function (i.e. the ability to save a visualisation as an image).
- Graphs can be customised, with options to change a plots colour, text and axes being provided. Additionally, the source code is open meaning extensions or more fundamental changes can be made.
- Various methods for obtaining help and support are available, including a dedicated support forum[5], published books [312, 313] and a Highcharts tag on Stack Overflow[6].
- Graphs are produced and displayed as vector graphics, meaning they are compatible with most modern Web and mobile browsers.
- The library can be hosted and managed locally. This avoids any dependence on external servers or a changing code base.
- Highcharts is released under the Creative Commons Attribution-NonCommercial 3.0 License. This means it can be freely redistributed and modified, providing that attribution is given and it is not used commercially.
- At the time of writing, Highcharts remains under active development with regular releases and updates.

Installing and using Highcharts is relatively straightforward. Highcharts graphs are encoded within a HyperText Markup Language (HTML) page and can be broadly split into three sections: the Highcharts dependencies; the visualisation data and properties; and the placement of the visualisation within a Web page. For example, Figure 5.13 shows the code required to produce Figure 5.12, which is a reproduction of the prototype visualisation (Figure 5.11) in Highcharts. This shows that the visualisation can be produced in a web-based format and fulfils requirement RQ5.

[5] `http://forum.highcharts.com/`
[6] `http://stackoverflow.com/tags/highcharts`

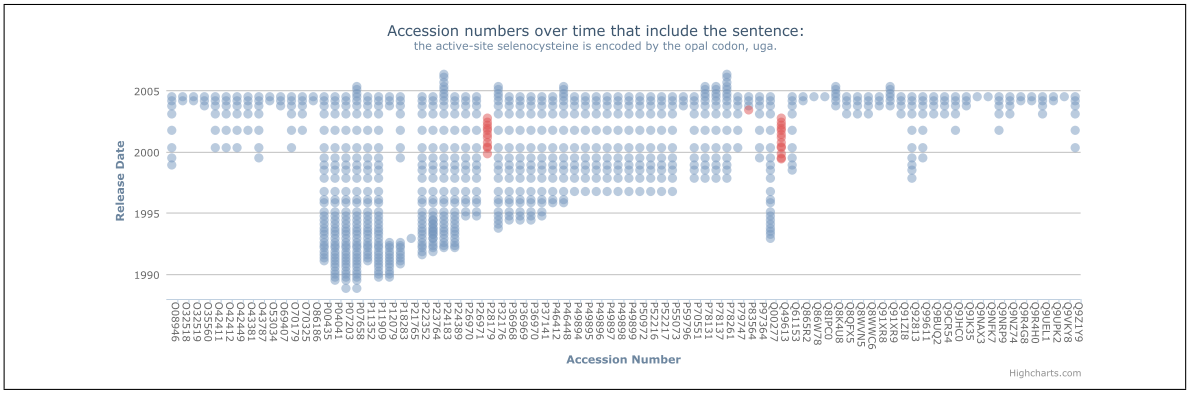


Figure 5.12: Visualising the propagation of the sentence "the active-site selenocysteine is encoded by the opal codon, uga." through the database.

At a minimum, Highcharts requires two dependencies – `highcharts.js` and `jquery.min.js` – which provide the necessary functions for producing a visualisation. These dependencies are JavaScript libraries and are included in the header of the HTML document. Additional dependencies, such as an export module, need to be included if Highcharts is to be installed locally in its entirety.

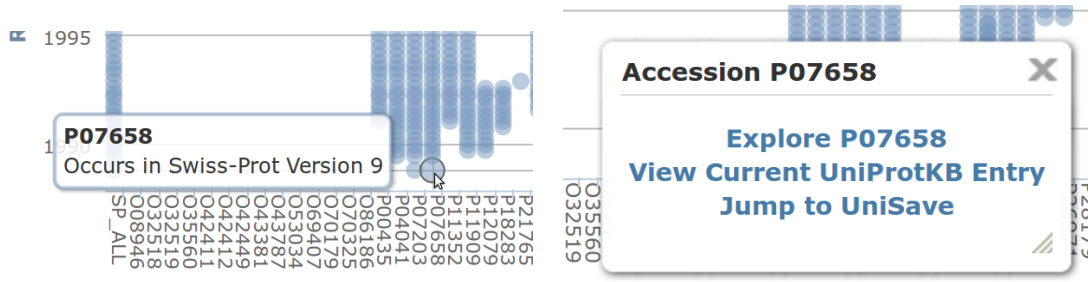
Similarly, the properties of the visualisation, and the actual data to be displayed, are also encoded in JavaScript in the head of the HTML document. The visualisation properties are stored in JavaScript Object Notation (JSON) format, with the actual data encoded in JavaScript arrays. This format allows properties to be easily customised to control how visualisations are produced and displayed. For example, a subtitle can be included in a visualisation by adding a `subtitle` property. Although the data points are represented as JavaScript arrays, Highcharts provides support for parsing data in XML, Comma Separated Values (CSV) and JSON files into the necessary format, as well as being able to deal with data stored in an external database. The actual placement and size of the visualisation is controlled by the `div` tag, which is added to the body of the HTML document.

The code for producing Figure 5.12 contains the `plotOptions` and `tooltip` properties which allow information to be displayed when either clicking or hovering over a data point, as illustrated in Figure 5.14. The information that is displayed is fully customisable and is used to show which accession and database version any given data points corresponds to.

Another interactive feature provided by Highcharts is the ability to zoom into a graph.

```
<html><head>
<script type="text/javascript" src="jquery.min.js"></script>
<script type="text/javascript" src="highcharts.js"></script>
<script type="text/javascript">
$(document).ready(function() {
    chart = new Highcharts.Chart({
      chart: {
        renderTo: 'container',
        defaultSeriesType: 'scatter',
        zoomType: 'xy'
      },
      title: {
        text: 'Accession numbers over time that include the sentence:'
      },
      subtitle: {
        text: 'the active-site selenocysteine is encoded by the opal codon, uga.'
      },
      xAxis: {
        categories: ['O08946', 'O32518', 'O32519', 'O35560', 'O42411', ... ],
        labels: {rotation: 90, align: 'left'},
        title: {text: 'Accession Number'}
      },
      yAxis: {
        type: 'datetime',
        dateTimeLabelFormats: {year: '%Y'},
        title: { enable: true, text: 'Release Date' }
      },
      tooltip: {
        formatter: function() {
            return this.x + 'Occurs in ' + getVersion(this.series.name, this.y);
        },
        crosshairs: [true, true]
      },
      plotOptions: {
        series: {
          cursor: 'pointer',
          point: {
            events: {
              click: function() { return getDataPointInfo(this.x, this.y); }
              }
            }
          },
          marker: {radius: 5, states: { hover: { enabled: true } } }
        }
      },
      series: [{name: 'Swiss-Prot', color: 'rgba(119, 152, 191, .5)', data: [null,
              ['O08946', Date.UTC(1998, (12 - 1), 1)]], ... }]
      );
    });
</script></head><body>
<div id="container" style="width: 100%; height: 100%"></div>
</body></html>
```

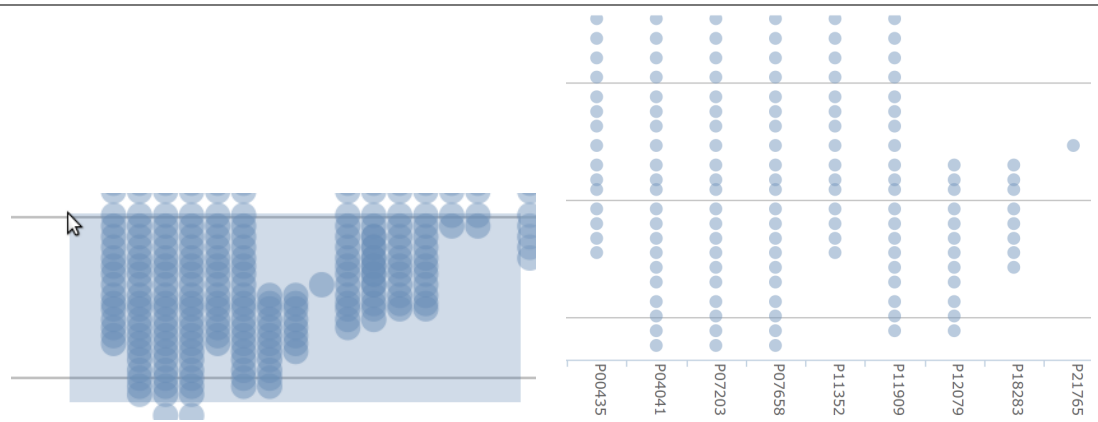
Figure 5.13: JavaScript and HTML code used to produce the Highcharts graph shown in Figure 5.12. Sections of the code are omitted due to space restrictions.



(a) Hovering (b) Clicking

Figure 5.14: Illustrating interactive features of Highcharts: (a) hovering over a data point (as defined by `tooltip`) and (b) clicking on a data point (as defined by `plotOptions`).

This zoom functionality, illustrated in Figure 5.15, allows sections of a graph to be analysed in greater detail and is especially beneficial when analysing dense graphs. When zoomed into a graph, a "reset zoom" button is displayed that allows the original visualisation to be shown. Additionally, all normal interactivity and functionality, such as tooltips, are still available when viewing a zoomed graph.



(a) Selecting zoom area (b) Zoomed in

Figure 5.15: Illustrating the zoom function in Highcharts: (a) A zoom area is selected by left clicking and dragging. (b) The visualisation is then reproduced to show just the selected area.

These interactive features help alleviate the issues encountered with static based graphs and allow us to fulfil requirement RQ6. However, the disjoint nature of early Swiss-Prot and TrEMBL releases can make graphs initially appear misleading. Specifically, several TrEMBL releases can be released between two Swiss-Prot releases which, when visualised, can make it appear that the sentence is constantly being removed and then re-added; i.e. graphs can exhibit *striping*.

One approach to overcome the issue of striping is binning. However, this would lose a major level of granularity as bins covering a six month period would be required to cover all Swiss-Prot releases. An alternative approach is to show points for all possible Swiss-Prot releases down the left side and all possible TrEMBL releases down the right. This approach is shown in Figure 5.16, which helps alleviate the issue of striping without a loss of data from the visualisation.

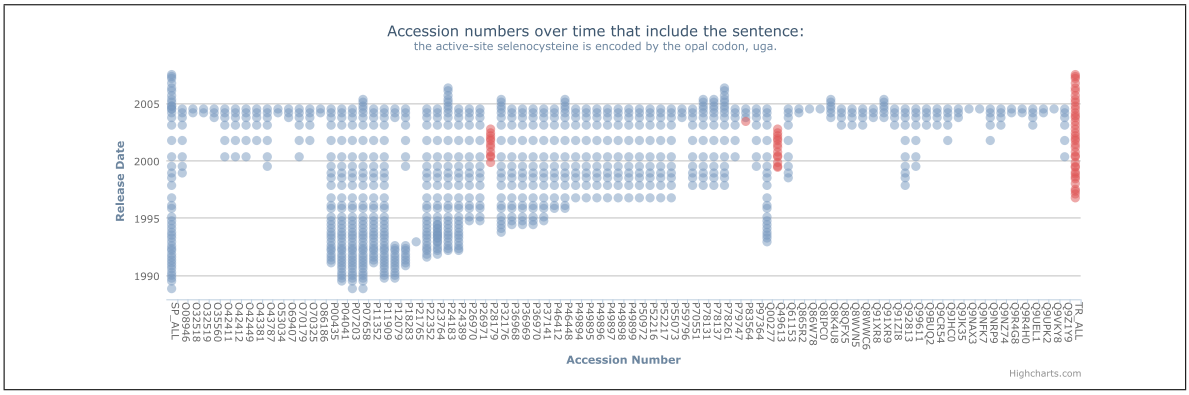


Figure 5.16: Same as Figure 5.12, except that all possible versions of Swiss-Prot and TrEMBL are shown at either end of the graph. This approach is used to alleviate the issue of striping.

Data points in Figure 5.16 represent a sentence occurring in a particular accession number for a given database version. However, accessions in UniProtKB can become merged, resulting in a single primary accession and one or more secondary accessions. Within Figure 5.16, no distinction between primary and secondary accessions is made meaning graphs can become redundant. This redundancy is highlighted in Figure 5.17 by displaying secondary accessions with more transparency than primary accessions.

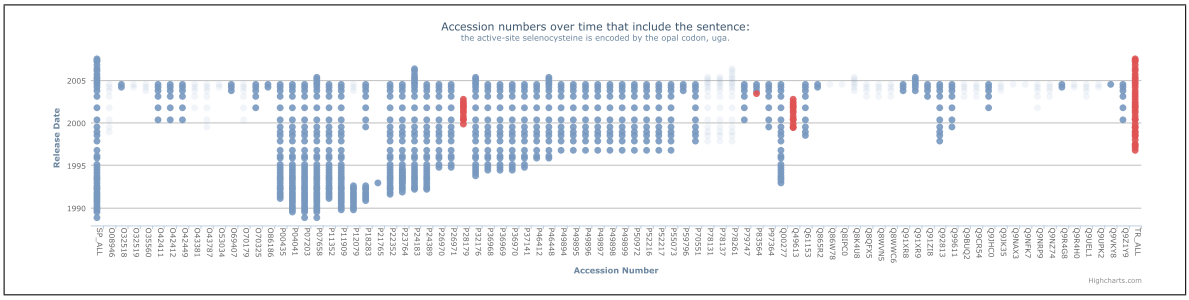


Figure 5.17: The same visualisation as Figure 5.16, although a distinction is made between primary and secondary accessions. Secondary accessions are shown with more transparency than primary accessions.

Although secondary accessions could simply be removed from each visualisation, it would become misleading. For example, in Figure 5.17 the sentence is removed from

the primary accession P12079 before being re-added to the entry when it becomes a secondary entry (i.e. is merged with another UniProtKB entry). Similarly a merged accession which retains its original annotation would be removed from the visualisation following the merge making it appear that the sentence was removed from the accession rather than just becoming merged. Therefore, visualisations will be shown with both primary and secondary accessions.

5.3 Discussion

The continuing increase in the amount of raw data requiring annotation has resulted in biological databases increasing the amount of annotations being reused. Although annotation reuse is not a problem *per se*, as a single annotation can be used to accurately describe many sequences sharing a common feature or property, its provenance is not always adequately documented. This could be problematic; if an annotation that has been propagated to multiple entries is later found to be inaccurate or erroneous, then are entries containing this annotation also affected? If so, is the annotation updated or have the entries become unsynchronised?

As textual annotations are mostly composed of unstructured free text, identifying an annotations provenance is not straightforward. We approach this by using sentences as the smallest unit of traceable annotation. By recording all entries a sentence occurs in, along with each database version, then we hypothesise that the propagation history of each sentence can be inferred.

From a dataset of sentence reuse the provenance can be suggested by extracting the entry (or entries) that occur in the oldest database version. However, by applying a suitable visualisation to the dataset it is possible that additional information can be identified. For example, an analysis of Wikipedia using the History Flow tool identified various patterns of conflict and cooperation between users [305]. Using these patterns it is possible that events such as "edit wars" and vandalism could be automatically detected. Without the aid of the History Flow tool visualisation it is unlikely that these patterns would have been identified.

The identification of the revision patterns in Wikipedia was made possible due to a visualisation that allowed page revisions over time to be shown and explored. If similar patterns also exist in sentence reuse, then a visualisation clearly showing the propagation of a sentence over time is required. A number of existing visualisations, including the History Flow tool, were analysed for this purpose. However no single existing approach was deemed entirely suitable.

The most common issue with the existing visualisations was regarding their production, with various levels of manual intervention being required. Another key issue

encountered by certain visualisations was the inability to correctly handle the disjoint nature of TrEMBL and Swiss-Prot releases. Although there were limitations, each of the surveyed visualisations provided a view of the annotation space that allowed sentence propagation to be either fully, or partially, identified.

The analysis of these existing visualisations identified a number of advantageous features and properties. For example, the usage of colour and the alignment of data points can aid with the interpretation of a visualisation. From this analysis seven requirements were derived, covering the main features we believe a visualisation must incorporate to adequately depict sentence reuse.

These requirements formed the basis of our developed visualisation (VIPeR), which shares a number of similarities with a scatter plot. These similarities are not unintentional; the widespread usage of scatter plots suggests that they are analytically beneficial and are well understood by users. By exploiting this familiarity we hope to increase the intuitiveness of VIPeR, and thus satisfy RQ1. VIPeR can also be produced with minimal data, meaning that RQ2 is also satisfied.

In addition to the core features of VIPeR, RQ6 deemed that interactive features, such as zooming and tooltips, were necessary. To provide these features, Highcharts was chosen for the implementation of VIPeR, which allowed RQ5, RQ6 and RQ7 to be satisfied. Highcharts provides a suitable platform for visualising sentence propagation, offering benefits such as the ability to easily alter and extend graphs. For example, a graph to represent the occurrences of a sentence over time, as shown in Figure 5.18, can be easily produced.

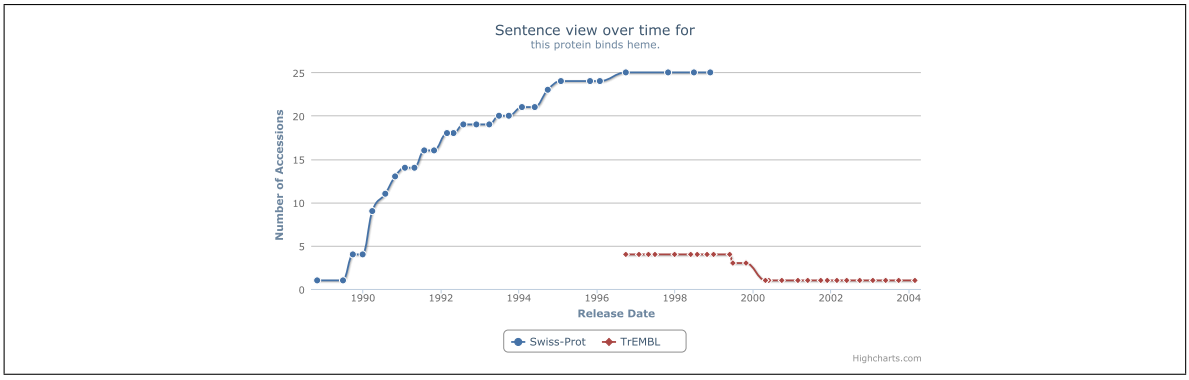


Figure 5.18: Illustrating the flexibility of Highcharts by visualising the frequency of the sentence “this protein binds heme.” in UniProtKB over time.

Throughout the development of VIPeR we discussed its suitability and intuitiveness, and identified a number of requirements that have been satisfied. However, to evaluate if RQ3 and RQ4 have been met, we need to apply VIPeR to a variety of sentence datasets and determine if the provenance and propagation can be inferred. Further, does the visualisation allow sentence propagation to be explored, offering additional value that could not otherwise be achieved without the aid of visualisation? This evaluation is performed in the following chapter, which applies VIPeR to textual annotation in UniProtKB.

6

Inferring the Provenance and Subsequent Propagation of Annotations in the UniProt Knowledgebase

Contents

Introduction

The evidence attached to textual annotations varies between biological databases. For example, whilst most databases make a distinction between manual and automated annotations, it is rare for the source, or *provenance*, of an annotation to be made apparent. Without the ability to determine the provenance of an annotation, users are restricted in how they can analyse and assess annotations; users have to accept an annotation "as is".

Determining the provenance of a given textual annotation is not straightforward. Annotations are often subject to reuse, being copied and pasted between entries and external databases as a matter of protocol. This is further complicated as annotations are mostly composed of unstructured free text.

Whilst annotation reuse can result in an entire textual annotation being propagated between entries, it is more common for a subsection of the annotation to be copied. These subsections are distinguished by the syntactic rules of natural language; annotations can be split into individual sentences. The propagation of these sentences is often verbatim, meaning that sentences can be tracked and used as *annotation markers*.

In the previous chapter VIPeR, a visualisation tool, was developed. By exploiting sentence reuse, VIPeR aims to allow the provenance, and subsequent propagation, of an annotation to be identified. This requires archives for the annotation as the occurrences of a sentence over time is required to infer its provenance. UniProtKB is a well established database with over twenty years of historical data, making it an ideal resource to evaluate the effectiveness of VIPeR.

VIPeR shows the occurrences of an individual sentence throughout a database and is reliant upon the correct extraction of sentences. To extract sentences from UniProtKB, BANE (the parsing framework developed in Chapter 4) is extended to allow entire sentences to be extracted and stored (Section 6.1). Whilst the extraction of sentences from all versions of UniProtKB provides an abundance of data, it does not imply that sentences will be suitable as annotation markers. Therefore, levels of sentence reuse in UniProtKB are analysed. This analysis identifies a significant quantity of sentence reuse, providing confidence that sentences can be utilised as annotation markers

(Section 6.2).

Applying VIPeR to individual sentences allows the provenance of a sentence to be identified. Visualisations also allow the propagation of a sentence throughout the database to be inferred (Section 6.3). By analysing the propagation of these sentences, a number of sentences exhibiting patterns that were irregular or unexpected were identified. It was hypothesised that these patterns could provide indicators for low quality or erroneous annotations.

To evaluate this hypothesis, sentences adhering to each pattern were identified and extracted. In total, over $85,000$ sentences followed at least one of the identified patterns. The analysis of these sentences suggests that propagation patterns can provide indicators for low quality annotation whilst one pattern, the *missing origin*, could also indicate erroneous annotation (Section 6.4).

Over $8,000$ sentences were identified as following the missing origin pattern. A protocol was developed to allow these sentences to be analysed, with approximately 50% of the analysed sentences being classified as either erroneous or inconsistent (Section 6.5). Finally a discussion and summary of the visualisations application, identified patterns and results obtained is presented (Section 6.6).

6.1 Sentence Extraction

The process of extracting sentences from UniProtKB involves extending BANE[1] to handle sentence extraction. Compared to the extraction of words, the extraction of sentences from a text is not as straightforward and is especially problematic within the biomedical domain [136]. A naïve approach would extract a sequence of words, the first beginning with a capital and ending with a full stop. However, in practice there are numerous exceptions and deviations from this rule. For example, Ribonucleic acid is often abbreviated as RNA and could be preceded with "e.g." to illustrate an example; such a sentence would be incorrectly parsed as two sentences. Conversely, sentences may not be separated correctly if one was to begin with a lower case letter. For example, any sentence that begins with tRNA; an abbreviation seen more commonly than its expansion "Transfer RNA".

In addition to the underlying linguistic properties of a sentence, consideration also has to be given to the presentation of textual annotations in UniProtKB. For example, annotations have evolved over time; originally all annotations were in upper-case, whilst later versions have become more structured, with the introduction of topic blocks. There have also been technical changes to the annotation, such as the introduction of copyright as identified in Section 4.2. Examples of annotations from UniProtKB that highlight a number of linguistic and presentational cases that have to be accounted for are shown in Figure 6.1.

Given these difficulties, a suite of Java libraries, named LingPipe, was utilised [314]. LingPipe provides a variety of textual processing features, including various rule models for identifying sentence boundaries within a piece of text. These models include the MEDLINE model which includes rules designed specifically for handling biomedical text [315]. For example, this model will determine that "...correlation. p-53 was..." contains a sentence boundary as, although p is a lower-case character, it is preceded by a full stop and is directly followed by a hyphen [316]. LingPipe was incorporated into BANE to extract whole sentences as opposed to just individual words.

[1]BANE is the parsing framework developed to extract words from UniProtKB annotation, as previously described in Section 4.1

```
(1)
CC   -!- TISSUE SPECIFICITY: mRNA found twofold higher in leaves and stems
CC       than in roots.

(2)
CC   -!- SEQUENCE CAUTION:
CC       Sequence=AAA40109.1; Type=Erroneous initiation;

(3)
CC   -!- MISCELLANEOUS: Plants lacking XTH8 exhibit up to 50\% growth
CC       reduction when they reach maturity. Lower level of XTH8 transcript
CC       detected in Tanginbozu, a GA-deficient semidwarf mutant, and
CC       higher level detected in Slender rice 1 (slr1), a GA-insensitive
CC       mutant showing a constitutive GA-response phenotype. -!-
CC       SIMILARITY: Belongs to the glycosyl hydrolase 16 family. XTH group
CC       2 subfamily.

(4)
CC   -!- IN EUKARYOTES THERE ARE TWO ISOZYMES: A CYTOPLASMIC ONE AND A
CC       MITOCHONDRIAL ONE.
```

Figure 6.1: Four annotations highlighting linguistic and presentational features taken into consideration when extracting sentences. (1) A sentence beginning with a lower-case letter. (2) A topic block which is not terminated with a full stop. (3) A topic block indicator becoming merged with the text. (4) A topic block indicator without a corresponding title and the entire annotation being in upper-case.

For words, a simple data model was sufficient to allow a Zipfian analysis to be performed. For tracing and displaying provenance information, however, we require the ability to perform richer queries to enable, for example, the extraction of all entries containing a given sentence, over a range of database versions. We implemented this using MySQL [317], because of its ubiquity, and good interoperability with both Java (through JDBC) and Highcharts (though PHP).

The overall extraction process, as summarised in Figure 6.2, involves:

1. Downloading and extracting complete datasets of historical versions of UniProtKB, in flat file format, from the UniProt FTP server[2].
2. Extracting comment lines from these flat files using BANE. Comment lines are extracted based upon the information given in the UniProtKB user manual [223].
3. Removing topic headings, the "CC" identifier and copyright and licence statements. Over time, annotations in UniProtKB have become more structured with the addition of topic headings (e.g. "subcellular location" and "function") in the comments lines, which were removed to maintain sentence integrity.
4. Extracting a list of all the sentences from each entry's comment lines using LingPipe.
5. Storing extracted sentences in the MySQL database, stating the entry it appears in and for which database version.

The first three of these steps are identical to those used previously for extracting words, therefore many of the integrity checks can be reused. Previously, errors in parsing were identified through visual inspection of Zipfian graphs; namely the identification of copyright. Applying the power-law model to sentences, as previously shown in Figure 4.7, does not exhibit any significant kinks or irregularities. This suggests that no significant errors in parsing have occurred. However, more detailed tests are required to check that sentences added or removed between entry versions are correctly detected.

[2] `ftp.uniprot.org`

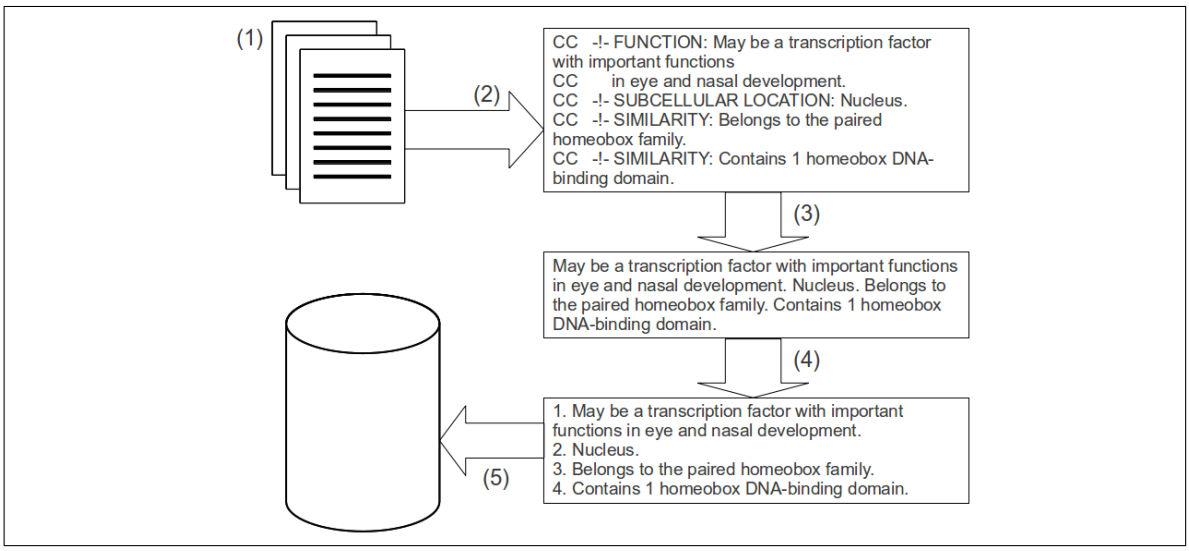


Figure 6.2: Outline view of the data extraction process. (1) Initially a complete dataset for a given database version in flat file format is downloaded. (2) The comment lines (lines beginning with 'CC', the comment indicator) are then extracted. (3) Comment blocks and properties (as defined in the UniProtKB manual [223]) and the 'CC' identifier are removed. (4) Sentences are then extracted, using LingPipe. (5) Finally, all of the identified sentences are added to the MySQL database.

To perform these tests, the UniSave tool [318], made available by UniProt, was utilised. UniSave shares similarities with the UNIX diff utility; it allows the differences between two entry versions to be compared. An example UniSave output is shown in Figure 6.3. UniSave only highlights changes; sections of the entry that remain unchanged are hidden, with their locations relative to the changed text indicated by ellipses.

Making use of UniSave, sentences were manually checked against a random selection of 50 entries and versions to ensure sentences were correctly parsed. For example, checking the sentences updated in Figure 6.3 includes a number of newly added sentences (lines 11-14 and 29-30) and sentences which have been replaced (lines 20-27). A convenient byproduct of using UniSave was that additional checks could also be performed. For example, Figure 6.3 shows the difference between Swiss-Prot Versions 2010_07 and 2010_09; there was no change to the entry between Swiss-Prot Versions 2010_07 and 2010_08. This is evident by the update to the entry version, as shown by the change in lines 1 and 2 (numbering for UniSave entry versions and UniProtKB database versions are independent). Using this information, a check can be easily performed to ensure that the sentences parsed for Version 2010_08 are identical to those parsed for Version 2010_07.

```
1   - DT   15-JUN-2010, entry version 77.
2   + DT   10-AUG-2010, entry version 78.
...
3   + RN   [10]
4   + RP   INTERACTION WITH TRIM11, UBIQUITINATION, AND DEVELOPMENTAL STAGE.
5   + RX   PubMed=18628401; DOI=10.1101/gad.471708;
6   + RA   Tuoc T.C., Stoykova A.;
7   + RT   "Trim11 modulates the function of neurogenic transcription factor Pax6
8   + RT   through ubiquitin-proteosome system.";
9   + RL   Genes Dev.  22:1972-1986(2008).
...
10  - CC   -!- SUBUNIT: Interacts with MAF and MAFB.
11  + CC   -!- SUBUNIT: Interacts with MAF and MAFB. Interacts with TRIM11; this
12  + CC       interaction leads to ubiquitination and proteasomal degradation,
13  + CC       as well as inhibition of transactivation, possibly in part by
14  + CC       preventing PAX6 binding to consensus DNA sequences.
...
15  - CC       and pancreas.  At day 9 of mouse embryonic development, expressed
16  - CC       in the telencephalon, diencephalon, neural tube, optic vesicle and
17  - CC       pancreas.  Throughout development, expression continues in the
18  - CC       dorsal and ventral pancreas.  In newborn animals, becomes
19  - CC       restricted to endocrine cells of the islets of Langerhans.
20  + CC       and pancreas.  At 9 dpc, expressed in the telencephalon,
21  + CC       diencephalon, neural tube, optic vesicle and pancreas.  Throughout
22  + CC       development, expression continues in the dorsal and ventral
23  + CC       pancreas.  Expressed during cortical neurogenesis from 11 to 18
24  + CC       dpc.  High levels in the early radial glial progenitors from 11 to
25  + CC       14 dpc and gradually decrease thereafter (at protein level).
26  + CC       During corticogenesis, the protein level declines faster than that
27  + CC       of the mRNA, due to proteasomal degradation.  In newborn animals,
28  + CC       becomes restricted to endocrine cells of the islets of Langerhans.
29  + CC   -!- PTM: Ubiquitinated by TRIM11, leading to ubiquitination and
30  + CC       proteasomal degradation.
...
31  + DR   ProteinModelPortal; P63015; -.
...
32  + DR   GO; GO:0010843; F:promoter binding; IDA:MGI.
...
33  - DR   GO; GO:0043565; F:sequence-specific DNA binding; IEA:InterPro.
...
34  - KW   Nucleus; Paired box; Transcription; Transcription regulation.
35  + KW   Nucleus; Paired box; Transcription; Transcription regulation;
36  + KW   Ubl conjugation.
```

Figure 6.3: An example of the UniSave view, shown for Swiss-Prot entry P63015, which highlights changes between Versions 2010_07 and 2010_09. UniSave illustrates additions by a green line starting with a "+" sign (e.g. line 2) with deletions indicated by a red line starting with a "-" sign (e.g. line 1).

Once all of the sentences are extracted from every entry within a given database version, then the set of sentences obtained will be referred to as the *total* number of sentences within a database version. From this set of sentences, some will occur multiple times (i.e. the set of total sentences is redundant). Taking each sentence from this redundant set only once (i.e. extracting the distinct sentences) results in a set of non-redundant *unique* sentences. Finally, within a set of unique sentences, some sentences will occur only a single time within a database version; that is they are *singleton* sentences. These definitions will be used throughout this thesis, and can be summarised as:

- **Total sentences** – A redundant set of all sentences in a database version.
- **Unique sentences** – A non-redundant set of all sentences in a database version.
- **Singleton sentences** – A set of sentences that occur only a single time within an entire database version.

Having extracted sentences from all UniProtKB entries, for all available versions, then questions about the overall statistical properties can be answered. Critically, for the developed visualisations to be effective, there needs to be substantial sentence reuse; that is, the number of singleton sentences as a percentage of unique sentences needs to be low. We are now in the position to answer such a question and analyse how much copying and pasting of textual annotation occurs in UniProtKB.

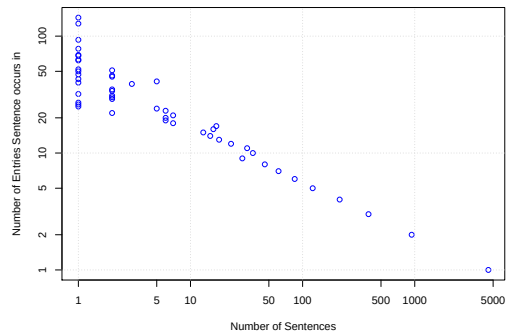
6.2 Sentences as Annotation Markers

The curation process implemented by UniProtKB (described in Section 2.4) means that sentences are subject to reuse. This reuse was confirmed by the application of the developed power-law model to sentences in UniProtKB and suggests that sentence reuse is increasing over time. However, how suitable are sentences as annotation markers? To assess this suitability, sentence reuse and the distribution of sentences is analysed.

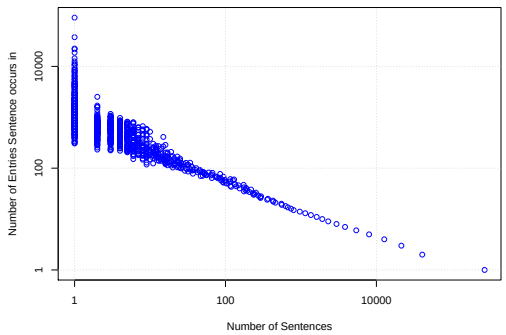
Whilst the application of the power-law model to sentences (discussed in Section 3.4) provides a unique view of sentence reuse within UniProtKB, a more detailed understanding of how sentences are reused and the relationship between the number of sentences and entries is required. We can explore this reuse in a number of different ways. For example, we analyse the total number of unique and singleton sentences in UniProtKB over time to identify the number of sentences which are not subjected to reuse and how the size of the corpus changes over time. We also analyse how reuse impacts the annotation within entries over time by exploring the number of sentences contained within an entry on average as well as the number of entries which actually contain no annotation.

We start this analysis by showing the distribution of sentences for four versions of UniProtKB in Figure 6.4. Within these figures, each point represents the number of sentences that occur in a particular number of entries. For example, in Figure 6.4d the bottom rightmost point indicates that there are approximately $1,000$ singleton sentences within UniProtKB/TrEMBL Version 2012_05, whilst the upper-left most points relates to the most commonly occurring sentence, which occurs in over seven million TrEMBL entries[3]. These visualisations show that both databases have increasing levels of reuse over time, with TrEMBL exhibiting much higher levels of reuse compared to Swiss-Prot. Additionally, the distribution of sentences is much more regular within Swiss-Prot than compared to TrEMBL and, comparatively, very few sentences in TrEMBL occur only a single time.

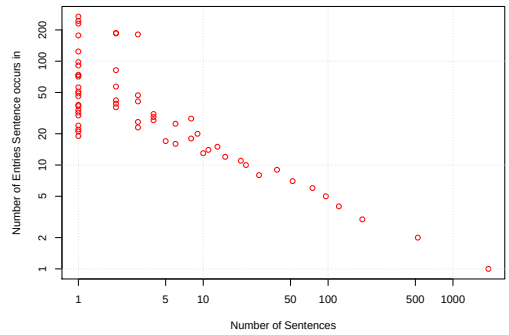
[3]The sentence "the sequence shown here is derived from an embl/genbank/ddbj whole genome shotgun (wgs) entry which is preliminary data." occurs ~ 7.2 million times.



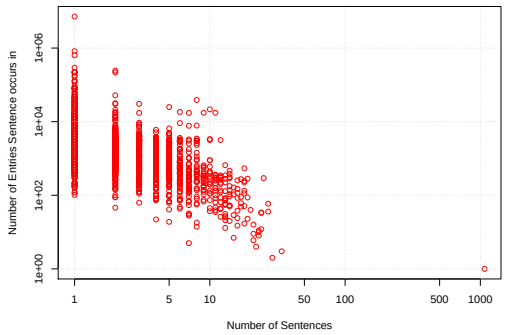
(a) Swiss-Prot Version 9



(b) UniProtKB/Swiss-Prot Version 2012_05



(c) TrEMBL Version 1



(d) UniProtKB/TrEMBL Version 2012_05

Figure 6.4: The distribution of sentences in four versions of UniProtKB. A point on the graph represents the number of entries that a given number of sentences occurs in. For example, the bottom right point in each graph will represent the number of sentences that occur in only a single entry (i.e. the number of singleton sentences in a database version).

Whilst Figure 6.4 highlights the rise of sentence reuse, the increase of data points between Swiss-Prot and TrEMBL versions is evidence of new sentences being added to UniProtKB over time. The growth of total sentences, as shown in Figure 6.5, is significant; in UniProtKB Version 2012_05, Swiss-Prot contained almost five million total sentences (Swiss-Prot Version 9 contained 15,773 sentences), whilst TrEMBL contained almost 27 million sentences (TrEMBL Version 1 contained 12,334 sentences). As also exhibited in Figure 6.5, the growth of TrEMBL is irregular and disjointed, with fluctuations often occurring between versions. For example, in fourteen instances, there is a decrease in the total number of sentences between TrEMBL versions. This growth of sentences within UniProtKB generally fits with the growth of the database as a whole, as previously shown in Figure 2.9; the number of entries correlates with the number of sentences.

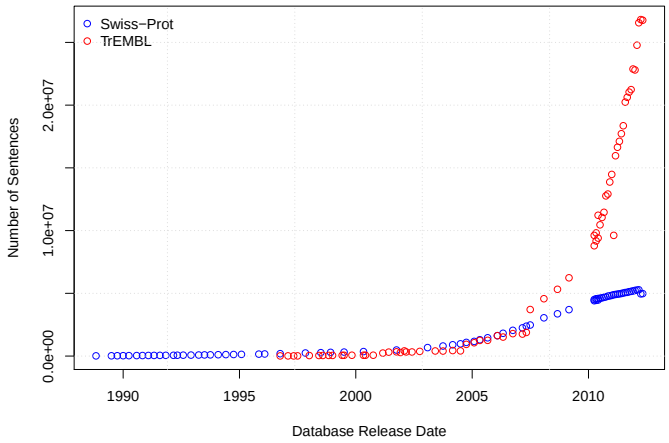


Figure 6.5: The total number of sentences in Swiss-Prot and TrEMBL entries over time.

Figures 2.9 and 6.5 allow the growth of UniProtKB to be analysed over time. However, the distribution of sentences, as shown in Figure 6.4, only allows individual database versions to be analysed. How, then, can the change in distribution be visualised between database versions? One possible approach is to show the average number of sentences appearing in entries for various versions of UniProtKB. This analysis, as shown in Figure 6.6, indicates that the number of sentences within an average Swiss-Prot entry is increasing over time[4]. The most recent version of Swiss-Prot has,

[4]Only entries containing textual annotation are considered in this calculation.

on average, approximately six sentences per entry compared to Swiss-Prot Version 9, which had approximately two sentences per entry; a threefold increase over twenty years. Conversely, TrEMBL has had a number of fluctuations over time, but has typically remained at an average of between two and three sentences per entry.

