DELFT UNIVERSITY OF TECHNOLOGY

MASTERS THESIS

Ontology Integration for Biomedical Data

Author:
Ana OPREA

Supervisor:
Dr. Christoph LOFI

A thesis submitted in fulfillment of the requirements
for the degree of Master of Science

in the

Web Information Systems Group
Software Technology

June 24, 2023

"I was born not knowing and have had only had a little time to change that here and there."

Richard Feynman

DELFT UNIVERSITY OF TECHNOLOGY

Abstract

Electrical Engineering, Mathematics and Computer Science
Software Technology

Master of Science

Ontology Integration for Biomedical Data

by Ana OPREA

Gene similarity has been an area of great interest in numerous fields for decades, as it can provide insights into the evolutionary relationships among different species. This knowledge is particularly useful for advancing biotechnologies, discovering new drugs and treatments for various issues and improving the characteristics in breed crops or animals. DNA sequencing enables gene annotation, which facilitates the identification of similarities between genes. Similar genes from different species are interesting candidates for studying gene functional similarity. Gene ontology (GO) provides a standardized vocabulary for gene annotation, which is considered to be the ground truth when describing their properties. Nevertheless, an interesting source for gathering additional information about genes can be the plethora of biomedical articles accessible online. These are the pillars on which is foundered the incentive of this paper - an endeavor to investigate how using graph theory could benefit scientists in transferring knowledge about gene functionalities between different plants. We present an overview of the methodology and the design of the system we used in order to convey to what extent using subgraphs similarity based on annotated data proves to yield results similar to those already established as ground truth for the model plant Arabidopsis Thaliana and its counterpart, Solanum Lycopersicum, as well as our conclusions and discussions regarding the quality of the datasets used throughout this research.

Thesis Committee
Chair: Assistant Prof. Dr. Christoph Lofi, TU Delft
TU Delft University supervisor: Assistant Prof. Dr. Christoph Lofi, TU Delft
Committee Member: Assistant Prof. Dr. Jana Weber, TU Delft

Acknowledgements

This master thesis symbolizes the bittersweet ending of my student era, at least for now. It is not confined only to research, deadlines, emails and coding, but rather it is part of the growth I have experienced in the past year.

I am grateful I embarked on this journey alongside my supervisor, Dr. Christoph Lofi, who helped me stir the wheel when the waves grew larger, who let me sail in my own rhythm and who has been a cheerful and encouraging presence, whether we talked about research or anything else.

Mom and Dad, you deserve half of the recognition for me being able to successfully deliver this thesis. As always, this could not have been done without you.

I have arrived here because countless wonderful things synchronized and many people met me at the right time. Nikos, you are one of them, I dedicate this happy ending to you too.

Family, friends, people who happened to be "around here" while I was doing all this work - thank you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Context presentation

The study of genetics and genomics is currently one of the most dynamic and rapidly progressing fields in science. Positioned at the intersection of biology and informatics, it generates vast amounts of data that require expert analysis and interpretation. In this context, gene similarity emerges as a highly captivating and worthy area of investigation, serving as a fundamental component for numerous core researches such as: **comparative genomics**, **biomedical data integration** and **transfer knowledge** from one biomedical entity to another [Bayat, 2002]. Gene similarity refers to the degree of similarity between a pair of genes at their nucleotide level, by comparing their nucleotide sequences.

Studying gene similarity across species offers numerous advantageous applications in domains such as evolutionary biology, pharmaceuticals or agriculture. With respect to evolutionary relationships between different organisms, gene similarity is employed to investigate the likelihood of two distinct genes sharing a common ancestor [Altenhoff and Dessimoz, 2012]. In essence, the higher the level of similarity between two different genes, the less likely they are to have developed independently and to have arrived at similar DNA sequences by mere coincidence.

Furthermore, gene similarity serves as a valuable avenue in the realm of drug discovery by enabling us to target those genes responsible in diseases and to understand how they span across different organism [Spreafico et al., 2020, Schlicker et al., 2006]. In addition, gene similarity plays a crucial role in our understanding of viral evolution, of the mechanisms beneath pathogens adapting to different hosts and of the development of antiviral strategies.[Shackelton and Holmes, 2004].

Lastly, gene similarity plays a significant role in agriculture. For instance, it offers insights to breeders seeking to enhance desirable characteristics in their crops or livestock, such as disease resistance, stress tolerance or nutritional content [Dennis et al., 2008]. Moreover, by delving into the similarities or differences between genes, researches can gain more knowledge into the processes and mechanisms of the domestication of livestock or crops [Hufford et al., 2012].

We have decided to align our efforts with **Genetwister** by talking to a representative and understanding a potential avenue to explore in relation to an issue which could also render valuable insights for their work. Genetwister is a Dutch company in the field of biotechnology that focuses on bioinformatics of agricultural, horticultural and ornamental plants, which serves its customers by examining ways to modify their crops accordingly to their requests. [*Genetwister* n.d.]

Two topics that gained our interest were the concept of transfer knowledge of functional similarity between plant species through **orthology** relationships and the application of **Natural Language Processing** (NLP) to augment biomedical datasets.

For numerous genes in model plants, scientists have a clear understanding of their role in different organisms or cells. They can employ the functional similarity between different genes in order to transfer knowledge about their functions. This proves to be a relevant aspect as this can empower researchers to leverage insights about genes roles in different biological processes and molecular functions.

The similarity of genes is determined through DNA sequencing. However, it is important to note that genes undergo evolutionary changes, which entail that their DNA sequences also change over time. The transfer of knowledge from a model plant species to another crop species in terms of their gene functionalities is based upon the gene orthology relationships: genes that posses similar DNA sequences have higher chances to perform similar functions in the two species they belong to. Therefore, scientists have gathered information about the sequences of numerous plant species in publicly available datasets of annotations. Additionally, there are datasets that provide information on the orthology relationships between genes across different species. These are the datasets which served as the foundation for our research efforts.