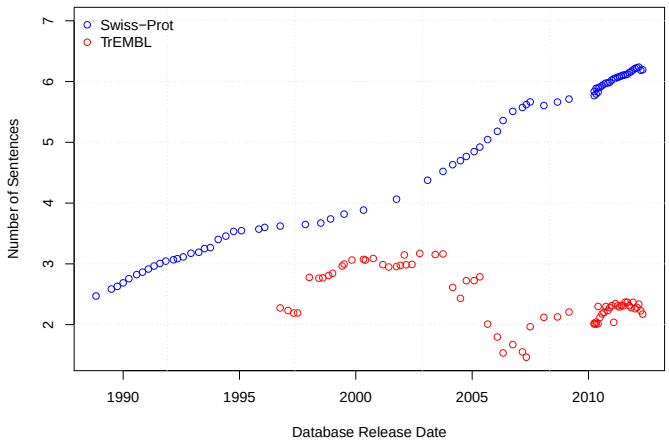
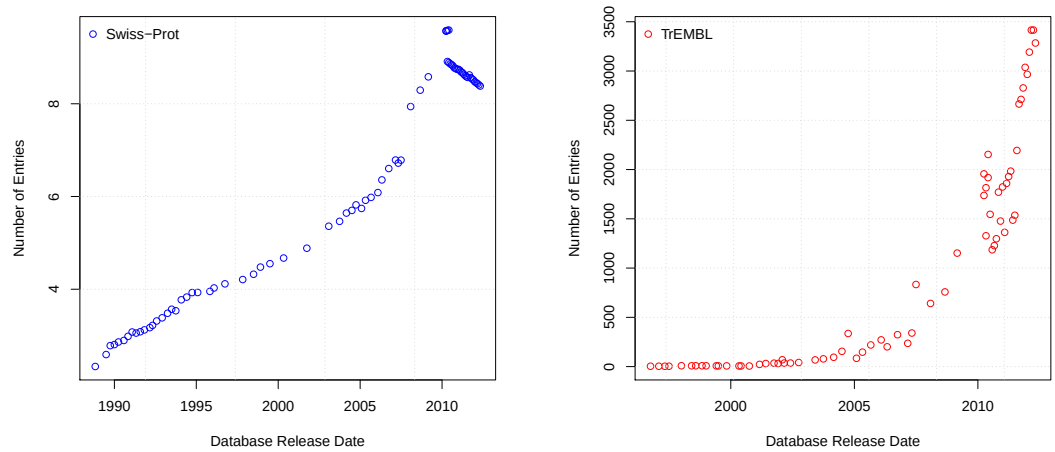


Figure 6.6: The number of sentences that are contained within the annotation of Swiss-Prot and TrEMBL entries on average over time.

This increase in the number of sentences within the textual annotation of Swiss-Prot entries over time fits with one of the goals of UniProtKB, which is to attach as much information as possible to each protein entry [216]. Given the exponential growth of entries in UniProtKB, it is likely that the increase of sentences within textual annotation has only become possible through sentence reuse. This is supported by Figure 6.7, which shows that the average number of entries that each unique sentence occurs in is generally increasing for Swiss-Prot and TrEMBL, to a current average of approximately 8 and $3,500$, respectively. Later versions of Swiss-Prot are an exception to this trend, as they have started to show a steady decline in reuse. Whilst this decline coincides with the change in release cycle of UniProtKB, it also appears that it is caused by a change to the annotation policy in Swiss-Prot. After 2010 only those entries containing sequences with experimental annotation were added to Swiss-Prot; previously ortholog sequences from complete genomes that were automatically annotated were frequently included.

These results suggest that, whilst the total textual annotation is increasing for en-



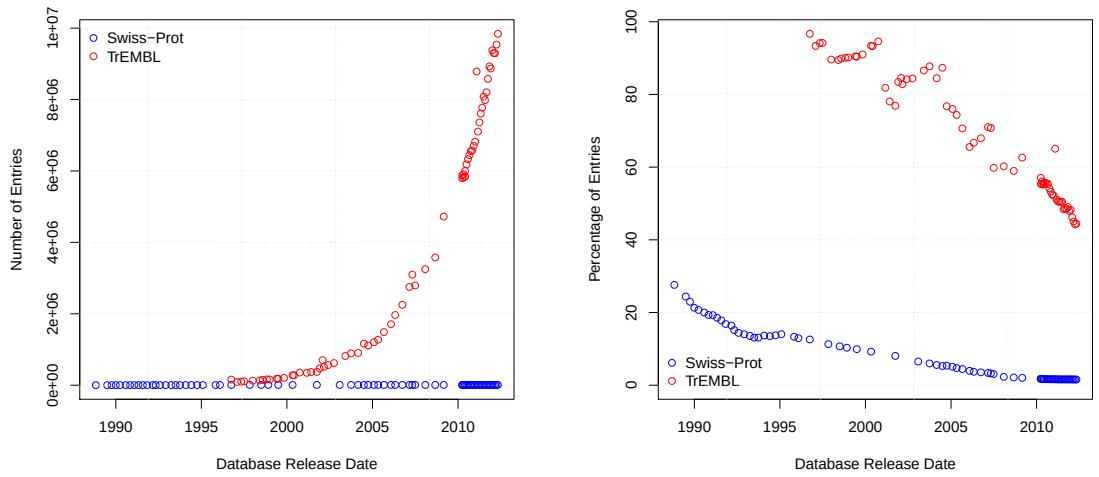
(a) Swiss-Prot (b) TrEMBL

Figure 6.7: The number of Swiss-Prot and TrEMBL entries that an average sentence occurs in over time.

tries on average, it is driven by sentence reuse. Another factor affecting the quantity of sentence reuse could be UniProtKB attempting to reduce the number of entries that remain without any textual annotation. Figure 6.8a shows the number of these UniProtKB entries over time, with the overall percentage shown in Figure 6.8b. The overall percentage is decreasing; only 1.5% of entries in UniProtKB/Swiss-Prot Version 2012_05 contain no textual annotation, compared to 45% of entries in TrEMBL. Both of these show significant improvements over time – initially Swiss-Prot had 27.6% of entries without any textual annotation in 1988 whilst TrEMBL had 96.7% in 1996. From these results, it is concluded that, in addition to the increase of the overall database size, the percentage of entries with annotation is increasing; these two factors both contribute to the increasing reuse of sentences.

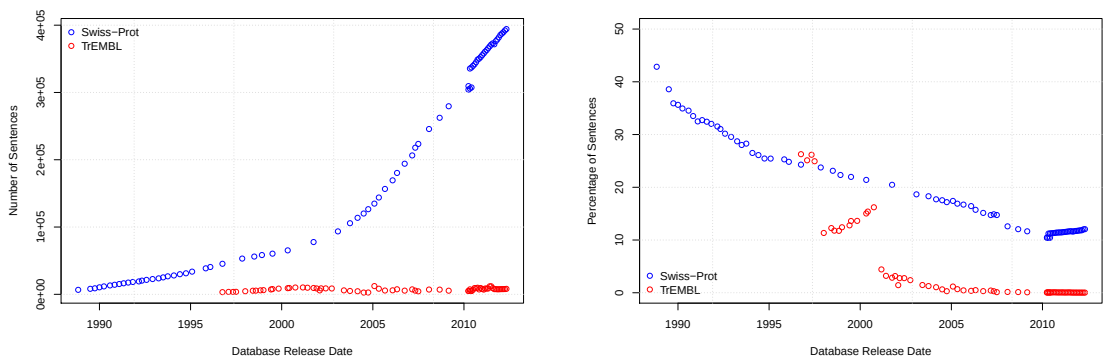
These results are based on the total number of sentences in UniProtKB, which includes redundant sentences. Removing this redundancy results in an analysis of the unique sentences within Swiss-Prot and TrEMBL, which is shown in Figure 6.9a. Figure 6.9a shows an increase of unique sentences in Swiss-Prot, due to new sentences being regularly added to the annotation corpus. Conversely, TrEMBL shows no overall trend, with fluctuations between versions suggesting that its corpus is volatile.

Whilst the number of unique sentences in Swiss-Prot is growing, the overall percent-



(a) Total number (b) Percentage

Figure 6.8: The number of Swiss-Prot and TrEMBL entries without any textual annotation over time.

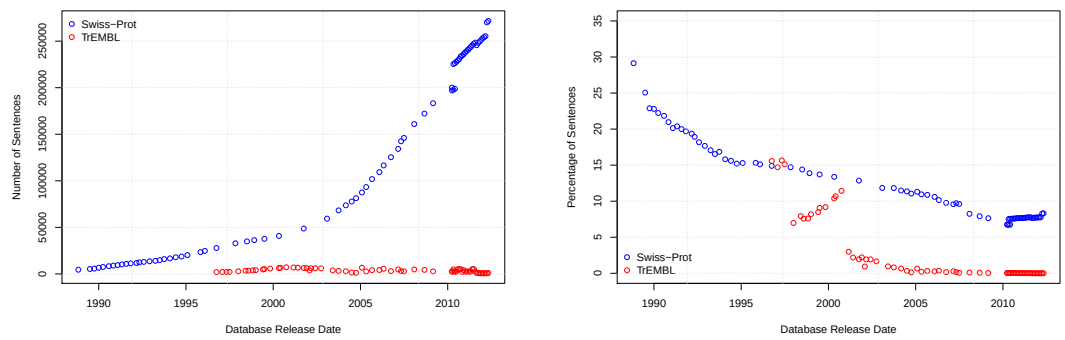


(a) Total Number (b) Percentage

Figure 6.9: The number of unique sentences in Swiss-Prot and TrEMBL over time.

age of unique sentences in both TrEMBL and Swiss-Prot, as shown in Figure 6.9b, is decreasing. Swiss-Prot shows a steady decline, with a small increase in later versions that coincides with the change in release cycle and annotation procedure. The decrease in TrEMBL shows two significant jumps, as previously seen and discussed in Section 4.4, which relate to changes in the automated curation process. These figures provide further evidence that sentence reuse in both databases is on the rise. For example, within UniProtKB/TrEMBL Version 2012_05 there are over 22 million entries, containing approximately 26.7 million sentences, $8,131$ of which are unique; i.e. the unique sentence corpus of TrEMBL is 0.03% of the total sentence corpus.

Finally, within a database version there will be a number of sentences which occur once, and only once, within a database version; that is, they are singleton sentences. The number of singleton sentences is shown in Figure 6.10a with the percentage shown in Figure 6.10b. Although there are less singleton sentences than unique sentences within a database version, they both follow an almost identical pattern of decrease over time. In UniProtKB Version 2012_05 there are $389,558$ and $7,760$ ($\sim$ 7% and $\sim$ 0.03%) unique sentences and $255,349$ and 735 ($\sim$ 5% and $\sim$ 0.003%) singleton sentences in Swiss-Prot and TrEMBL, respectively.



(a) Total Number (b) Percentage

Figure 6.10: The number of singleton sentences in Swiss-Prot and TrEMBL over time.

These results, as summarised in Table 6.1, show that sentence reuse is common in UniProtKB, with a steady introduction of new sentences to the annotation corpus over time; this suggests that sentences can be used as annotation markers. If this is true, then can the provenance, and subsequent propagation, of a sentence be inferred?

This can be explored through the application of VIPeR to individual sentences.

Figure	Summary
6.4	Illustrates the distribution of sentences throughout the Swiss-Prot and TrEMBL databases, providing an overview of how frequently each sentence occurs.
6.5	The total number of sentences represents how many sentences are contained across all entries within each database version.
6.6	The number of sentences within each database entry on average gives a broad illustration of annotation depth within database entries.
6.7	How many entries an average sentence appears in over time broadly shows how generic an average sentence is.
6.8	The number of database entries without any entries gives an indication of annotation coverage over time.
6.9	Unique sentences represent the size of the annotation corpus.
6.10	Singleton sentences represent how many sentences within the corpus are not subjected to reuse.

Table 6.1: Summarising the information that can be drawn from each of the seven figures presented in this section.

6.3 Inferring Provenance and Exploring Annotation Propagation

The analysis of sentence reuse performed in the previous section provides confidence that sentences can be used as annotation markers. Here VIPeR, the visualisation approach developed in Chapter 5, is applied to a variety of sentences to determine if the provenance of an annotation can be identified. The visualisations chosen as exemplars are based on their features and clarity in highlighting a given issue or pattern.

Figure 6.11 shows the corresponding visualisation for two sentences: "it is uncertain whether met-1 or met-4 is the initiator." and "the active-site selenocysteine is encoded by the opal codon, uga.". In both cases the visualisation shows that the sentences originated in Swiss-Prot Version 9. In the visualisations Swiss-Prot Version 9 is the earliest possible database version, as Swiss-Prot Versions 1-8 and 10 were never archived. Therefore, it is possible that these two sentences actually originated in an earlier database version.

Figure 6.11a shows that the sentence "it is uncertain..." originated in a single entry, with the accession number P01011. This visualisation has, therefore, allowed the likely provenance of the sentence to be inferred; entry P01011 can be defined as the *root entry* for this particular sentence. The sentence "it is uncertain..." has remained in this root entry throughout the history of the database, with it still remaining in UniProtKB Version 2012_05, where it appears in a further 48 UniProtKB entries. Since the sentence first appeared in the database over twenty years ago its reuse has slowly increased, as illustrated in Figure 6.12.

This increase does not distinguish between primary and secondary accessions; all entries are treated as primary. This approach is used as UniProtKB entries may become merged, demerged or deleted, as previously discussed in Section 2.4. Whilst the inclusion of all accession points can make graphs more dense, their exclusion would make many graphs appear misleading. For example, the entry Q13703 was originally in TrEMBL until it was merged with entry P01011. At this point, Q13703 only contained a single annotation ("belongs to the serpin family."), which was also present within P01011. If secondary accessions were removed from these visualisations, then

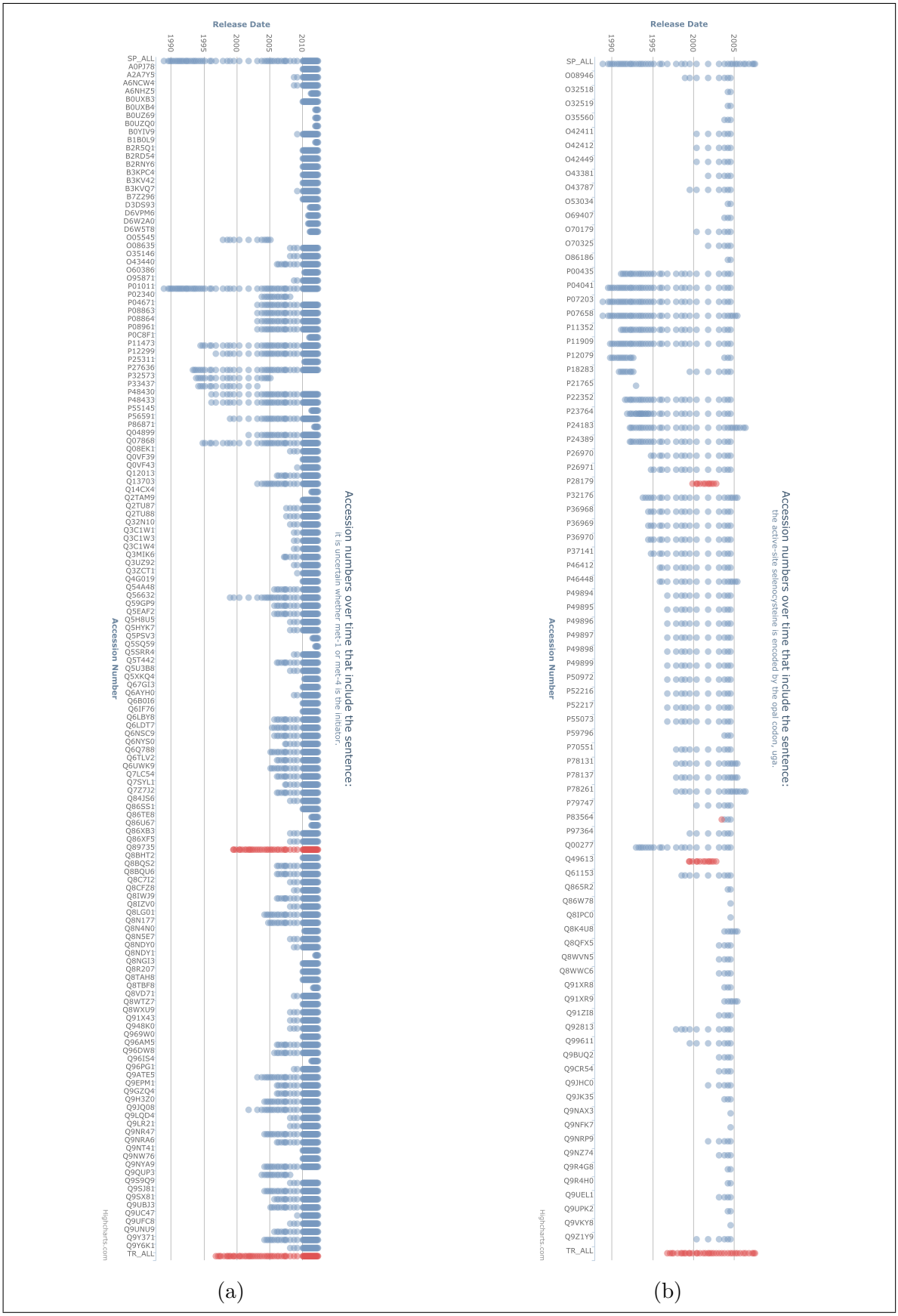


Figure 6.11: Visualisation for the sentences (a) “it is uncertain whether met-1 or met-4 is the initiator.” and (b) “the active-site selenocysteine is encoded by the opal codon, uga.”.

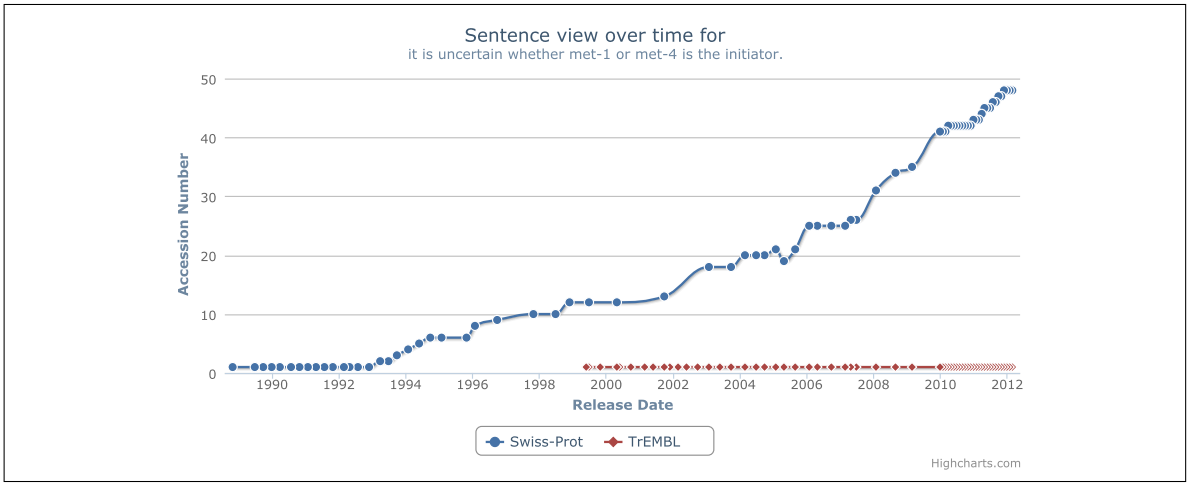


Figure 6.12: The number of UniProtKB entries that the sentence "it is uncertain whether met-1 or met-4 is the initiator." appears in over time.

the corresponding visualisation for the sentence "belongs to the serpin family." would show that the sentence had been removed from this accession, rather than merged. In total, the root entry P01011 has become merged with 11 other entries.

Unlike Figure 6.11a, the provenance for the sentence "the active-site. . . ", as visualised in Figure 6.11b, shows that it originates in two entries (P07658 and P07203). It is possible, given the UniProt curation protocol, that a sentence may be introduced within multiple entries for the same database release. However, prior to 2010, UniProt made a distinction between minor and major releases. As the visualisations only show major releases, it is possible that a finer level of granularity could be identified between minor releases. Whilst it is possible to access minor release data via UniSave, UniProt do not archive minor releases on their FTP server. Given the large number of minor releases and the subsequent volumes of data that would have to be parsed from the UniSave website, minor data was excluded.

From Figure 6.11b, it can be inferred that the origin entries for the sentence "the active-site selenocysteine is encoded by the opal codon, uga." are P07658 and P07203. Overall, this sentence appeared in 84 unique entries and, at its peak, was found in a total of 52 Swiss-Prot entries (Swiss-Prot Version 44), as illustrated in Figure 6.13. After Swiss-Prot Version 44, the sentence was removed from the majority of entries and by UniProtKB Version 8 it was removed entirely from the database.

This result shows that, whilst sentences may be added to the database, they are also

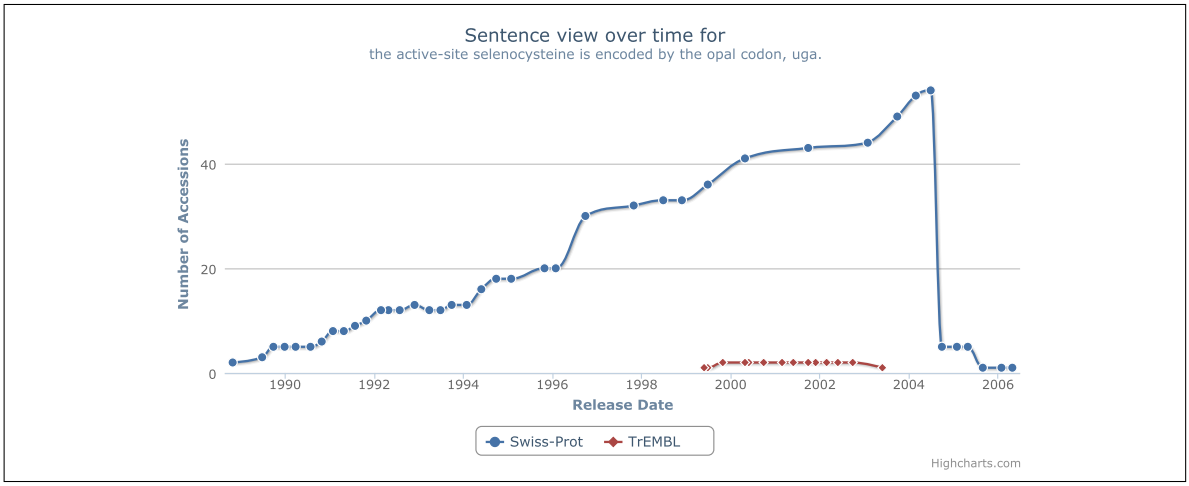


Figure 6.13: The number of UniProtKB entries that the sentence "the active-site selenocysteine is encoded by the opal codon, uga." appears in over time.

eligible for removal. In total, there has been 611,080 unique sentences across all versions of UniProtKB, with UniProtKB Version 2012_05 containing 395,564 unique sentences. A total of 215,516 unique sentences have been removed from UniProtKB.

The removal of a sentence can happen for a number of reasons, such as the correction of spelling and grammatical mistakes or because the biological information was found to be incorrect. The sentence "the active-site selenocysteine is encoded by the opal codon, uga.", was likely removed from the database due to formatting changes. Investigating the entries where the sentence was removed between Swiss-Prot Version 44 and 45 in UniSave[5] shows that the selenocysteine information was moved from the textual annotation to the feature table. The latest version of the UniProtKB user manual also includes details on how selenocysteine information is encoded within the feature table.

It is not publicly documented when this change in formatting procedure was introduced. However, it can be assumed that the procedure was introduced between Swiss-Prot Version 44 and 45, given this sentence was removed from 72 entries between these two versions. There does not appear to be any other clear reason for this sharp decline. Interestingly, the sentence still remained in nine entries after this point, until they were eventually updated to the new format. It appears that these entries were initially missed from the format update.

After this update, any sentences that remain in the database should technically have

[5]For example, in P07203 (`http://www.uniprot.org/uniprot/P07203?version=52&version=47`)

been removed. In this case, the sentence "the active-site..." remained in a number of entries after it was removed from the two origin entries. Sentences where the origin sentence has been removed could indicate sentences which incorrectly remain in the database; that is, sentences which follow the *missing origin propagation pattern*.

In addition to the missing origin pattern, a number of other propagation patterns can be identified by analysing the visualisation for the sentence "the active-site...", as shown in Figure 6.11b:

- **Reappearing entry** – In entries P18283 and P12079, the sentence is initially removed, only for it to be re-added after a number of versions have elapsed.
- **Transient appearance** – In a number of entries, such as P21765, the sentence only appears in a single database version. It is removed from the subsequent release.
- **Originating in TrEMBL** – Although not shown in Figure 6.11b, there are sentences that originate in TrEMBL, before subsequently being propagated into Swiss-Prot entries.

VIPeR allows us to infer the provenance of a sentence. Additionally, inspection of these graphs has led to the discovery of a set of propagation patterns. These patterns are unexpected; why, for instance, should a sentence appear in only a single version of UniProtKB, or should a sentence disappear in an originating entry, but remain in an apparently derived entry? These patterns need to be examined further, exploring how frequently each pattern occurs within the database and what quality information can be drawn from them.

6.4 Identifying and Analysing Propagation Patterns

In the previous section, four propagation patterns were identified through the examination of sentence visualisations. Within this section these patterns are analysed as possible indicators for quality and correctness. If these patterns do hold analytical value, then a significant number of sentences adhering to each pattern will exist within the database.

For a sentence to be identified as adhering to a pattern, it only has to be exhibited in a single entry. For example, a sentence identified as reappearing may occur in ten entries, but only reappear within one of these entries; being removed and then re-added to a single entry is enough to classify the sentence as reappearing.

6.4.1 Transient sentences

A sentence is defined as transient when it appears within an entry for only a single database version. For example, Figure 6.14 shows the visualisation for the sentence "this is a conceptual translation; a frameshift was introduced in position 81 to produce this orf.", which only appears in two entries (P22788 and P27826) for a single database version (Swiss-Prot Version 40). As this sentence does not appear in either of these entries for another release of Swiss-Prot, it is classified as transient.

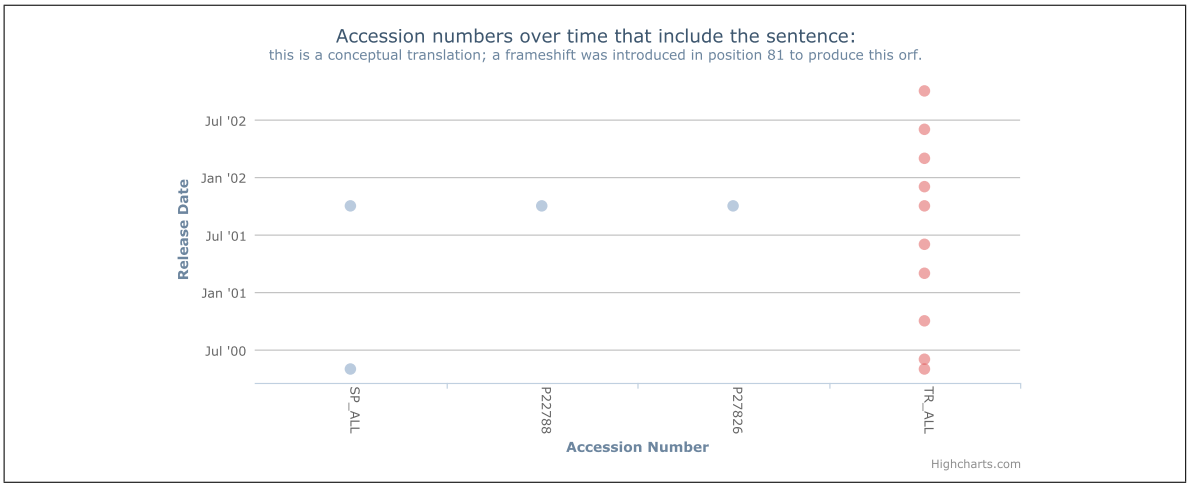


Figure 6.14: An example of a sentence ("this is a conceptual translation; a frameshift was introduced in position 81 to produce this orf.") that follows the transient propagation pattern.

Programmatically identifying transient sentences is straightforward. The number of database versions that a sentence appears in an entry for is counted, with those sentences existing in only a single version being extracted. A total of 68,042 transient sentences were identified, 25,582 of which exist in UniProtKB Version 2012_05.

Transient sentences may occur for a number of reasons. For example, the sentence "this is a conceptual translation; a frameshift was introduced in position 81 to produce this orf." shown in Figure 6.14 was replaced in P27826 and P22788 by the sentence "ref.3 sequence differs from that shown due to a frameshift in position 81 that produces two separate orfs."[6]. In this case, the sentence was replaced by a more specific and detailed sentence.

A further example can be seen in Figure 6.11b for the sentence "the active-site selenocysteine is encoded by the opal codon, uga.". This sentence was transient in six entries. Five of these cases occurred in Swiss-Prot Version 44 when the sentence was moved to the feature table, as previously discussed. The remaining instance was in P21765 for Swiss-Prot Version 24, which was replaced by "the active-site is not encoded by the opal codon uga but by ugc."[7]. This replacement indicates that the knowledge in the original annotation is now considered *erroneous*. The definition of an erroneous annotation used within this thesis follows that of UniProt [319]: An erroneous annotation is one that is out of sync with respect to the biological knowledge (it may be that the original information is incorrect, rather than the annotation).

These two examples suggest that transient sentences could provide indications for erroneous and low quality annotations. However, these indications can only be detected after the sentences has already been modified; if it remains in the database, then it is no longer transient.

Within UniProtKB Version 2012_05 there are 25,582 potentially transient sentences. Many of these sentences however will remain in the next and subsequent versions of UniProtKB, i.e. will not become transient. Therefore potential transience is not in itself an indicator of low quality. This pattern does, however, fit with previous research that links annotation quality to stability [171]; annotations that have persisted over

[6] `http://www.uniprot.org/uniprot/P27826?version=16&version=21`
[7] `http://www.uniprot.org/uniprot/P21765?version=6&version=8`

many release cycles provide greater confidence and likelihood in their correctness.

Therefore, using this information it can be concluded that an annotation introduced within one entry update is more likely to be volatile than those which have remained over numerous releases. Unlike the other patterns which will be discussed, the presence of potentially transient sentences should not be seen as an indicator of low quality, given that all sentences are new at some point in time.

6.4.2 Originating in TrEMBL

Given the curation protocol implemented by UniProtKB, many sentences will exist in TrEMBL that originated from Swiss-Prot. However, there are cases of sentences existing within Swiss-Prot entries that appear to have originated from TrEMBL. An example of this is the sentence "inactivated by cyanide.", which originated in eleven TrEMBL entries, as shown in Figure 6.15. This sentence appeared in Swiss-Prot after three of the entries were moved from TrEMBL into Swiss-Prot.

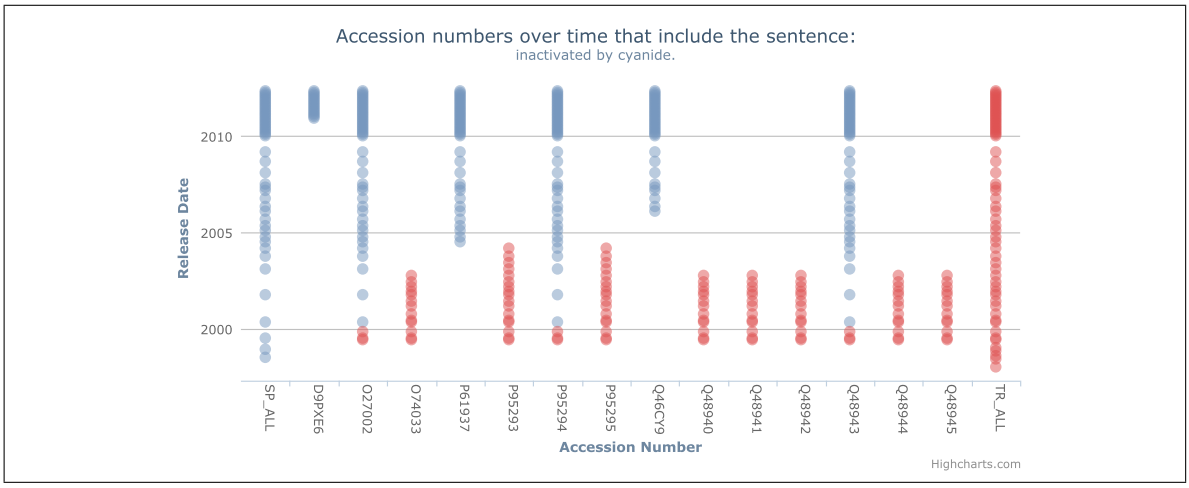


Figure 6.15: An example of a sentence ("inactivated by cyanide.") that originates in TrEMBL, but ends up in Swiss-Prot. In this case, a number of the TrEMBL entries are merged into Swiss-Prot.

Identifying sentences that originate in TrEMBL requires all of the origin entries that a sentence appears in to be extracted. If all of these entries appear in TrEMBL, then subsequent versions are checked to see if the sentence appears in any Swiss-Prot entries; those that do are extracted.

In total, over 8,500 sentences were identified in Swiss-Prot that originated in TrEMBL.

This is a surprising observation; annotations in Swiss-Prot are considered manually reviewed and curated, whilst TrEMBL annotations can be generated based upon information from Swiss-Prot annotations [134].

Sentences can move from TrEMBL into Swiss-Prot in two ways: the entire entry moves from TrEMBL into Swiss-Prot, with the sentence retained; or a sentence is propagated independently. Figure 6.15 illustrates the former of these possibilities, with three origin entries being moved into Swiss-Prot, whilst Figure 6.16 shows an example of the latter possibility. Although these sentences are not manually curated, they will undergo manual review when being included within Swiss-Prot.

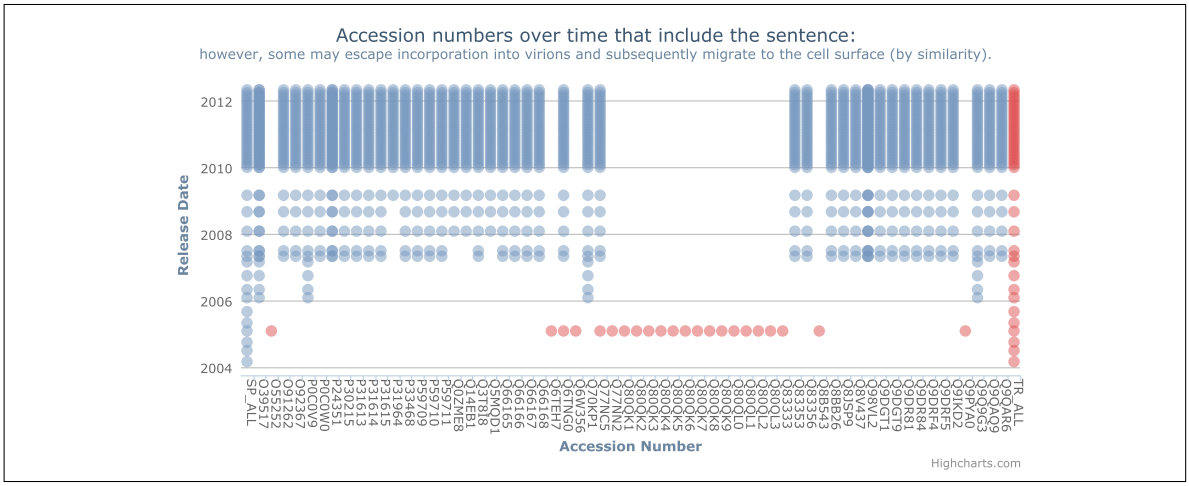


Figure 6.16: An example of a sentence ("however, some may escape incorporation into virions and subsequently migrate to the cell surface (by similarity).") that originates in TrEMBL, but ends up in Swiss-Prot. In this case, none of the origin entries were merged into Swiss-Prot.

One possible explanation for a number of sentences appearing in Swiss-Prot is that, for a period of approximately two years, some annotations in TrEMBL underwent manual annotation [320]. Therefore, a number of sentences propagated from TrEMBL will have been manually curated.

Given that the quality of annotations in TrEMBL is typically of lesser quality than those in Swiss-Prot, it would be of interest to perform a quality analysis between these sentences and those originating directly from Swiss-Prot. However, as previously discussed in Chapter 4, a quality analysis of small sets of annotations is problematic. This result does, however, highlight that annotation provenance should be clearly documented and available to users, especially given that research has suggested that

users often assume annotations are of a consistent quality [321].

6.4.3 Reappearing sentences

A reappearing sentence is defined as one which is removed from an entry and then, after a number of database releases, is re-added to the same entry. An example of a reappearing sentence is shown in Figure 6.17. In this case, the sentence was removed from entry P06229 for seven versions.

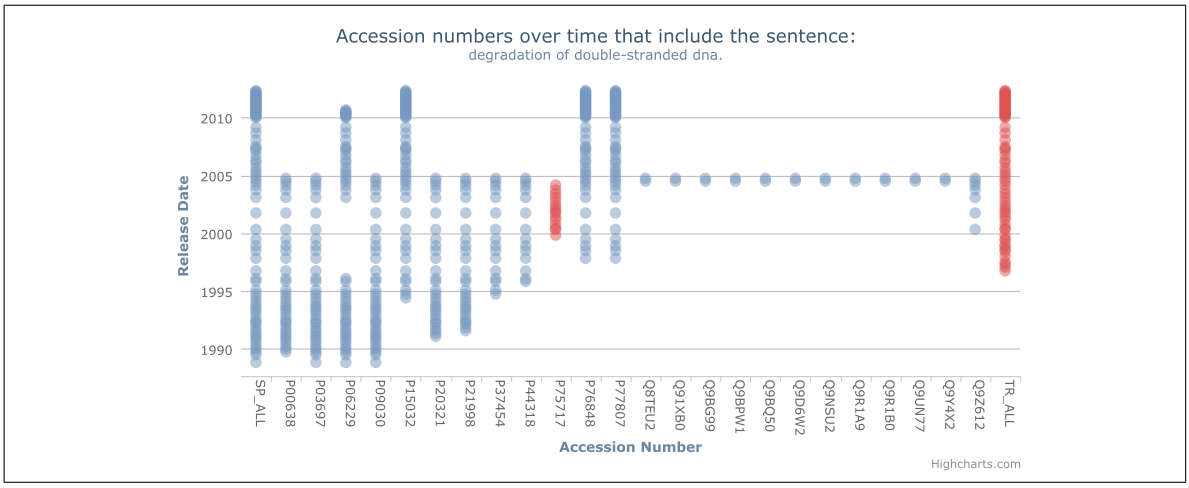


Figure 6.17: An example of a sentence ("degradation of double-stranded dna.") that follows the reappearing propagation pattern. The sentence is removed from P06229, but is re-added to the entry after approximately seven years.

Identifying reappearing sentences is less straightforward than for other patterns. For each entry a sentence appears in, the first and last versions are identified and the difference between these versions is calculated. Generally, this should equal the number of releases. For example, if a sentence originates in Swiss-Prot Version 15 and is last seen in Swiss-Prot Version 35, then it should appear in a total of 20 versions; if it does not, then it is a reappearing entry. However, this approach also has to take into account changes in version numbering, release cycles and entries that move from TrEMBL into Swiss-Prot. In total, 15,587 reappearing sentences were identified in UniProtKB.

In Figure 6.17, the sentence "degradation of double-stranded dna." was removed from P06229 for seven versions. During this time it was replaced with the sentence "degradation of double-stranded and singel-stranded [sic] dna"[8], before reverting back to the

[8] `http://www.uniprot.org/uniprot/P06229?version=7&version=6`

original sentence ("degradation of double-stranded dna.")[9].

In Figure 6.11b there are two examples of this pattern; the sentence "the active-site selenocysteine is encoded by the opal codon, uga." is removed from both entry P18283[10] and P12079[11] in Swiss-Prot Version 24. In these entries, the sentence "the active-site selenocysteine is encoded by the opal codon, uga." was replaced with "the active-site selenocysteine is encoded by the opal codon, uga (by similarity).", with the corresponding visualisation for this sentence shown in Figure 6.18. The usage of "by similarity" suggests that the information is based on sequence similarity. Interestingly, the sentence "the active-site selenocysteine is encoded by the opal codon, uga (by similarity)." follows the "missing origin" pattern.

The original sentence ("the active-site selenocysteine is encoded by the opal codon, uga.") is re-added to P18283 and P12079 after seven and eleven years, respectively. In the latest version of P12079, which has become merged with P11352, a comment has been added that states "sequence was originally thought to originate from human."[12]. Looking at the history of both entries this confusion appears to have led to the uncertainty about the selenocysteine annotation. The sentence reappears in P12079 when it is merged with P11352. There is no clear indication in P18283 why the sentence was reinstated. In the latest version of both these entries, the encoding of selenocysteine is documented in the feature table.

Sentences exhibiting this pattern appear to indicate a conflict in the underlying evidence and some uncertainty as to the correct annotation. The impact of this pattern is similar to transient sentences; they highlight the importance of annotation stability and provenance.

6.4.4 Missing origin

Of the identified patterns, the missing origin pattern holds the most promise as being an indicator for erroneous annotation. Missing origin sentences are those that exists in

[9] `http://www.uniprot.org/uniprot/P06229?version=25&version=17`
[10] `http://www.uniprot.org/uniprot/P18283?version=5&version=6`
[11] `http://www.uniprot.org/uniprot/P12079?version=6&version=7`
[12] `http://www.uniprot.org/uniprot/P11352.txt?version=128`

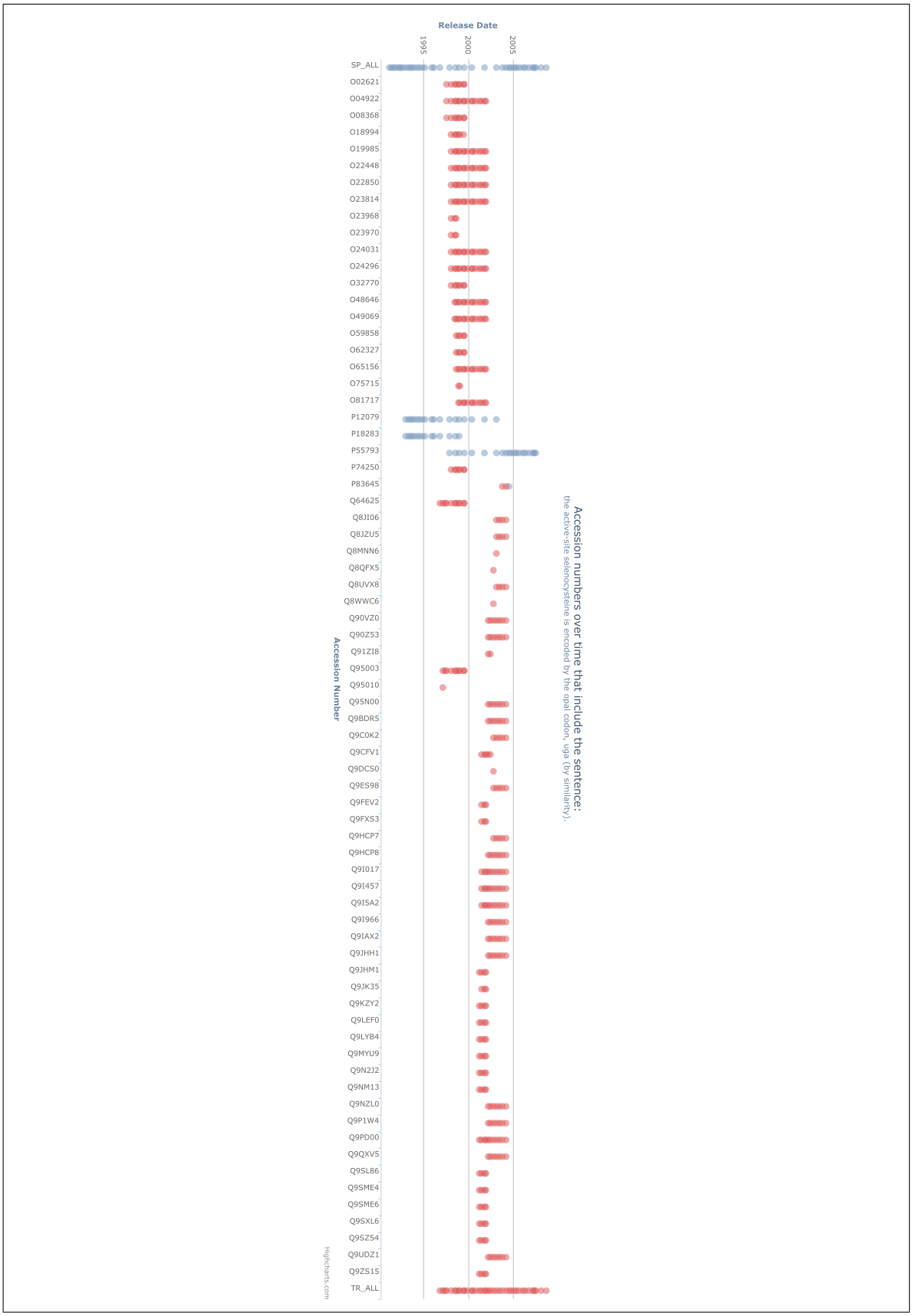


Figure 6.18: Visualisation of the sentence “the active-site selenocysteine is encoded by the opal codon, uga (by similarity).”.

the database after they have been removed from the entries where they originated. Figure 6.19 illustrates a sentence ("this methionine-rich region is probably important for copper tolerance in bacteria (by similarity).") that originated in two entries and, after its removal from the origin entry, remained within the database in another (secondary) entry.

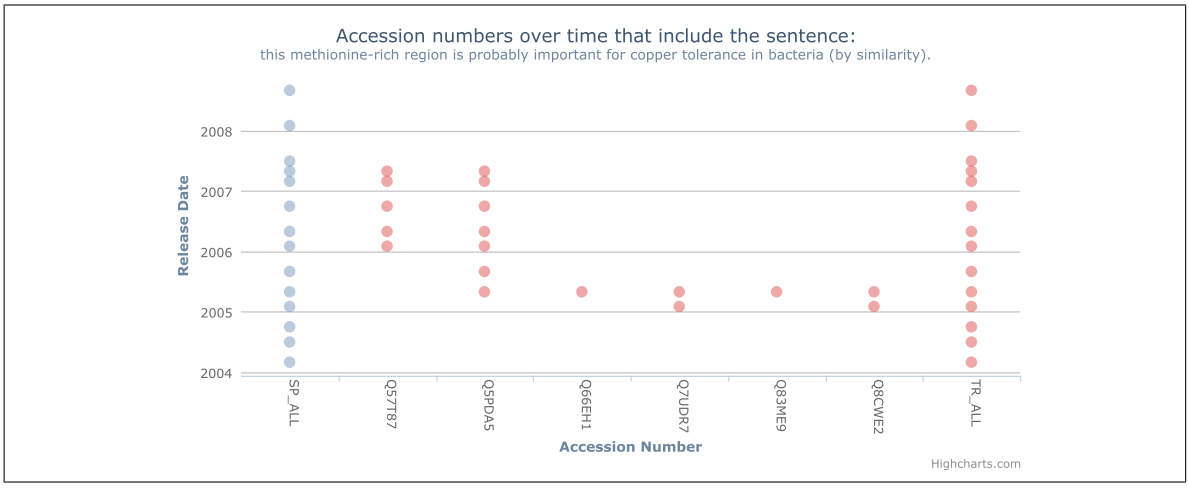


Figure 6.19: An example of a sentence "this methionine-rich region is probably important for copper tolerance in bacteria (by similarity).", which follows the missing origin pattern.

Sentences adhering to the pattern can be identified by taking two sets: the origin entry, or entries, and the final entries that the sentence occurs within. If the intersection of these sets is empty, then the sentence follows the missing origin pattern. Using this approach, a total of $8,355$ sentences were identified.

The removal of the sentence "this methionine-rich region is probably important for copper tolerance in bacteria (by similarity)." from the origin entry Q7UDR7, as shown in Figure 6.19, coincided with the removal of a substantial amount of other information[13], including the sentence "this methionine-rich...", in UniProtKB/TrEMBL Version 5. In UniProtKB/TrEMBL Version 11 a significant amount of annotation was also removed from the secondary entry Q5PDA5[14], with the majority of the annotation, including "this methionine-rich...", being the same as that removed from Q7UDR7 earlier. This suggests that the sentence "this methionine-rich..." was erroneous and could have been removed six database releases (two years) earlier.

[13] `http://www.uniprot.org/uniprot/Q7UDR7?version=10&version=14`
[14] `http://www.uniprot.org/uniprot/Q5PDA5?version=23&version=20`

Similar to other analyses, the pattern may identify sentences that have been removed for other reasons, such as formatting changes. As previously seen, in Figure 6.11b the sentence was removed from the origin entries due to moving selenocysteine information from the textual annotation to the feature table in UniProtKB entries. Therefore, this was not biologically erroneous in these nine entries. However, it clearly should have been moved to the feature table in all entries for consistency. This highlights how missing the propagation of textual annotation can lead to inconsistencies between entries.

Changes in annotation are typically made to reflect an update in knowledge; in light of new knowledge a previous annotation may now be erroneous with respect to current knowledge. Given that annotations propagate, any updates to an original annotation should also be propagated. However, over 8,000 sentences have been identified which may, or may have, incorrectly remained in the database with 3,835 of these sentences remaining in UniProtKB Version 2012_05.

6.4.5 Propagation patterns summary

Overall, numerous sentences were identified for each pattern, as summarised in Table 6.2. In total, over 85,000 sentences followed at least one of the identified patterns, with over 35,000 sentences remaining in UniProtKB Version 2012_05; in other words, approximately 9% of the unique sentences in UniProtKB Version 2012_05 follow one of the identified patterns.

Pattern	Number of sentences	Number in just UniProtKB Version 2012_05
Missing Origin	8,355	3,835
Reappearing Sentence	15,587	7,011
Transient appearance	68,042	25,582
Originating in TrEMBL	8,649	5,330

Table 6.2: The number of sentences that adhere to each pattern, for all versions of UniProtKB and for those just in UniProtKB Version 2012_05. To place these results in context, there have been a total of 611,080 unique sentences, with UniProtKB Version 2012_05 containing 394,233 unique sentences.

The identification of these patterns mostly highlight the importance of annotation stability and provenance. However, the examples explored for the missing origin pattern

suggest that the identified sentences could be erroneous. This hypothesis is explored in the following section.

6.5 Error Detection

As an annotation is subject to propagation, then any changes to the original annotation should percolate to all entries that are affected. Those that do not are said to follow the missing origin pattern and could indicate erroneous annotation. A missing origin sentence is defined as one which:

1. Initially occurs within an *origin* entry (or entries).
2. Later appears in an additional entry (or entries); i.e. a *secondary* entry.
3. Is removed or changed in the origin entry.
4. Remains unchanged within the secondary entry for a subsequent database release (or releases).

In the previous section, over $8,000$ sentences in UniProtKB were identified as following this pattern. We first define a *classification* scheme for these sentences.

6.5.1 Defining classifications

As identified in Section 6.4, a sentence may be removed from the database for a number of reasons. These reasons can be categorised into five possible classifications:

- **Erroneous** — The sentence in the secondary entry is inaccurate or incorrect given updates to the origin entry. Details may have been added or removed from the annotation, making it out of sync with the current biological knowledge. For example, a sentence that has been removed entirely may be deemed erroneous.
- **Inconsistent** — Although the sentence in the origin entry has been updated, it has not changed the biological information contained within the annotation. For example, a sentence where a grammatical error has been corrected.
- **Accurate** — The sentence in the secondary entry is accurate. Either the sentences appear identical by coincidence or the updates to the origin are not valid

in the secondary entry. Therefore both annotations have become independent. For example, expression information may not be relevant in different organisms for the same gene.

- **Too many results** — The sentence is very heavily reused within UniProtKB and is deemed infeasible to analyse. The more entries that a sentence occurs within, the more troublesome it becomes to classify individually. Sentences that occur in over 100 entries are classified as "too many results".

- **Possibly erroneous** — Some sentences will not carry enough evidence, or contain conflicting information, to allow a classification to be assigned with confidence. In these case, sentences are classified as "possibly erroneous".

To determine the classification of a particular sentence, various factors need to be considered. As the interpretation of biological data could vary between users, we defined a series of steps that are used to classify a sentence.

6.5.2 Classification protocol

There are four main decisions when evaluating the classification of a textual annotation: deciding if analysing the sentence is feasible; determining if the sentence was propagated between the entries; deciding whether the update to the origin was relevant to the secondary entry; and deciding whether the update affected the meaning of the textual annotation in the secondary entry. These decisions are summarised as a decision tree in Figure 6.20.

The development of the decision tree provides the foundations of the classification protocol, which details the steps required to classify a given sentence. The entire protocol involves six key stages:

1. Determine how many entries the sentence has propagated to. A sentence occurring in over 100 entries is infeasible to analyse (Figure 6.20, Question 1).

2. Using VIPeR, identify both the *origin* and *secondary* entries that the sentence occurs in.

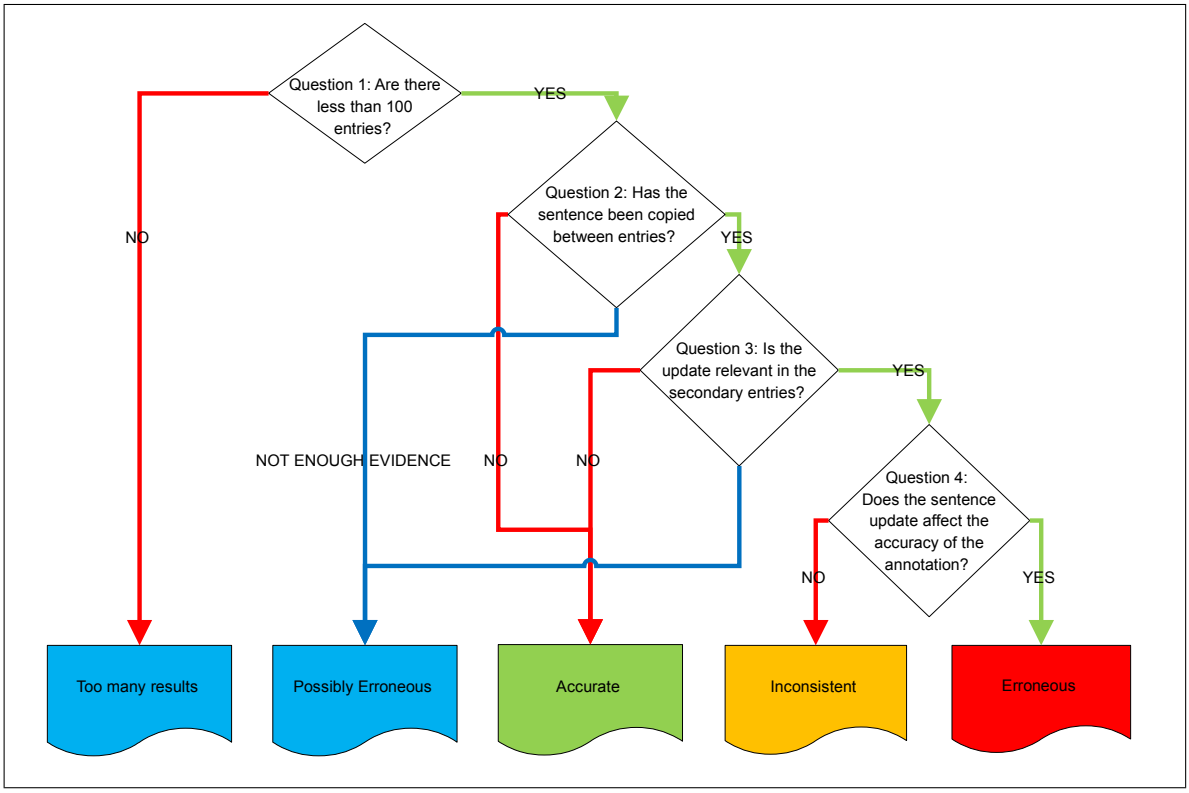


Figure 6.20: Decision tree summarising the protocol used to determine the classification of sentences. There are four main questions within the protocol that lead to a sentence being classified into one of five possible classifications.

3. Using the UniSave tool, analyse the context of the sentence within the origin and secondary entries at the time that the sentence was initially added to the secondary entry. Does this context suggest the sentence was propagated between the entries (Figure 6.20, Question 2)?

4. Determine the context for when the sentence was updated or deleted in the origin entry, then determine the context of the sentence in the secondary entry at the time when the sentence was deleted from the origin entry.

5. Is the update in the origin sentence relevant to the secondary entry (Figure 6.20, Question 3)?

6. Does the update in the origin entry affect the accuracy of the secondary entry (Figure 6.20, Question 4)?

The definition of the classification protocol aims to reduce inconsistencies in the classification of a sentence and encourage reproducible results. Having a process by which

to classify sentences means that the identified sentences can now be analysed.

6.5.3 Protocol application

Applying the protocol to a sentence can lead to one of five possible classifications. To illustrate the application of the classification protocol, a sentence classified as erroneous is provided as a worked example, showing the outcome for each stage of the classification process.

Additional illustrations, covering each of the remaining four possible classifications, are also provided; however these abstract from the overall protocol, only providing detail relevant for their classification.

6.5.4 Protocol application: Erroneous

The sentence "may have an essential function in lipopolysaccharides biosynthesis.", as shown in Figure 6.21, is classified as erroneous. For each numbered step of the protocol, the determined outcome with associated evidence is explained below:

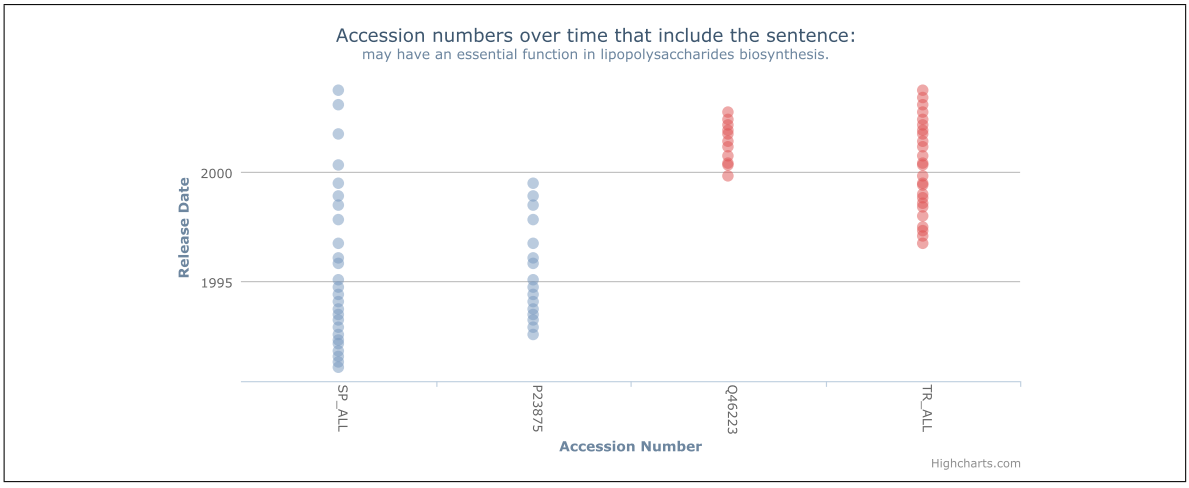


Figure 6.21: Visualising the propagation of the sentence "may have an essential function in lipopolysaccharides biosynthesis.", which was classified as erroneous.

1. The sentence appears in a total of two entries. As it appears in less than 100 entries, it is feasible to analyse.

2. The sentence originates in a single entry (P23875) and propagates to a single secondary entry (Q46223).

3. The context of the sentence in the origin entry when the sentence was added to the secondary entry is shown in Figure 6.22. The context of the sentence when added to the secondary entry Q46223 is shown in Figure 6.23.

   At this point, there is significant overlap between the two entries. For example, description (identified by the "DE" line) and pathway information contained the origin entry is also added to the secondary entry (Figure 6.23 Lines 4 and 17). This suggests that the sentence was propagated between the two entries.

4. The removal of the sentence from the origin entry, along with other changes to the entry, is shown in Figure 6.24. The secondary entry remains unchanged during this period.

5. As no updates were made to the secondary entry, then it appears that the removal of the sentence should also have been applied to the secondary entry.

6. All information relating to lipopolysaccharides is removed from the origin entry with a cautionary topic stating that it was initially believed to have a function in lipopolysaccharides biosynthesis being introduced (Figure 6.24, lines 45–46). This questions the accuracy of the secondary entry and the sentence is therefore classified as erroneous.

The sentence is eventually removed from the secondary entry, along with all other comments, after TrEMBL Version 22, as shown in Figure 6.25. References to lipopolysaccharides biosynthesis were also removed from the keyword list; the only reference to lipopolysaccharides biosynthesis remaining within the entry is in the title of a referenced article. This provides further confidence that the sentence was erroneous in the secondary entry and that it could have been removed ten database releases (three years) earlier.

6.5.5 Protocol application: Too many results

Figure 6.26 shows the visualisation for the sentence "contains 1 immunoglobulin-like v-type domain". The sentence occurs in a total of $1,603$ primary accessions ($3,048$ when including secondary accessions) and originates in a total of 62 primary accessions (79

```
ID   KDTB_ECOLI     STANDARD;      PRT;   159 AA.
AC   P23875;
DT   01-NOV-1991 (Rel. 20, Created)
DT   01-NOV-1991 (Rel. 20, Last sequence update)
DT   01-NOV-1997 (Rel. 35, Last annotation update)
DE   LIPOPOLYSACCHARIDE CORE BIOSYNTHESIS PROTEIN KDTB.
GN   KDTB.
OS   Escherichia coli.
OC   Bacteria; Proteobacteria; gamma subdivision; Enterobacteriaceae;
OC   Escherichia.
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=K12;
RX   MEDLINE; 91236744.
RA   CLEMENTZ T., RAETZ C.R.H.;
RT   "A gene coding for 3-deoxy-D-manno-octulosonic-acid transferase in
RT   Escherichia coli. Identification, mapping, cloning, and sequencing.";
RL   J. Biol. Chem. 266:9687-9696(1991).
RN   [2]
RP   SEQUENCE FROM N.A.
RC   STRAIN=K12 / MG1655;
RX   MEDLINE; 94316500.
RA   SOFIA H.J., BURLAND V., DANIELS D.L., PLUNKETT G. III, BLATTNER F.R.;
RT   "Analysis of the Escherichia coli genome. V. DNA sequence of the
RT   region from 76.0 to 81.5 minutes.";
RL   Nucleic Acids Res. 22:2576-2586(1994).
RN   [3]
RP   CHARACTERIZATION.
RC   STRAIN=K12;
RX   MEDLINE; 92250420.
RA   RONCERO C., CASADABAN M.J.;
RT   "Genetic analysis of the genes involved in synthesis of the
RT   lipopolysaccharide core in Escherichia coli K-12: three operons in
RT   the rfa locus.";
RL   J. Bacteriol. 174:3250-3260(1992).
CC   -!- FUNCTION: MAY HAVE AN ESSENTIAL FUNCTION IN LIPOPOLYSACCHARIDES
CC       BIOSYNTHESIS.
CC   -!- PATHWAY: LIPOPOLYSACCHARIDE CORE BIOSYNTHESIS.
DR   EMBL; M60670; AAA24044.1; -.
DR   EMBL; M86305; AAA03746.1; -.
DR   EMBL; U00039; AAB18611.1; -.
DR   EMBL; AE000441; AAC76658.1; -.
DR   PIR; JU0468; JU0468.
DR   PIR; S27562; S27562.
DR   ECOGENE; EG11190; KDTB.
DR   PFAM; PF01467; Cytidylyltransf; 1.
KW   Lipopolysaccharide biosynthesis.
```

Figure 6.22: Flat file view for Version 38 of Swiss-Prot entry P23875. Copyright and Sequence information has been removed.

```
...
1   - DT   01-AUG-1998 (TrEMBLrel.  07, Last annotation update)
2   - DE   KDO-TRANSFERASE.
3   + DT   01-NOV-1999 (TrEMBLrel.  12, Last annotation update)
4   + DE   LIPOPOLYSACCHARIDE CORE BIOSYNTHESIS PROTEIN KDTB.
...
5   - RA   GIRJES A.A.;
6   - RL   Submitted (APR-1994) to the EMBL/GenBank/DDBJ databases.
7   - RN   [3]
8   - RP   SEQUENCE FROM N.A.
9   - RC   STRAIN=KOALA TYPE I;
10  - RA   GLASSICK T., GIFFARD P., TIMMS P.;
11  - RL   Syst.  Appl.  Microbiol.  19:457-464(1996).
12  - RN   [4]
13  - RP   SEQUENCE FROM N.A.
14  - RC   STRAIN=KOALA TYPE I;
...
15  + CC   -!- FUNCTION: MAY HAVE AN ESSENTIAL FUNCTION IN LIPOPOLYSACCHARIDES
16  + CC       BIOSYNTHESIS.
17  + CC   -!- PATHWAY: LIPOPOLYSACCHARIDE CORE BIOSYNTHESIS.
...
18  - KW   Transferase.
19  + KW   Lipopolysaccharide biosynthesis.
...
```

Figure 6.23: UniSave view for entry Q46223, showing the differences between TrEMBL Versions 10 and 12. There was no change to the entry in TrEMBL Version 11.

when including secondary accessions). As it appears in over 100 entries, it is classified as "too many results".

With the sentence appearing in a large number of accessions, any classification would prove difficult. For example, the context of the sentence is unlikely to be consistent between all of the origin entries. Additionally, the visualisation becomes heavily populated and dense. Although this can be alleviated with the interactive features, such as zooming, it becomes cumbersome to navigate, unlike those graphs showing less than 100 entries.

6.5.6 Protocol application: Possibly erroneous

The sentence "ring cleavage of 2,3-dihydroxybiphenyl.", as shown in Figure 6.27, is removed from the origin entry[15] in Swiss-Prot Version 11 and appears in the secondary entries approximately ten years later[16]. Whilst the context between these two entries is deemed similar (the same cofactor information is also added, for example), the large gap between the releases makes it highly doubtful that the sentence was copied between these entries. The sentence is therefore classified as possibly erroneous.

[15] http://www.uniprot.org/uniprot/P08695?version=3&version=2
[16] http://www.uniprot.org/uniprot/P72325?version=4&version=2

```
1   - ID   KDTB_ECOLI STANDARD; PRT; 159 AA.
2   + ID   COAD_ECOLI STANDARD; PRT; 159 AA.
...
3   - DT   01-NOV-1997 (Rel.  35, Last annotation update)
4   - DE   LIPOPOLYSACCHARIDE CORE BIOSYNTHESIS PROTEIN KDTB.
5   - GN   KDTB.
6   + DT   30-MAY-2000 (Rel.  39, Last annotation update)
7   + DE   PHOSPHOPANTETHEINE ADENYLYLTRANSFERASE (EC 2.7.7.3) (PANTETHEINE-
8   + DE   PHOSPHATE ADENYLYLTRANSFERASE) (PPAT) (DEPHOSPHO-COA
9   + DE   PYROPHOSPHORYLASE).
10  + GN   COAD OR KDTB.
...
11  - RA   CLEMENTZ T., RAETZ C.R.H.;
12  + RA   Clementz T., Raetz C.R.H.;
...
13  - RA   SOFIA H.J., BURLAND V., DANIELS D.L., PLUNKETT G. III, BLATTNER F.R.;
14  + RA   Sofia H.J., Burland V., Daniels D.L., Plunkett G. III, Blattner F.R.;
...
15  - RP   CHARACTERIZATION.
16  + RP   GENETIC CHARACTERIZATION.
...
17  - RA   RONCERO C., CASADABAN M.J.;
18  + RA   Roncero C., Casadaban M.J.;
...
19  - CC   -!- FUNCTION: MAY HAVE AN ESSENTIAL FUNCTION IN LIPOPOLYSACCHARIDES
20  - CC       BIOSYNTHESIS.
21  - CC   -!- PATHWAY: LIPOPOLYSACCHARIDE CORE BIOSYNTHESIS.
22  + RN   [4]
23  + RP   SEQUENCE OF 1-10, AND CHARACTERIZATION.
24  + RX   MEDLINE; 99410451.
25  + RA   Geerlof A., Lewendon A., Shaw W.V.;
26  + RT   "Purification and characterization of phosphopantetheine
27  + RT   adenylyltransferase from Escherichia coli.";
28  + RL   J. Biol.  Chem.  274:27105-27111(1999).
29  + RN   [5]
30  + RP   X-RAY CRYSTALLOGRAPHY (1.8 ANGSTROMS).
31  + RX   MEDLINE; 99221637.
32  + RA   Izard T., Geerlof A.;
33  + RT   "The crystal structure of a novel bacterial adenylyltransferase
34  + RT   reveals half of sites reactivity.";
35  + RL   EMBO J. 18:2021-2030(1999).
36  + CC   -!- FUNCTION: REVERSIBLY TRANSFERS AN ADENYLYL GROUP FROM ATP TO 4'-
37  + CC       PHOSPHOPANTETHEINE, YIELDING DEPHOSPHO-COA (DPCOA) AND
38  + CC       PYROPHOSPHATE.
39  + CC   -!- CATALYTIC ACTIVITY: ATP + PANTETHEINE 4'-PHOSPHATE = DIPHOSPHATE +
40  + CC       DEPHOSPHO-COA.
41  + CC   -!- PATHWAY: FOURTH STEP IN COENZYME A (COA) BIOSYNTHESIS.
42  + CC   -!- SUBUNIT: HOMOHEXAMER.
43  + CC   -!- SUBCELLULAR LOCATION: CYTOPLASMIC.
44  + CC   -!- SIMILARITY: BELONGS TO THE COAD FAMILY.
45  + CC   -!- CAUTION: WAS ORIGINALLY THOUGHT TO HAVE AN ESSENTIAL FUNCTION IN
46  + CC       LIPOPOLYSACCHARIDES BIOSYNTHESIS.
...
47  + DR   PDB; 1B6T; 19-APR-00.
...
48  + DR   INTERPRO; IPR001980; -.
49  + DR   INTERPRO; IPR001994; -.
...
50  - KW   Lipopolysaccharide biosynthesis.
51  - SQ   SEQUENCE 159 AA; 17837 MW; E6F8F948 CRC32;
52  + DR   PRINTS; PR01020; LPSBIOSNTHSS.
53  + KW   Transferase; Nucleotidyltransferase; Coenzyme A biosynthesis;
54  + KW   3D-structure.
55  + SQ   SEQUENCE 159 AA; 17837 MW; C4D7B8715A061B91 CRC64;
...
```

Figure 6.24: UniSave view showing the differences in entry P23875 between Swiss-Prot Versions 38 and 39.

```
1   - DT   01-MAY-2000 (TrEMBLrel.  13, Last annotation update)
2   - DE   Lipopolysaccharide core biosynthesis protein KDTB.
3   - OS   Chlamydia psittaci (Chlamydophila psittaci).
4   + DT   01-MAR-2003 (TrEMBLrel.  23, Last annotation update)
5   + DE   KDO-transferase.
6   + GN   ORF2.
7   + OS   Chlamydia pneumoniae (Chlamydophila pneumoniae).
...
8   - OX   NCBI_TaxID=83554;
9   + OX   NCBI_TaxID=83558;
...
10  - RC   STRAIN=KOALA TYPE I;
11  + RC   STRAIN=Koala type I;
12  + RA   Girjes A.A.;
13  + RL   Submitted (APR-1994) to the EMBL/GenBank/DDBJ databases.
14  + RN   [3]
15  + RP   SEQUENCE FROM N.A.
16  + RC   STRAIN=Koala type I;
17  + RA   Glassick T., Giffard P., Timms P.;
18  + RT   "Outer-membrane protein-2 gene-sequences indicate that Chlamydia-
19  + RT   pecorum and Chlamydia-pneomoniae cause infections in Koalas.";
20  + RL   Syst.  Appl.  Microbiol.  19:457-464(1996).
21  + RN   [4]
22  + RP   SEQUENCE FROM N.A.
23  + RC   STRAIN=Koala type I;
...
24  - CC   -!- FUNCTION: MAY HAVE AN ESSENTIAL FUNCTION IN LIPOPOLYSACCHARIDES
25  - CC       BIOSYNTHESIS.
26  - CC   -!- PATHWAY: LIPOPOLYSACCHARIDE CORE BIOSYNTHESIS.
...
27  - KW   Lipopolysaccharide biosynthesis.
28  + KW   Transferase.
...
```

Figure 6.25: UniSave view for entry Q46223, showing the differences between TrEMBL Versions 22 and 23.

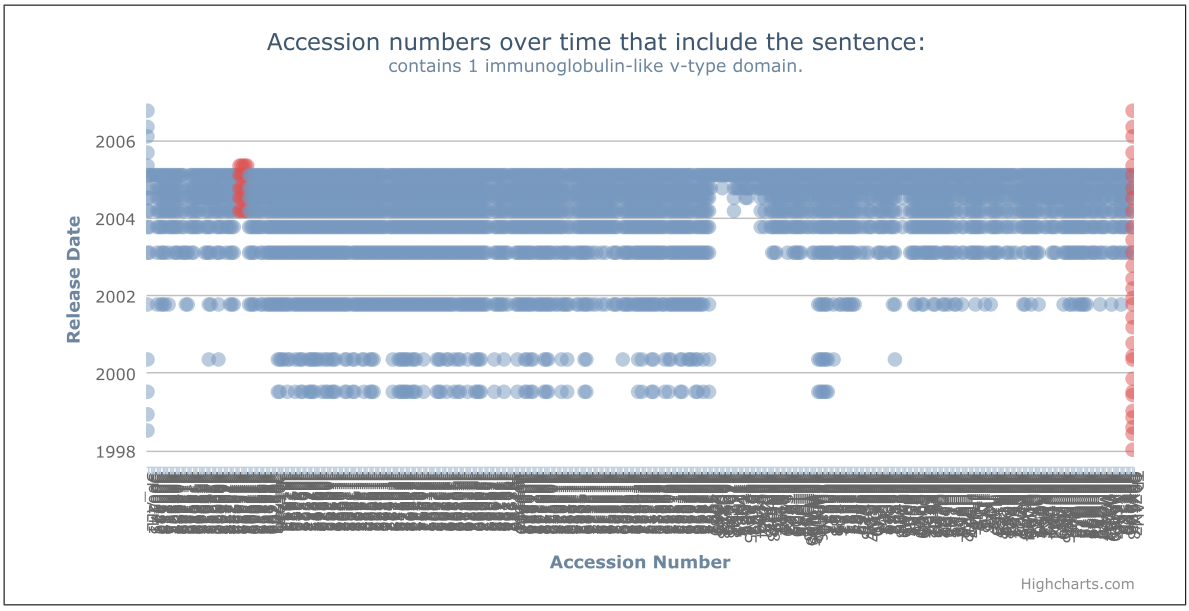


Figure 6.26: Visualising the propagation of the sentence “contains 1 immunoglobulin-like v-type domain”, which was classified as having too many results.

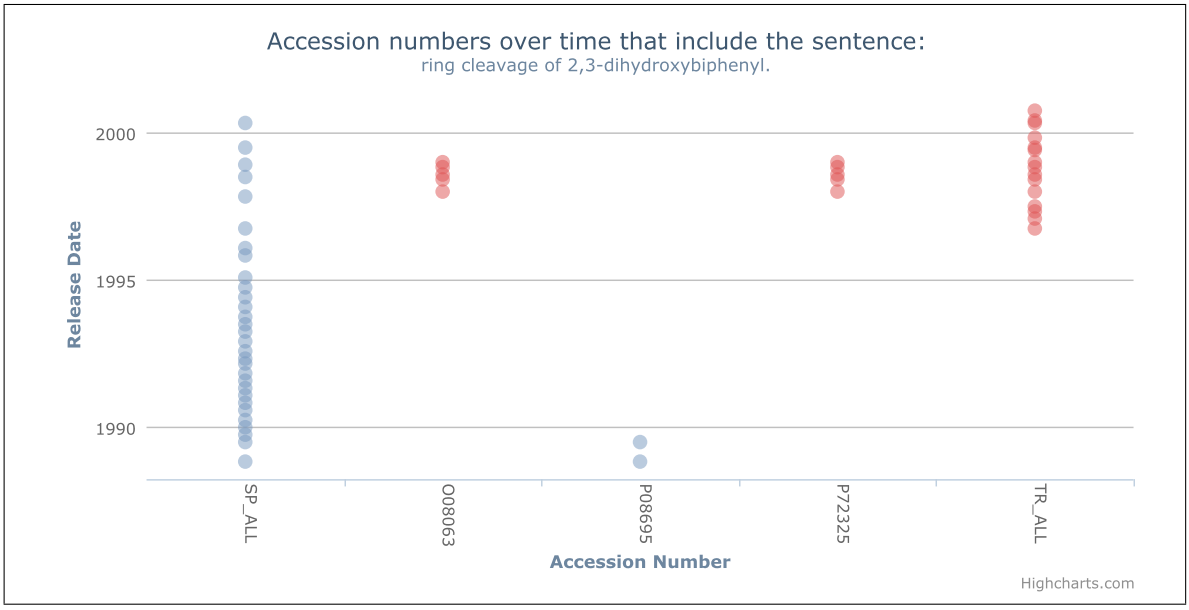


Figure 6.27: Visualising the propagation of the sentence "phosphorylates ppp1r12a.", which was classified as possibly erroneous.

6.5.7 Protocol application: Accurate

The sentence "involved in tumorigenesis.", originates in a single entry and remains in a secondary entry in UniProtKB Version 2012_05, as shown in Figure 6.28. The sentence was contained, and removed, from the disease topic block in the origin entry[17], whilst the sentence was added to the function topic block in the secondary entry[18]. Additionally, the secondary entry relates to the organism *Rhizobium radiobacter* (*Agrobacterium tumefaciens*) and cites a paper titled "An Agrobacterium catalase is a virulence factor involved in tumorigenesis"; the origin entry relates to *Homo sapiens* and has no citations in common. This suggests that the sentence was not copied between these entries, and was added independently. Therefore, the sentence is classified as "accurate".

6.5.8 Protocol application: Inconsistent

The sentence "bind preferentially single-stranded dna and unwind double stranded dna.", as shown in Figure 6.29, originates in seven entries and propagates to a single secondary entry. Within these origin entries, the sentence was altered to grammatically correct the first word; "bind" was replaced with "binds"[19]. This does not affect

[17] `http://www.uniprot.org/uniprot/P35125?version=84&version=83`
[18] `http://www.uniprot.org/uniprot/Q9R708?version=44&version=39`
[19] `http://www.uniprot.org/uniprot/P17741?version=7&version=6`

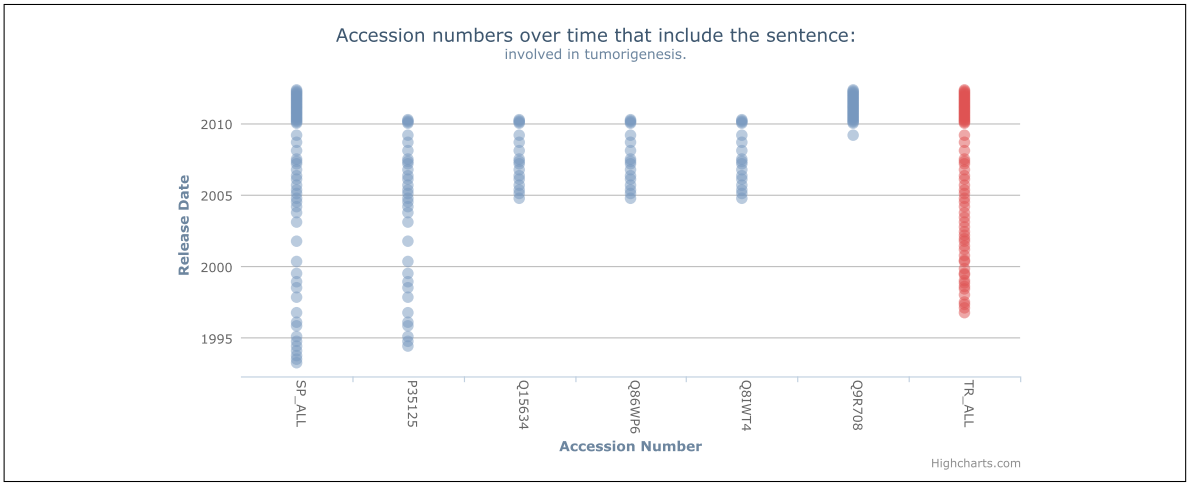


Figure 6.28: Visualising the propagation of the sentence "involved in tumorigenesis.", which was classified as accurate.

the biological knowledge of the annotation in the secondary entry, so it is therefore classified as inconsistent.