1.2 Dataset augmentation using NLP techniques

The importance and the benefits of using of Natural language processing (NLP) techniques cannot be emphasized enough in today's various disciplines. NLP, a subdisciplinary field of Artificial Intelligence, is tightly related to linguistics and it focuses on integrating knowledge, algorithms and techniques from computer science in order to enable systems to interact in a similar fashion that humans do with natural language. Whether it involves tasks such as text mining scientific literature to extract relevant information, curating biomedical ontologies or assisting machine learning models in analyzing biomedical data, NLP is the foundational pillar in all instances.

In our endeavour, we considered turning to an approach which tackles text mining scientific articles because, despite being a core feature in order to assess the similarity between various genomes, gene annotation can be time-consuming and resource-intensive. Typically, one approach to perform text mining is to use named entity recognition (NER) algorithms to identify genes and proteins mentioned in text, followed by the extraction of interactions between them through the application of additional NLP algorithms. Our goal was to enhance the existing database, which initially comprised only datasets of annotations for two plant species (Arabidopsis Thaliana and Solanum Lycopersicum), and asses the extent to which this strategy could support us in creating a framework for investigating orthologous relationships between their genomes.

1.3 Research Questions

The research questions we answered in this paper are:

1. How faithfully can we reproduce the similarity of orthologous genes, as established as the ground truth knowledge bases belonging to InParanoid [*InParanoiDB* n.d.], using only datasets publicly available for the plant species Arabidopsis Thaliana, respectively Solanum Lycopersicum, consisting of data such as genes, transcripts and annotations describing the properties and particularities of those genes and transcripts?

2. Does including information extracted from text mining biomedical papers be of use for enriching datasets related to ortholog genes?

3. How much does including provenance knowledge of the extracted information from the previously mentioned mined biomedical papers be of use for enriching datasets related to ortholog genes in such a way that the similarity scores improve?

1.4 Contributions

We summarize our contributions as follows:

1. Creating a methodology for computing similarity scores based on subgraph similarities for transcripts belonging to two different plant species.

2. Defining a system for amassing the descriptive annotations and the transcripts together and creating links and relationships between them.

3. Leveraging insights about the quality of the knowledge bases involved in the study.

4. Providing an evaluation of our system and discussing the limitations of this study with emphasis on the influences of aggregating text-mined entities to our database.

1.5 Thesis Structure

This paper is organized as follows: in Chapter 2 we explain the elementary notions in the land of genetics and genomics for a clearer depiction of the theory behind our research questions; in Chapter 3 we present related work in the field of using graph theory in order to study the relationships between genes, respectively related work in the field of using NLP techniques for dataset augmentation and the latest advances in the field of AI for discovering protein structures; in Chapter 4 we talk about the knowledge bases involved in our study; in Chapter 5 we unravel the methodology we envisioned, the steps we followed and the design choices we approved in order to achieve the desideratum; in Chapter 6 we account for the insights leveraged about the quality of the datasets; in Chapter 7 we discuss the evaluation of our results outputted by our methodology, and, lastly, we dedicate Chapter 8 for our conclusions and final remarks.

Chapter 2

Gene Fundamentals

This chapter is dedicated to familiarize the reader with the elementary notions that are necessary in order to understand the biology stance behind our study. We provide definitions for the core concepts, followed by an overview which highlights how each of these concepts are tied to one another.

2.1 Vocabulary

Definition 2.1.1 (Model Plant) *Model plants are plant species on which extensive studies have been made due to the fact that significant advances about plant growth and development are made by focusing on their characteristics.*

[Meinke et al., 1998]

Definition 2.1.2 (Arabidopsis Thaliana) *Arabidopsis thaliana is a small flowering plant that is widely used as a model organism in plant biology. Arabidopsis is a member of the mustard (Brassicaceae) family, which includes cultivated species such as cabbage and radish. Arabidopsis is not of major agronomic significance, but it offers important advantages for basic research in genetics and molecular biology.*

[*Tair - About Arabidopsis* n.d.]

Definition 2.1.3 (Solanum Lycopersicum) *Solanum Lycopersicum is the plant species widely known as tomato.*

Definition 2.1.4 (Homology) *Homology is a relation between a pair of genes that share a common ancestor. All pairs of genes in the figure above are homologous to each other.*

[Altenhoff and Dessimoz, 2012]

Definition 2.1.5 (Ortholog genes) *Two ortholog genes are two genes from two different species that derive from a single gene in the last common ancestor of the species.*

[Sonnhammer and Koonin, 2002]

Definition 2.1.6 (Paralog genes) *Two paralog genes are two genes from two different species that derive from a single gene which was duplicated within the genome.*

[Sonnhammer and Koonin, 2002]

Definition 2.1.7 (Comparative genomics) *Comparative genomics is the direct comparison of complete genetic material of one organism against that of another to gain a better understanding of how species evolved and to determine the function of genes and non-coding regions in genomes.*

[Sivashankari and Shanmughavel, 2007]

Definition 2.1.8 (DNA sequencing) *DNA sequencing refers to the general laboratory technique for determining the exact sequence of nucleotides, or bases, in a DNA molecule.*

[*National Human Genome Institute* n.d.(a)]

Definition 2.1.9 (DNA annotation) *Genome annotation is the process of deriving the structural and functional information of a protein or gene from a raw data set using different analysis, comparison, estimation, precision, and other mining techniques. Genome annotation is essential because the sequencing of the genome or DNA generates sequence information without its functional role. After the genome is sequenced, it must be annotated to bring more logical information about its structural features and functional roles.*

[Harbola et al., 2022]

Definition 2.1.10 (Transcription) *Transcription, as related to genomics, is the process of making an RNA copy of a gene's DNA sequence. This copy, called messenger RNA (mRNA), carries the gene's protein information encoded in DNA.*

[*National Human Genome Institute* n.d.(b)]

Definition 2.1.11 (Genome Duplication) *Duplication, as related to genomics, refers to a type of mutation in which one or more copies of a DNA segment (which can be as small as a few bases or as large as a major chromosomal region) is produced. Duplicates occur in all organisms.*

[*National Human Genome Institute* n.d.(c)]

Definition 2.1.12 (Gene Product) *A protein molecule that is the product of the expression of a gene, through which the gene influences development or metabolism.*

[*Mouse Genome Informatics* n.d.]