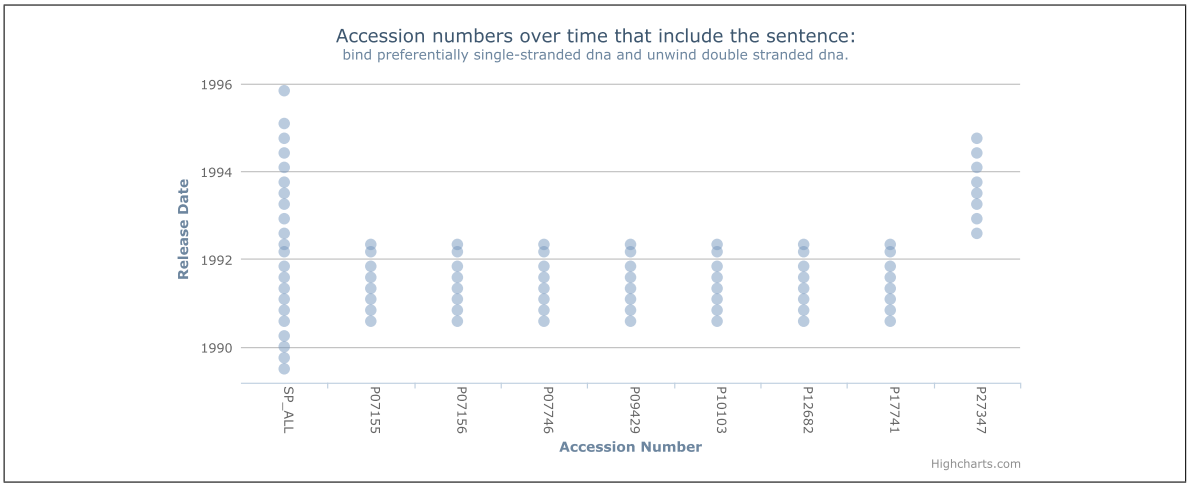


Figure 6.29: Visualising the propagation of the sentence "bind preferentially single-stranded dna and unwind double stranded dna.", which was classified as inconsistent.

The sentence is eventually removed from the secondary entry after seven versions[20], when the sentence is removed entirely as the entry undergoes significant updates and changes.

[20] `http://www.uniprot.org/uniprot/P27347?version=6&version=5`

6.5.9 Protocol application: Results

The analysis and classification of a sentence is labour intensive. To maximise the number of sentences analysed, the analysis was performed in conjunction with Matthew Collison, a domain expert[21]. This collaboration also aided with the testing and refinement of the developed protocol.

A subset of the $8,355$ missing origin sentences were analysed; sentences were sorted by length with every hundredth sentence being extracted. This approach aimed to remove sentence length bias, based on the assumption that longer sentences are more likely to have a greater information content and have been propagated. Given this assumption, sentences with 20 characters or less were discarded. This resulted in 65 sentences being extracted, with the classification results of these sentences summarised in Table 6.3.

Classification	Erroneous	Inconsistent	Accurate	Too Many Results	Possibly Erroneous
Absolute	16	11	20	5	13
Percentage	24.6%	16.9%	30.8%	7.7%	20.0%
Potentially Erroneous	$2,057$	$1,414$	$2,571$	643	$1,671$

Table 6.3: The classification results of the 65 sentences analysed, controlling for sentence length bias (i.e. every 100th sentence over 20 characters in length).

A total of 27 sentences were identified as erroneous or inconsistent; approximately 42% of the analysed sentences. The sentences classified as inconsistent were mostly due to grammatical inconsistencies and were often corrected in the secondary entries after a number of versions.

The number of sentences classified as "possibly erroneous" is similar to the number of "inconsistent" sentences. The thirteen "possibly erroneous" sentences mostly arose when trying to determine if a sentence was propagated between entries. These results suggest that the curation process is asynchronized and that these inconsistencies could be overcome, or substantially reduced, if formal provenance were available. Only a small number of sentences were deemed infeasible to analyse, whilst almost a third of sentences were classified as "accurate".

[21]Matthew Collison is an EngD student, studying the role of the gut microbiome in health and disease. Matthew holds degrees in Neuroscience and Physiological sciences.

Of the 65 analysed sentences, almost half remain in UniProtKB Version 2012_05. This subset consists of 32 sentences, and is summarised in Table 6.4. Nine sentences that remain in UniProtKB Version 2012_05 are classified as either erroneous or inconsistent.

Classification	Erroneous	Inconsistent	Accurate	Too Many Results	Possibly Erroneous
Absolute	4	5	12	1	10
Percentage	12.5%	15.6%	37.5%	3.1%	31.3%
Potentially Erroneous	479	599	1,438	120	1,198

Table 6.4: The classification of results for the subset of sentences analysed, controlling for sentence length bias, that remain in UniProtKB Version 2012_05.

In addition to the subset of sentences based on sentence length, an additional 57 sentences were analysed during the development and refinement of the protocol. This analysis also included sentences under 20 characters long. In total, 122 sentences were analysed; approximately 1.5% of the 8,355 identified sentences. These results are summarised in Table 6.5.

Classification	Erroneous	Inconsistent	Accurate	Too Many Results	Possibly Erroneous
Absolute	36	29	28	15	14
Percentage	29.5%	23.8%	23.0%	12.3%	11.5%
Potentially Erroneous	2,465	1,986	1,918	1,027	959

Table 6.5: The classification results for all of the analysed sentences (122 in total).

Although a significant percentage of sentences are accurate, these results show that the missing origin pattern can be used to detect erroneous and inconsistent annotation. The complete set of analysed sentences, with classifications, is provided in Appendix A.

To evaluate these classifications, the UniProtKB help desk were contacted with a detailed breakdown of three sentences. The three sentences chosen were all classified as erroneous, with two being historical and one remaining in the latest database version. For the two historical sentences, the help desk confirmed that if the sentence was to be re-added to the entry it would now be considered incorrect. Information relating to the final sentence, in the UniProtKB Version 2012_05 database, was deemed to not be rich enough to determine whether the sentence is accurately contained within the secondary entry. However, this analysis raises a sensible question, that should be addressed as knowledge increases.

6.6 Discussion

The exponential growth of biological data has resulted in an inevitable reliance on automated methods for the production of textual annotation. These methods involve the propagation of annotation between database entries; sentences are effectively copied between entries as a matter of protocol. This process can see sections, or sometimes whole annotations, from one entry being copied to other entries without change.

Analysing this reuse in UniProtKB showed an increase in both Swiss-Prot and TrEMBL, resulting in the annotation corpus becoming more replicated; over time, the average number of entries each sentence appears in is increasing. For example, in UniProtKB Version 2012_05, the percentage of unique sentences is $< 7\%$ for Swiss-Prot and $< 0.03\%$ for TrEMBL. The reuse of knowledge is not just restricted to textual annotation, with similar patterns being identified in high-throughput experiments; many experiments are based on a very small amount of experimental data [322]. These patterns are likely to continue as manual curation remains a labour intensive bottleneck.

Whilst high levels of reuse are expected in automated methods, manual curation is also showing increased levels of reuse. This is due, in part, to annotations becoming standardised and being used to enforce levels of quality control. As UniProtKB matures, sentences are increasingly following the form of nanopublications [323], where each sentence contains an independent segment of biological knowledge. Sentences structured in this manner allow, and encourage, the reuse of sentences in a manner similar to ontologies. This structure has enabled the reduction in unannotated entries in UniProtKB, whilst also increasing the average number of sentences within textual annotation.

However, unlike ontologies, changes to a sentence are independent and will not percolate automatically. Therefore, erroneous annotations can propagate within a single database and, potentially, to external databases. For example, an annotation regarding acetylchloine had incorrectly spread throughout the Biomolecular Interaction Network Database (BIND) database [324], whilst annotations regarding the Histone arginine demethylase JMJD6 protein were found to be incorrect within the UniProtKB database [319]. Whilst these may not be errors in the annotation, but in the under-

lying source of the annotation, the denormalised nature of annotation means not all occurrences may be corrected; it is for this reason that provenance should be made clear to a user.

Utilising VIPeR, developed in Chapter 4, the provenance of an annotation can be realised for any individual sentence. Identifying provenance was only achievable given that UniProtKB make available all major historical versions of Swiss-Prot and TrEMBL. Users are typically only interested in the most recent and up-to-date biological data available, yet this work highlights the added value and importance of being able to scour archival data; database features such as UniSave should be a requirement rather than a luxury.

Provenance is inferred by identifying the first UniProtKB entry that a sentence appears in, with all subsequent entries representing the sentences' propagation. For individual sentences, this inference is not necessarily accurate. For example, a sentence may originate in an entry outside of UniProtKB or within a minor release. Further, the appearance of a sentence in multiple entries may be an independent event, with no relationship between the entries. However, the curation process and levels of reuse identified would argue against this often being the case. More formal tracking of provenance within the database curation process would help to alleviate this difficulty.

VIPeR also appears beneficial for the identification of propagation patterns, which hold promise as quality and correctness indicators. For example, a sentence which adheres to the "reappearing entry" pattern could be considered more dubious, as its inclusion (or exclusion) within an entry is not definitive. These patterns were identified through manual inspection of graphs when analysing sentence provenance. Further work could be undertaken to perform a comprehensive search to identify any additional propagation patterns.

Analysing sentences adhering to the missing origin pattern resulted in a number of erroneous annotations being identified, including some that remain in UniProtKB Version 2012_05. As acknowledged earlier, these results are somewhat subjective. Therefore, the UniProt help desk checked our conclusions for three cases; in two cases these were correct, and in the third they claim that there is a lack of biological knowledge to draw a definitive conclusion. These results suggest that propagation patterns could aid in

the discovery of erroneous annotations, and act as a mechanism to increase confidence into an annotation's quality.

The structure and features of UniProtKB made it an ideal resource to perform this analysis. A clear extension is to apply VIPeR to other databases, allowing the propagation and provenance to be identified. As previously discussed, it is plausible that annotations propagate *between* databases. For example, the InterPro database is used in the production of TrEMBL [325], whilst the neXtProt database integrates the annotation in UniProtKB/Swiss-Prot as a primary source, as well as incorporating data from a number of other sources such as The Gene Ontology (GO) and Ensembl [70]. With over $1,500$ active biological databases, if cross-database propagation does indeed occur, then the provenance map could be vast, and using this approach it is plausible that the "true" provenance and propagation of an annotation could be identified. VIPeR was developed in a manner that will allow any textual resource to be compared, which is explored in the following section (Section 7.3).

7

Provenance, Propagation and Quality of Annotations in Biological Databases

Contents

Introduction

In previous sections of this thesis two approaches were developed that allow textual annotations to be explored. Specifically, these approaches are a quality metric (QUALM) which is based on word distribution (Chapter 3), and a visualisation tool (VIPeR) that allows the provenance and propagation of an annotation to be inferred (Chapter 5). To assess their suitability, both approaches were applied to The UniProt Knowledgebase (UniProtKB) (Chapters 4 and 6).

The analysis performed on UniProtKB suggests that both approaches hold value. Both approaches rely solely on simple text analyses of annotations and should, therefore, be reasonably generic and applicable to any biological database containing significant amounts of textual annotation. Within this chapter we extend the analysis of both approaches to the textual annotation in a variety of biological databases. This analysis will also allow the generality of the approaches to be evaluated.

The previous analyses performed on UniProtKB were thorough and required a detailed understanding of the database. This was necessary to establish the value of the measures used, due to the lack of an explicit gold standard dataset. In this chapter, we present an initial and shallower analysis of a number of additional biological databases. This also allows us to investigate the propagation of sentences between databases as well as within them.

Prior to this analysis we identify a number of suitable databases (Section 7.1). For each of the identified databases, we apply QUALM to their textual annotation and present a selection of the power-law graphs, including a single graph showing the α value for each database version over time (Section 7.2). Following this, the provenance and propagation of each database is analysed, including an analysis of cross-database propagation (Section 7.3). The chapter then concludes with a discussion of the results from these two analyses (Section 7.4).

7.1 Identifying Biological Databases

As previously discussed in Section 2.1 there are over $1,500$ active biological databases covering a variety of areas and specialisms. Therefore, we need to identify a manageable subset of suitable databases for further analysis.

Databases which do not make available historical versions were excluded from consideration, as historical data is required for analysing annotation provenance and propagation. Similarly, databases which contain only minimal amounts of textual annotation are not considered, as meaningful conclusions cannot be drawn from QUALM when applied to small corpora of words.

While a suitable database will provide both adequate amounts of textual annotation and archived versions, the annotation should be presented in a format that can be parsed with relative ease. Specifically, data should be formatted consistently across all historical versions, with the ability to obtain data in bulk.

Although a number of databases fulfil these requirements, the neXtProt, InterPro, PRINTS, TIGRFAMs and PROSITE databases were chosen for further analysis. These databases were chosen as they have dependencies on other databases or are utilised by external databases (or both). By choosing these databases, we also increase the likelihood of cross-database propagation and analyse databases of different levels of maturity. A brief overview of each of these databases is provided below:

neXtProt

neXtProt [70] is a relatively new database that is maintained at the Swiss Institute of Bioinformatics (SIB). Initiated in 2011, the sole focus of neXtProt is on human proteins, with the aim of being the central hub for all human protein information. To achieve this aim, the database incorporates data from various sources and is built as a participative platform to leverage knowledge from the scientific community; the core corpus of neXtProt is based on human proteins from Swiss-Prot, whilst the integration of data is often done in collaboration with groups identified as having the relevant expertise.

As the name suggests, neXtProt has a number of similarities to Swiss-Prot. One such similarity is the aim of containing only high-quality information. However, unlike many databases, neXtProt provides a classification system that categorises data based on its quality into gold, silver or bronze. The evidence used to determine the category is documented in the metadata of the database entry.

All historical releases of neXtProt are made available on its FTP server[1].

PROSITE

PROSITE [326], like Swiss-Prot, was developed by Amos Bairoch at the SIB. The PROSITE database consists of sequence patterns, or motifs, that are conserved in protein sequences and can be used to help infer information about a sequence, such as which protein family it belongs to and its possible function.

PROSITE is composed of two flat files: "PROSITE.DAT", which is the data file containing protein patterns; and "PROSITE.DOC", which contains associated documentation for each protein pattern. Each PROSITE entry is assigned a unique identifier, and contains a pointer to the relevant documentation entry, which provides biological information that can be inferred by the pattern. This separation means that only the documentation file is required for our analysis.

Whilst the first release of PROSITE was in 1989, the earliest archived version available on its FTP server[2] is Version 8, which was released in 1991. Between 1991 and 2010, there was a total of 12 releases, with Version 20 being released in late 2010. Following Version 20, the frequency of PROSITE releases increased resulting in a total of 89 archived versions being available for download.

PRINTS

PRINTS [327] is a database that was created in 1991 by Teresa Attwood and is currently maintained at Manchester University. PRINTS, like PROSITE, is a database which contains sequence motifs. However, entries in PRINTS are known as fingerprints, as they are composed of multiple motifs, unlike entries in PROSITE which contain only single motifs. All PRINTS entries are manually

[1] `ftp://ftp.nextprot.org/pub`
[2] `ftp://ftp.expasy.org/databases/prosite/`

curated and provide cross-references to the equivalent PROSITE entries, if they exist.

Although there have been almost 40 releases of PRINTS, only 16 archived releases are available on its FTP server[3].

TIGRFAMs

The TIGRFAMs [328] database, first released in 2001, is a collection of protein families which are designed to assist with the prediction of protein function. Each protein family entry is described using manually curated multiple sequence alignments and Hidden Markov Models (HMMs). Each TIGRFAMs entry also contains a textual annotation section with additional supporting information, such as GO annotations and references to relevant Pfam and InterPro entries.

TIGRFAMs is maintained and hosted by the J. Craig Venter Institute and has a total of 13 releases, which are available on its FTP[4] server.

InterPro

InterPro [329] is an integrative database developed and maintained at the European Bioinformatics Institute (EBI). The database integrates information regarding protein families, domains and functional sites from eleven member databases, including PROSITE, PRINTS and TIGRFAMs. Each InterPro entry contains a description, or abstract, which is often supplemented with references to relevant literature.

The first release of InterPro was in 2000, with a further 35 versions having since been released. All historical versions are archived on the InterPro FTP[5] server.

In order to analyse these databases the correct extraction of sentences and words from the textual annotation is required. As previously discussed in Section 4.1, BANE was developed to extract both sentences and words from UniProtKB annotation. Therefore, BANE was extended to handle the extraction of annotation from these five additional databases.

[3] `ftp://ftp.bioinf.man.ac.uk/pub/prints`
[4] `ftp://ftp.jcvi.org/pub/data/TIGRFAMs/`
[5] `ftp://ftp.ebi.ac.uk/pub/databases/interpro/`

Like UniProtKB, the PRINTS, TIGRFAMs and PROSITE databases provide entries in flat file format. Therefore, for these databases BANE only had to be extended to handle the different line formats. BANE had to be extended further to handle the neXtProt and InterPro databases as they are made available in XML format. To gain confidence that these extensions correctly handle the extraction of data from each database, the tests previously discussed in Section 4.1 were re-performed, while a random subset of entries for each database were manually checked against the parsed outputs.

7.2 Analysing Annotation Quality

Previously, in Chapter 3, we applied QUALM to textual annotation in UniProtKB in order to evaluate its suitability. Although limitations with QUALM were identified, this analysis suggested that QUALM holds promise as an indicator of annotation quality. Within this section we extend the analysis to textual annotation in the neXtProt, InterPro, PRINTS, TIGRFAMs, and PROSITE databases. The analysis of these additional databases will further test the suitability of the power-laws model fitting and the usability of α as a measure of quality.

We start by initially analysing the oldest of these five databases, PROSITE.

7.2.1 PROSITE

Out of the five databases PROSITE has the most archived versions available. The α values derived for each of the 89 archived versions are shown in Figure 7.1, while Figure 7.2 shows the underlying power-law graphs for four PROSITE versions.

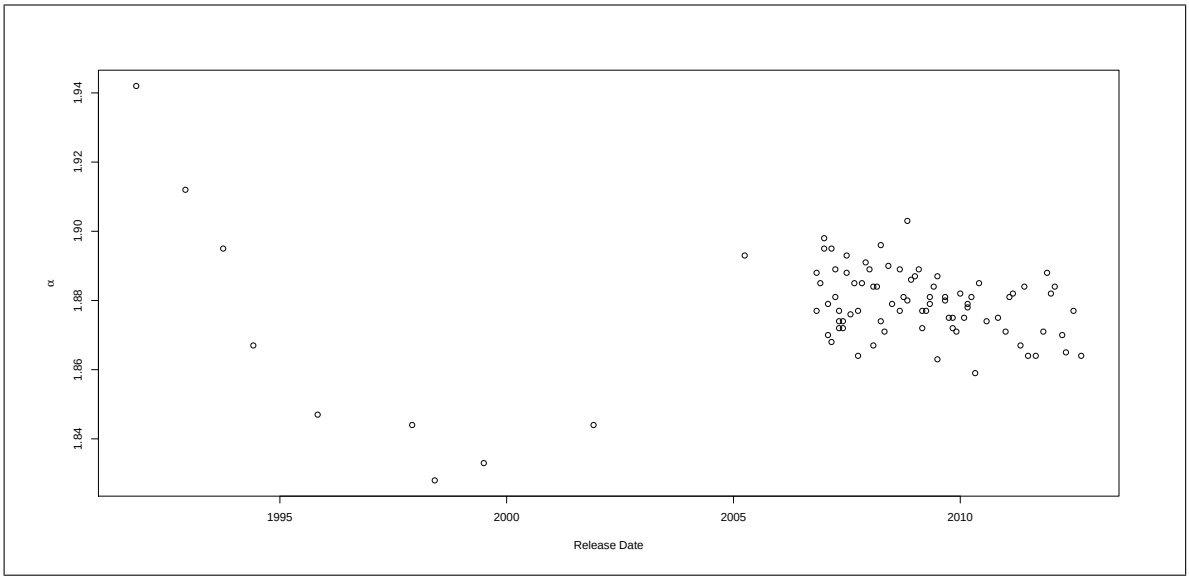
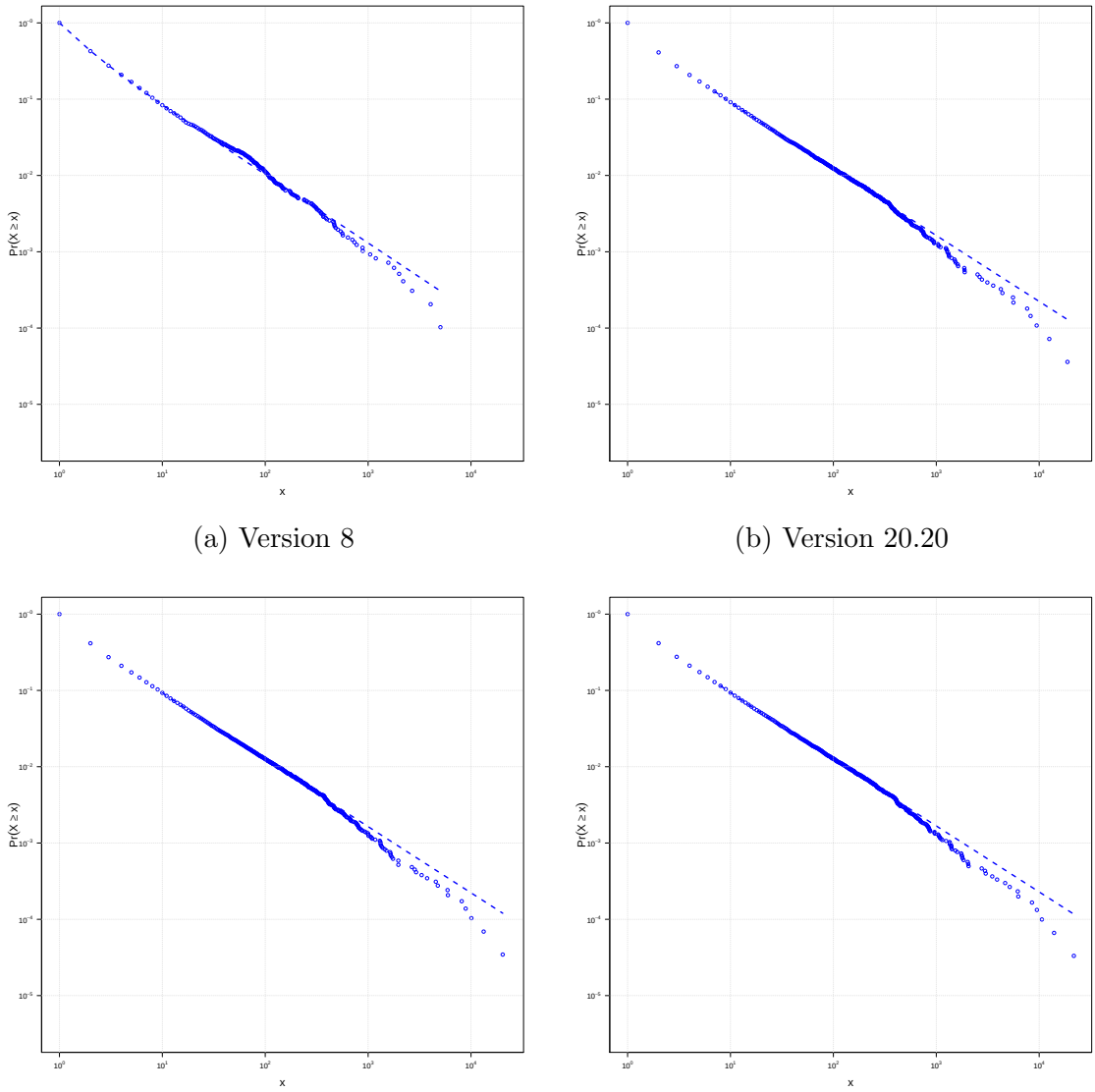


Figure 7.1: α for each archived version of PROSITE over time.

Figure 7.1 suggests that there has been only minor differences in α over the history of PROSITE. The highest α value obtained was for the first archived version of PROSITE (Version 8; $\alpha \approx 1.94$), while the lowest α was seen ten years later (Version 14; $\alpha \approx 1.83$). Following this release the α values increase until the change in the PROSITE



(a) Version 8

(b) Version 20.20

(c) Version 20.60

(d) Version 20.85

Figure 7.2: The power-law model applied to four versions of PROSITE.

release cycle, which results in numerous small fluctuations between versions, which have an average $\alpha \approx 1.88$.

Overall the α values only vary slightly between PROSITE versions, with the difference between the most extreme values being ~ 0.1. This is also reflected in the four power-law graphs, shown in Figure 7.2, with the slope and behaviour of the graph remaining proportional over time, even though the amount of annotation is increasing; over $83,000$ words ($\sim 10,000$ unique) are contained in PROSITE Version 8, whilst PROSITE Version 20.85 contains over $350,000$ words ($\sim 30,000$ unique).

Compared to most sequence databases, the growth of PROSITE has been relatively modest. The most recent version (20.85) of PROSITE contains just over $1,650$ documentation entries, which is over $1,100$ more documentation entries than the first archived version, which contained only 530 documentation entries. Since the change of release cycle in 2006, PROSITE has been averaging a total of three new documentation entries per release.

The documentation entries in PROSITE contain a substantial amount of textual annotation. For example, Version 8 averaged 157 words per entry, whilst Version 20.85 averages 213 words per documentation entry. This is over four times the amount of textual annotation contained within an average Swiss-Prot entry in UniProtKB/Swiss-Prot Version 2012_05. Additionally, PROSITE documentation entries are presented as descriptions, similar to an abstract from an academic paper, rather than being formatted into specific topic blocks like Swiss-Prot annotation.

7.2.2 PRINTS

PRINTS is the second oldest database we have chosen to analyse and has a total of 16 archived versions available. The α values obtained for each archived version are shown in Figure 7.3, with the power-law graph for four versions shown in Figure 7.4.

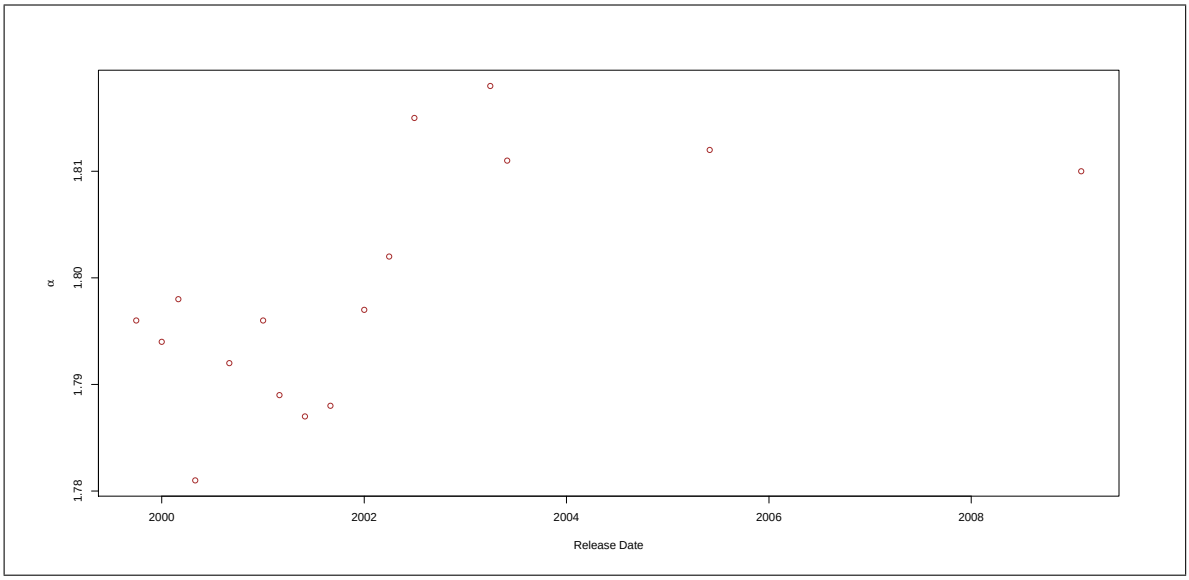
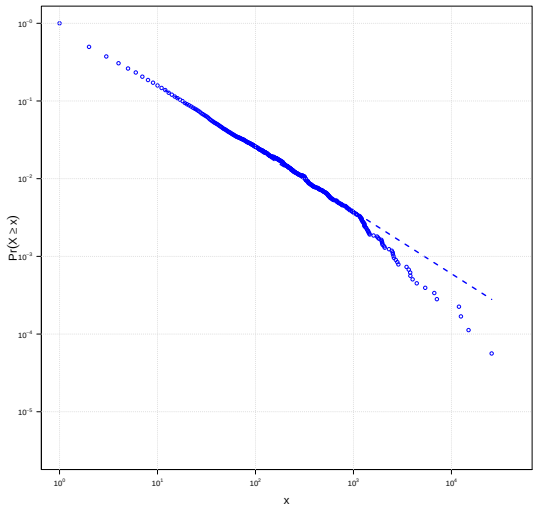


Figure 7.3: α for each archived version of PRINTS over time.

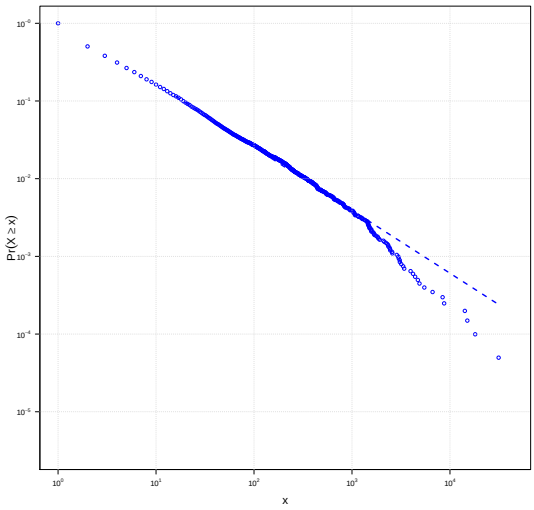
Like PROSITE, the α values obtained from PRINTS all fall within a small range. This range is even smaller than PROSITE, with the highest obtained α value being approximately 1.82 (Version 36), and the lowest being just above 1.78 (Version 27). This suggests that the annotation quality in PRINTS has remained at an almost constant level for over ten years. Although only small, the actual α values obtained from PRINTS show an increase over time.

Analysing the power-law model graphs, as shown in Figure 7.4, also shows strong similarities to PROSITE; while the underlying data increases, the power-law grows proportionally. PRINTS has grown steadily, with the number of entries increasing by 62% over ten years. Specifically, PRINTS Version 24 had a total of $1,210$ entries compared to the $1,950$ entries within PRINTS Version 39.

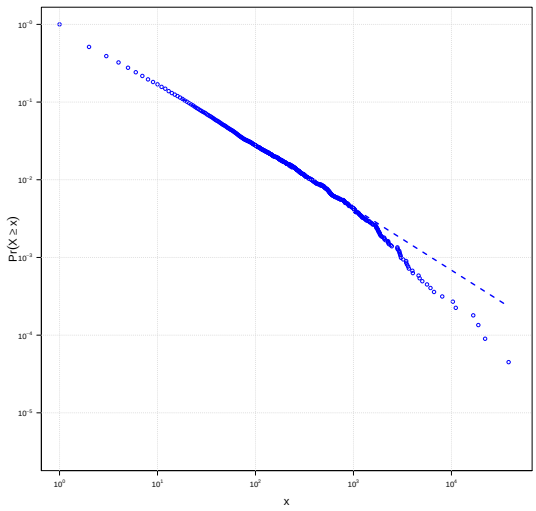
As well as an increase in the number of PRINTS entries there has also been a rise in the average amount of textual annotation per entry. Specifically, PRINTS Version 24 had under $400,000$ total words ($\sim 17,000$ unique) compared to the latest version which has over $675,000$ total words ($\sim 24,500$ unique), which corresponds to an increase



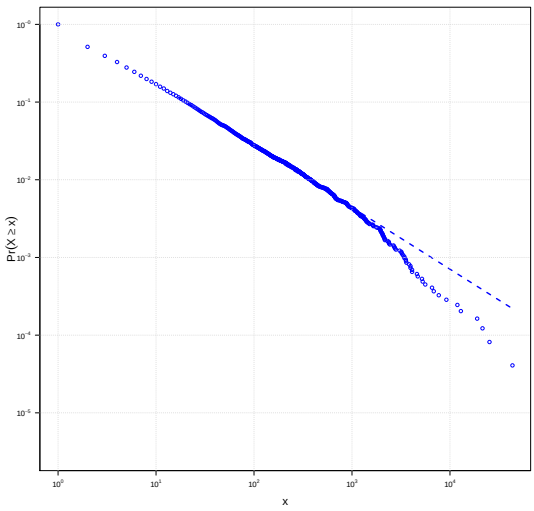
(a) Version 24



(b) Version 29



(c) Version 34



(d) Version 39

Figure 7.4: The power-law model applied to four versions of PRINTS.

of 18 words per entry on average (PRINTS Version 24 averages 328 words per entry, while PRINTS 39 averages 346 words per entry).

7.2.3 TIGRFAMs

There has been a total of 13 TIGRFAMs releases since the first version was released in early 2001. However, the release of TIGRFAMs Version 11 that is available on its FTP server does not contain TIGRFAMs entries, but rather the seed alignment for each entry, meaning that this version cannot be parsed. The α values for the 12 versions we are able to parse are shown in Figure 7.5, with a subset of four power-law graphs shown in Figure 7.6.

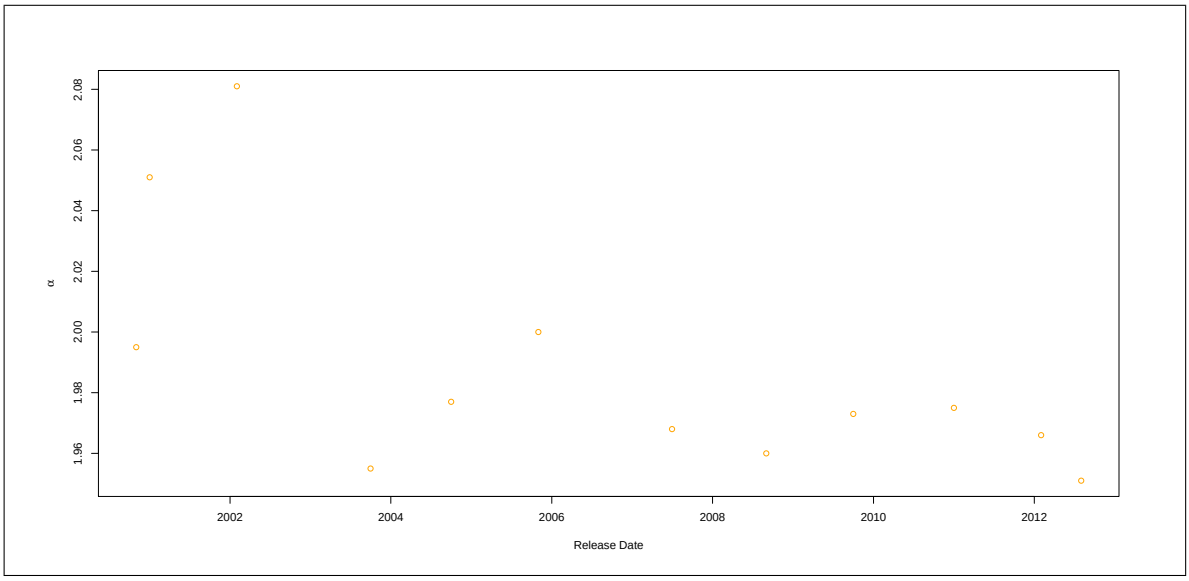
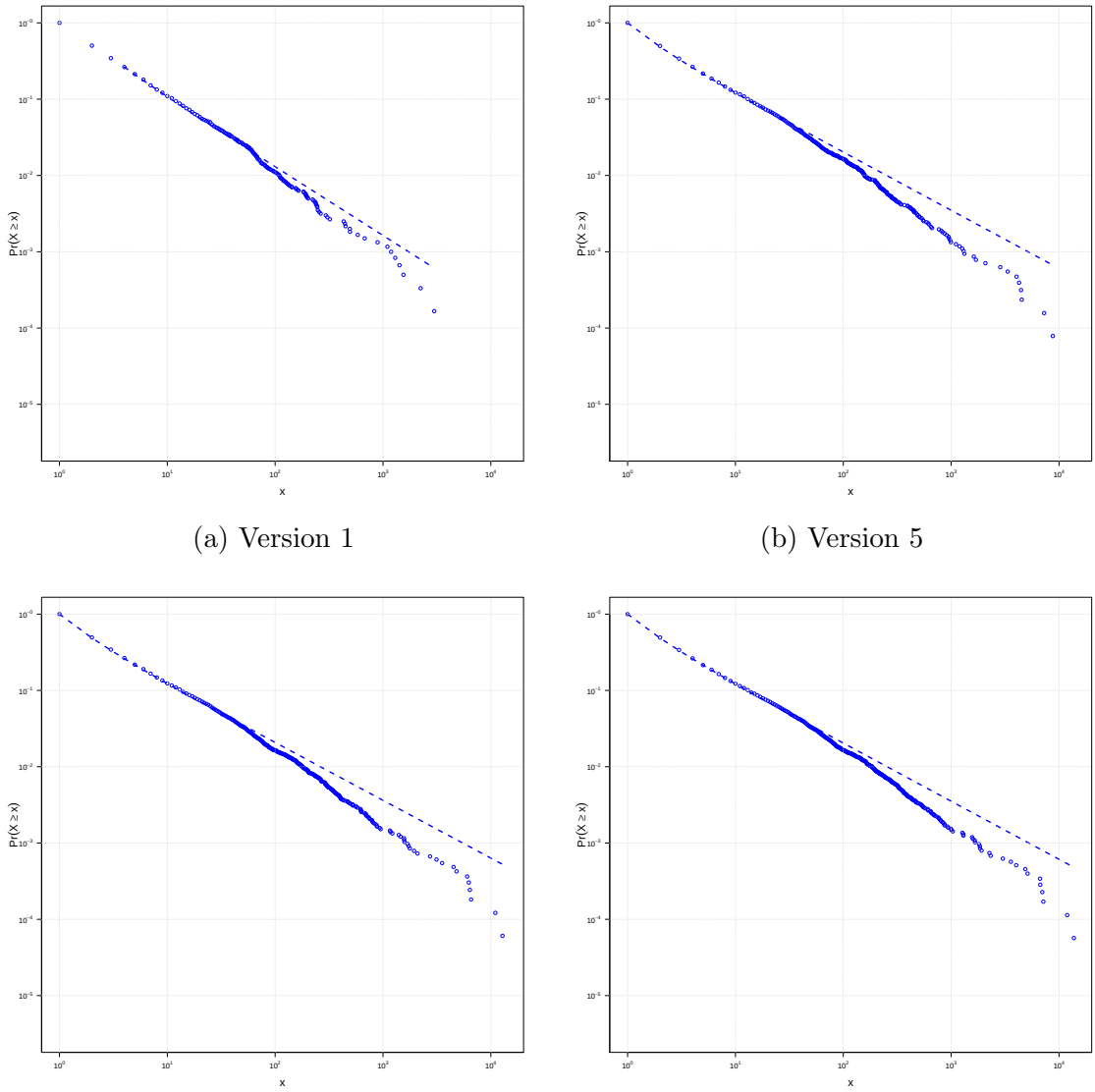


Figure 7.5: α for each archived version of TIGRFAMs over time.

Over the history of the TIGRFAMs database the α values have decreased from an initial value of ~ 2 to an α value of ~ 1.95. Although the α values for TIGRFAMs are higher than those obtained from both PROSITE and PRINTS, they all show only small changes to the observed α values over time; the α values for TIGRFAMs all fall within ~ 0.15 of each other.

The TIGRFAMs annotation corpus also shares similar levels of growth to PROSITE and PRINTS, with the total number of words in the database having increased by over $200,000$. The latest version of TIGRFAMs contains $\sim 255,000$ words ($\sim 17,500$ unique), compared to the $\sim 52,500$ words ($\sim 6,000$ unique) in Version 1. This corre-



(a) Version 1

(b) Version 5

(c) Version 9

(d) Version 13

Figure 7.6: The power-law model applied to four versions of TIGRFAMs.

sponds to an average of 47 words per entry in Version 1 and an average of 60 words per entry in Version 13.

However, unlike PROSITE and PRINTS, the underlying power-law graphs for TIGRFAMs show that a two slope behaviour is starting to be exhibited over time. This development is not as prominent as the two slopes exhibited in Swiss-Prot, likely due to a slower growth. TIGRFAMs initial release contained just over $1,100$ entries, which has increased to $\sim 4,250$ in its latest version, while over the same period the total number of entries in Swiss-Prot rose from $\sim 95,000$ to $\sim 535,000$. This suggests that the two slope behaviour will become more evident as the database continues to grow.

7.2.4 InterPro

With the exception of Version 17, all historical versions of InterPro are archived on its FTP server. In total there are 36 releases available for analysis, with the α values for these releases shown in Figure 7.7. We also show the power-law graphs for four evenly spaced versions in Figure 7.8.

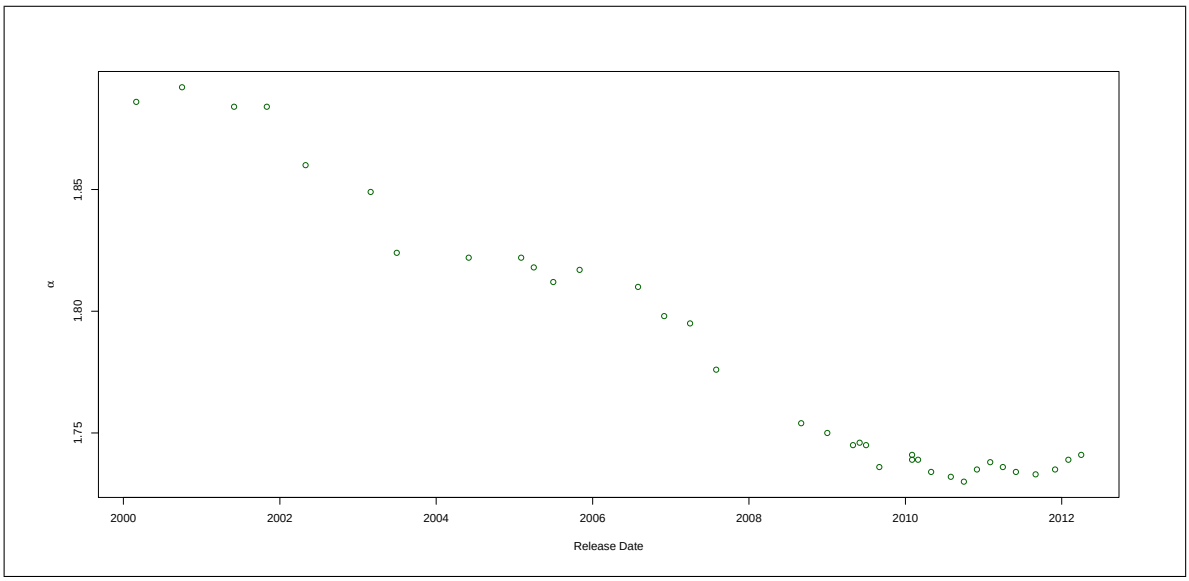
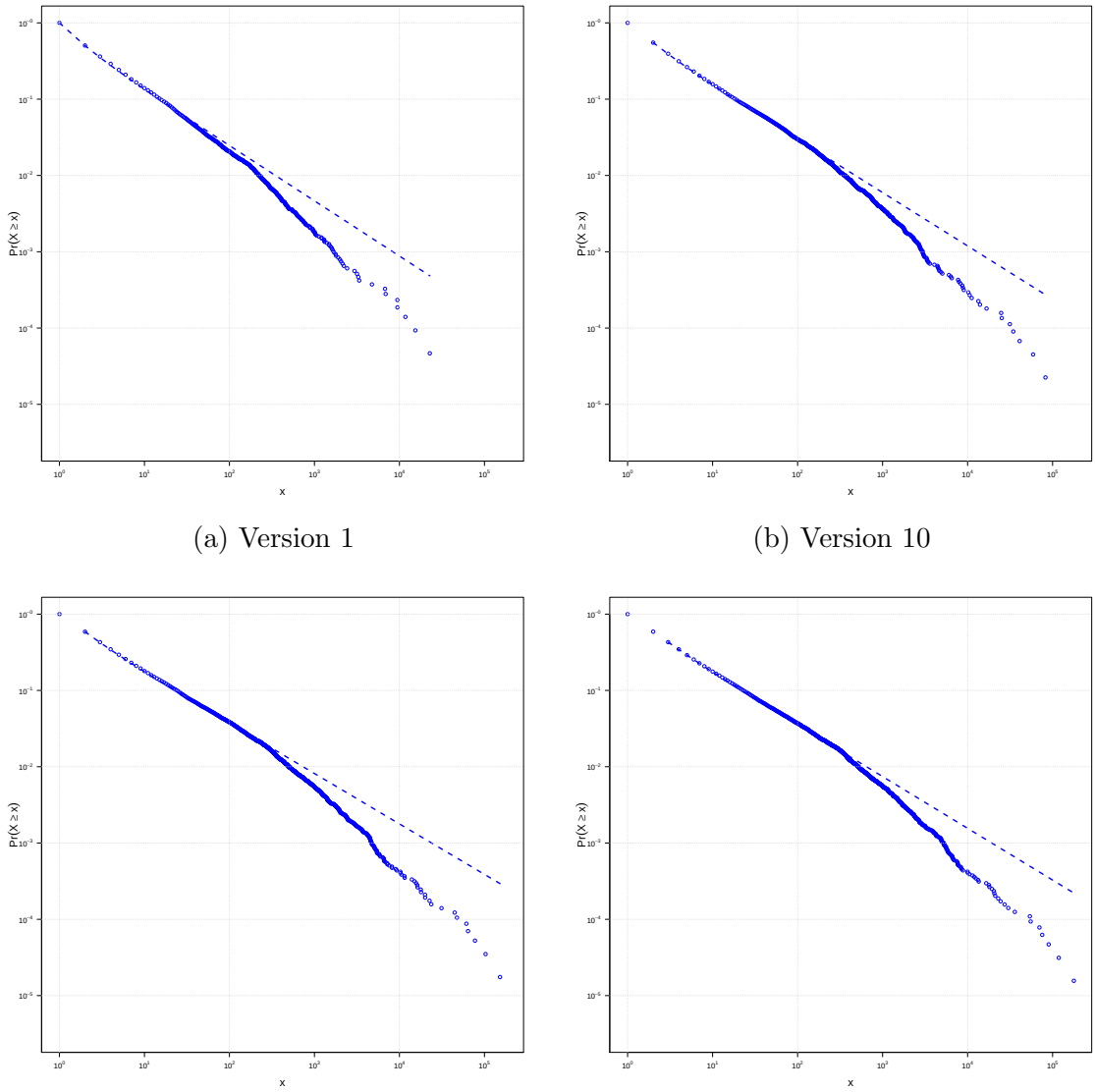


Figure 7.7: α for each archived version of InterPro over time.

Figure 7.7 shows that the obtained α values decline over time, with later versions of InterPro having α values ~ 0.15 less than those obtained from the initial versions. Although there is an overall decline in α values, later versions have shown little change since 2008, having stabilised at $\alpha \approx 1.74$. Although this overall decline is more sub-



(a) Version 1

(b) Version 10

(c) Version 20

(d) Version 36

Figure 7.8: The power-law model applied to four versions of InterPro.

stantial than PRINTS, PROSITE and TIGRFAMs, it is still not as significant as the reduction seen in Swiss-Prot and TrEMBL.

The corresponding power-law graphs, as shown in Figure 7.8, show that InterPro exhibits a two slope behaviour, which is apparent from its first release. Interestingly, the gradients of these slopes are not increasing at noticeable rates. This behaviour is unlike other resources exhibiting two slopes; in these databases two slope behaviour is generally developed over time as the database matures.

InterPro has shown significant growth since its first version, with the number of entries increasing almost eight-fold over ten years. Specifically, the first version of InterPro contained just under $3,000$ entries, which has grown to over $23,000$ in InterPro Version 37. The amount of annotation is also growing at a similar rate, with the total of $\sim 350,000$ words in InterPro Version 1 ($\sim 21,500$ unique) increasing to $\sim 2,800,000$ words in InterPro Version 37 ($\sim 64,000$ unique).

7.2.5 neXtProt

The final database we have chosen to analyse is neXtProt. Being a relatively new database, neXtProt has fewer releases than the other analysed databases, with only eight versions available. The obtained α values for these eight versions are shown in Figure 7.9, with Figure 7.10 showing the corresponding power-law graphs for four neXtProt releases.

These neXtProt versions cover just under one year of data. Given this, we would expect the α value to not change significantly between versions. If we exclude the initial version, then this is true, as the obtained α values are all close to ~ 1.8. However, the first neXtProt version has an α value ~ 0.5 higher than the remaining versions.

The corresponding power-law graphs, as shown in Figure 7.10, show no major differences between each neXtProt version. However, the later three versions exhibit a slight kink in their tails (visible between 10^4 and 10^5), which is not featured in the first version (Figure 7.10a). However, all of the power-law graphs exhibit a two slope behaviour, although this is less pronounced than in other analysed databases, such as

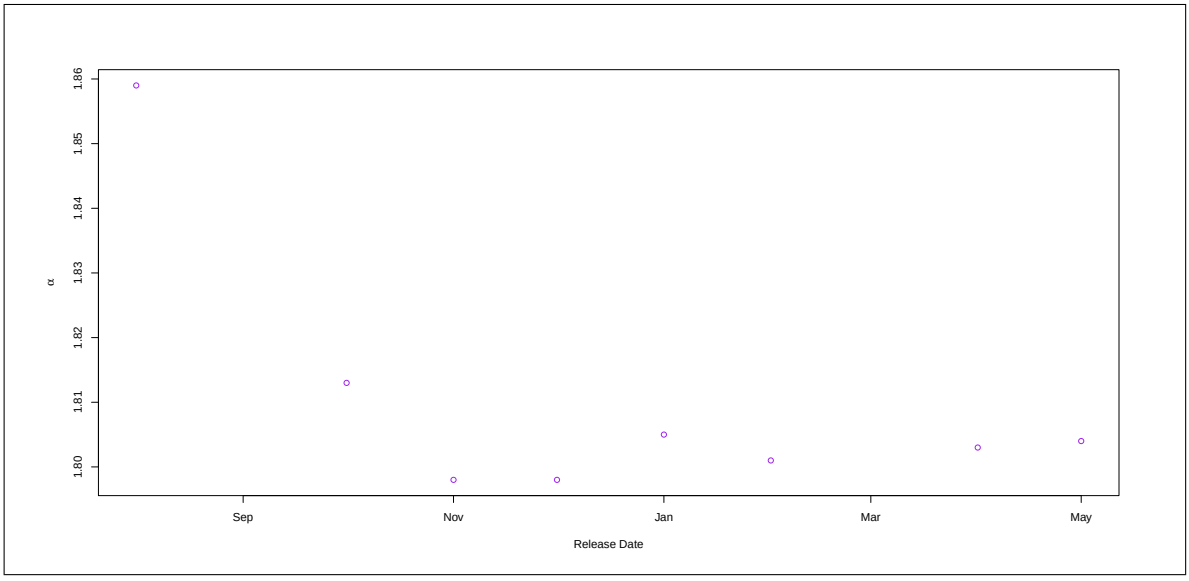


Figure 7.9: α for each archived version of neXtProt over time.

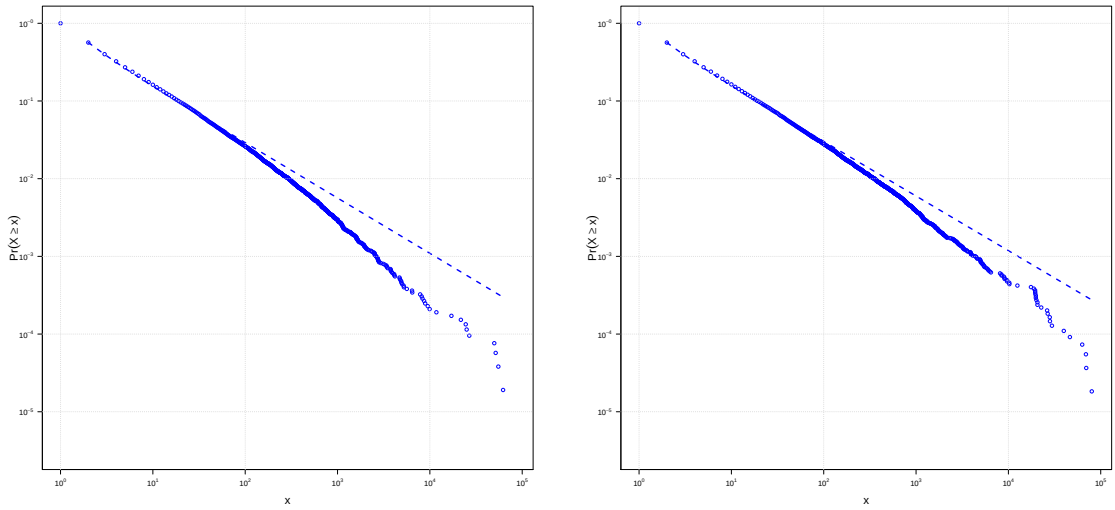
InterPro.

Although the analysed neXtProt versions span just under a year, there is a reasonable growth in textual annotation. Specifically, the first version of neXtProt has a total of $\sim 1,325,000$ words ($\sim 52,500$ unique), which grows to $\sim 2,000,000$ words ($\sim 56,500$ unique) in the latest version. However, unlike other databases, the number of entries within neXtProt remains relatively stable with just over $20,000$ records in all database versions.

Another unique feature of neXtProt is that it provides a classification system for annotation quality; annotations are classified as either gold or silver[6]. By distinguishing between annotation based on this classification we can extract two datasets for each version. Figure 7.11 shows the power-law graph for two neXtProt versions which differentiate between gold and silver annotation.

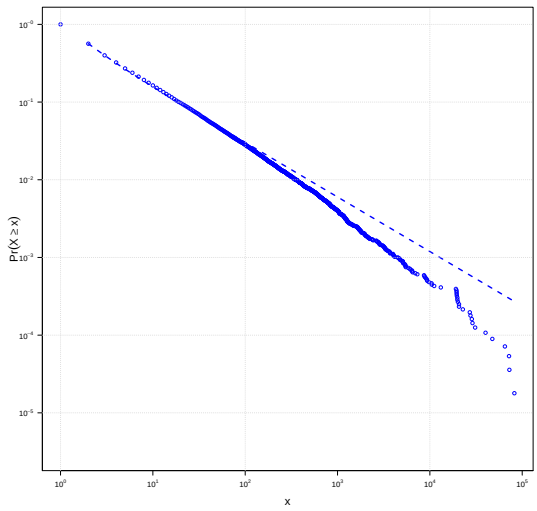
Figure 7.11 shows a similar pattern to the power-law graphs comparing Swiss-Prot and TrEMBL. Specifically, the gold annotation corpus acts as a more mature dataset, with the silver dataset exhibiting higher levels of reuse. In addition to this, the two datasets show signs of divergence over time. This conclusion is also reflected in the obtained α values; the gold datasets have α values of 1.8 and 1.7, whilst α values of 1.6 and 1.4 were obtained from the silver datasets.

[6]Data which neXtProt classify as bronze is not included in the database.

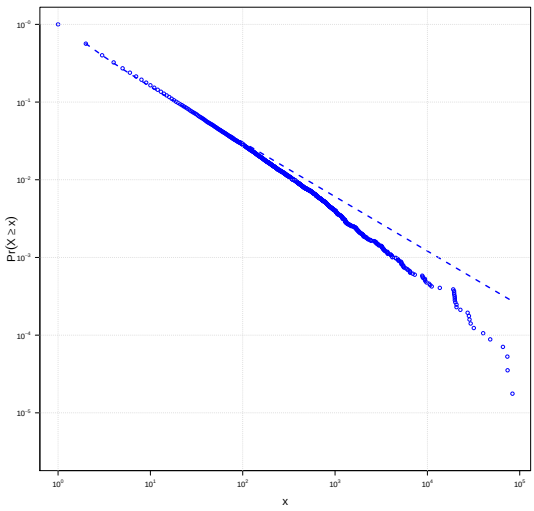


(a) Version 2011-08-23

(b) Version 2011-11-11

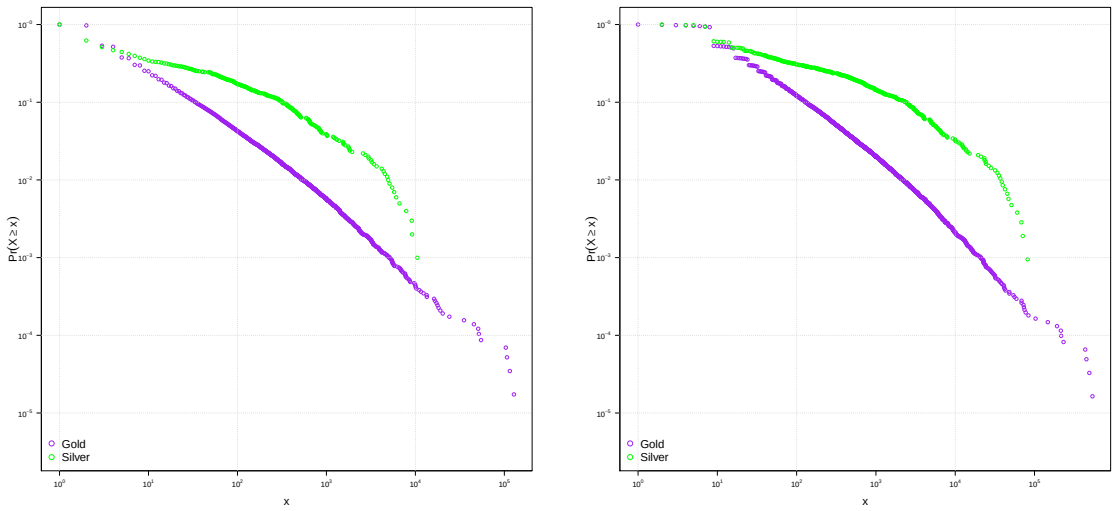


(c) Version 2012-01-13



(d) Version 2012-05-07

Figure 7.10: The power-law model applied to four versions of neXtProt.



(a) Version 2011-10-20 (b) Version 2012-05-07

Figure 7.11: The power-law model applied to gold and silver annotation from two versions of neXtProt.

It is also interesting to note that the first version of neXtProt contained no annotation that was classified as silver. The exclusion of silver annotation likely explains the significant difference in the obtained α value between the first and second versions. We also suspect it is responsible for the development of the kink exhibited in the tail of later neXtProt releases.

7.2.6 Summary

In total QUALM has been applied to seven databases, covering over 300 individual database releases. To allow these results to be easily compared, we can combine all obtained α values into a single graph, as shown in Figure 7.12.

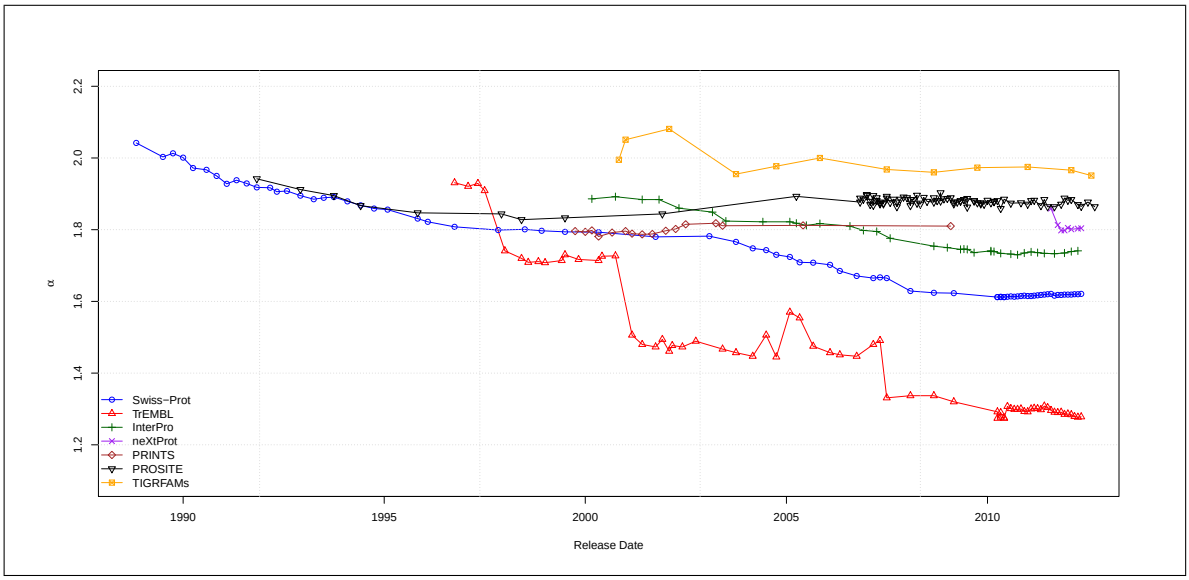


Figure 7.12: Graph combining the α values from Swiss-Prot and TrEMBL with the five newly analysed databases.

This graph shows that the latest versions of TIGRFAMs and PROSITE have the highest α values, whilst TrEMBL has the lowest by a substantial margin. This figure also illustrates the rate at which α values change over time within the context of other databases. For example, TrEMBL has declined more drastically than any other database, whilst the α values for InterPro have dropped below those for PRINTS and PROSITE, having initially been higher.

Over time, we have seen the amount of textual annotation within each database rising. We summarise the word statistics for the first and latest release of each analysed database in Table 7.1. This table shows that the annotation corpus within each database is growing, with an increase of unique words over the lifetime of each database. The largest corpus is contained within Swiss-Prot, which has over $333,500$ unique words. However, the Swiss-Prot corpus has higher levels of reuse than many databases; the average word in Swiss-Prot is reused 97 times compared to databases such as PROSITE where a word is only reused 12 times on average.

	First Archived Version			Latest Archived Version		
	Total	Unique	Average	Total	Unique	Average
PROSITE	83,213	9,745	157	351,083	30,137	213
PRINTS	396,583	17,776	328	676,400	24,514	347
TIGRFAMs	52,572	6,006	47	255,677	17,602	60
InterPro	356,707	21,438	119	2,810,767	63,962	121
neXtProt	1,326,535	52,445	66	2,016,741	56,663	100
Swiss-Prot	203,315	10,745	24	32,309,446	333,528	60
TrEMBL	114,363	6,181	1	330,437,593	12,080	15

Table 7.1: Summary of the word statistics for the five databases, including those for Swiss-Prot and TrEMBL. The columns represent the total number of words in each database, the number of unique words and the average number of words per entry, for both the earliest and latest databases analysed.

This growth in annotation has resulted in the average number of words per entry rising for each database, even though new entries are continually being added to the majority of the databases. The PRINTS and PROSITE databases have the highest average amount of annotation per entry, with TrEMBL having the least.

7.3 Inferring Sentence Provenance and Propagation

As with QUALM, we used UniProtKB as the basis for our analysis of inferring the provenance and propagation of textual annotation. This analysis resulted in three main conclusions: sentences can be used as annotation markers; the visualisation (VIPeR) developed in Chapter 5 provides a mechanism to infer the provenance and propagation of sentences; and a sentence may follow one or more propagation patterns. Within this section we extend our analysis to include sentences from the neXtProt, InterPro, PRINTS, TIGRFAMs and PROSITE databases.

Following the extraction of all sentences from each database we can initially investigate the number of unique and singleton sentences. This data, as shown in Table 7.2, shows that the reuse of sentences within UniProtKB is more prolific than in the other databases. For example, there is a total of 22,940 sentences in the latest version of PROSITE with the majority being unique (21,902), whilst in TrEMBL there are over 26 million sentences, with just over 8,000 being unique. Further, the number of singleton sentences in the newly analysed databases is also higher than UniProtKB, with the vast majority of the PROSITE and TIGRFAMs corpora consisting of sentences that only exist within a single entry. Table 7.2 also shows the total number of unique sentences obtained across all historical database versions and highlights that a number of sentences have been removed from the corpus of each database over time.

	Total Sentences	Unique Sentences	Singleton Sentences	Total Unique
Swiss-Prot	3,304,681	394,233	255,349	531,206
TrEMBL	26,706,421	8,131	735	49,665
InterPro	139,624	71,755	57,628	100,874
neXtProt	158,929	101,822	90,875	110,607
PROSITE	22,940	21,902	21,356	29,127
PRINTS	27,987	16,953	14,356	17,858
TIGRFAMs	13,360	12,155	11,481	13,373

Table 7.2: Table showing the total number of sentences, unique (i.e. distinct) sentences and singleton sentences contained within the latest version of each analysed database. Additionally, we show the total number of unique sentences over the lifetime of the entire database.

The previous analysis of sentence reuse in UniProtKB identified four propagation pat-

terns: missing origin; reappearing entry; transient; and originating in TrEMBL. Although sentence reuse in the newly analysed databases is lower than UniProtKB, the removal of unique sentences makes it plausible that these databases will also contain sentences which follow a propagation pattern. Indeed, the application of the missing origin and transient propagation patterns to these databases identified a number of sentences that adhere to these two patterns, as summarised in table 7.3.

Database Name	Missing Origin	Transient	Possibly Transient
UniProtKB	8,355	42,460	25,582
InterPro	2,689	4,094	1,293
neXtProt	35	5,148	773
PROSITE	132	2,644	21
PRINTS	81	206	363
TIGRFAMs	17	563	63

Table 7.3: Table summarising the number of sentences following the transient and missing origin propagation patterns for each database. Sentences classified as possibly transient are those which appear a single time in the latest version of the database.

As expected, UniProtKB contains the highest number of sentences following the transient and missing origin propagation patterns. However, as a percentage of the total unique sentences, the InterPro database has a higher percentage of its corpus that follows the missing origin pattern (4% in InterPro, 2% in UniProtKB). The number of missing origin sentences in the remaining databases is very low, with neXtProt and TIGRFAMs containing just 35 and 17 sentences, respectively. Additionally, the number of transient sentences in these databases is higher than the number of missing origin sentences, but still relatively low.

To explore sentence propagation within these databases we can simply extend VIPeR. As specified in its requirements, incorporating new databases into VIPeR should be relatively straightforward. Visually this change will show data points for each database in a new colour. For example, we show a sentence from TIGRFAMs that follows the missing origin propagation pattern in Figure 7.13.

This figure shows that VIPeR has been extended to incorporate the TIGRFAMs database with no loss of interactive features or changes to the layout and appearance[7]. Additional examples of sentences which follow the missing origin pattern are

[7]Visualisations not including UniProtKB entries have the striping (i.e. all possible Swiss-Prot and TrEMBL entries) removed.

shown in Figures 7.14 and 7.15, with Figure 7.16 showing an example of a sentence exhibiting the transient pattern.

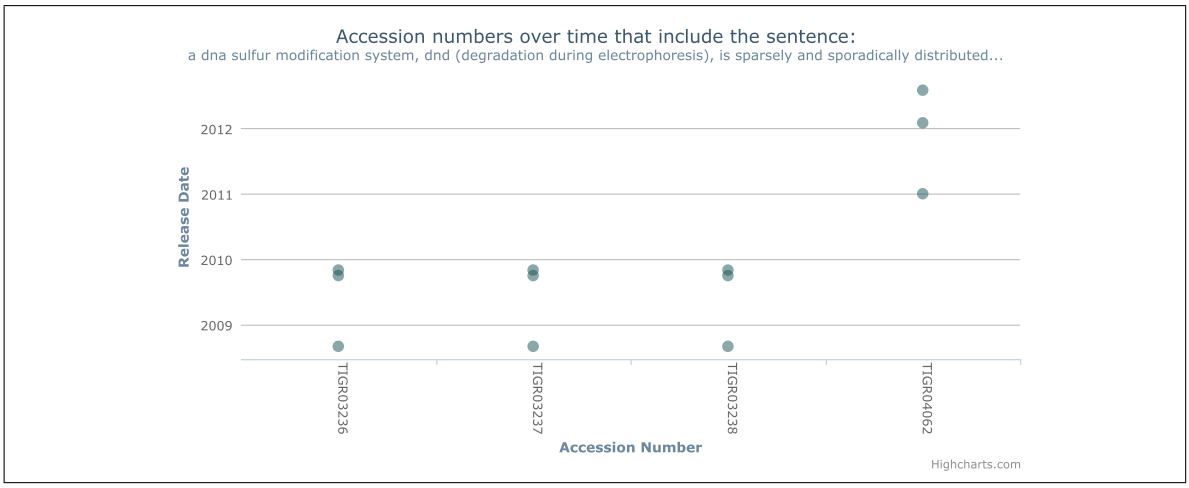


Figure 7.13: Visualisation of the sentence "a dna sulfur modification system, dnd (degradation during electrophoresis), is sparsely and sporadically distributed among the bacteria." which follows the missing origin propagation pattern. The sentence originates in three TIGRFAMs entries and remains in a single TIGRFAMs entry.

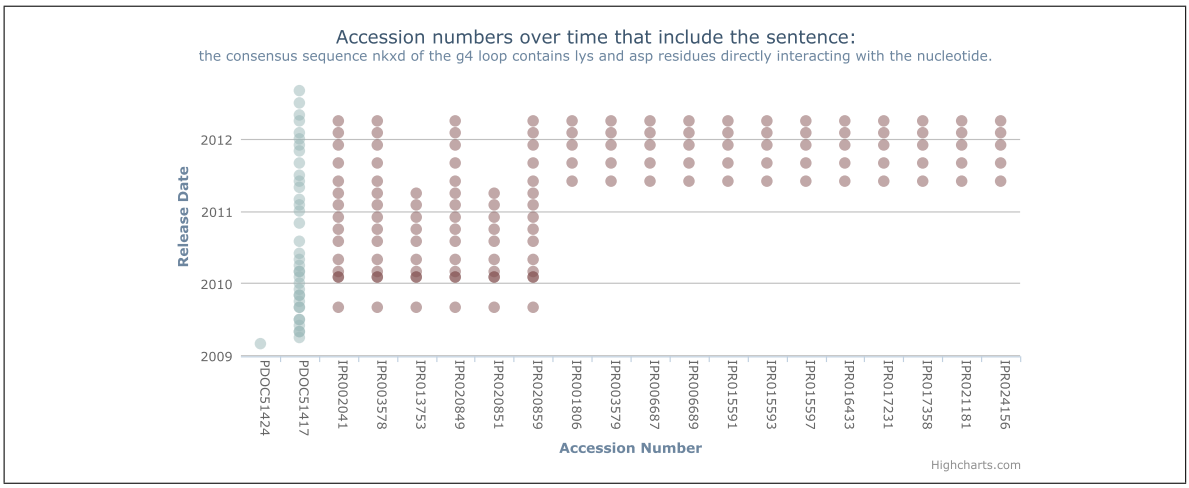


Figure 7.14: Visualisation of the sentence "the consensus sequence nkxd of the g4 loop contains lys and asp residues directly interacting with the nucleotide." which follows the missing origin propagation pattern. The sentence originates in PROSITE entry PDOC51424 but ends up being propagated to an additional PROSITE entry and 18 InterPro entries.

These three figures show that VIPeR has successfully incorporated various databases, with data points for each database being assigned a new colour. These figures also highlight that a sentence can appear in multiple databases simultaneously: within Figure 7.14 the sentence occurs in PROSITE and InterPro; in Figure 7.15 the sentence occurs in both InterPro and PRINTS entries; whilst in Figure 7.16 the sentence occurs

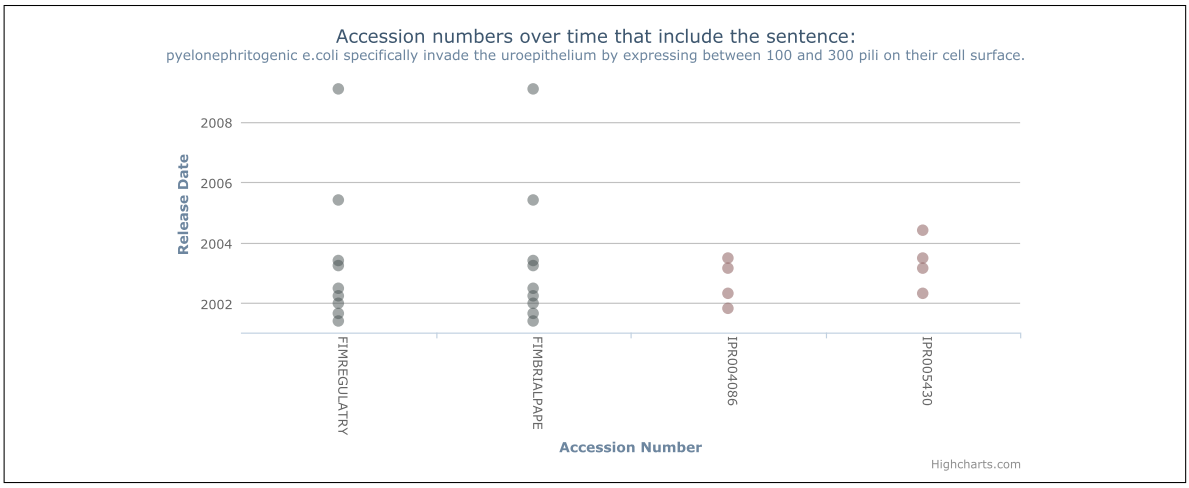


Figure 7.15: Visualisation of the sentence "pyelonephritogenic e.coli specifically invade the uroepithelium by expressing between 100 and 300 pili on their cell surface." which follows the missing origin propagation pattern. The sentence actually originates in PRINTS before being propagated to InterPro. Within InterPro the entry originates in entry IPR004086 and is propagated to entry IPR005430 and remains in this entry for a single entry after it is removed from the origin.

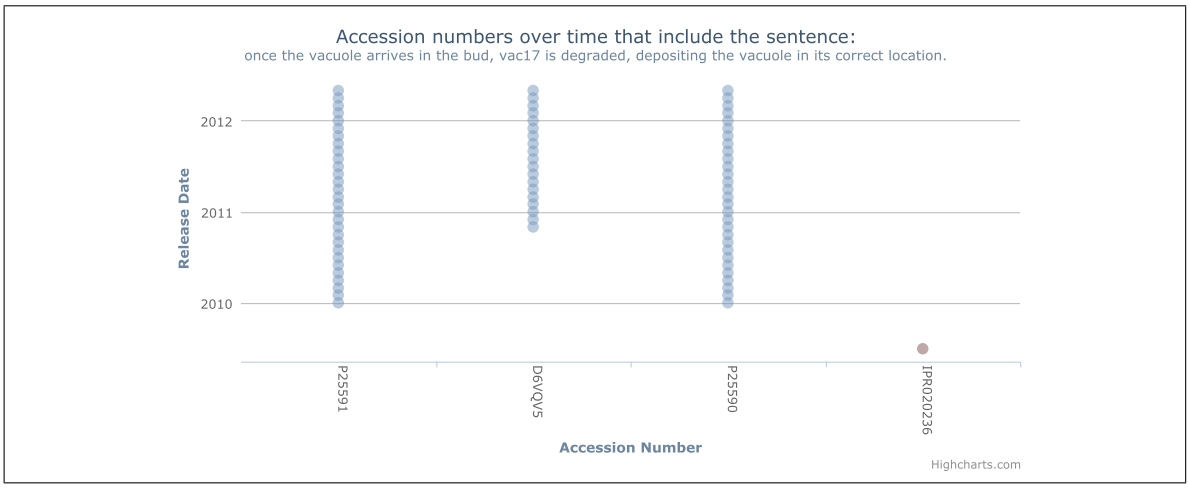


Figure 7.16: Visualisation of the sentence "once the vacuole arrives in the bud, vac17 is degraded, depositing the vacuole in its correct location." which occurs within an InterPro entry for only a single version; i.e. it is transient. Additionally, the sentence is later found in three Swiss-Prot entries.

in a single InterPro entry and three Swiss-Prot entries.

Although four propagation patterns were identified from the UniProtKB analysis, we have only extracted sentences which follow the transient and missing origin propagation patterns. The reappearing entries pattern, which identifies sentences that are removed and then re-added to an entry, was not analysed as it is computationally intensive to calculate and has less analytical value than the missing origin pattern. Additionally, we did not analyse the originating in TrEMBL pattern as it was deemed specific to UniProtKB. However, these graphs suggests that there are sentences which are propagated between external databases.

Between the seven databases a total of over $850,000$ unique sentences have been extracted over their lifetimes, with $\sim 70,000$ being contained in more than one database (i.e. the seven databases share a corpus of $780,000$ sentences). To explore the distribution of these $780,000$ sentences we take each sentence and check if it has appeared within each individual database. The results from this analysis are shown in Table 7.4.

These results show that UniProtKB and neXtProt have the most sentences in common, with over half of the sentences in neXtProt being shared with UniProtKB. This is not unexpected as neXtProt obtains a substantial amount of its data from UniProtKB. With the exception of neXtProt, the majority of sentences within the remaining databases are unique to each database. However, each database does have a significant number of shared sentences.

After neXtProt, the database with the most shared sentences is InterPro. As previously discussed, InterPro is an integrative resource and has over $3,000$ sentences shared with each of the PRINTS, PROSITE and TIGRFAMs databases, with less than 500 shared with UniProtKB. These results show that if a sentence is shared, it is generally only between two resources; with the exception of the 151 sentences shared between neXtProt, InterPro and UniProtKB, which is due to the significant overlap between neXtProt and UniProtKB. There are only a handful of sentences shared between three or more databases.

These results confirm that sentences are propagated between external databases. While cross-database propagation is advantageous, allowing knowledge to be shared between

Database Combination	Total Sentences
UniProtKB	526,435
neXtProt; UniProtKB	83,868
InterPro	82,968
neXtProt	26,539
PROSITE	23,182
PRINTS	10,064
TIGRFAMs	9,661
InterPro; PRINTS	7,751
InterPro; PROSITE	5,790
InterPro; TIGRFAMs	3,681
InterPro; UniProtKB	435
InterPro; neXtProt; UniProtKB	151
PROSITE; UniProtKB	71
InterPro; PROSITE; UniProtKB	26
neXtProt; PROSITE; UniProtKB	20
InterPro; PRINTS; UniProtKB	20
InterPro; neXtProt; PROSITE; UniProtKB	19
InterPro; PRINTS; PROSITE	14
TIGRFAMs; UniProtKB	14
InterPro; TIGRFAMs; UniProtKB	9
InterPro; neXtProt; PRINTS; UniProtKB	4
neXtProt; TIGRFAMs; UniProtKB	3
InterPro; neXtProt; TIGRFAMs; UniProtKB	2
InterPro; TIGRFAMs; PROSITE	2
InterPro; neXtProt; PRINTS; PROSITE; UniProtKB	1
InterPro; PRINTS; PROSITE; UniProtKB	1
PRINTS; UniProtKB	1
PRINTS; PROSITE	1
PRINTS; TIGRFAMs	1

Table 7.4: Table summarising the distribution of all unique sentences shared between the analysed databases.

resources and curators, it exacerbates the issue of identifying the true provenance and propagation of a sentence. For example, Figure 7.16 shows that a sentence in Swiss-Prot was previously seen in the InterPro database – was it propagated from InterPro into Swiss-Prot? If so, then it technically follows the missing origin pattern, as it has been removed from the root entry.

The scale of cross-database propagation between the analysed databases is not as significant as initially expected, with very few sentences appearing in three or more databases. However, this analysis covers only a small subset of the 1,500 active databases. Extending this analysis to cover the majority of these databases would likely identify further cross-database propagation.

7.4 Discussion

Although there are over $1,500$ active biological databases with varying features and specialisations, they all share the common property of having some form of annotation. As the quality and correctness of annotation will inevitable vary between these databases, it is of both interest and importance for a user to be able to evaluate an annotation. We have previously presented two techniques that allow textual annotation to be explored and have used these tools to perform an in-depth analysis of The UniProt Knowledgebase (UniProtKB). Within this chapter we extended these analyses to the neXtProt, InterPro, PRINTS, TIGRFAMs and PROSITE databases.