2.2 Overview

In the field of plant biology, certain model species have been extensively studied, an example which would precisely highlight this fact being the infamous species Arabidopsis Thaliana, often called Arabidopsis, and popularly known as the thale cress or the mouse-ear cress. Arabidopsis belongs to the mustard family and shares its plant group with species such as cabbage, broccoli, and radish. Model plants are especially valuable due to the elaborate research conducted on them which can be of immense use in the field of comparative genomics. Research in this domain often involves the annotation of various genes, which entails identifying the features and functions of genes within their respective organisms. Gene annotation plays a pivotal role in understanding the genetic basis of any organism.

What will be considered the corner stone in this study will be the notion of orthologous genes. These are genes that can be traced back to having a common ancestor which diverged subsequently at a moment in time. By means of knowledge transfer via orthology relationships from one (plant) species to another, researchers can make predictions about the functionality of the latter species based on the functionality of

the former. Sometimes, it can be the case that one gene from a species can have multiple orthologous genes in another species, which is a result of a genome duplication over time, as explained in the paper of Altenhoff and Dessimoz, 2012.

Furthermore, an essential observation is that genes can have several different transcripts. In our study, we will look at the similarity of genes through their transcripts.

Chapter 3

Related Work

In this chapter, we outline comprehensive literature reviews of the research conducted in the field of discovering orthologous genes using graph theory, along with notable studies that focus on the utilization of Natural Language Processing techniques for gene dataset augmentation.

3.1 Graph theory for gene similarity

The information presented by Altenhoff and Dessimoz, 2012 in their research study about inferring the relation of orthology or paralogy between genes creates an appropriate context for the discussion of the methods employed by scientists within the domain of comparative genomics. Most orthology inference techniques are divided into two prevalent groups:

- graph-based methods
- tree-based methods

Techniques belonging to the former category usually consider graphs where genes or proteins are nodes and evolutionary relationships between them are depicted as edges. On the other hand, techniques belonging to the latter category are based upon gene/species tree reconciliation, which is the process of annotating all splits of a particular gene tree either as duplication or speciation with respect to its phylogeny.

The graph-based methods tackle the graph-construction phase by considering pairs of genomes at a time. Typically, similarity scores of different sequences are used as an indicator for gene closeness on the phylogenetic scale, therefore the inference of orthology between genome pairs can be computed efficiently using dynamic programming [Smith, Waterman, et al., 1981] or heuristics, such as BLAST [Altschul et al., 1997]. An interesting remark addressed in the paper is that pairwise comparisons between genes is not as robust as comparisons between multiple organisms. This strategy often helps researches in correcting and identifying misleading predictions. Henceforth, clustering of genes into orthologous groups can yield better results. The paper covers several grouping techniques, as shown in Table 3.1 from the original paper, of which we briefly summarize the main ones below:

- Tatusov, Koonin, and Lipman, 1997 coined the concept of cluster orthologous groups (COGs) which are triangles computed on triplets of connected genes which are subsequently merged together if they share a common face until every possible merging has been completed.

- Li, Stoeckert, and Roos, 2003 identified the groups of orthologs using Markov Clustering by simulating a random walk on the orthology graph with edges being weighted with respect to their similarity scores, therefore the grouping stage outputting probabilities for two genes to be part of the same group. The orthology graph is partitioned according to these probabilities so that genes arriving at the same partition are belonging to the same orthologous group.

- Dessimoz et al., 2005 proposed a different grouping approach by isolating cliques (the fully connected graphs) in their graph. This is computationally expensive as it is an NP-complete problem, nevertheless it has the advantage of leveraging a sound outcome due to the high consistency required to form a graph where all genes are orthologous to one another.

Method	In-Paralogs	Based on	Grouping Strategy	Database	Extra	Avaiable Algo/DB	Reference
COG	Yes	BLAST Scores	Merged adjacent triangles of BeTs	COG/KOG		X/X	Tatusov, Koonin, and Lipman, 1997
BBH	No	BLAST Scores	n.a.	n.a.		-/-	Overbeek et al., 1999
Inparanoid	Yes	BLAST Scores	Only between pairs of species	Inparanoid		X/X	Remm, Storm, and Sonnhammer, 2001 Östlund et al., 2010
RSD	No	ML distance estimates	n.a.	RoundUp		X/X	DeLuca et al., 2006 Wall, Fraser, and Hirsh, 2003
OMA	Yes	ML distance estimates	Every pair is ortholog	OMA Browser	Detects differential gene loss	-/X	Dessimoz et al., 2005 Altenhoff et al., 2010
OrthoMCL	Yes	BLAST Scores	MLC clusters	OrthoMCL-DB		X/X	Li, Stoeckert, and Roos, 2003 Chen et al., 2006
EggNOG	Yes	BLAST Scores	Merged adjacent triangles of BeTs	EggNOG	Computed at several levels of taxonomic tree	-/X	Muller et al., 2010 Jensen et al., 2007
OrthoDB	Yes	Smith Waterman Scores	Merged adjacent triangles of BeTs	OrthoDB	Computed at any level of taxonomic level	-/X	Kriventseva et al., 2007
COCO-CL	Yes	MSA-induced scores	Hierarchical clusters	n.a.		X/-	Jothi et al., 2006
OrthoInspector	Yes	BLAST Scores	Only between pairs of species	OrthoInspector		X/X	Linard et al., 2011

TABLE 3.1: Table presents the main graph-based strategies involved in orthology inference as presented by Altenhoff and Dessimoz, 2012

The tree-based methods assume tree reconciliation once it is known that all the branchings of a gene tree over time have been resolved as either events of speciation or duplication, therefore it becomes very simple to deduct if a pair of genes can be orthologous or paralogous.