These databases were chosen as they cover a range of properties, features and specialisms. For example, the neXtProt and InterPro databases are integrative databases, whilst the PRINTS, PROSITE and TIGRFAMs databases provide features that allow information about unknown proteins to be inferred. Although our analyses of these databases are essentially preliminary, as we do not explore our results in substantial detail, they provide further evaluation of the utility of our developed tools.

Initially we applied QUALM to each of the five databases and extracted α values for each available archived version. Relating the results for the latest version of each database to Zipf's principle of least effort suggests that each database places the least effort onto the curator, rather than the reader. The only exceptions to this are early versions of Swiss-Prot and TIGRFAMs, which register α values ≥ 2. As also seen in Swiss-Prot and TrEMBL, most of the databases show a decrease in the α values obtained over time, although this decrease is less significant than observed in UniProtKB. The one exception to this trend is the PRINTS database which shows a small increase in obtained α value over time.

One of the issues previously identified with QUALM was the inability to handle graphs exhibiting two slopes. Analysing the underlying power-law graphs for each database showed that a number of versions exhibit two slopes. Performing a goodness-of-fit test for each database version resulted in p-values of ≤ 0.1 being obtained for those database versions exhibiting two slopes.

This lack of confidence in the obtained α values means we cannot link directly to Zipf's

principle of least effort for the majority of database versions. However, as seen in the UniProtKB analysis, the metric still appears to provide a reasonable approximation of the underlying data. For example, an analysis of the neXtProt database shows that the gold dataset is of better quality than silver, whilst the database as a whole has a higher α value than for Swiss-Prot and TrEMBL.

The second section of this chapter extended this analysis to sentence reuse. We identified that annotation reuse varies significantly between databases, with the lowest levels of reuse exhibited in the PROSITE database, with only 2% of its sentences being reused. The TIGRFAMs and PRINTS databases, like PROSITE, also have relatively low levels of sentence reuse (6% & 15%, respectively) likely due to the types of data that they annotate. Specifically, these databases produce information about protein families and associated patterns. As these databases produce both the raw data and its corresponding annotation, there is no dependency or pressure from external data requiring annotation, unlike databases such as UniProtKB. This means curators can dedicate more time and resources to individual database entries, with the resulting annotation being similar that of an abstract from an academic paper.

The neXtProt corpus also has low levels of sentence reuse (11%), which is surprising as over half of its corpus is shared with UniProtKB. However, as the sole focus of neXtProt is on human proteins, then annotation from UniProtKB will mostly come from *Homo sapiens* entries which will include some of the oldest and best curated UniProtKB entries. Additionally, as *Homo sapiens* are a model organism they are well-studied with a fully sequenced genome, meaning the number of entries in neXtProt remains relatively constant. Therefore neXtProt only has to focus on improving the existing set of entries, with annotation also being provided by external groups who will provide unique information from their own research and expertise.

Even though some of these databases have very low levels of sentence reuse, they each have a number of sentences which follow the missing origin and transient propagation patterns. Although these identified sentences were not analysed in detail we suspect that they are indicators of low quality and erroneous annotation; based on this previous analysis of UniProtKB, up to 50% of the missing origin sentences could be erroneous.

Perhaps the most significant observation from this section is the evidence that an-

notations are propagated between external databases. The neXtProt and InterPro databases were central to this cross-database propagation with over half of the neXtProt corpus being shared with UniProtKB and over 3,000 sentences shared between the remaining databases and InterPro. Additionally, a greater level of granularity could have been achieved within this analysis by distinguishing between sentences in Swiss-Prot and TrEMBL. However, to avoid a high number of propagation permutations we opted to only distinguish by UniProtKB.

To analyse this cross-database propagation, we extended VIPeR, which was developed in Chapter 5. This analysis proved that VIPeR could easily incorporate new databases and handle sentences which appear in multiple databases, providing confidence that the requirement of being generic has been fulfilled (RQ3 and RQ4).

Although multiple databases can be shown in a single visualisation, a number of potential problems became identifiable from this analysis. For example, we previously identified the issue of striping, which is caused by the unsynchronised releases of early Swiss-Prot and TrEMBL versions. We overcome this by showing all possible versions of Swiss-Prot and TrEMBL down the side of each graph. Within graphs that show points from a number of different databases then showing all possible releases for each databases can become insufficient and potentially misleading. Additionally, although VIPeR provides features such as zooming, it is problematic to visualise particularly large datasets, with Web browsers struggling to render the visualisation.

The ability to visualise sentence propagation between databases is highly beneficial and greatly outweighs these issues, which are relatively minor. For example, Figures 7.14, 7.15 and 7.16 show three sentences which appear in multiple databases and highlight the issue of identifying the true provenance of a sentence. In each example the sentence appears to originate in a single database and then propagate to an external database – without VIPeR this behaviour would be problematic to identify and analyse.

These figures also raise a further question: should the missing origin propagation pattern cover the propagation of sentences between databases? For example, the sentence in Figure 7.16 originates in a single InterPro entry before propagating to three Swiss-Prot entries. Within each individual database the sentence does not follow the

missing origin pattern, but if the true provenance of the sentence is from InterPro, then it should be identified as following the missing origin propagation pattern. However, as is possible within a single database, a sentence can appear in multiple entries coincidentally. Therefore, inferring the propagation and provenance of sentences requires additional care when multiple databases are involved.

8

General Discussions and Future Research

Contents

Introduction

Since the first release of Atlas in 1965, there has been an explosion of biological data. There are currently over $1,500$ active biological databases, such as GenBank and UniProtKB which contain millions of data entries. Many of these databases also contain some annotation in the form of unstructured free text. Within this thesis, we explored ways in which this textual annotation can be analysed.

Specifically, we developed two tools, QUALM and VIPeR, which were initially used to perform a detailed analysis of UniProtKB. This analysis tested the suitability and effectiveness of the tools, with a more general analysis also being performed on the neXtProt, PRINTS, PROSITE, TIGRFAMs and InterPro databases. Although these tools have proven to be analytically beneficial, we encountered a number of issues and limitations. Within this chapter we discuss these limitations and identify possible improvements and extensions that could be incorporated into these tools (Section 8.1).

Whilst our research has taken a step towards addressing the lack of tools for assessing textual annotation, there are various features and procedures that can be implemented by annotation curators and biological databases to aid the assessment of annotation quality. We discuss a number of features that have assisted our analyses, as well highlighting improvements that could help enhance the annotation landscape (Section 8.2).

8.1 QUALM and VIPeR: Limitations and Improvements

Although we are drawing the thesis to a close, there are a various refinements and extensions that could be incorporated into our tools and analyses. For example, extending or refining QUALM so that it can adequately handle datasets exhibiting two slopes would be of significant benefit.

Currently, when fitting a power-law, we only consider values larger than $x_{\min}$, meaning a number of data points are essentially discarded from consideration. One possible refinement for this would be to also discard values *above* a certain value (i.e. $x_{\max}$). Alternatively, a refinement could be introduced that ensures $x_{\min}$ is calculated such that only the second slope is considered. Whilst both of these approaches would involve discarding a significant amount of data, they could provide a more accurate α value. However, these techniques could be used in conjunction to calculate two separate regression lines for the head and the tail, as illustrated in Figure 8.1a.

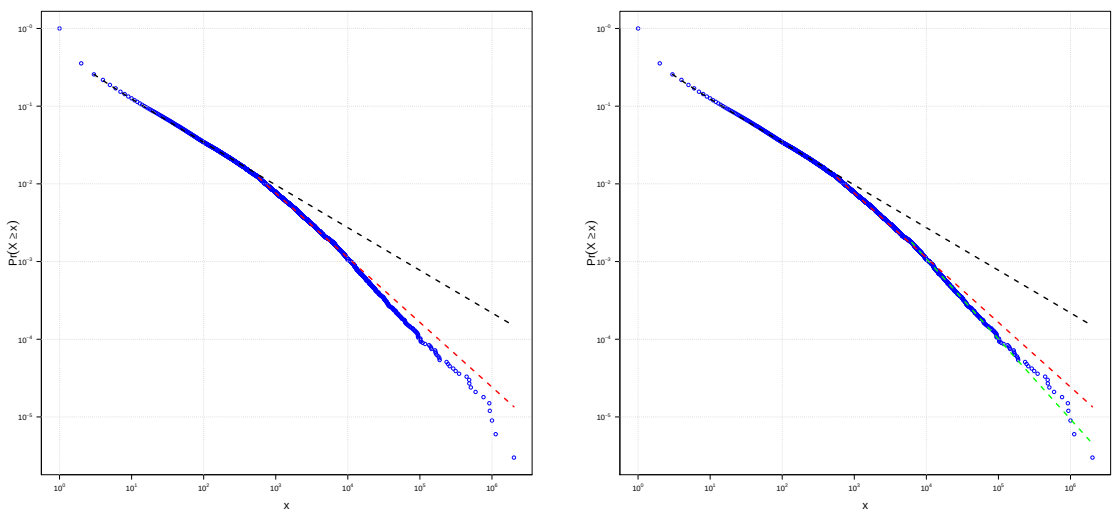


Figure 8.1: (a) Fitting two regression lines to UniProtKB/Swiss-Prot Version 2012_05 and (b) fitting three regression lines.

The two regression lines shown in Figure 8.1a provide a more suitable fit to the power-law than achieved with a single line. However, as we illustrate in Figure 8.1b, the introduction of a third regression line offers an overall better fit than achieved in

Figure 8.1a. This raises questions about the number of regression lines that could be applied to a graph and, if graphs have a differing number of lines, can they be equally compared? Further questions that would need to be considered include: how would a single α value from multiple regression lines be calculated and would multiple regression lines change the analytical value of α?

Currently, as a two slope behaviour becomes evident, the plausibility of the α values accurately reflecting the underlying data is ruled out. Therefore, in these cases, we cannot link directly to Zipf's principle of least effort. However, even if we hypothetically could use the α values, we are unsure how meaningful a direct comparison would be. For example, applying QUALM to the first seven chapters of this thesis, as shown in Figure 8.2, returns an α value of 2. This implies that this thesis is of better quality than Great Expectations ($\alpha = 1.82$).

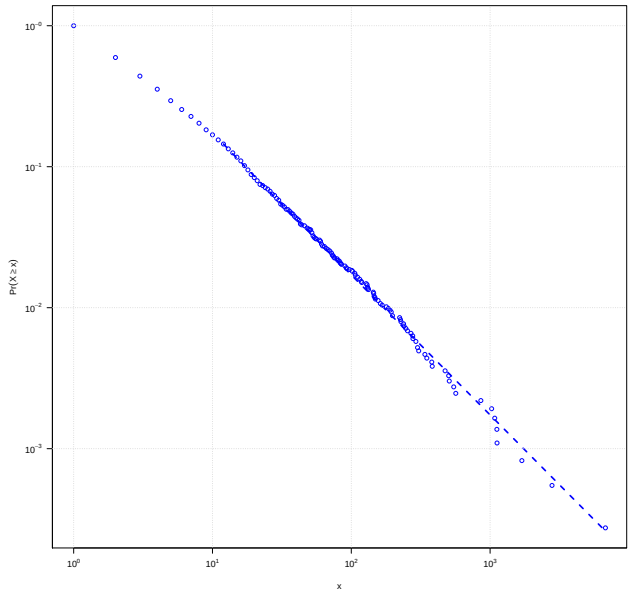


Figure 8.2: Power-law graph for the first seven chapters of this thesis.

However, it would be beneficial to determine if α values obtained from different biological resources can be directly compared. For example, if such a comparison is meaningful, then their α values would suggest that the latest version of TIGRFAMs ($\alpha \approx 1.95$) is of higher quality than PROSITE ($\alpha \approx 1.86$), which is in turn better quality than PRINTS ($\alpha \approx 1.81$). There is no clear way to gain confidence in this conclusion; we need more explicit gold standard datasets to allow us to consider making

such claims.

Like QUALM, we also identified limitations with VIPeR. The most prominent issue is the inability to handle very large datasets, which became more noticeable when trying to analyse sentences that occur in multiple databases. Additionally, increasing the number of databases analysed simultaneously compounds the issue of striping; overcoming striping with dummy data is not intuitive for more than two databases.

Handling large datasets is a problem encountered by many visualisation approaches. We help alleviate this issue with a zooming functionality, but this issue could be alleviated further by reducing the number of years that are shown within a single visualisation. For example, we could allow the user to easily adjust which years are shown. Alternatively, for sentences occurring in multiple databases, we could provide an abstract view to show only a single data series for each database to represent if the sentence occurs in the database or not.

Whilst the usage of tooltips when hovering over data points reduces the issues of multiple striping, we could also introduce a distinct point, such as a red cross, to represent when a sentence is removed from an entry. However, it is possible that this approach would become misleading for those sentences which follow the reappearing propagation pattern.

The reappearing propagation pattern was one of four. However, as these were manually identified based on an analysis of sentences in UniProtKB, it is possible that there are others. Clearly, it would be useful to have a more automated system for identifying further patterns. Additionally, there are still analyses that could be performed on the existing patterns. For example, improving the efficiency of the algorithm used to extract reappearing sentences would allow us to apply this to all of the analysed databases.

We could also aim to classify all of the sentences identified as following the missing origin propagation pattern. Unfortunately, we were limited in the number of sentences analysed as the classification of a sentence is labour and time intensive. To alleviate this, we could look to introduce an automated method for determining the context of a sentence. We could implement this method in a number of ways, such as by

checking if the sentence occurs in the same topic block in each entry, or if the sentence is surrounded by the same sentences in all entries.

Both the propagation patterns and VIPeR are dependent upon sentence reuse, with the amount of sentence reuse increasing over time. Our analysis of these sentences only considered sentences which are identical. This was intentional, not only to reflect that sentences are copied verbatim as a matter of protocol, but also due to its simplicity and increased likelihood that the inferred provenance and propagation is correct. However, it would be of interest to extend this work to consider sentences which are semantically similar.

This could have a number of benefits. For example, we identified a number of spelling and grammatical errors within sentences, such as "it probably replaces ef-tu for the insertion of selenocysteine directd [sic] by the uga codon.", which was contained in the Swiss-Prot entry P14081 for 18 versions until the incorrect spelling of "directed" was corrected. As this change has no biological significance, we could consider these two sentences as equal, which would arguably make the inference of provenance and propagation more accurate.

We could also aim to identify and explore sentence evolution. For example, a sentence which is removed from an entry may have been replaced by an updated sentence with additional information. By identifying updates to individual sentences, we could potentially analyse how the information and structure of annotation changes over time. It may also be possible to determine if other occurrences of the sentence exhibit the same changes, or evolve differently.

The identification of semantically similar sentences would also allow users to browse a database via its annotation, rather than by just its raw data. For example, a user could search for entries which contain an annotation similar to a given sentence. We have experimented with a concordance view, that showed the context of the sentence in each database entry. The identification of similar sentences could be done using techniques such as inverse document frequency and n-gram models. However, such extensions all come with a computational cost, which is not insignificant given the quantity of sentences analysed.

Although we have mostly discussed QUALM and VIPeR individually, we could use them in a complementary manner. For example, we could combine α values for a database with information about sentences in a particular entry, such as if they follow a propagation pattern. Using this information we could attempt to grade the annotation within a given entry.

We could present this information to a user by overlaying the textual annotation on a databases website with specific colours. Essentially, we could have three confidence levels for an annotation (high, medium and low) which would be represented as either green, amber or red. This could be presented in a number of ways, with two possible examples shown in Figure 8.3; either the entire text is highlighted or a small coloured circle appears after each sentence.

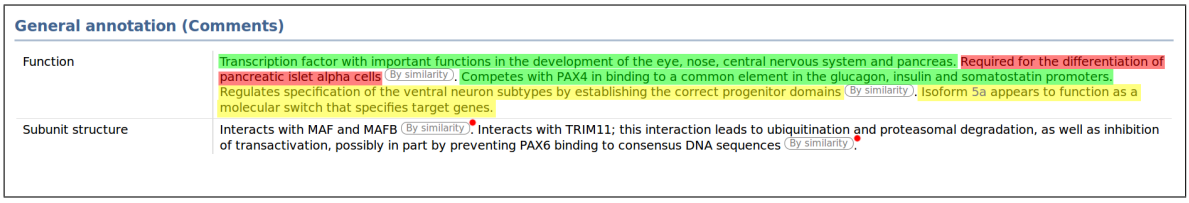


Figure 8.3: Features of a hypothetical browser plug-in that augments a databases entry view. We show two examples for a given UniProtKB entry, with the first example (for annotation in the function topic block) highlighting each sentence with a colour to indicate the confidence in its quality and correctness. The second example (for annotation in the subunit structure topic block) is less obtrusive with a small coloured circle being added after each sentence to indicate its quality. In both cases, additional information could be provided when a user clicks on, or hovers over, a sentence.

The actual implementation of this would be relatively straightforward and could be achieved by developing a browser plug-in or by providing a proxy website. Although each approach has its benefits, a major advantage is that we already know which sentences will occur in a given database entry. This would mean that the sentences could be easily identified within an HTML document, regardless of formatting. This approach would also allow incorrectly parsed sentences to be identified, as they would not be highlighted.

This implementation could be extended further to allow users to generate a visualisation for any sentence in VIPeR directly from a databases website. This implementation would mean QUALM and VIPeR are more easily accessible and provide users with methods for gaining confidence in an annotation directly from the databases website.

8.2 Improving the Annotation Landscape

During the work presented in this thesis we identified a number of features and properties that helped us explore textual annotation. Within this section, we discuss each of these identified features as well as other properties that could change how annotations are analysed or produced.

8.2.1 History is not just for historians

The provision of textual annotation is an aim of many databases. Within UniProtKB, the amount of textual annotation has been increasing over time, with our analysis showing that this has helped reduce the number of entries containing no annotation. Being able to draw this conclusion was only possible as UniProtKB have made available the majority of their historical versions. Without the availability of historical data VIPeR would be of extremely limited analytical value.

Although many users are only concerned with accessing the most recent and up-to-date data, our analyses have highlighted that being able to delve into historical data can provide invaluable insights. For example, being able to analyse the evolution of a database would allow users to establish what impact certain changes had, or how historical issues were overcome. To quote George Santayana "Those who cannot remember the past are condemned to repeat it".

However, if a database makes available its historical data it is generally only provided as a full database dump, meaning many of the rich features for traversing the data are unavailable. An exception to this is UniProtKB, who enable navigation of their historical data with UniSave. UniSave was invaluable when trying to classify sentences following the missing origin propagation pattern; without this tool the analysis would have taken substantially longer.

Ideally, historical data should be made available in a consistent format. For example, UniProtKB still make available recent releases in flat file format, even though internally the format is obsolete, which enabled us to use a single version of BANE for all historical versions.

8.2.2 Annotating annotation

UniSave allows changes between entry versions to be easily analysed and has similarities to revision control systems, such as subversion. However, unlike these systems, it does not provide any information about why changes have occurred. If a change log was available, then it would be a trivial matter to identify why an annotation was added or removed.

An example of a database with such a revision log is Pfam, since it started to replace its annotation with Wikipedia articles. By utilising Wikipedia, the features of a wiki are automatically inherited. For example, it is possible to view which users have contributed to an article and obtain statistics regarding the popularity of the article. Additionally, Wikipedia is a well-studied resource allowing, for example, various quality analyses to be drawn upon.

By recording the reasons for each revision within a publicly available change log, it is possible that more information could be extracted about the biological knowledge. The provision of such metadata is a requisite of many version control systems and is often of benefit to users within a programming environment; such advantages would likely translate to textual annotation.

8.2.3 A little provenance goes a long way

In total, our analysis covered seven databases. By exploiting the historical data of these databases we were able to infer the provenance and propagation of sentences both within each database as well as *between* the databases. However, this inference requires that each database be analysed, and is limited by the inability to definitively determine the provenance of a sentence; we can only ever be sure of an annotations provenance if it is formally documented by the database.

Unlike textual annotation, many databases do provide formal provenance for other forms of annotation and data. For example, in UniProtKB, the source database used to obtain GO annotation is provided, whilst links to the corresponding nucleotide sequences and their translations used to generate the protein sequence are also documented. If similar evidence was also provided for the textual annotation, then our

analyses would have been simpler and, probably, more powerful. For example, classifying sentences following the missing origin propagation pattern could be easily scaled, whilst no sentences would have to be classified as "too many results". Further, such evidence would allow our work to be extended to more structural annotation, such as GO annotations.

UniProtKB are taking steps towards addressing this, having discussed plans to extend their evidence codes to provide more granularity about the source of annotations and the methods used in their production [232]. For example, textual annotations that are produced automatically by rule based systems, such as High-quality Automated and Manual Annotation of Proteins (HAMAP) and Statistical Automatic Annotation System (SAAS), are acknowledged with links to the rule attached to the annotation. This is a key feature that can help users assess the source of an annotation.

8.2.4 Bad annotation is good annotation

Although databases may attach evidence and provenance to an annotation, this in itself does not provide an indication of annotation quality and correctness. Although it is rare for databases to provide a public quality score for their annotation, we have seen that the neXtProt database makes a distinction between gold, silver and bronze annotation. This assessment of quality can be invaluable to users, who can easily assess the confidence they have in the annotation based on its quality score.

Although neXtProt has three confidence levels, annotations which are classified as bronze quality are excluded from the public version of the database. From a data analysis point of view, this is unfortunate. With a lack of an explicit gold standard dataset the inclusion of low quality annotation would enable the assessment of the confidence in an annotation and the development of future quality metrics, by providing a baseline for comparison. If databases stated the quality of their annotations, then low quality annotation could still be incorporated.

8.2.5 Exploit the past to enhance the future

Following the discussion of these database features, there are a number of requirements that we believe biological databases should strive towards providing:

RQ1 Databases should make available their history in a consistent format and, ideally, in a manner that allows this history to be easily accessed and searched.

RQ2 Incorporate all of the available textual annotation into the database, irrespective of its perceived quality.

RQ3 Textual annotation should be provided with an associated confidence or quality score.

RQ4 Each revision to a database entry should be recorded within a change log and made available with the historical data.

RQ5 The original source of textual annotation should be formally acknowledged, with the ability to uniquely identify and reference individual statements. If applicable, the formal provenance should also include external databases.

It is quite possible that many databases already incorporate these requirements but do not make them publicly accessible. Not releasing this data may be done to avoid confusion; for example, the inclusion of low quality annotation may make users incorrectly perceive the database as being of low quality. Alternatively, as we have previously discussed, producing textual annotation requires skilled curators and an investment of time and money. With the ongoing increase of data, manual resources are already strained, which could be further exacerbated if curators were required to implement all of our suggestions. From a database point of view, how beneficial would users find such information, and would it be worth the investment?

Without further information, it is difficult to assess these features. For example, how well utilised are tools such as UniSave? One possible indication is to analyse the number of citations that the corresponding UniSave paper [318] has amassed. Published in 2006, the UniSave paper has been cited a total of 19 times; this is much

less than the corresponding UniRef [212] and UniProt [215] papers published around the same time, which have had 330 and 771 citations, respectively. Although a flawed measure, this suggests that a only subset of UniProtKB users are aware of or utilise UniSave.

Despite this, UniSave is a tool of clear benefit and we are hopeful that our work has helped emphasise its importance. It is an example of one of the many features and ongoing refinements of UniProtKB which helps make it a world-leading biological database. Many of the features we have identified are already incorporated within UniProtKB, suggesting that our recommendations are not unreasonable. Over time, it will be of interest to see how many other databases begin to introduce such features and how many open their doors to welcome annotations from external contributors. For databases yet to explore these changes, it is likely that their best textual annotation is yet to come.

However, if all biological databases implemented formal provenance, provided confidence scores and made available detailed history, then how would our work be impacted? These features would allow our analyses to scale, allowing many more databases to be analysed. Further, we could develop more accurate and rich visualisations and extend our analyses to other forms of structured data. The grading of annotation would provide multiple gold standard datasets, allowing us to refine both QUALM and other quality metrics. We could also automate the classification of sentences, with all of our analyses having reproducible results. In short, these features would allow our analyses to obtain even more detailed and wider-reaching results.

Whilst our work has provided new foundations for assessing textual annotation, it has also contributed to raising a number of important questions about the provenance, propagation and quality of biological annotation. Taken together these contributions should help to ensure that the knowledge that we create now will continue to contribute to the knowledge that we gain in the future.

A

Complete Sentence Classifications

Table A.1: All of the analysed sentences, and their corresponding classification. Sentences have been stored in lowercase to allow for case insensitive comparison.

Sentence	Classification
belongs to the 40s cdc5-associated complex (or cwf complex), a spliceosome sub-complex reminiscent of a late-stage spliceosome composed of the u2, u5 and u6 snrnas and at least brr2, cdc5, cwf2, cwf3, cwf4, cwf5, cwf6, cwf7, cwf8, cwf9, cwf10, cwf11, cwf12, cwf13, cwf14, cwf15, cwf16, cwf17, cwf18, cwf19, cwf20, cwf21, cwf22, cwf23, cwf24, cwf25, cwf26, cwf27, cwf28, ist3, lea1, msl1, prp5, prp10, prp12, prp17, prp22, sap61, sap62, sap114, sap145, slu7, smb1, smd1, smd3, smf1, smg1 and syf2.	Inconsistent
the light chain is composed of three structural domains: a large globular n-terminal domain which may be involved in binding to kinesin heavy chains, a central alpha-helical coiled-coil domain that mediates the light chain dimerization; and a small globular c-terminal which may play a role in regulating mechanochemical activity or attachment of kinesin to membrane-bound organelles (by similarity).	Erroneous
the biological conversion of cellulose to glucose generally requires three types of hydrolytic enzymes: 1) endoglucanases which cut internal beta-1,4-glucosidic bonds; 2) exocellobiohydrolases that cut the dissaccharide cellobiose from the nonreducing end of the cellulose polymer chain; 3) beta-1,4-glucosidases which hydrolyze the cellobiose and other short cello-oligosaccharides to glucose.	Inconsistent

Continued on next page

Sentence	**Classification**
in the hair cortex, hair keratin intermediate filaments are embedded in an interfilamentous matrix, consisting of hair keratin-associated protein (krtap), which are essential for the formation of a rigid and resistant hair shaft through their extensive disulfide bond cross-linking with abundant cysteine residues of hair keratins.	Inconsistent
the beta subunit of voltage-dependent calcium channels contributes to the function of the calcium channel by increasing peak calcium current, shifting the voltage dependencies of activation and inactivation, modulating g protein inhibition and controlling the alpha-1 subunit membrane targeting (by similarity).	Erroneous
interacts with the c-terminal of peptidylglycine alpha-amidating monooxygenase (pam) and may act as part of a signal transduction system linking the catalytic domains of pam in the lumen of the secretory pathway to cytosolic factors regulating the cytoskeleton and signal transduction pathways.	Erroneous
the modification is dependent on dna and is involved in the regulation of various important cellular processes such as differentiation, proliferation, and tumor transformation and also in the regulation of the molecular events involved in the recovery of cell from dna damage (by similarity).	Erroneous
adenosylhomocysteine is a competitive inhibitor of s-adenosyl-l-methinine-dependent methyl transferase reactions; therefore adenosylhomocysteinase may play a key role in the control of methylations via regulation of the intracellular concentration of adenosylhomocysteine (by similarity).	Inconsistent

Continued on next page

Sentence	**Classification**
component of the multisynthetase complex which is comprised of a bifunctional glutamyl-prolyl-trna synthetase, the monospecific isoleucyl, leucyl, glutaminyl, methionyl, lysyl, arginyl, and aspartyl-trna synthetases as well as three auxiliary proteins, p18, p48 and p43 (by similarity).	Erroneous
self; 2; ebi-311928, ebi-311928; p03949:abl-1; 4; ebi-311928, ebi-2315883; q17539:c01b10.8; 5; ebi-311928, ebi-311920; q95qi7:daf-3; 2; ebi-311928, ebi-326363; q09248:dnc-2; 2; ebi-311928, ebi-316282; q09975:lys-8; 2; ebi-311928, ebi-313861; q21831:snfc-5; 2; ebi-311928, ebi-360213;	Erroneous
the n-terminal of the protein extends into the stroma where it is involved with adhesion of granal membranes and photoregulated by reversible phosphorylation of its threonine residues; both are believed to mediate the distribution of excitation energy between photosystems i and ii.	Inconsistent
the modification is dependent on dna and is involved in the regulation of various important cellular processes such as differentiation, proliferation, and tumor transformation and also in the regulation of the molecular events involved in the recovery of cell from dna damage.	Erroneous
the iicd domains contain the sugar binding site and the transmembrane channel; the iia domain contains the primary phosphorylation site (the donor is phospho-hpr); iia transfers its phosphoryl group to the iib domain which finally transfers it to the sugar (by similarity).	Too Many Results

Continued on next page

Sentence	**Classification**
adenosylhomocysteine is a competitive inhibitor of s-adenosyl-l-methinine-dependent methyl transferase reactions; therefore adenosylhomocysteinase may play a key role in the control of methylations via regulation of the intracellular concentration of adenosylhomocysteine.	Inconsistent
this delta-9 desaturase is a terminal component of the liver microsomal stearyl-coa desaturase system, that utilizes o(2) and electrons from reduced cytochrome b(5) to catalyze the insertion of a double bond into a spectrum of fatty acyl-coa substrates (by similarity).	Inconsistent
in the absence of mercury merr represses transcription by binding tightly to the mer operator region; when mercury is present the dimeric complex binds a single ion and becomes a potent transcriptional activator, while remaining bound to the mer site (by similarity).	Erroneous
chemotactic-signal tranducers respond to changes in the concentration of attractants and repellents in the environment, transduce a signal from the outside to the inside of the cell, and facilitate sensory adaptation through the variation of the level of methylation.	Inconsistent
activated by tyrosine-phosphorylation in response to either integrin clustering induced by cell adhesion or antibody cross-linking, or via g-protein coupled receptor (gpcr) occupancy by ligands such as bombesin or lysophosphatidic acid, or via ldl receptor occupancy.	Erroneous

Continued on next page

Sentence	**Classification**
laminin is a complex glycoprotein, consisting of three different polypeptide chains (alpha, beta, gamma), which are bound to each other by disulfide bonds into a cross-shaped molecule comprising one long and three short arms with globules at each end (by similarity).	Erroneous
psi is a plastocyanin-ferredoxin oxidoreductase, converting photonic excitation into a charge separation, which transfers an electron from the donor p700 chlorophyll pair to the spectroscopically characterized acceptors a0, a1, fx, fa and fb in turn (by similarity).	Erroneous
involved in protection of chromosomal dna from damage under nutrient-limited and oxidative stress conditions.	Inconsistent
belongs to the cold-shock domain (csd) family.	Too Many Results
p35415:prm; 1; ebi-86215, ebi-133215;	Erroneous
composed of 14 different subunits.	Possibly Erroneous
proteins that associate with the core dimer include three families of regulatory subunits b (the r2/b/pr55/b55, r3/b”/pr72/pr130/pr59 and r5/b’/b56 families), the 48 kda variable regulatory subunit, viral proteins, and cell signaling molecules (by similarity).	Inconsistent
type i restriction and modification enzymes are complex, multifunctional systems which require atp, s-adenosyl methionine and mg(2+) as cofactors and, in addition to their endonucleolytic and methylase activities, are potent dna-dependent atpases (by similarity).	Inconsistent
3-beta-hydroxy-delta(5)-steroid + nad(+) = 3-oxo-delta(5)-steroid + nadh (acts on 3-beta-hydroxyandrost-5-en-17-one to form androst-4-ene-3,17-dione and on 3-beta-hydroxypregn -5-en-20-one to form progesterone).	Accurate

Continued on next page

Sentence	**Classification**
udp-n-acetyl-d-glucosamine + n-acetyl-beta-d-glucosaminyl-1,2-alpha-d-mannosyl-1,3(6)-(n-acetyl-beta-d-glucosaminyl-1,2-alpha-d-mannosyl,1,6(3))-beta-d-mannosyl-1,4-n-acetyl-beta-d-glucosaminyl-r = udp + n-acetyl-beta-d-glucosaminyl-1,2-(n-acetyl-beta-d-glucosaminyl-1,6)-1,2-alpha-d-mannosyl-1,3(6) -(n-acetyl-beta-d-glucosaminyl-1,2-alpha-d-mannosyl-1,6(3))-beta-d-mannosyl-1,4-n-acetyl-beta-d-glucosaminyl-r.	Erroneous
in e.coli rnase h participare in dna replication; it helps to specify the origin of genomic replication by suppressing initiation at origins other than the locus oric; along with the 5'-3' exonuclease of pol1, it removes rna primers from the okazaki fragments of lagging strand symthesis; and it defines the origin of replication for cole1-type plasmids by specific cleavage of an rna preprimer.	Inconsistent
thoracic aortic aneurysms and dissections are primarily associated with a characteristic histologic appearance known as 'medial necrosis' or 'erdheim cystic medial necrosis' in which there is degeneration and fragmentation of elastic fibers, loss of smooth muscle cells, and an accumulation of basophilic ground substance.	Erroneous
component of the cleavage and polyadenylation specificity factor (cpsf) complex that play a key role in pre-mrna 3'-end formation, recognizing the aauaaa signal sequence and interacting with poly(a) polymerase and other factors to bring about cleavage and poly(a) addition (by similarity).	Inconsistent

Continued on next page

Sentence	**Classification**
there are two operons: the xylcab operon is responsible for the upper metabolic pathway from toluene to aromatic carboxylic acids, & the xyldlefg operon is required for the lower catabolic pathway from aromatic carboxylic acids to compounds that enter the trycarboxylic acid cycle.	Erroneous
hh is characterized by abnormal intestinal iron absorption and progressive increase of total body iron, which results in midlife in clinical complications including cirrhosis, cardiopathy, diabetes, endocrine dysfunctions, arthropathy, and susceptibility to liver cancer.	Inconsistent
prp is found in high quantity in the brain of humans and animals infected with the degenerative neurological diseases kuru, creutzfeldt-jacob disease (cjd), gerstmann-straussler syndrome (gss), scrapie, bovine spongiform encephalopathy (bse), etc. to other prp.	Accurate
involved in the atp-dependent selective degradation of cellular proteins, the maintenance of chromatin structure, the regulation of gene expression, the stress response, and ribosome biogenesis (by similarity).	Erroneous
coup (chicken ovalbumin upstream promoter) transcription factor binds to the ovalbumin promoter and, in cunjunction with another protein (s300-ii) stimulates initiation of transcription.	Inconsistent
the lys-124 ubiquitination also modulates the formation of double-strand breaks during meiosis and is a prerequisite for and dna-damage checkpoint activation (by similarity).	Erroneous
the export to cytoplasm depends on the interaction with a 14-3-3 chaperone protein and is due to its phosphorylation at ser-259 and ser-498 by camk (by similarity).	Erroneous

Continued on next page

Sentence	**Classification**
the sigma factor is an initiation factor that promotes attachment of the rna polymerase to specific initiation sites and then is released (by similarity).	Too Many Results
hydrolysis of 1,4-alpha-d-glucosidic linkages in polysaccharides so as to remove successive maltose units from the non-reducing ends of the chains.	Accurate
the resulting products may subsequently be converted to the corresponding alcohols that are incorporated into lignins (by similarity).	Erroneous
involved in the initial immune cell clustering during inflammatory response and may regulate chemotactic activity of chemokines.	Inconsistent
s-adenosyl-l-methionine + magnesium protoporphyrin = s-adenosyl-l-homocysteine + magnesium protoporphyrin monomethyl ester.	Erroneous
component of the coat surrounding the cytoplasmic face of coated vesicles located at the golgi complex (by similarity).	Accurate
hsp82 is an essential protein that is required by cells in higher concentrations for growth at higher temperatures.	Accurate
monoubiquitinated on lys-147; may give a specific tag for epigenetic transcriptional activation (by similarity).	Erroneous
probably a dodecamer composed of six biotin-containing alpha subunits and six beta subunits (by similarity).	Possibly Erroneous
organized into a structure (processome or rna degradosome) containing a number of rna-processing enzymes.	Inconsistent
involved in the formation of the nuclear envelope and of the transitional endoplasmic reticulum (ter).	Inconsistent
this methionine-rich region is probably important for copper tolerance in bacteria (by similarity).	Erroneous

Continued on next page

Sentence	**Classification**
they have identical ligand binding properties but different coupling properties with g proteins.	Possibly Erroneous
3-carboxy-2-hydroxy-4-methylpentanoate + nad(+) = 3-carboxy-4-methyl-2- oxopentanoate + nadh.	Accurate
this is a conceptual translation; two frameshifts had to be introduced to produce this orf.	Erroneous
component of the infraciliary lattice (icl) and the ciliary basal bodies (by similarity).	Possibly Erroneous
catalyzes the methylation of c-11 in precorrin-4 to form precorrin-5 (by similarity).	Possibly Erroneous
on the 2d-gel the determined pi of this unknown protein is: 6.2, its mw is: 28 kda.	Accurate
heterodimer of a p110 (catalytic) and a p85 (regulatory) subunit (by similarity).	Accurate
this viral protein may be involved in the regulation of the complement cascade.	Inconsistent
two forms; long (shown here) and short; are produced by alternative splicing.	Inconsistent
assembles at the inner surface of the cytoplasmic membrane (by similarity).	Too Many Results
1-aminocyclopropane-1-carboxylate + o2 = ethylene + hcn + co(2) + 2 h(2)o.	Accurate
bind preferentially single-stranded dna and unwind double stranded dna.	Inconsistent
involved in the regulation of hydrogenase expression (by similarity).	Erroneous
may have an essential function in lipopolysaccharides biosynthesis.	Erroneous

Continued on next page

Sentence	**Classification**
rch(2)nh(2) + h(2)o + acceptor = rcho + nh(3) + reduced acceptor.	Accurate
subunit 1 binds to the primer-template junction (by similarity).	Inconsistent
to immunoglobulin and major histocompatibility complex domain.	Too Many Results
isoform 3: membrane; multi-pass membrane protein (potential).	Possibly Erroneous
the beta subunit seems to be encoded by a multigene family.	Erroneous
atp + adenylylsulfate = adp + 3'-phosphoadenylylsulfate.	Inconsistent
an aryl sulfate + a phenol = a phenol + an aryl sulfate.	Erroneous
peptidyl-l-amino acid + h(2)o = peptide + l-amino acid.	Possibly Erroneous
in the c-terminus to yeast sla2 and c.elegans zk370.3.	Erroneous
mediates e2-dependent ubiquitination (by similarity).	Accurate
villin is a ca(2+)-regulated actin-binding protein.	Inconsistent
atp + undecaprenol = adp + undecaprenyl phosphate.	Accurate
aminoacyl-peptide + h(2)o = amino acid + peptide.	Inconsistent
to the calcitonin and to the secretin receptors.	Erroneous
heterodimer of an alpha chain and a beta chain.	Too Many Results
requires ca2+ and mn2+ ions for full activity.	Inconsistent
contains 1 immunoglobulin-like v-type domain.	Too Many Results
belongs to family 13 of glycosyl hydrolases.	Too Many Results
acts as a transglycosylase (by similarity).	Erroneous
nuclear effector molecule (by similarity).	Possibly Erroneous
involved in carbon catabolite repression.	Erroneous
q9vy42:cg1461; 1; ebi-194476, ebi-127720;	Erroneous
contains 6 ldl-receptor class b domains.	Erroneous
ring cleavage of 2,3-dihydroxybiphenyl.	Possibly Erroneous
not expected to have protease activity.	Accurate

Continued on next page

Sentence	**Classification**
secreted in hemolymph (by similarity).	Accurate
interacts with rad51 (by similarity).	Accurate
endplasmic reticulum membrane bound.	Accurate
associated with the plasma membrane.	Accurate
does not have a catalytic activity.	Possibly Erroneous
belongs to the eae/invasin family.	Erroneous
interacts with cyclin g in vitro.	Possibly Erroneous
self; 1; ebi-190958, ebi-190958;	Possibly Erroneous
binds 1 nickel ion per monomer.	Accurate
binds 1 magnesium per subunit.	Inconsistent
clavulanic acid biosynthesis.	Accurate
belongs to the ycf50 family.	Accurate
inhibited by acetazolamide.	Erroneous
involved in tumorigenesis.	Accurate
acetyltransferase enzyme.	Possibly Erroneous
phosphorylates ppp1r12a.	Possibly Erroneous
detected at low levels.	Accurate
interacts with trim28.	Accurate
contacts protein l19.	Erroneous
interacts with gcn5.	Accurate
may self-associate.	Accurate
secreted in milk.	Too Many Results
heme-thiolate.	Accurate
adipocytes.	Accurate
nadp.	Accurate
nuclear.	Too Many Results
p.	Too Many Results
25.	Too Many Results
1.	Too Many Results

Continued on next page

Sentence	**Classification**
3.	Too Many Results
2.	Too Many Results
venom.	Inconsistent
roots.	Inconsistent
leaf.	Inconsistent

Bibliography

[1] E. J. Richardson and M. Watson, "The automatic annotation of bacterial genomes," *Briefings in Bioinformatics*, vol. 14, pp. 1–12, Jan. 2013.

[2] M. Magrane and U. Consortium, "UniProt knowledgebase: a hub of integrated protein data.," *Database : the journal of biological databases and curation*, vol. 2011, Mar. 2011.

[3] B. Strasser, "Collecting, Comparing, and Computing Sequences: The Making of Margaret O. Dayhoff's Atlas of Protein Sequence and Structure, 1954–1965," *Journal of the history of biology*, vol. 43, pp. 623–660, Dec. 2010.

[4] National Human Genome Research Institute, "Human Genome Project Completion: Frequently Asked Questions," Apr. 2003. `http://www.genome.gov/11006943` [Online. Accessed 2014-01-15].

[5] O. McCrimmon, "2013 Release: NHGRI celebrates 10th anniversary of the Human Genome Project." `http://www.genome.gov/27553526` [Online. Accessed 2014-01-15], Apr. 2013.

[6] National Center for Biotechnology Information, "GenBank Release Notes." `http://www.ncbi.nlm.nih.gov/genbank/statistics` [Online. Accessed 2014-02-15], Feb. 2014.

[7] R. J. Robbins, "Biological databases: A new scientific literature," *Publishing Research Quarterly*, vol. 10, pp. 3–27, Mar. 1994.

[8] J. D. Wren and A. Bateman, "Databases, data tombs and dust in the wind," *Bioinformatics*, vol. 24, pp. 2127–2128, Oct. 2008.

[9] P. G. Higgs and T. K. Attwood, *Bioinformatics and Molecular Evolution*. Wiley-Blackwell, 1 ed., Feb. 2005.

[10] D. Howe, M. Costanzo, P. Fey, T. Gojobori, L. Hannick, W. Hide, D. P. Hill, R. Kania, M. Schaeffer, S. St Pierre, S. Twigger, O. White, and S. Yon Rhee, "Big data: The future of biocuration," *Nature*, vol. 455, pp. 47–50, Sept. 2008.

[11] L. D. Stein, "Integrating biological databases.," *Nature reviews. Genetics*, vol. 4, pp. 337–345, May 2003.

[12] P. Gaudet, A. Bairoch, D. Field, S.-A. A. Sansone, C. Taylor, T. K. Attwood, A. Bateman, J. A. Blake, C. J. Bult, J. M. Cherry, R. L. Chisholm, G. Cochrane, C. E. Cook, J. T. Eppig, M. Y. Galperin, R. Gentleman, C. A. Goble, T. Gojobori, J. M. Hancock, D. G. Howe, T. Imanishi, J. Kelso, D. Landsman, S. E. Lewis, I. Karsch Mizrachi, S. Orchard, B. F. Ouellette, S. Ranganathan, L. Richardson, P. Rocca-Serra, P. N. Schofield, D. Smedley, C. Southan, T. W.

Tan, T. Tatusova, P. L. Whetzel, O. White, C. Yamasaki, and BioDBCore Working Group, "Towards BioDBcore: a community-defined information specification for biological databases.," *Database : the journal of biological databases and curation*, vol. 2011, Jan. 2011.

[13] Wikipedia, "Biological database — Wikipedia, the free encyclopedia." `http://en.wikipedia.org/wiki/Biological_database` [Online. Accessed 2013-03-31], Mar. 2013.

[14] International Society for Biocuration, "International society for biocuration mission statement." `http://biocurator.org/mission.shtml` [Online. Accessed 2013-04-03], Apr. 2013.

[15] D. Natale, U. Shankavaram, M. Galperin, Y. Wolf, L. Aravind, and E. Koonin, "Towards understanding the first genome sequence of a crenarchaeon by genome annotation using clusters of orthologous groups of proteins (COGs)," *Genome Biology*, vol. 1, no. 5, 2000.

[16] M. Hoebeke, H. Chiapello, J. F. Gibrat, J. Garnier, and P. Bessières, "Annotation and databases: Status and prospects," in *Database Annotation in Molecular Biology: Principles and Practice* (A. M. Lesk, ed.), pp. 1–21, John Wiley & Sons, Ltd, Sept. 2005.

[17] I. Karsch-Mizrachi and B. F. F. Ouellette, "The GenBank sequence," in *Bioinformatics: a practical guide to the analysis of genes and proteins* (A. D. Baxevanis and B. F. Ouellette, eds.), vol. 43, pp. 45–63, John Wiley & Sons, Inc., 2004.

[18] A. Lesk, *Introduction to Bioinformatics*. Oxford University Press, USA, 3 ed., June 2008.

[19] O. Carugo and S. Pongor, "The evolution of structural databases," *Trends in Biotechnology*, vol. 20, pp. 498–501, Dec. 2002.

[20] B. J. Strasser, "Collecting and experimenting: the moral economies of biological research, 1960s–1980s," *Preprints of the Max-Planck Institute for the History of Science*, vol. 310, pp. 105–123, 2006.

[21] M. Dayhoff, R. Eck, M. Chang, and M. Sochard, *Atlas of Protein Sequence and Structure.*, vol. 1. National Biomedical Research Foundation, Silver Spring, MD., 1965.

[22] M. Schneider, A. Bairoch, C. H. Wu, and R. Apweiler, "Plant protein annotation in the UniProt knowledgebase.," *Plant physiology*, vol. 138, pp. 59–66, May 2005.

[23] D. M. Bolser, P.-Y. Chibon, N. Palopoli, S. Gong, D. Jacob, V. D. Del Angel, D. Swan, S. Bassi, V. González, P. Suravajhala, S. Hwang, P. Romano, R. Edwards, B. Bishop, J. Eargle, T. Shtatland, N. J. Provart, D. Clements, D. P. Renfro, D. Bhak, and J. Bhak, "MetaBase – the wiki-database of biological databases," *Nucleic Acids Research*, vol. 40, pp. D1250–D1254, Jan. 2012.

[24] D. Landsman, R. Gentleman, J. Kelso, and B. F. Francis Ouellette, "DATABASE: A new forum for biological databases and curation.," *Database : the journal of biological databases and curation*, vol. 2009, p. bap002, Jan. 2009.

[25] Oxford University Press, "Oxford Journals | Life Sciences | Database." `http://database.oxfordjournals.org/` [Online. Accessed 2013-03-30], Mar. 2013.

[26] R. M. Waterhouse, F. Tegenfeldt, J. Li, E. M. Zdobnov, and E. V. Kriventseva, "OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs," *Nucleic Acids Research*, vol. 41, pp. D358–D365, Jan. 2013.

[27] A. Chatr-aryamontri, B.-J. Breitkreutz, S. Heinicke, L. Boucher, A. Winter, C. Stark, J. Nixon, L. Ramage, N. Kolas, L. O'Donnell, *et al.*, "The BioGRID interaction database: 2013 update," *Nucleic acids research*, vol. 41, no. D1, pp. D816–D823, 2013.

[28] C. Stark, B.-J. Breitkreutz, A. Chatr-Aryamontri, L. Boucher, R. Oughtred, M. S. Livstone, J. Nixon, K. Van Auken, X. Wang, X. Shi, *et al.*, "The BioGRID interaction database: 2011 update," *Nucleic acids research*, vol. 39, no. suppl 1, pp. D698–D704, 2011.

[29] B.-J. Breitkreutz, C. Stark, T. Reguly, L. Boucher, A. Breitkreutz, M. Livstone, R. Oughtred, D. H. Lackner, J. Bähler, V. Wood, *et al.*, "The BioGRID interaction database: 2008 update," *Nucleic acids research*, vol. 36, no. suppl 1, pp. D637–D640, 2008.

[30] C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, "BioGRID: a general repository for interaction datasets," *Nucleic acids research*, vol. 34, no. suppl 1, pp. D535–D539, 2006.

[31] M. D. Brazas, D. S. Yim, J. T. Yamada, and B. F. F. Ouellette, "The 2011 bioinformatics links directory update: more resources, tools and databases and features to empower the bioinformatics community," *Nucleic Acids Research*, vol. 39, pp. W3–W7, July 2011.

[32] C. Discala, X. Benigni, E. Barillot, and G. Vaysseix, "DBcat: a catalog of 500 biological databases.," *Nucleic acids research*, vol. 28, pp. 8–9, Jan. 2000.

[33] Wikipedia, "List of biological databases — Wikipedia, the free encyclopedia." `http://en.wikipedia.org/wiki/List_of_biological_databases` [Online. Accessed 2013-03-30], Mar. 2013.

[34] X. M. Fernández-Suárez and M. Y. Galperin, "The 2013 nucleic acids research database issue and the online molecular biology database collection," *Nucleic Acids Research*, vol. 41, pp. D1–D7, Jan. 2013.

[35] M. Y. Galperin, "The molecular biology database collection: 2006 update," *Nucleic Acids Research*, vol. 34, pp. D3–D5, Jan. 2006.

[36] M. Y. Galperin, "The molecular biology database collection: 2007 update," *Nucleic Acids Research*, vol. 35, pp. D3–D4, Jan. 2007.

[37] G. R. Cochrane and M. Y. Galperin, "The 2010 nucleic acids research database issue and online database collection: a community of data resources," *Nucleic Acids Research*, vol. 38, pp. D1–D4, Jan. 2010.

[38] C. Burks, "Molecular biology database list," *Nucleic Acids Research*, vol. 27, pp. 1–9, Jan. 1999.

[39] A. D. Baxevanis, "The molecular biology database collection: an online compilation of relevant database resources," *Nucleic Acids Research*, vol. 28, pp. 1–7, Jan. 2000.

[40] A. D. Baxevanis, "The molecular biology database collection: an updated compilation of biological database resources," *Nucleic Acids Research*, vol. 29, pp. 1–10, Jan. 2001.

[41] A. D. Baxevanis, "The molecular biology database collection: 2002 update," *Nucleic Acids Research*, vol. 30, pp. 1–12, Jan. 2002.

[42] A. D. Baxevanis, "The molecular biology database collection: 2003 update," *Nucleic Acids Research*, vol. 31, pp. 1–12, Jan. 2003.

[43] M. Y. Galperin, "The molecular biology database collection: 2004 update," *Nucleic Acids Research*, vol. 32, pp. D3–D22, Jan. 2004.

[44] M. Y. Galperin, "The molecular biology database collection: 2005 update," *Nucleic Acids Research*, vol. 33, pp. D5–D24, Jan. 2005.

[45] M. Y. Galperin, "The molecular biology database collection: 2008 update," *Nucleic Acids Research*, vol. 36, pp. D2–D4, Jan. 2008.

[46] M. Y. Galperin and G. R. Cochrane, "Nucleic acids research annual database issue and the NAR online molecular biology database collection in 2009," *Nucleic Acids Research*, vol. 37, pp. D1–D4, Jan. 2009.

[47] M. Y. Galperin and G. R. Cochrane, "The 2011 nucleic acids research database issue and the online molecular biology database collection," *Nucleic Acids Research*, vol. 39, pp. D1–D6, Jan. 2011.

[48] M. Y. Galperin and X. M. Fernández-Suárez, "The 2012 nucleic acids research database issue and the online molecular biology database collection," *Nucleic Acids Research*, vol. 40, pp. D1–D8, Jan. 2012.

[49] Oxford University Press, "Oxford Journals | Life Sciences | Nucleic Acids Research | SUBMITTING TO THE ANNUAL DATABASE ISSUE." `http://www.oxfordjournals.org/our_journals/nar/for_authors/msprep_database.html` [Online. Accessed 2013-02-28], Apr. 2013.

[50] The Internet Archive, "Internet archive: Digital library of free books, movies, music & wayback machine." http://archive.org/index.php [Online. Accessed 2013-04-03 22:04:13], Apr. 2013.

[51] Infobiogen, "Informations - Annonces - Alertes." http://morissardjerome.free.fr/infobiogen/www.infobiogen.fr/annonces/annonces.html [Online. Accessed 2013-02-28], Jan. 2006.

[52] C. Chandras, T. Weaver, M. Zouberakis, D. Smedley, K. Schughart, N. Rosenthal, J. M. Hancock, G. Kollias, P. N. Schofield, and V. Aidinis, "Models for financial sustainability of biological databases and resources," *Database*, vol. 2009, pp. bap017+, Jan. 2009.

[53] A. Bairoch, "SWISS-PROT funding crisis, help us !." http://web.expasy.org/docs/crisis96/help-sprot.html [Online. Accessed 2013-02-28], Apr. 1996.

[54] N. Williams, "Unique protein database imperiled," *Science*, vol. 272, pp. 946+, May 1996.

[55] Y. Kodama, M. Shumway, R. Leinonen, and International Nucleotide Sequence Database Collaboration, "The sequence read archive: explosive growth of sequencing data.," *Nucleic acids research*, vol. 40, Jan. 2012.

[56] D. Lipman, P. Flicek, S. Salzberg, M. Gerstein, and R. Knight, "Closure of the NCBI SRA and implications for the long-term future of genomics data storage.," *Genome biology*, vol. 12, Mar. 2011.

[57] The National Center for Biotechnology Information, "NCBI To Discontinue Sequence Read Archive and Peptidome. NLM Technical Bulletin. 2011 Jan-Feb." http://www.nlm.nih.gov/pubs/techbull/jf11/jf11_ncbi_reprint_sra.html [Online. Accessed 2013-02-28], Feb. 2011.

[58] S. International, "Submit a letter of support for EcoCyc." http://bioinformatics.ai.sri.com/ptools/ecocyc-letters-of-support2.shtml [Online. Accessed 2013-07-10].

[59] L. B. Ellis and D. Kalumbi, "Financing a future for public biological data.," *Bioinformatics*, vol. 15, pp. 717–722, Sept. 1999.

[60] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe, "KEGG for integration and interpretation of large-scale molecular data sets.," *Nucleic acids research*, vol. 40, pp. D109–D114, Jan. 2012.

[61] M. Kanehisa, "KEGG: Kyoto Encyclopedia of Genes and Genomes." http://www.genome.jp/kegg/docs/plea.html [Online. Accessed 2013-04-08], May 2011.

[62] M. Baker, "Databases fight funding cuts," *Nature*, vol. 489, p. 19, Sept. 2012.

[63] L. Ji, T. Barrett, O. Ayanbule, D. B. Troup, D. Rudnev, R. N. Muertter, M. Tomashevsky, A. Soboleva, and D. J. Slotta, "NCBI peptidome: a new repository for mass spectrometry proteomics data.," *Nucleic acids research*, vol. 38, pp. D731–D735, Jan. 2010.

[64] National Center for Biotechnology Information, "NCBI - Gone." `http://www.ncbi.nlm.nih.gov/peptidome` [Online. Accessed 2013-04-03], Apr. 2013.

[65] P. J. Kersey, J. Duarte, A. Williams, Y. Karavidopoulou, E. Birney, and R. Apweiler, "The international protein index: An integrated database for proteomics experiments," *Proteomics*, vol. 4, pp. 1985–1988, July 2004.

[66] J. Griss, M. Martín, C. O'Donovan, R. Apweiler, H. Hermjakob, and J. A. A. Vizcaíno, "Consequences of the discontinuation of the international protein index (IPI) database and its substitution by the UniProtKB "complete proteome" sets.," *Proteomics*, vol. 11, pp. 4434–4438, Nov. 2011.

[67] J. D. Wren, "404 not found: the stability and persistence of URLs published in MEDLINE," *Bioinformatics*, vol. 20, pp. 668–672, Mar. 2004.

[68] J. D. Wren, "URL decay in MEDLINE – a 4-year follow-up study," *Bioinformatics*, vol. 24, pp. 1381–1385, June 2008.

[69] J. L. Markley, H. Akutsu, T. Asakura, M. Baldus, R. Boelens, A. Bonvin, R. Kaptein, A. Bax, I. Bezsonova, M. R. Gryk, J. C. Hoch, D. M. Korzhnev, M. W. Maciejewski, D. Case, W. J. Chazin, T. A. Cross, S. Dames, H. Kessler, O. Lange, T. Madl, B. Reif, M. Sattler, D. Eliezer, A. Fersht, J. Forman-Kay, L. E. Kay, J. Fraser, J. Gross, T. Kortemme, A. Sali, T. Fujiwara, K. Gardner, X. Luo, J. Rizo-Rey, M. Rosen, R. R. Gil, C. Ho, G. Rule, A. M. Gronenborn, R. Ishima, J. Klein-Seetharaman, P. Tang, P. van der Wel, Y. Xu, S. Grzesiek, S. Hiller, J. Seelig, E. D. Laue, H. Mott, D. Nietlispach, I. Barsukov, L.-Y. Y. Lian, D. Middleton, T. Blumenschein, G. Moore, I. Campbell, J. Schnell, I. J. J. Vakonakis, A. Watts, M. R. Conte, J. Mason, M. Pfuhl, M. R. Sanderson, J. Craven, M. Williamson, C. Dominguez, G. Roberts, U. Günther, M. Overduin, J. Werner, P. Williamson, C. Blindauer, M. Crump, P. Driscoll, T. Frenkiel, A. Golovanov, S. Matthews, J. Parkinson, D. Uhrin, M. Williams, D. Neuhaus, H. Oschkinat, A. Ramos, D. E. Shaw, C. Steinbeck, M. Vendruscolo, G. W. Vuister, K. J. Walters, H. Weinstein, K. Wüthrich, and S. Yokoyama, "In support of the BMRB.," *Nature structural & molecular biology*, vol. 19, pp. 854–860, Sept. 2012.

[70] L. Lane, G. Argoud-Puy, A. Britan, I. Cusin, P. D. Duek, O. Evalet, A. Gateau, P. Gaudet, A. Gleizes, A. Masselot, C. Zwahlen, and A. Bairoch, "neXtProt: a knowledge platform for human proteins," *Nucleic Acids Research*, vol. 40, pp. D76–D83, Jan. 2012.

[71] I. M. Keseler, A. Mackie, M. Peralta-Gil, A. Santos-Zavaleta, S. Gama-Castro, C. Bonavides-Martínez, C. Fulcher, A. M. Huerta, A. Kothari, M. Krummenacker, *et al.*, "EcoCyc: fusing model organism databases with systems biology," *Nucleic acids research*, vol. 41, no. D1, pp. D605–D612, 2013.

[72] The UniProt Consortium, "Databases cross-referenced in UniProtKB." http://www.uniprot.org/docs/dbxref [Online. Accessed 2013-04-03], Apr. 2013.

[73] E. Gasteiger, E. Jung, and A. Bairoch, "SWISS-PROT: connecting biomolecular knowledge via a protein database.," *Current issues in molecular biology*, vol. 3, pp. 47–55, July 2001.

[74] R. Cyganiak and A. Jentzsch, "The Linking Open Data cloud diagram." http://www.lod-cloud.net/ [Online. Accessed 2013-04-03], Sept. 2011.

[75] Dublin Core Metadata Initiative, "DCMI Metadata Basics." http://www.dublincore.org/metadata-basics [Online. Accessed 2014-09-14], 2014.

[76] N. Press, "Understanding metadata," *National Information Standards*, vol. 20, 2004.

[77] S. Weibel, J. Kunze, C. Lagoze, and M. Wolf, "Dublin core metadata for resource discovery," *Internet Engineering Task Force RFC*, vol. 2413, no. 222, p. 132, 1998.

[78] J. Wieczorek, D. Bloom, R. Guralnick, S. Blum, M. Döring, R. Giovanni, T. Robertson, and D. Vieglais, "Darwin Core: An evolving community-developed biodiversity data standard," *PLoS One*, vol. 7, no. 1, p. e29715, 2012.

[79] J. Bolleman, A. Gateau, S. Gehant, and N. Redaschi, "Provenance and evidence in UniProtKB," *arXiv preprint arXiv:1012.1660*, 2010.

[80] EMBL-EBI, "about:webservices [EMBL-EBI Web Services]." http://www.ebi.ac.uk/Tools/webservices [Online. Accessed 2014-09-14], 2014.

[81] The UniProt Consortium, "UniProt." http://beta.sparql.uniprot.org/ [Online. Accessed 2014-09-14], 2014.

[82] J. C. Wooley and H. S. Lin, *On the Nature of Biological Data*, ch. 3, pp. 35–56. Washington DC: National Academies Press, 2005.

[83] A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C. A. Ball, H. C. Causton, T. Gaasterland, P. Glenisson, F. C. Holstege, I. F. Kim, V. Markowitz, J. C. Matese, H. Parkinson, A. Robinson, U. Sarkans, S. Schulze-Kremer, J. Stewart, R. Taylor, J. Vilo, and M. Vingron, "Minimum information about a microarray experiment (MIAME)-toward standards for microarray data.," *Nature genetics*, vol. 29, pp. 365–371, Dec. 2001.

[84] L. T. Corporation, " Welcome | Ion Community Home ." http://ioncommunity.lifetechnologies.com [Online. Accessed 2013-07-10].

[85] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool.," *Journal of molecular biology*, vol. 215, pp. 403–410, Oct. 1990.

[86] S. Brenner, "Life sentences: Ontology recapitulates philology," *SCIENTIST-PHILADELPHIA-*, vol. 16, no. 6, pp. 12–12, 2002.

[87] C. Kanz, P. Aldebert, N. Althorpe, W. Baker, A. Baldwin, K. Bates, P. Browne, A. van den Broek, M. Castro, G. Cochrane, K. Duggan, R. Eberhardt, N. Faruque, J. Gamble, F. G. Diez, N. Harte, T. Kulikova, Q. Lin, V. Lombard, R. Lopez, R. Mancuso, M. McHale, F. Nardone, V. Silventoinen, S. Sobhany, P. Stoehr, M. A. Tuli, K. Tzouvara, R. Vaughan, D. Wu, W. Zhu, and R. Apweiler, "The EMBL nucleotide sequence database," *Nucleic Acids Research*, vol. 33, pp. D29–D33, Jan. 2005.

[88] G. Rustici, N. Kolesnikov, M. Brandizi, T. Burdett, M. Dylag, I. Emam, A. Farne, E. Hastings, J. Ison, M. Keays, N. Kurbatova, J. Malone, R. Mani, A. Mupo, R. P. Pereira, E. Pilicheva, J. Rung, A. Sharma, Y. A. Tang, T. Ternent, A. Tikhonov, D. Welter, E. Williams, A. Brazma, H. Parkinson, and U. Sarkans, "ArrayExpress update – trends in database growth and links to data analysis tools," *Nucleic Acids Research*, vol. 41, pp. D987–D990, Jan. 2013.

[89] M. Magrane, "UniProt: Quick tour | EBI Train online." `http://www.ebi.ac.uk/training/online/outline_print/2051/all` [Online. Accessed 2013-02-10], 2013.

[90] Collins English Dictionary, "Definition of annotation." `http://www.collinsdictionary.com/dictionary/english/annotation` [Online. Accessed 2013-02-10], 2013.

[91] A. Bairoch and R. Apweiler, "The SWISS-PROT protein sequence data bank and its new supplement TREMBL," *Nucleic Acids Research*, vol. 24, pp. 21–25, Jan. 1996.

[92] P. W. Lord, J. R. Reich, A. Mitchell, R. D. Stevens, and C. A. Goble, "PRECIS: an automated pipeline for producing concise reports about proteins," in *Bioinformatics and Bioengineering Conference, 2001. Proceedings of the IEEE 2nd International Symposium on*, pp. 57–64, IEEE, Nov. 2001.

[93] B. Alberts, D. Bray, K. Hopkin, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Essential Cell Biology.* Garland Publishing Inc, 3rd revised edition ed., 2009.

[94] J. G. Sgouros and R. M. Twyman, "Gene finding," in *Bioinformatics: Genes, Proteins & Computers* (C. Orengo, D. T. Jones, and J. M. Thornton, eds.), pp. 18–28, BIOS Scientific Publishers, 2002.

[95] L. Stein, "Genome annotation: from sequence to biology.," *Nature reviews. Genetics*, vol. 2, pp. 493–503, July 2001.

[96] U. Hinz and UniProt Consortium, "From protein sequences to 3D-structures and beyond: the example of the UniProt knowledgebase.," *Cellular and molecular life sciences : CMLS*, vol. 67, pp. 1049–1064, Apr. 2010.

[97] F. Jungo, L. Bougueleret, I. Xenarios, and S. Poux, "The UniProtKB/Swiss-Prot Tox-Prot program: A central hub of integrated venom protein data," *Toxicon*, vol. 60, pp. 551–557, Sept. 2012.

[98] M. Yandell and D. Ence, "A beginner's guide to eukaryotic genome annotation," *Nature Reviews Genetics*, vol. 13, pp. 329–342, Apr. 2012.

[99] P. Stothard and D. S. Wishart, "Automated bacterial genome analysis and annotation," *Current Opinion in Microbiology*, vol. 9, pp. 505–510, Oct. 2006.

[100] P. E. Bourne and J. McEntyre, "Biocurators: Contributors to the world of science," *PLoS Comput Biol*, vol. 2, pp. e142+, Oct. 2006.

[101] The Reference Genome Group of the Gene Ontology Consortium, "The Gene Ontology's Reference Genome Project: A Unified Framework for Functional Annotation across Species," *PLoS Comput Biol*, vol. 5, pp. e1000431+, July 2009.

[102] P. McQuilton, S. E. St Pierre, J. Thurmond, and FlyBase Consortium, "FlyBase 101–the basics of navigating FlyBase.," *Nucleic acids research*, vol. 40, Jan. 2012.

[103] P. McQuilton, "Opportunities for text mining in the FlyBase genetic literature curation workflow," *Database*, vol. 2012, Jan. 2012.

[104] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The protein data bank," *Nucleic Acids Research*, vol. 28, pp. 235–242, Jan. 2000.

[105] K. Burkhardt, B. Schneider, and J. Ory, "A biocurator perspective: annotation at the Research Collaboratory for Structural Bioinformatics Protein Data Bank," *PLoS Computational Biology*, vol. 2, pp. e99+, Oct. 2006.

[106] E. Curry, A. Freitas, and S. O'Riáin, "The role of Community-Driven data curation for enterprises," in *Linking Enterprise Data* (D. Wood, ed.), pp. 25–47, Springer US, 2010.

[107] H. Berman, K. Henrick, and H. Nakamura, "Announcing the worldwide Protein Data Bank," *Nat Struct Mol Biol*, vol. 10, p. 980, Dec. 2003.

[108] The UniProt Consortium, "Uniprot staff (full-time and part-time)." `http://www.uniprot.org/help/uniprotstaff` [Online. Accessed 2013-02-24], 2013.

[109] D. O. Inglis, M. B. Arnaud, J. Binkley, P. Shah, M. S. Skrzypek, F. Wymore, G. Binkley, S. R. Miyasato, M. Simison, and G. Sherlock, "The Candida genome database incorporates multiple Candida species: multispecies search and analysis tools with curated gene and protein information for Candida albicans and Candida glabrata," *Nucleic Acids Research*, vol. 40, pp. D667–D674, Jan. 2012.

[110] Candida Genome Database, "*Candida* Genome Database Staff." `http://www.candidagenome.org/staff.shtml` [Online. Accessed 2013-03-13], Mar. 2013.

[111] R. P. Huntley, E. C. Dimmer, and R. Apweiler, "Practical Applications of the Gene Ontology Resource," in *Problem Solving Handbook in Computational Biology and Bioinformatics*, pp. 319–339, Springer, 2011.

[112] L. Ding, A. Sabo, N. Berkowicz, R. R. Meyer, Y. Shotland, M. R. Johnson, K. H. Pepin, R. K. Wilson, and J. Spieth, "EAnnot: A genome annotation tool using experimental evidence," *Genome research*, vol. 14, no. 12, pp. 2503–2509, 2004.

[113] M. A. F. Noor, K. J. Zimmerman, and K. C. Teeter, "Data sharing: How much doesn't get submitted to GenBank?," *PLoS Biol*, vol. 4, pp. e228+, July 2006.

[114] T. Kulikova, P. Aldebert, N. Althorpe, W. Baker, K. Bates, P. Browne, A. van den Broek, G. Cochrane, K. Duggan, R. Eberhardt, N. Faruque, M. Garcia-Pastor, N. Harte, C. Kanz, R. Leinonen, Q. Lin, V. Lombard, R. Lopez, R. Mancuso, M. McHale, F. Nardone, V. Silventoinen, P. Stoehr, G. Stoesser, M. A. Tuli, K. Tzouvara, R. Vaughan, D. Wu, W. Zhu, and R. Apweiler, "The EMBL nucleotide sequence database," *Nucleic Acids Research*, vol. 32, pp. D27–D30, Jan. 2004.

[115] American Society for Biochemistry and Molecular Biology, "The Journal of Biological Chemistry: Instructions for Authors." `http://www.jbc.org/site/misc/ifora.xhtml` [Online. Accessed 2013-02-28], 2013.

[116] Nature Publishing Group, "Availability of data & materials : authors & referees @ npg." `http://www.nature.com/authors/policies/availability.html` [Online. Accessed 2013-02-28], 2013.

[117] Oxford University Press, "Oxford Journals | Life Sciences | Nucleic Acids Research | Preparing and Submitting Your Manuscript." `http://www.oxfordjournals.org/our_journals/nar/for_authors/msprep_submission.html` [Online. Accessed 2013-02-28], 2013.

[118] European Bioinformatics Institute, "SPIN: Welcome to SPIN." `http://www.ebi.ac.uk/swissprot/Submissions/spin/` [Online. Accessed 2013-02-28], 2013.

[119] European Bioinformatics Institute, "SPIN Help Pages." `http://www.ebi.ac.uk/swissprot/Submissions/spin_help/spin_help.html` [Online. Accessed 2013-02-28], 2013.

[120] The UniProt Consortium, "Where do the UniProtKB protein sequences come from?." `http://www.uniprot.org/faq/37` [Online. Accessed 2013-03-17], 2011.

[121] R. D. Finn and J. Tate, "No, seriously, we've made a release." `http://xfam.wordpress.com/2011/04/01/no-seriously-weve-made-a-release/` [Online. Accessed 2013-02-27], 2011.

[122] R. D. Finn, J. Mistry, J. Tate, P. Coggill, A. Heger, J. E. Pollington, O. L. Gavin, P. Gunasekaran, G. Ceric, K. Forslund, L. Holm, E. L. L. Sonnhammer, S. R. Eddy, and A. Bateman, "The pfam protein families database," *Nucleic Acids Research*, vol. 38, pp. D211–D222, Jan. 2010.

[123] Pfam, "Pfam: Help." `http://pfam.sanger.ac.uk/help#tabview=tab7` [Online. Accessed 2013-02-27], 2013.

[124] C. G. Elsik, K. C. Worley, L. Zhang, N. V. Milshina, H. Jiang, J. T. Reese, K. L. Childs, A. Venkatraman, C. M. Dickens, G. M. Weinstock, and R. A. Gibbs, "Community annotation: procedures, protocols, and supporting tools.," *Genome research*, vol. 16, pp. 1329–1333, Nov. 2006.

[125] D. L. Hartl, "Fly meets shotgun: shotgun wins.," *Nature genetics*, vol. 24, pp. 327–328, Apr. 2000.

[126] E. Pennisi, "Ideas fly at gene-finding jamboree.," *Science (New York, N.Y.)*, vol. 287, pp. 2182–2184, Mar. 2000.

[127] P. Dehal, Y. Satou, R. K. Campbell, J. Chapman, B. Degnan, A. De Tomaso, B. Davidson, A. Di Gregorio, M. Gelpke, D. M. Goodstein, N. Harafuji, K. E. M. Hastings, I. Ho, K. Hotta, W. Huang, T. Kawashima, P. Lemaire, D. Martinez, I. A. Meinertzhagen, S. Necula, M. Nonaka, N. Putnam, S. Rash, H. Saiga, M. Satake, A. Terry, L. Yamada, H.-G. Wang, S. Awazu, K. Azumi, J. Boore, M. Branno, S. Chin-bow, R. DeSantis, S. Doyle, P. Francino, D. N. Keys, S. Haga, H. Hayashi, K. Hino, K. S. Imai, K. Inaba, S. Kano, K. Kobayashi, M. Kobayashi, B.-I. Lee, K. W. Makabe, C. Manohar, G. Matassi, M. Medina, Y. Mochizuki, S. Mount, T. Morishita, S. Miura, A. Nakayama, S. Nishizaka, H. Nomoto, F. Ohta, K. Oishi, I. Rigoutsos, M. Sano, A. Sasaki, Y. Sasakura, E. Shoguchi, T. Shin-i, A. Spagnuolo, D. Stainier, M. M. Suzuki, O. Tassy, N. Takatori, M. Tokuoka, K. Yagi, F. Yoshizaki, S. Wada, C. Zhang, P. D. Hyatt, F. Larimer, C. Detter, N. Doggett, T. Glavina, T. Hawkins, P. Richardson, S. Lucas, Y. Kohara, M. Levine, N. Satoh, and D. S. Rokhsar, "The draft genome of ciona intestinalis: Insights into chordate and vertebrate origins," *Science*, vol. 298, pp. 2157–2167, Dec. 2002.

[128] M. Riley, T. Abe, M. B. Arnaud, M. K. Berlyn, F. R. Blattner, R. R. Chaudhuri, J. D. Glasner, T. Horiuchi, I. M. Keseler, T. Kosuge, H. Mori, N. T. Perna, G. Plunkett, K. E. Rudd, M. H. Serres, G. H. Thomas, N. R. Thomson, D. Wishart, and B. L. Wanner, "Escherichia coli K-12: a cooperatively developed annotation snapshot–2005.," *Nucleic acids research*, vol. 34, pp. 1–9, Jan. 2006.

[129] T. Itoh, T. Tanaka, R. A. Barrero, C. Yamasaki, Y. Fujii, P. B. Hilton, B. A. Antonio, H. Aono, R. Apweiler, R. Bruskiewich, T. Bureau, F. Burr, A. C. de Oliveira, G. Fuks, T. Habara, G. Haberer, B. Han, E. Harada, A. T. Hiraki, H. Hirochika, D. Hoen, H. Hokari, S. Hosokawa, Y. Hsing, H. Ikawa, K. Ikeo, T. Imanishi, Y. Ito, P. Jaiswal, M. Kanno, Y. Kawahara, T. Kawamura, H. Kawashima, J. P. Khurana, S. Kikuchi, S. Komatsu, K. O. Koyanagi, H. Kubooka, D. Lieberherr, Y.-C. Lin, D. Lonsdale, T. Matsumoto, A. Matsuya, W. R. McCombie, J. Messing, A. Miyao, N. Mulder, Y. Nagamura, J. Nam, N. Namiki, H. Numa, S. Nurimoto, C. O'Donovan, H. Ohyanagi, T. Okido, S. OOta, N. Osato, L. E. Palmer, F. Quetier, S. Raghuvanshi, N. Saichi, H. Sakai, Y. Sakai, K. Sakata, T. Sakurai, F. Sato, Y. Sato, H. Schoof, M. Seki, M. Shibata, Y. Shimizu, K. Shinozaki, Y. Shinso, N. K. Singh, B. Smith-White, J.-i. Takeda, M. Tanino, T. Tatusova, S. Thongjuea, F. Todokoro, M. Tsugane, A. K. Tyagi, A. Vanavichit, A. Wang, R. A. Wing, K. Yamaguchi, M. Yamamoto, N. Yamamoto, Y. Yu, H. Zhang, Q. Zhao, K. Higo, B. Burr, T. Gojobori, and

T. Sasaki, "Curated genome annotation of oryza sativa ssp. japonica and comparative genome analysis with arabidopsis thaliana," *Genome Research*, vol. 17, p. 000, Jan. 2007.

[130] H. Ohyanagi, T. Tanaka, H. Sakai, Y. Shigemoto, K. Yamaguchi, T. Habara, Y. Fujii, B. A. Antonio, Y. Nagamura, T. Imanishi, K. Ikeo, T. Itoh, T. Gojobori, and T. Sasaki, "The rice annotation project database (RAP-DB): hub for oryza sativa ssp. japonica genome information," *Nucleic Acids Research*, vol. 34, pp. D741–D744, Jan. 2006.

[131] V. Curwen, E. Eyras, T. D. Andrews, L. Clarke, E. Mongin, S. M. Searle, and M. Clamp, "The Ensembl automatic gene annotation system.," *Genome research*, vol. 14, pp. 942–950, May 2004.

[132] E. Camon, M. Magrane, D. Barrell, V. Lee, E. Dimmer, J. Maslen, D. Binns, N. Harte, R. Lopez, and R. Apweiler, "The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology," *Nucleic Acids Research*, vol. 32, pp. D262–D266, Jan. 2004.

[133] W. A. Baumgartner, K. B. Cohen, L. M. Fox, G. Acquaah-Mensah, and L. Hunter, "Manual curation is not sufficient for annotation of genomic databases.," *Bioinformatics (Oxford, England)*, vol. 23, pp. i41–48, July 2007.

[134] The UniProt Consortium, "Update on activities at the universal protein resource (UniProt) in 2013," *Nucleic Acids Research*, vol. 41, pp. D43–D47, Jan. 2013.

[135] A. M. Cohen and W. R. Hersh, "A survey of current work in biomedical text mining," *Briefings in Bioinformatics*, vol. 6, pp. 57–71, Mar. 2005.

[136] L. Hunter and K. B. Cohen, "Biomedical language processing: what's beyond PubMed?," *Molecular cell*, vol. 21, pp. 589–594, Mar. 2006.

[137] U.S. National Library of Medicine, "PubMed: MEDLINE®; Retrieval on the World Wide Web Fact Sheet." `http://www.nlm.nih.gov/pubs/factsheets/pubmed.html` [Online. Accessed 2013-03-12], March 2010.

[138] National Center for Biotechnology Information, "Home - PubMed - NCBI." `http://www.ncbi.nlm.nih.gov/pubmed/` [Online. Accessed 2013-02-03], 2013.

[139] G. D. Bader, D. Betel, and C. W. V. Hogue, "BIND: the Biomolecular Interaction Network Database," *Nucleic Acids Research*, vol. 31, pp. 248–250, Jan. 2003.

[140] I. Donaldson, J. Martin, B. de Bruijn, C. Wolting, V. Lay, B. Tuekam, S. Zhang, B. Baskin, G. Bader, K. Michalickova, T. Pawson, and C. Hogue, "PreBIND and textomy - mining the biomedical literature for protein-protein interactions using a support vector machine," *BMC Bioinformatics*, vol. 4, pp. 11+, Mar. 2003.

[141] C. Alfarano, C. E. Andrade, K. Anthony, N. Bahroos, M. Bajec, K. Bantoft, D. Betel, B. Bobechko, K. Boutilier, E. Burgess, K. Buzadzija, R. Cavero, C. D'Abreo, I. Donaldson, D. Dorairajoo, M. J. Dumontier, M. R. Dumontier,

V. Earles, R. Farrall, H. Feldman, E. Garderman, Y. Gong, R. Gonzaga, V. Grytsan, E. Gryz, V. Gu, E. Haldorsen, A. Halupa, R. Haw, A. Hrvojic, L. Hurrell, R. Isserlin, F. Jack, F. Juma, A. Khan, T. Kon, S. Konopinsky, V. Le, E. Lee, S. Ling, M. Magidin, J. Moniakis, J. Montojo, S. Moore, B. Muskat, I. Ng, J. P. Paraiso, B. Parker, G. Pintilie, R. Pirone, J. J. Salama, S. Sgro, T. Shan, Y. Shu, J. Siew, D. Skinner, K. Snyder, R. Stasiuk, D. Strumpf, B. Tuekam, S. Tao, Z. Wang, M. White, R. Willis, C. Wolting, S. Wong, A. Wrong, C. Xin, R. Yao, B. Yates, S. Zhang, K. Zheng, T. Pawson, B. F. Ouellette, and C. W. Hogue, "The Biomolecular Interaction Network Database and related tools 2005 update," *Nucleic acids research*, vol. 33, pp. D418–424, Jan. 2005.

[142] Swiss Institute of Bioinformatics, "UniProtKB/Swiss-Prot Release 2013_03 statistics." `http://web.expasy.org/docs/relnotes/relstat.html` [Online. Accessed 2013-03-12], March 2013.

[143] The UniProt Consortium, "Current Release Statistics ⟨ Uniprot ⟨ EMBL-EBI." `http://www.ebi.ac.uk/uniprot/TrEMBLstats` [Online. Accessed 2013-03-12], March 2013.