Figure 3.1 provides a sound example. It is considered that the most likely tree reconciliation involves the least number of gene duplication or losses. However straightforward the initial premise is, there are several issues put forward. A first problem was the uncertainty often associate with different species. A second problem would be the requirement of rooting both the gene and the species trees, whereas most of the models of sequence evolution do not allow to infer the rooting of the reconstructed gene tree.

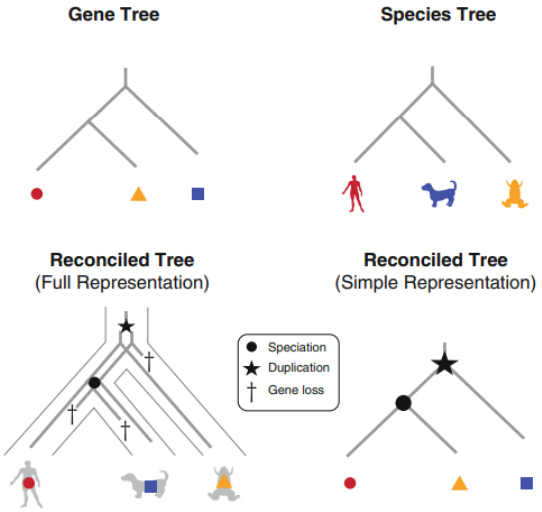


FIGURE 3.1: Figure presents schematic tree/species reconciliation as exemplified in the original paper

Another comprehensive study which provides an overview of various methods for computing semantic similarity between different biological concepts, including genes based on their annotations, is the one presented by Pesquita et al., 2009. According to the authors, ontologies have become a common schema for describing entities in the biomedical field, possibly the most noteworthy one for our discussion being The Gene Ontology (GO) [Consortium, 2004].

Despite the fact that genes can be directly compared via their sequence alignment, the same is not valid for their functional aspects. The semantic similarity applied on the GO annotations of gene products is a venture point for describing their functional similarity. The schema described by GO for representing the gene products has three independent direct acyclic graphs (DAGs) that correspond to: molecular function, biological process and cellular component. The nodes in the graphs signify terms that describe components of gene products, while the edges associate terms between one another, most frequently by relationships such as *"is a"* or *"part of"*, an example which would clarify this point being illustrated in Figure 3.2.

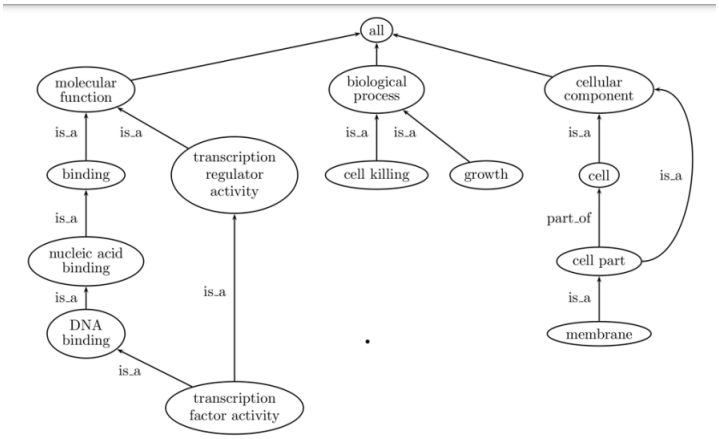


FIGURE 3.2: Figure presents GO graph showing the molecular function, biological process and cellular component aspects, as presented in the paper Pesquita et al., 2009

The paper explains the methods to quantify the semantic similarity based on GO DAGs: by comparing terms or by comparing gene products. The former type is further divided into two groups: by considering edges as data sources, or the nodes. The edge-based approach rely on counting the number of edges in the graph path between two terms. The measure employed in order to quantify the similarity is the *distance* between two terms, by either looking at the shortest path or at the average between all possible path between the two respective terms, if there are more than one. However, several issues emerge using this strategy. In contrast, the node-based approach rely on the information encoded in the terms themselves, either investigating the number of common annotations or by the relevance of the information content.

The latter type, comparing gene products, has multiple sub-categories, as presented in Figure 3.3. Due to the fact that gene product functions are described by molecular function terms, participate in various biological processes within multiple cellular components, it is required that sets of terms are compared in order to assess the semantic similarity.

The pairwise method computes similarity between the annotations of two genes, in some cases considering all pairwise combinations of terms, while in other cases reckoning only the best pairs. Conversely, the groupwise method can be regarded as using only direct annotations via set similarity techniques, using subgraph similarities for gene products depicted as subgraphs of GO corresponding to all their annotations, or using vector similarity measures for gene products represented in vector spaces.

The study mentions the use of gene coexpression data in the studies conducted by Sevilla et al., 2005 and by Wang et al., 2004 in order to test similarity measurements.

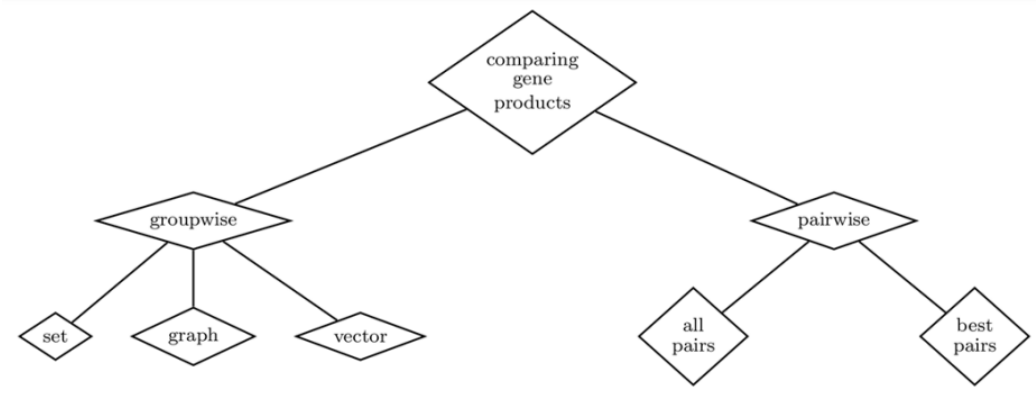


FIGURE 3.3: Figure presents the principal methods for comparing gene products, as presented in the paper Pesquita et al., 2009

3.2 NLP for biomedical datasets augmentation

This section provides an extensive literature review of the use of NLP algorithms and methods in order to mine for gene-related entities in biomedical scientific papers, which are an exceedingly rich source of information. The challenge of dataset augmentation has long been a critical concern, particularly in instances where data acquisition is prohibitively costly or infeasible. As human curation is not a sustainable option to extract information from scientific literature due to the large number

of papers to examine, [Singhal et al., 2016], machine learning approaches have often been used to circumvent this issue by creating synthetic data. In addition to conventional methods, NLP presents a promising avenue for gathering and labeling supplementary observations from the biomedical literature. Abstracts can be an acceptable target for mining, as the advantage they posses in the detriment of full-text paper is reducing computational time while preserving in a concise form the prevalent data.

One study presenting stimulating work in this area was the one of Rindflesch et al., 1999. The NLP system described in the paper, EDGAR (Extraction of Drugs, Gene and Relations) is tailored to carry out these extractions for these entities related to cancer from the biomedical literature, while the authors state that their technology could be easily employed to many other areas of biomedicine.

In comparison to previous work done in automated understanding of biomedical literature, EDGAR focuses on extracting factual assertions, a problem more complex than the discovery of descriptive terms in a paper. The factual assertions are composed of genes, cells and drugs, as they are in a relationship to one another, an example which would precisely highlight this fact being the example extracted from the original paper from Figure 3.4:

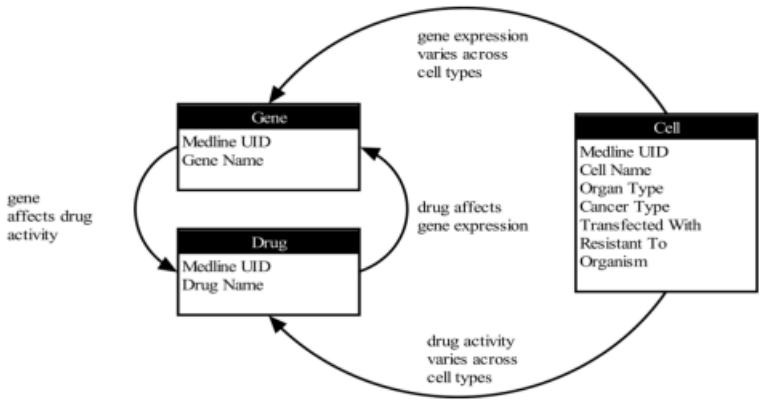


FIGURE 3.4: Figure presents the entities and relationships involved, as presented in the original paper

The connections between genes, cells and drugs can be inferred from their relationships to other drugs, cells and genes. The first in the entire extraction pipeline is the semantic interpretation: identifying terms in the text of MEDLINE abstracts, followed by the identification of relationships with respect to the interaction of gene expression and drug sensitivity in particular cell types. The processing of each sentence from the abstract begins with a stochastic tagger responsible with resolving part-of-speech ambiguities.

A real application of the EDGAR system is explained in the paper as following: a PubMed query to generate 383 abstracts related to anti-tumor drug resistance was used in order to feed these abstracts to EDGAR in batches. The outputs of EDGAR were further processed in order to create document vectors of the entities depicted, which were later on used to perform hierarchical clustering. By inspecting the dendrogram obtained over the 383 abstracts, hypothesis with regard to the relationships between terms were raised without reading one single abstract, and yet these were validated after scrutinizing the abstracts. The authors mention that such conclusions would have not been supported only by the examination of titles.

Another paper addressing this topic is the review literature study of Conceição and Couto, 2021. The study is centered around researches done on the construction of biological networks using text-mined information related to cancer.

As there is a plethora of clinical reports with respect to cancer, a category which encompasses many different types of cancer with their own particularities to examine, an evident challenge arises regarding investigating unstructured data.

The authors mention Jurca et al., 2016 for their large scale analysis performed on PubMed abstracts for the named entity extraction, followed by the relation extraction using the co-occurrence method. The study aimed to form a hypothesis with respect to cancer biomarkers and to identify those genes which were the most intensely studied across countries. In order to identify and select the relations with high frequency among abstracts, they considered genes and nodes and their relations as edges and further obtained a connected component based on ten or more abstracts. By evaluating the closeness and betweenness of the present genes, ten of them were selected as the most important.