[144] T. U. Consortium, "The Universal Protein Resource (UniProt) in 2010," *Nucl. Acids Res.*, vol. 38, pp. D142–148, January 2010.

[145] F. Rinaldi, S. Clematide, S. Hafner, G. Schneider, G. Grigonytė, M. Romacker, and T. Vachon, "Using the OntoGene pipeline for the triage task of BioCreative 2012," *Database*, vol. 2013, pp. bas053+, Jan. 2013.

[146] M. Krallinger, R. A.-A. A. Erhardt, and A. Valencia, "Text-mining approaches in molecular biology and biomedicine.," *Drug discovery today*, vol. 10, pp. 439–445, Mar. 2005.

[147] P. K. Shah and P. Bork, "LSAT: learning about alternative transcripts in MEDLINE," *Bioinformatics*, vol. 22, pp. 857–865, Apr. 2006.

[148] T. A. Thanaraj, S. Stamm, F. Clark, J.-J. J. Riethoven, V. Le Texier, and J. Muilu, "ASD: the alternative splicing database.," *Nucleic acids research*, vol. 32, Jan. 2004.

[149] A. L. Veuthey, A. Bridge, J. Gobeill, P. Ruch, J. McEntyre, L. Bougueleret, and I. Xenarios, "Application of text-mining for updating protein post-translational modification annotation in UniProtKB," *BMC Bioinformatics*, vol. 14, pp. 104+, Mar. 2013.

[150] J. B. Bowes, K. A. Snyder, C. James-Zorn, V. G. Ponferrada, C. J. Jarabek, K. A. Burns, B. Bhattacharyya, A. M. Zorn, and P. D. Vize, "The Xenbase literature curation process," *Database*, vol. 2013, Jan. 2013.

[151] P. Radivojac, W. T. Clark, T. R. R. Oron, A. M. Schnoes, T. Wittkop, A. Sokolov, K. Graim, C. Funk, K. Verspoor, A. Ben-Hur, G. Pandey, J. M. Yunes, A. S. Talwalkar, S. Repo, M. L. Souza, D. Piovesan, R. Casadio, Z. Wang, J. Cheng, H. Fang, J. Gough, P. Koskinen, P. Törönen, J. Nokso-Koivisto,

L. Holm, D. Cozzetto, D. W. Buchan, K. Bryson, D. T. Jones, B. Limaye, H. Inamdar, A. Datta, S. K. Manjari, R. Joshi, M. Chitale, D. Kihara, A. M. Lisewski, S. Erdin, E. Venner, O. Lichtarge, R. Rentzsch, H. Yang, A. E. Romero, P. Bhat, A. Paccanaro, T. Hamp, R. Kaßner, S. Seemayer, E. Vicedo, C. Schaefer, D. Achten, F. Auer, A. Boehm, T. Braun, M. Hecht, M. Heron, P. Hönigschmid, T. A. Hopf, S. Kaufmann, M. Kiening, D. Krompass, C. Landerer, Y. Mahlich, M. Roos, J. Björne, T. Salakoski, A. Wong, H. Shatkay, F. Gatzmann, I. Sommer, M. N. Wass, M. J. Sternberg, N. Skunca, F. Supek, M. Bošnjak, P. Panov, S. Džeroski, T. Smuc, Y. A. Kourmpetis, A. D. van Dijk, C. T. J. Braak, Y. Zhou, Q. Gong, X. Dong, W. Tian, M. Falda, P. Fontana, E. Lavezzo, B. Di Camillo, S. Toppo, L. Lan, N. Djuric, Y. Guo, S. Vucetic, A. Bairoch, M. Linial, P. C. Babbitt, S. E. Brenner, C. Orengo, B. Rost, S. D. Mooney, and I. Friedberg, "A large-scale evaluation of computational protein function prediction.," *Nature methods*, vol. 10, pp. 221–227, Mar. 2013.

[152] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.," *Nature genetics*, vol. 25, pp. 25–29, May 2000.

[153] The Gene Ontology, "Guide to GO Evidence Codes." `http://www.geneontology.org/GO.evidence.shtml#comp-assigned` [Online. Accessed 2013-10-24], 2013.

[154] A. J. Bridge, D. Poggioli, and C. O'Donovan, "UniRule – automatic annotation in UniProtKB." Presentation at Biocuration 2010, 10 November 2010.

[155] Collins English Dictionary, "Definition of quality." `http://www.collinsdictionary.com/dictionary/english/quality` [Online. Accessed 2013-08-10], 2013.

[156] B. Ewing, L. Hillier, M. C. Wendl, and P. Green, "Base-calling of automated sequencer traces usingPhred. I. Accuracy assessment," *Genome research*, vol. 8, no. 3, pp. 175–185, 1998.

[157] B. Ewing and P. Green, "Base-calling of automated sequencer traces usingPhred. II. error probabilities," *Genome research*, vol. 8, no. 3, pp. 186–194, 1998.

[158] D. L. Tabb, "What's driving false discovery rates?," *Journal of proteome research*, vol. 7, no. 01, pp. 45–46, 2007.

[159] J. Rizkallah and D. D. Sin, "Integrative approach to quality assessment of medical journals using impact factor, eigenfactor, and article influence scores," *PloS one*, vol. 5, no. 4, p. e10204, 2010.

[160] E. C. Dimmer, R. P. Huntley, Y. Alam-Faruque, T. Sawford, C. O'Donovan, M. J. Martin, B. Bely, P. Browne, W. Mun Chan, R. Eberhardt, M. Gardner, K. Laiho, D. Legge, M. Magrane, K. Pichler, D. Poggioli, H. Sehra, A. Auchincloss, K. Axelsen, M.-C. C. Blatter, E. Boutet, S. Braconi-Quintaje, L. Breuza,

A. Bridge, E. Coudert, A. Estreicher, L. Famiglietti, S. Ferro-Rojas, M. Feuermann, A. Gos, N. Gruaz-Gumowski, U. Hinz, C. Hulo, J. James, S. Jimenez, F. Jungo, G. Keller, P. Lemercier, D. Lieberherr, P. Masson, M. Moinat, I. Pedruzzi, S. Poux, C. Rivoire, B. Roechert, M. Schneider, A. Stutz, S. Sundaram, M. Tognolli, L. Bougueleret, G. Argoud-Puy, I. Cusin, P. Duek-Roggli, I. Xenarios, and R. Apweiler, "The UniProt-GO annotation database in 2011.," *Nucleic acids research*, vol. 40, pp. D565–D570, Jan. 2012.

[161] L. du Plessis, N. Škunca, and C. Dessimoz, "The what, where, how and why of gene ontology – a primer for bioinformaticians," *Briefings in bioinformatics*, vol. 12, no. 6, pp. 723–735, 2011.

[162] T. J. J. Buza, F. M. M. McCarthy, N. Wang, S. M. M. Bridges, and S. C. C. Burgess, "Gene Ontology annotation quality analysis in model eukaryotes.," *Nucleic Acids Res*, January 2008.

[163] F. M. McCarthy, C. R. Gresham, T. J. Buza, P. Chouvarine, L. R. Pillai, R. Kumar, S. Ozkan, H. Wang, P. Manda, T. Arick, *et al.*, "AgBase: supporting functional modeling in agricultural organisms," *Nucleic acids research*, vol. 39, no. suppl 1, pp. D497–D506, 2011.

[164] D. Peddinti, E. Memili, and S. C. Burgess, "Proteomics-based systems biology modeling of bovine germinal vesicle stage oocyte and cumulus cell interaction," *PloS one*, vol. 5, no. 6, p. e11240, 2010.

[165] M. Van Bel, S. Proost, E. Wischnitzki, S. Movahedi, C. Scheerlinck, Y. Van de Peer, and K. Vandepoele, "Dissecting plant genomes with the PLAZA comparative genomics platform," *Plant physiology*, vol. 158, no. 2, pp. 590–600, 2012.

[166] GO Consortium, "Guide to GO Evidence Codes." `http://www.geneontology.org/GO.evidence.shtml` [Online. Accessed 2011-10-17], 2011.

[167] M. F. Rogers and A. Ben-Hur, "The use of Gene Ontology evidence codes in preventing classifier assessment bias," *Bioinformatics*, vol. 25, pp. 1173–1177, May 2009.

[168] D. Pal and D. Eisenberg, "Inference of protein function from protein structure," *Structure*, vol. 13, pp. 121–130, January 2005.

[169] C. Jones, A. Brown, and U. Baumann, "Estimating the annotation error rate of curated GO database sequence annotations," *BMC Bioinformatics*, vol. 8, no. 1, 2007.

[170] N. Škunca, A. Altenhoff, and C. Dessimoz, "Quality of Computationally Inferred Gene Ontology Annotations," *PLoS Comput Biol*, vol. 8, pp. e1002533+, May 2012.

[171] A. Gross, M. Hartung, T. Kirsten, and E. Rahm, "Estimating the quality of ontology-based annotations by considering evolutionary changes," in *DILS '09: Proceedings of the 6th International Workshop on Data Integration in the Life Sciences*, (Berlin, Heidelberg), pp. 71–87, Springer-Verlag, 2009.

[172] E. L. Clarke, S. Loguercio, B. M. Good, A. I. Su, *et al.*, "A task-based approach for Gene Ontology evaluation," *Journal of biomedical semantics*, vol. 4, no. Suppl 1, p. S4, 2013.

[173] L. R. Kalankesh, R. Stevens, and A. Brass, "The language of gene ontology: a Zipf's law analysis," *BMC Bioinformatics*, vol. 13, no. 1, pp. 127+, 2012.

[174] I. I. I. Artamonova, G. Frishman, M. S. S. Gelfand, and D. Frishman, "Mining sequence annotation databanks for association patterns.," *Bioinformatics*, vol. 21, Nov. 2005.

[175] P. W. Lord, R. D. Stevens, A. Brass, and C. A. Goble, "Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation," *Bioinformatics*, vol. 19, pp. 1275–1283, July 2003.

[176] B. T. Adler, K. Chatterjee, L. De Alfaro, M. Faella, I. Pye, and V. Raman, "Assigning trust to Wikipedia content," in *Proceedings of the 4th International Symposium on Wikis*, p. 26, ACM, 2008.

[177] H. Zeng, M. A. Alhossaini, L. Ding, R. Fikes, and D. L. McGuinness, "Computing trust from revision history," tech. rep., DTIC Document, 2006.

[178] D. Anthony, S. W. Smith, and T. Williamson, "The quality of open source production: Zealots and Good Samaritans in the case of Wikipedia," *Rationality and Society*, 2007.

[179] J. E. Blumenstock, "Size matters: word count as a measure of quality on wikipedia," in *Proceedings of the 17th international conference on World Wide Web*, pp. 1095–1096, ACM, 2008.

[180] M. Hu, E.-P. Lim, A. Sun, H. W. Lauw, and B.-Q. Vuong, "Measuring article quality in wikipedia: models and evaluation," in *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pp. 243–252, ACM, 2007.

[181] K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of documentation*, vol. 28, no. 1, pp. 11–21, 1972.

[182] S. Robertson, "Understanding inverse document frequency: on theoretical arguments for IDF," *Journal of documentation*, vol. 60, no. 5, pp. 503–520, 2004.

[183] J. Ramos, "Using tf-idf to determine word relevance in document queries," in *Proceedings of the First Instructional Conference on Machine Learning*, 2003.

[184] H. M. Laughlin, "SMOG grading-a new readability formula," *Journal of Reading*, vol. 12, no. 8, 1969.

[185] G. H. McLaughlin, "SMOG Readability Calculator." `http://www.harrymclaughlin.com/SMOG.htm` [Online. Accessed 2014-02-10], 2008.

[186] R. Flesch, "A new readability yardstick," *Journal of Applied Psychology*, vol. 32, no. 3, pp. 221–233, 1948.

[187] R. Flesch, *How to write plain English.* New York: Harper and Row, 1979.

[188] P. Fitzsimmons, B. Michael, J. Hulley, and G. Scott, "A readability assessment of online Parkinson's disease information," *The journal of the Royal College of Physicians of Edinburgh*, vol. 40, no. 4, pp. 292–296, 2010.

[189] K. A. Schriver, "Evaluating text quality: The continuum from text-focused to reader-focused methods," *Professional Communication, IEEE Transactions on*, vol. 32, no. 4, pp. 238–255, 1989.

[190] M. Corporation, "Test your document's readability." `http://office.microsoft.com/en-gb/word-help/test-your-document-s-readability-HP010148506.aspx` [Online. Accessed 2014-02-10].

[191] J. Hartley, *Evaluation of Human Work.* CRC Press, 3rd edition ed., 2005.

[192] I. M. Keseler, M. Skrzypek, D. Weerasinghe, A. Y. Chen, C. Fulcher, G.-W. Li, K. C. Lemmer, K. M. Mladinich, E. D. Chow, G. Sherlock, *et al.*, "Curation accuracy of model organism databases," *Database*, vol. 2014, p. bau058, 2014.

[193] M. Green and P. Karp, "Genome annotation errors in pathway databases due to semantic ambiguity in partial EC numbers," *Nucleic acids research*, vol. 33, no. 13, pp. 4035–4039, 2005.

[194] A. M. Schnoes, S. D. Brown, I. Dodevski, and P. C. Babbitt, "Annotation error in public databases: Misannotation of molecular function in enzyme superfamilies," *PLoS Comput Biol*, vol. 5, pp. e1000605+, December 2009.

[195] W. R. Gilks, B. Audit, D. De Angelis, S. Tsoka, and C. A. Ouzounis, "Modeling the percolation of annotation errors in a database of protein sequences," *Bioinformatics*, vol. 18, pp. 1641–1649, December 2002.

[196] C. H. Wu, H. Huang, L.-S. L. Yeh, and W. C. Barker, "Protein family classification and functional annotation," *Computational Biology and Chemistry*, vol. 27, no. 1, pp. 37–47, 2003.

[197] W. R. Gilks, B. Audit, D. de Angelis, S. Tsoka, and C. A. Ouzounis, "Percolation of annotation errors through hierarchically structured protein sequence databases," *Mathematical Biosciences*, vol. 193, pp. 223–234, Feb. 2005.

[198] P. Buneman, A. Chapman, J. Cheney, and S. Vansummeren, "A provenance model for manually curated data," in *Provenance and Annotation of Data*, pp. 162–170, Springer, 2006.

[199] S. Miles, "Electronically querying for the provenance of entities," in *Provenance and Annotation of Data*, pp. 184–192, Springer, 2006.

[200] P. Buneman and W.-C. Tan, "Provenance in databases," in *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, pp. 1171–1173, ACM, 2007.

[201] M. Gamble and C. Goble, "Quality, trust, and utility of scientific data on the web: Towards a joint model," in *Proceedings of the 3rd International Web Science Conference*, p. 15, ACM, 2011.

[202] P. Buneman, A. Chapman, and J. Cheney, "Provenance management in curated databases," in *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*, pp. 539–550, ACM, 2006.

[203] S. Magliacane, "Reconstructing provenance," in *The Semantic Web–ISWC 2012*, pp. 399–406, Springer, 2012.

[204] D. Metzler, Y. Bernstein, W. B. Croft, A. Moffat, and J. Zobel, "The recap system for identifying information flow," in *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 678–678, ACM, 2005.

[205] D. Metzler, Y. Bernstein, W. B. Croft, A. Moffat, and J. Zobel, "Similarity measures for tracking information flow," in *Proceedings of the 14th ACM international conference on Information and knowledge management*, CIKM '05, (New York, NY, USA), pp. 517–524, ACM, 2005.

[206] M. Bendersky and W. B. Croft, "Finding text reuse on the web," in *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pp. 262–271, ACM, 2009.

[207] iParadigms, LLC, "Turnitin — Originality Check, Online Grading & Peer Review." `http://www.turnitin.com` [Online. Accessed 2014-09-14], 2014.

[208] Dupli Checker, "Plagiarism Checker — Free Online Software For Plagiarism Detection." `http://www.duplichecker.com/` [Online. Accessed 2014-09-14], 2014.

[209] Y. Nakamura, G. Cochrane, and I. Karsch-Mizrachi, "The international nucleotide sequence database collaboration," *Nucleic acids research*, vol. 41, no. D1, pp. D21–D24, 2013.

[210] The UniProt Consortium, "Reorganizing the protein space at the universal protein resource (UniProt)," *Nucleic Acids Research*, vol. 40, pp. D71–D75, Jan. 2012.

[211] R. Leinonen, F. G. Diez, D. Binns, W. Fleischmann, R. Lopez, and R. Apweiler, "UniProt archive," *Bioinformatics*, vol. 20, no. 17, pp. 3236–3237, 2004.

[212] B. E. Suzek, H. Huang, P. McGarvey, R. Mazumder, and C. H. Wu, "UniRef: comprehensive and non-redundant UniProt reference clusters," *Bioinformatics*, vol. 23, no. 10, pp. 1282–1288, 2007.

[213] The UniProt Consortium, "UniMES." `http://www.uniprot.org/help/unimes` [Online. Accessed 2013-10-24], 2013.

[214] The UniProt Consortium, "UniProt release 11.0." `http://www.uniprot.org/news/2007/05/29/release` [Online. Accessed 2013-10-24], 2007.

[215] C. H. Wu, R. Apweiler, A. Bairoch, D. A. Natale, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, R. Mazumder, C. O'Donovan, N. Redaschi, and B. Suzek, "The universal protein resource (UniProt): an expanding universe of protein information.," *Nucleic acids research*, vol. 34, pp. D187–D191, Jan. 2006.

[216] R. Apweiler, A. Bairoch, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, D. A. Natale, C. O'Donovan, N. Redaschi, and L. L. Yeh, "UniProt: the Universal Protein knowledgebase," *Nucleic Acids Research*, vol. 32, pp. D115–D119, Jan. 2004.

[217] Y. L. Yip, N. Lachenal, V. Pillet, and a.-L. Veuthey, "Retrieving mutation-specific information for human proteins in UniProt/Swiss-Prot Knowledgebase," *Journal of Bioinformatics and Computational Biology*, vol. 05, pp. 1215–1231, Dec. 2007.

[218] I. Solt, D. Tikk, and U. Leser, "Species identification for gene name normalization," *BMC Bioinformatics*, vol. 11, no. Suppl 5, pp. P5+, 2010.

[219] The UniProt Consortium, "About UniProt." `http://www.uniprot.org/help/about` [Online. Accessed 2013-02-24], 2013.

[220] The UniProt Consortium, "Uniprot release 1.0." `http://www.uniprot.org/news/2003/12/15/release` [Online. Accessed 2013-02-10], Dec. 2003.

[221] A. Bairoch and B. Boeckmann, "The SWISS-PROT protein sequence data bank, recent developments.," *Nucleic acids research*, vol. 21, pp. 3093–3096, July 1993.

[222] A. Bairoch, "Serendipity in bioinformatics, the tribulations of a Swiss bioinformatician through exciting times!," *Bioinformatics (Oxford, England)*, vol. 16, pp. 48–64, Jan. 2000.

[223] The UniProt Consortium, "UniProt Knowledgebase User Manual." `http://web.expasy.org/docs/userman.html` [Online. Accessed 2013-02-10], 2011.

[224] S. Altairac, "The beginnings of a database. An interview with Prof. Amos Bairoch," *Protéines à la "Une"*, vol. 18, Aug. 2006.

[225] J. Blondeau, "SIB establishes new group and announces new head of Swiss-Prot." `http://www.isb-sib.ch/news-a-events/news/344-sib-swiss-institute-of-bioinformatics-establishes-new-group-for-functional-characterisation-of-human-proteins-and-announces-new-head-of-swiss-prot.html` [Online. Accessed 2013-02-24], June 2009.

[226] A. Bairoch and R. Apweiler, "The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000," *Nucleic Acids Research*, vol. 28, pp. 45–48, Jan. 2000.

[227] W. Lathe, J. Williams, M. Mangan, and D. Karolchik, "Genomic data resources: challenges and promises," *Nature Education*, vol. 1, no. 3, 2008.

[228] K. A. Wetterstrand, "DNA Sequencing Costs." http://www.genome.gov/sequencingcosts/ [Online. Accessed 2013-01-15], Nov. 2012.

[229] R. Apweiler and T. Etzold, "ANNOUNCEMENT: Beta release of TREMBL, a supplement to SWISS-PROT." http://www.bio.net/mm/embl-db/1996-April/000577.html [Online. Accessed 2013-02-24], 1996.

[230] The UniProt Consortium, "How redundant are the UniProt databases?." http://www.uniprot.org/faq/33 [Online. Accessed 2013-03-17], 2012.

[231] The UniProt Consortium, "Why have some UniProtKB accession numbers been deleted? How can I track them?." http://www.uniprot.org/faq/11 [Online. Accessed 2013-03-17], March 2010.

[232] R. Apweiler, M. J. Martin, C. O'Donovan, M. Magrane, Y. Alam-Faruque, R. Antunes, D. Barrell, B. Bely, M. Bingley, D. Binns, *et al.*, "Ongoing and future developments at the Universal Protein Resource.," *Nucleic acids research*, vol. 39, no. Database issue, pp. D214–9, 2011.

[233] The UniProt Consortium, "UniProt release 2010_09." http://www.uniprot.org/news/2010/08/10/release [Online. Accessed 2013-10-24], August 2010.

[234] S. Poux, M. Magrane, and The UniProt Consortium, "Manual biocuration in UniProtKB/Swiss-Prot." Poster at the 6th International Biocuration Conference 2013, Churchill College, Cambridge, UK, April 2013.

[235] The UniProt Consortium, "UniProt Manual Curation SOP." http://www.uniprot.org/docs/sop_manual_curation.pdf [Online. Accessed 2013-10-24], June 2011.

[236] The UniProt Consortium, "How do we manually annotate a UniProtKB entry?." http://www.uniprot.org/faq/45 [Online. Accessed 2013-03-17], September 2011.

[237] The UniProt Consortium, "Automatic annotation program." www.uniprot.org/program/automatic_annotation [Online. Accessed 2013-03-17], 2013.

[238] A. Gattiker, K. Michoud, C. Rivoire, A. H. Auchincloss, E. Coudert, T. Lima, P. Kersey, M. Pagni, C. J. Sigrist, C. Lachaize, *et al.*, "Automated annotation of microbial proteomes in SWISS-PROT," *Computational biology and chemistry*, vol. 27, no. 1, pp. 49–58, 2003.

[239] I. Pedruzzi, C. Rivoire, A. H. Auchincloss, E. Coudert, G. Keller, E. De Castro, D. Baratin, B. A. Cuche, L. Bougueleret, S. Poux, *et al.*, "HAMAP in 2013, new developments in the protein family classification and annotation system," *Nucleic acids research*, vol. 41, no. D1, pp. D584–D589, 2013.

[240] T. Lima, A. H. Auchincloss, E. Coudert, G. Keller, K. Michoud, C. Rivoire, V. Bulliard, E. De Castro, C. Lachaize, D. Baratin, *et al.*, "HAMAP: a database of completely sequenced microbial proteome sets and manually curated microbial protein families in UniProtKB/Swiss-Prot," *Nucleic acids research*, vol. 37, no. suppl 1, pp. D471–D478, 2009.

[241] D. A. Natale, C. Vinayaka, and C. H. Wu, "Large-scale, classification-driven, rule-based functional annotation of proteins," *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics*, 2004.

[242] S. Vasudevan, C. Vinayaka, D. A. Natale, H. Huang, R. Y. Kahsay, and C. H. Wu, "Structure-guided rule-based annotation of protein functional sites in UniProt knowledgebase," in *Bioinformatics for Comparative Proteomics*, pp. 91–105, Springer, 2011.

[243] W. Fleischmann, A. Gateau, R. Apweiler, *et al.*, "A novel method for automatic functional annotation of proteins.," *Bioinformatics*, vol. 15, no. 3, pp. 228–233, 1999.

[244] E. Kretschmann, W. Fleischmann, and R. Apweiler, "Automatic rule generation for protein annotation with the C4. 5 data mining algorithm applied on SWISS-PROT," *Bioinformatics*, vol. 17, no. 10, pp. 920–926, 2001.

[245] The UniProt Consortium, "UniProt release 2010_04." `http://www.uniprot.org/news/2010/03/23/release` [Online. Accessed 2013-10-24], March 2010.

[246] G. K. Zipf, *Human Behaviour and the Principle of Least Effort*. Hafner Publishing Co Ltd, new issue of 1949 ed ed., 1949.

[247] G. K. Zipf, *The Psycho-Biology of Language*. Houghton Mifflin Company, first ed., 1935.

[248] G. Altmann, K. H. Best, L. Hřebíček, R. Köhler, O. Rottmann, G. Wimmer, and A. Ziegler, "Glottometrics 3. To Honor G. K. Zipf," *Glottometrics*, vol. 3, no. 3, pp. 1–155, 2002.

[249] G. K. Zipf, *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Martino Fine Books, June 2012.

[250] C. Tullo and J. R. Hurford, "Modelling zipfian distributions in language," in *Proceedings of Language Evolution and Computation Workshop/Course at ESSLLI* (S. Kirby, ed.), (Vienna), pp. 62–75, 2003.

[251] C. S. Gillespie, "Fitting heavy tailed distributions: the poweRlaw package," *Journal of Statistical Software*, forthcoming.

[252] Project Gutenbert, "Sense and Sensibility by Jane Austen." `http://www.gutenberg.org/ebooks/161` [Online. Accessed 2013-05-15], May 2013.

[253] Wikipedia, "Wikipedia, the free encyclopedia." `http://en.wikipedia.org/wiki/Main_Page` [Online. Accessed 2013-03-31], Mar. 2013.

[254] L. Q. Ha, E. I. Sicilia-garcia, J. Ming, and F. J. Smith, "Extension of zipf's law to words and phrases," in *Proceedings of the 19th International Conference on Computational Linguistics (COLING*, pp. 315–320, 2002.

[255] M. E. J. Newman, "Power laws, Pareto distributions and Zipf's law," *Contemporary Physics*, vol. 46, pp. 323–351, Sept. 2005.

[256] Project Gutenbert, "Great Expectations by Charles Dickens." `http://www.gutenberg.org/ebooks/1400` [Online. Accessed 2013-05-15], May 2013.

[257] L. Q. Ha, D. W. Stewart, P. Hanna, and F. J. Smith, "Zipf and type-token rules for the English, Spanish, Irish and Latin languages," *Web Journal of Formal, Computational & Congnitive Linguistics*, vol. 8, 2006.

[258] R. F. i Cancho, "The variation of Zipf's law in human language," *European Physical Journal B*, vol. 44, no. 2, pp. 249–257, 2005.

[259] W. Piotrowska and X. Piotrowska, "Statistical parameters in pathological text," *Journal of Quantitative Linguistics*, vol. 11, no. 1, pp. 133–140, 2004.

[260] L. Brillouin, *Science and Information Theory, Second Edition (Dover Phoenix Editions)*. Dover Publications, 2nd ed., September 2004.

[261] M. A. Serrano, A. Flammini, and F. Menczer, "Modeling statistical properties of written text," *PLoS ONE*, vol. 4, pp. e5372+, Apr. 2009.

[262] V. K. Balasubrahmanyan and S. Naranan, "Quantitative Linguistics and Complex System Studies," *Journal of Quantitative Linguistics*, vol. 3, no. 3, pp. 177–228, 1996.

[263] R. F. i Cancho, "Decoding least effort and scaling in signal frequency distributions," *Physica A: Statistical Mechanics and its Applications*, vol. 345, pp. 275–284, January 2005.

[264] A. Clauset, M. Young, and K. S. Gleditsch, "On the frequency of severe terrorist events," *Journal of Conflict Resolution*, vol. 51, pp. 58–87, Mar. 2007.

[265] S. Redner, "How popular is your paper? an empirical study of the citation distribution," *The European Physical Journal B*, vol. 4, pp. 131–134, Apr. 1998.

[266] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener, "Graph structure in the web," *Comput. Netw.*, vol. 33, pp. 309–320, June 2000.

[267] G. K. Zipf, "National unity and disunity: the nation as a bio-social organism," *Bloomington (IN): Princeton Press*, 1941.

[268] R. N. Mantegna, S. V. Buldyrev, A. L. Goldberger, S. Havlin, C. K. Peng, M. Simons, and H. E. Stanley, "Linguistic features of noncoding DNA sequences," *Physical Review Letters*, vol. 73, pp. 3169–3172, Dec. 1994.

[269] S. Bonhoeffer, A. V. M. Herz, M. C. Boerlijst, S. Nee, M. A. Nowak, and R. M. May, "Explaining "linguistic features" of noncoding DNA," *Science*, vol. 271, pp. 14–15, Jan. 1996.

[270] S. Bonhoeffer, A. V. Herz, M. C. Boerlijst, S. Nee, M. A. Nowak, and R. M. May, "No signs of hidden language in noncoding DNA.," *Physical review letters*, vol. 76, Mar. 1996.

[271] N. E. Israeloff, M. Kagalenko, and K. Chan, "Can Zipf distinguish language from noise in noncoding DNA?," *Physical review letters*, vol. 76, Mar. 1996.

[272] R. F. Voss, "Comment on "linguistic features of noncoding DNA sequences"," *Physical Review Letters*, vol. 76, p. 1978, Mar. 1996.

[273] W. Li, "Random texts exhibit Zipf's-law-like word frequency distribution," *IEEE Transactions on Information Theory*, vol. 38, pp. 1842–1845, Nov. 1992.

[274] R. F. i Cancho and B. Elvevåg, "Random texts do not exhibit the real Zipf's law-like rank distribution," *PLoS ONE*, vol. 5, pp. e9411+, March 2010.

[275] W. Li, "Zipf's law everywhere," *Glottometrics*, vol. 5, pp. 14–21, 2002.

[276] M. Mitzenmacher, "A brief history of generative models for power law and log-normal distributions," *Internet mathematics*, vol. 1, no. 2, pp. 226–251, 2004.

[277] V. Pareto, *Le Cours d'Economie Politique*. MacMillan, London, 1897.

[278] A. Ultsch and A. Ultsch, "Proof of Pareto's 80/20 law and Precise Limits for ABC-Analysis," *Data Bionics Research Group University of Marburg/Lahn, Germany*, pp. 1–11, 2002.

[279] L. A. Adamic and B. A. Huberman, "Zipf's law and the Internet," *Glottometrics*, vol. 3, no. 1, pp. 143–150, 2002.

[280] S. Salat and L. Bourdic, "Power laws for energy efficient and resilient cities," *Procedia Engineering*, vol. 21, pp. 1193–1198, Jan. 2011.

[281] A. Clauset, C. R. Shalizi, and M. E. J. Newman, "Power-law distributions in empirical data," *SIAM Review*, vol. 51, pp. 661+, Feb 2009.

[282] J. Alstott, E. Bullmore, and D. Plenz, "powerlaw: a Python package for analysis of heavy-tailed distributions," *PloS one*, vol. 9, no. 1, p. e85777, 2014.

[283] B. Efron and R. Tibshirani, *An introduction to the bootstrap*, vol. 57. CRC press, 1993.

[284] B. Hernández-Bermejo, V. Fairén, and A. Sorribas, "Power-law modeling based on least-squares minimization criteria," *Mathematical biosciences*, vol. 161, no. 1, pp. 83–94, 1999.

[285] E. P. White, B. J. Enquist, and J. L. Green, "On estimating the exponent of power-law frequency distributions," *Ecology*, vol. 89, no. 4, pp. 905–912, 2008.

[286] A. Prlić, A. Yates, S. E. Bliven, P. W. Rose, J. Jacobsen, P. V. Troshin, M. Chapman, J. Gao, C. H. Koh, S. Foisy, R. Holland, G. Rimša, M. L. Heuer, H. BrandstätterâĂŞMüller, P. E. Bourne, and S. Willis, "BioJava: an open-source framework for bioinformatics in 2012," *Bioinformatics*, vol. 28, pp. 2693–2695, Oct. 2012.

[287] J. E. Stajich, D. Block, K. Boulez, S. E. Brenner, S. A. Chervitz, C. Dagdigian, G. Fuellen, J. G. Gilbert, I. Korf, H. Lapp, H. Lehväslaiho, C. Matsalla, C. J. Mungall, B. I. Osborne, M. R. Pocock, P. Schattner, M. Senger, L. D. Stein, E. Stupka, M. D. Wilkinson, and E. Birney, "The Bioperl toolkit: Perl modules for the life sciences.," *Genome research*, vol. 12, pp. 1611–1618, Oct. 2002.

[288] H. Hermjakob, W. Fleischmann, and R. Apweiler, "Swissknife - 'lazy parsing' of SWISS-PROT entries.," *Bioinformatics*, vol. 15, pp. 771–772, Sept. 1999.

[289] UniProt Help Desk. `help@uniprot.org` Personal Communication, Feb. 2010.

[290] R. F. Cancho and R. V. Solé, "Two Regimes in the Frequency of Words and the Origins of Complex Lexicons: Zipf's Law Revisited," *Journal of Quantitative Linguistics*, vol. 8, no. 3, pp. 165–173, 2001.

[291] S. F. Rojas, A. Morgat, *et al.*, "Standardization in UniProtKB/Swiss-Prot." Poster at the 3rd International Biocuration Conference, 16 April 2009.

[292] M. J. Bell, C. S. Gillespie, D. Swan, and P. Lord, "An approach to describing and analysing bulk biological annotation quality: a case study using UniProtKB," *Bioinformatics*, vol. 28, pp. i562–i568, Sept. 2012.

[293] A. Bairoch and R. Apweiler, "The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1998," *Nucleic Acids Research*, vol. 26, pp. 38–42, Jan. 1998.

[294] U. Consortium *et al.*, "The universal protein resource (UniProt)," *Nucleic acids research*, vol. 36, no. suppl 1, pp. D190–D195, 2008.

[295] UniProt Help Desk. `help@uniprot.org` Personal Communication, October 2011.

[296] MAN Diesel & Turbo, "Thermo Efficiency System (TES) Reduces Fuel Costs and CO2," 2005.

[297] A. Berdanier, "Sankey.R." `https://gist.github.com/aaronberdanier/1423501` [Online. Accessed 2013-10-24], 2010.

[298] G. Doka, "Sankey Helper 2.4.1 by G.Doka." `http://www.doka.ch/sankey.htm` [Online. Accessed 2013-10-24], 2009.

[299] The Dia Developers, "Dia draws your structured diagrams: Free Windows, Mac OS X and Linux version of the popular open source program." `http://dia-installer.de/` [Online. Accessed 2013-10-03], July 2013.

[300] Microsoft Corporation, "Microsoft Visio 2013 – flowchart software - Office.com." `http://office.microsoft.com/en-gb/visio/` [Online. Accessed 2013-10-03], Dec. 2013.

[301] G. Csardi and T. Nepusz, "The igraph software package for complex network research," *InterJournal, Complex Systems*, vol. 1695, no. 5, 2006.

[302] M. E. Smoot, K. Ono, J. Ruscheinski, P.-L. Wang, and T. Ideker, "Cytoscape 2.8: new features for data integration and network visualization," *Bioinformatics*, vol. 27, no. 3, pp. 431–432, 2011.

[303] J. Köhler, J. Baumbach, J. Taubert, M. Specht, A. Skusa, A. Rüegg, C. Rawlings, P. Verrier, and S. Philippi, "Graph-based analysis and visualization of experimental results with ONDEX," *Bioinformatics*, vol. 22, pp. 1383–1390, June 2006.

[304] J. Heer and D. Boyd, "Vizster: Visualizing online social networks," in *Information Visualization, 2005. INFOVIS 2005. IEEE Symposium on*, pp. 32–39, IEEE, 2005.

[305] F. B. Viégas, M. Wattenberg, and K. Dave, "Studying cooperation and conflict between authors with *history flow* visualizations," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '04, (New York, NY, USA), pp. 575–582, ACM, 2004.

[306] C. T. Lopes, M. Franz, F. Kazi, S. L. Donaldson, Q. Morris, and G. D. Bader, "Cytoscape web: an interactive web-based network browser," *Bioinformatics*, vol. 26, no. 18, pp. 2347–2348, 2010.

[307] P. Nadhani, "FusionCharts - JavaScript Charts for the Grown-Ups." `http://www.fusioncharts.com/` [Online. Accessed 2013-10-03], July 2012.

[308] J.-D. Pogolotti, "pChart 2.0 - a PHP charting library." `http://www.pchart.net/` [Online. Accessed 2013-10-03], May 2013.

[309] A. Tse, "PlotKit." `http://www.liquidx.net/plotkit/` [Online. Accessed 2013-10-03], Aug. 2006.

[310] Google, "Google Charts — Google Developers." `http://developers.google.com/chart/` [Online. Accessed 2013-10-03], Apr. 2012.

[311] Highsoft AS, "Highcharts - Highcarts product." `http://www.highcharts.com/products/highcharts` [Online. Accessed 2013-12-12], Oct. 2013.

[312] C. Grandjean, *Instant Highcharts*. Packt Publishing Ltd, 2013.

[313] J. Kuan, *Learning Highcharts*. Packt Publishing, 2012.

[314] Alias-i, "LingPipe 4.1.0." `http://alias-i.com/lingpipe` [Online. Accessed 2014-02-10], Oct. 2008.

[315] M. Morris, "LingPipe: Sentence Extraction Tutorial." `http://alias-i.com/lingpipe/demos/tutorial/sentences/read-me.html` [Online. Accessed 2013-10-10], 2013.

[316] M. Morris and B. Carpenter, "MedlineSentenceModel (LingPipe API)." `http://http://alias-i.com/lingpipe/docs/api/com/aliasi/sentences/MedlineSentenceModel.html` [Online. Accessed 2013-10-10], 2013.

[317] Oracle Corporation, "MySQL :: The world's most popular open source database." `http://www.mysql.com/` [Online. Accessed 2013-07-10], 2013.

[318] R. Leinonen, F. Nardone, W. Zhu, and R. Apweiler, "UniSave: the UniProtKB Sequence/Annotation version database," *Bioinformatics*, vol. 22, pp. 1284–1285, May 2006.

[319] The UniProt Consortium, "UniProt release 2011_03." `http://www.uniprot.org/news/2011/03/08/release` [Online. Accessed 2013-02-25], 2011.

[320] C. O'Donovan, M. J. Martin, A. Gattiker, E. Gasteiger, A. Bairoch, and R. Apweiler, "High-quality protein knowledge resource: SWISS-PROT and TrEMBL," *Briefings in Bioinformatics*, vol. 3, no. 3, pp. 275–284, 2002.

[321] D. W. Ussery and P. F. Hallin, "Genome update: annotation quality in sequenced microbial genomes.," *Microbiology (Reading, England)*, vol. 150, pp. 2015–2017, July 2004.

[322] A. M. Schnoes, D. C. Ream, A. W. Thorman, P. C. Babbitt, and I. Friedberg, "Biases in the experimental annotations of protein function and their effect on our understanding of protein function space," *PLoS computational biology*, vol. 9, no. 5, p. e1003063, 2013.

[323] T. Kuhn and M. Krauthammer, "Underspecified scientific claims in nanopublications," *arXiv preprint arXiv:1209.1483*, 2012.

[324] BioModels.net Team, "Biomodels database." `http://www.ebi.ac.uk/biomodels-main/annotationtips` [Online. Accessed 2013-09-10], 2013.

[325] R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, E. Birney, M. Biswas, P. Bucher, L. Cerutti, F. Corpet, M. D. R. Croning, *et al.*, "The InterPro database, an integrated documentation resource for protein families, domains and functional sites," *Nucleic acids research*, vol. 29, no. 1, pp. 37–40, 2001.

[326] C. J. Sigrist, E. De Castro, L. Cerutti, B. A. Cuche, N. Hulo, A. Bridge, L. Bougueleret, and I. Xenarios, "New and continuing developments at PROSITE," *Nucleic acids research*, vol. 41, no. D1, pp. D344–D347, 2013.

[327] T. K. Attwood, P. Bradley, D. R. Flower, A. Gaulton, N. Maudling, A. Mitchell, G. Moulton, A. Nordle, K. Paine, P. Taylor, *et al.*, "PRINTS and its automatic supplement, prePRINTS," *Nucleic acids research*, vol. 31, no. 1, pp. 400–402, 2003.

[328] D. H. Haft, J. D. Selengut, and O. White, "The TIGRFAMs database of protein families," *Nucleic acids research*, vol. 31, no. 1, pp. 371–373, 2003.

[329] S. Hunter, P. Jones, A. Mitchell, R. Apweiler, T. K. Attwood, A. Bateman, T. Bernard, D. Binns, P. Bork, S. Burge, *et al.*, "InterPro in 2011: new developments in the family and domain prediction database," *Nucleic acids research*, vol. 40, no. D1, pp. D306–D312, 2012.