In this study of Kawashima, Bai, and Quan, 2017 unsupervised learning, text mining, and pattern mining techniques were applied to extract relationships between breast cancer and the associated genes from PubMed. The extracted genes were then utilized as data vectors for a clustering approach. These gene vectors were combined with a pre-existing list of genes associated with breast cancer. However, the clustering technique employed yielded a low F1 score, specifically below 0.14.

Although some studies have confirmed the feasibility of constructing accurate gene-gene networks using relations extracted from literature [Jurca et al., 2016], the paper covers essential limitations of text mining to consider:

1. informational bias due to the focus on specific terms
2. inclusion of errors resulted from the automatic extraction
3. the difficulty associated with entity extraction due to the variety of synonyms, abbreviations or acronyms they may appear under in textual data
4. the ambiguity associated with including homonyms when different entities have the same label

3.3 AI for protein structures

Metagenomics is a a field tightly coupled to comparative genomics and to the issue of finding orthologous genes. It is a scientific field dedicated to the exploration of proteins in various samples across the planet through the application of gene sequencing techniques. This field showcases a remarkable breadth and diversity of proteins, introducing billions of novel protein structures into databases for the first time. The latest advances in the field can be attributed to DeepMind in 2021, respectively to Meta AI in 2022.

Lin et al., 2023 presents in their research paper ESM Metagenomic Atlas by Meta AI, an atlas of more than 772 million metagenomic protein structures predicted by leveraging the power of language models.

According to the study, the utilization of language models in the creation of comprehensive protein structure views has the potential to accelerate atomic-level three-dimensional prediction by up to 60 times compared to existing state-of-the-art approaches. Protein sequences encode information not only about the chemical structure of the molecules, but also about how they fold into a three-dimensional shape according to the laws of physics. Scientists established that the study of these structures is intertwined with understanding the arrangement of amino acid building blocks within proteins. Consequently, structures can be inferred from the patterns observed in protein sequences.

To investigate these patterns, Meta employs evolutionary scale modelling (ESM), a methodology that harnesses artificial intelligence which is based upon a self supervised learning model, referred to as a masked language model, which has been trained on millions of protein sequences. It functions by predicting the missing parts of a protein sequence, akin to filling in the gaps in a sentence with the appropriate words.

The speed at which such predictions can be made is a crucial aspect as, compared to previous strategies, predicting hundreds of millions of protein structures was computationally expensive in terms of the used resources, sometimes spanning over many years.

The researches from DeepMind proposed AlphaFold, another AI system specialized in predicting the structure of proteins, as it is presented in their article [Jumper et al., 2021]. According to the paper, AlphaFold is a neural-network model capable to predict the three-dimensional structure of proteins based solely on their amino-acid sequences, even in cases where no similar structure is available, with almost experimental accuracy. It combines novel neural networks architectures and training procedures based on evolutionary, physical and geometrical constraints of protein structures. The algorithm architecture is using only supervised learning on the structures deposited in the Protein Data Bank. The principal features of the study are as follow:

- the principal component of the network, names Evoform, which views the prediction of the three-dimensional structure as a graph inference problem in 3D
- multiple sequence alignments (MSAs) that capture evolutionary relationships and correlations between them
- attention-based mechanisms that learn interactions between non-neighbouring nodes in a graph representations of the amino-acids involved

Chapter 4

Datasets

This chapter presents a detailed introduction of the datasets involved in our study. Essentially, we gathered our data from three sources:

- online portals - Phytozome's Biomart tool [*JGI Phytozome* n.d.]
- online repositories - i.e Ensemble Plants [*Ensembl* n.d.]
- Narrative Service, a text mining platform for biomedical entities presented in the paper of Kroll et al., 2023

4.1 Datasets of annotations

The most important components in our study, which will reveal to also for the most relevant limitations, were the datasets for the annotations for two different plant species. We employed public datasets containing the annotations for the model plant, Arabidopsis Thaliana, and Solanum Lycopersicum. These datasets were presented in the form of text files, with each row providing information about a specific transcript of a particular gene. The columns of the files represented different types of annotations, describing the properties and features of the genes and transcripts.

	# of genes	**# of annotations**	**# of used annotations**
Arabidopsis Thaliana	48457	12	7
Solanum Lycopersicum	34725	12	7

TABLE 4.1: Table summarizes the number of genes and the number of useful columns representing annotations, collected from the two datasets for the plant species

As presented in table 4.1, we worked with more than 48k transcripts of genes for Arabidopsis Thaliana and more than 34k transcripts of genes for Solanum Lycopersicum. Out of 12 columns with annotations for genes, in both cases, we used seven common columns and discarded the ones which did not did not appear in both files or which did not represent any significant information, such as row id.

Two relevant mentions on the nature of these datasets are:

- as it will be later described in Chapter 6, the datasets have numerous missing values

- every column represented a type of annotation and transcripts usually have several distinct values for an entry for the corresponding column

4.2 Dataset of ortholog genes

In addition to the two datasets for the annotations, we employed a public dataset containing information about the orthology relationships between the transcripts of the two different plant species. This dataset had more than 12k row entries depicting one or multiple transcripts and numerical values from the two different plants species. As we previously established that one gene from one species can have multiple orthologous genes from another plant species, we inferred that each row entry was linking several genes from one plant to their orthologous genes from the counterpart plant.

4.3 BioMart Datasets

The datasets for the synonyms of transcripts, whose purpose will be detailed thoroughly in Chapter 5, were obtained by using the Phytozome's tool - BioMart. With the use of Biomart we obtained two datasets, one for Arabidopsis Thaliana, and another one for Solanum Lycopersicum, under the form of text files. Each file contained three columns, one with transcripts, one with corresponding gene and one with the synonyms. A highly important remark, which will be extrapolated in Chapter 6, is that the synonym file for the Solanum Lycopersicum species was completely lacking any synonyms, accounting for an important impediment in our research which will be tackled in 7.3.

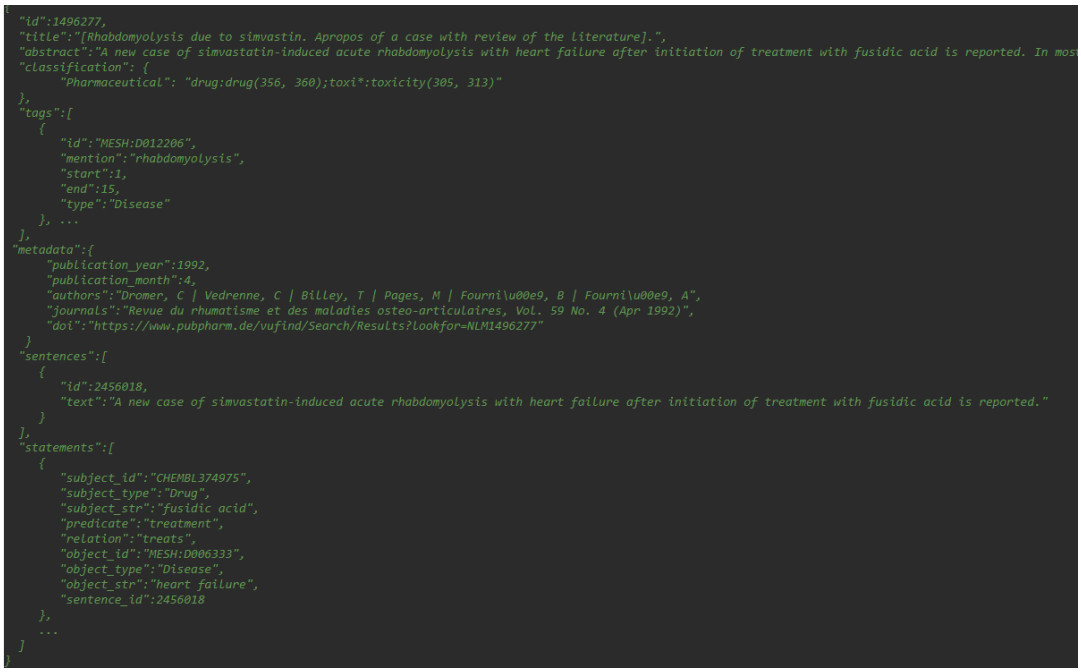
4.4 Dataset of text mined entities

The dataset obtained through the use of the text mining platform we employed in our research served as a tipping point in our study. It played an important role in determining the extent to which our model could benefit from newly aggregated information, as the quality and the integrity of data dictate the outcomes of the model.

The dataset was built upon the use of the BioMart datasets previously mentioned by scraping PubMed [*National Library of Medicine* n.d.]. PubMed is a free search engine which can be used in order to access more then 35 millions of publications for biomedical literature. The whole PubMed was scraped with this tool using the synonym vocabulary for the genes.

The scraping process involved in this task entails the extraction of terms from the titles and abstracts of all biomedical documents available on PubMed, followed by the identification of their synonymous expressions that the tool was previously provided with, as well as the relationships and predicates present in the sentences they are mentioned in. The end goal is to construct small-scale knowledge graphs by linking the various biomedical elements based on their relationship with one another as presented in the abstracts and titles.

The dataset, foundered on using our vocabulary, with the synonyms and names of all transcripts as they might appear in the scientific literature, structured under the form of JSON lines, proved to be a large one, having 16GB of data and containing around 500k PubMed documents. These are all the documents which contain **at least** one entry from the gene vocabulary. Figure 4.1 shows a snippet of our dataset.



```
{
 "id":1496277,
 "title":"[Rhabdomyolysis due to simvastin. Apropos of a case with review of the literature].",
 "abstract":"A new case of simvastatin-induced acute rhabdomyolysis with heart failure after initiation of treatment with fusidic acid is reported. In most
 "classification": {
      "Pharmaceutical": "drug:drug(356, 360);toxi*:toxicity(305, 313)"
 },
 "tags":[
   {
      "id":"MESH:D012206",
      "mention":"rhabdomyolysis",
      "start":1,
      "end":15,
      "type":"Disease"
   }, ...
 ],
"metadata":{
    "publication_year":1992,
    "publication_month":4,
    "authors":"Dromer, C | Vedrenne, C | Billey, T | Pages, M | Fournilu00e9, B | Fournilu00e9, A",
    "journals":"Revue du rhumatisme et des maladies osteo-articulaires, Vol. 59 No. 4 (Apr 1992)",
    "doi":"https://www.pubpharm.de/vufind/Search/Results?lookfor=NLM1496277"
 }
 "sentences":[
   {
      "id":2456018,
      "text":"A new case of simvastatin-induced acute rhabdomyolysis with heart failure after initiation of treatment with fusidic acid is reported."
   }
 ],
 "statements":[
   {
      "subject_id":"CHEMBL374975",
      "subject_type":"Drug",
      "subject_str":"fusidic acid",
      "predicate":"treatment",
      "relation":"treats",
      "object_id":"MESH:D006333",
      "object_type":"Disease",
      "object_str":"heart failure",
      "sentence_id":2456018
   },
   ...
 ]
}
```


FIGURE 4.1: Figure presents the information contained in one JSON line with respect to the PubMed documents with id 1496227

Each JSON line consisted of:

- a document id
- document title and abstract (if present)
- a section dedicated to the tags - the mined entities, with information such as their id, their type and where in the text they were discovered
- a section dedicated to the statements (if present) - an object entity and a subject entity, their types, the sentence id where they occur, the type of relation between them and the exact used predicate in the paper
- a section dedicated to sentences (if present) that incorporated relationships between tags with their id
- a section dedicated to metadata belonging to the document such as its publication year and month, its authors, the journal where it was featured and a link towards it

Chapter 5

Approach

In this chapter, we aim to provide the readers with a comprehensive and thorough account of the path we have undertaken to address our research questions. We delve into the methodology steps and details, thoroughly outlining the course we have followed, along with the various design choices and thresholds we encountered along the way.

5.1 Preliminaries

To alleviate potential confusion, we considered to be necessary to introduce two distinct terms that refer to the same concept, with the specific term utilized depending on its contextual usage. Hence, we hereby introduce the following terms:

1. **established gene similarity** - similarity between a pair of genes belonging to two different species, which is promulgated in the biological community to be the ground truth.
2. **computed gene similarity** - similarity between a pair of genes belonging to two different species as indicated by the results obtained through our computations.

As previously mentioned, the experiments were conducted and the implementation was carried out using two distinct plant species: Arabidopsis Thaliana, a well-annotated model plant, and Solanum Lycopersicum. It is important to note that the selection of these two plants was not within our discretion, but rather it was Genetwister who provided us with the necessary data to initiate our investigations or who pointed out which are the necessary tools in order to gather the data required. Furthermore, the format of the datasets was predetermined and we worked with them as they were presented to us: text files for the datasets discussed in 4.1 and 4.2, respectively JSON line files for the dataset of the text-mined literature presented in 4.4.

5.2 Datasets cleanup

The initial step involved preprocessing the data present in the gene annotations datasets and the orthology dataset. For the annotations datasets, the cleanup consisted of removing symbols such as "." from the entities names as they did not adhere to the conventions used in Neo4j for naming entities.

Regarding the orthology dataset, its original format was not suitable since it contained multiple transcripts from both Solanum Lycopersicum and Arabidopsis Thaliana genes on the same line. To address this, we extracted all possible pairs of

transcripts between the two species. In the end, we obtained a total of 25,441 pairs of orthologous genes.

5.3 System Design

The design choices of the system were primarily influenced by the characteristics of the datasets and the nature of the problem at hand. However, it is important to note that the initial level of flexibility incorporated into the system can have significant implications on the outcomes. These implications will be further discussed in Chapter 7.

Figure 5.1 reflects the initial stage of our work . Having two annotations datasets for the two plant species, each containing transcripts of genes and their corresponding annotations, our task was to determine the similarity of annotations between pairs of genes. To address this challenge, we opted to construct a comprehensive graph that encompasses all genes and annotations as interconnected nodes. By computing subgraph similarities, we aimed to identify common sets of annotations between genes, as depicted by the highlighted nodes in red in Figure 5.1: if the nodes they are linked two, representing corresponding annotations, would indeed represent the same entities.

Therefore, the desideratum was to use existing tools to incorporate the data provided by Genetwister and to leverage insights about the extent to which using the annotations could imply relations of orthology between different genes from different species.

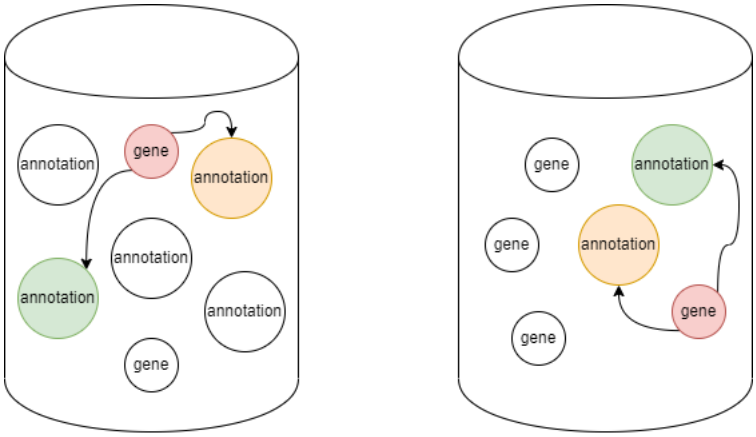


FIGURE 5.1: Figure presents two different annotation datasets for the two different place species

5.3.1 Graph Database

The first **design choice** we made was to employ a graph database as we considered the high degree of connectivity between transcripts of genes and their corresponding annotations. Given that the establishment of orthology relationships between genes relies on the assumption of similar annotations, a graph database proved to be an effective solution for handling subgraph similarities between nodes. Thus, this decision further reinforced the usefulness of having a graph database as a starting point in our endeavours.

Our selection of Neo4j as the database management system was not driven by any specific bias. However, the user-friendly nature of Neo4j and its visually appealing interface, which allowed us to easily explore subsets of nodes and their relationships, were significant factors that influenced our decision-making process.

In our graph database, each transcript and each annotation describing a transcript are created as nodes. In Neo4j, each node has a label and different properties. In our case, each transcript from both species is stored under the label called *Gene*, while all the other annotations appearing in the same column are assigned labels corresponding to the column name.

Transcripts are connected to their annotations through up to seven different types of relationships, with each relationship type determined by the type of node it connects to the transcript. We want to emphasize the aspect that not all transcripts have all seven types of annotations present in the graph database, an observation discussed in Chapter 8.

Moreover, each relationship can be customized to have different properties which can later account for influencing the properties of a subgraph containing those relationships. This brings us to our second **design choice** - assigning each relationship a property called *strength*, with a default value of one. By modifying this property during the calculation of subgraph similarities, we can implement a **weighted similarity function**.

5.3.2 Similarity Scores

As previously mentioned, the strategy we used in order to compute similarity similarity scores between transcripts of genes using their annotations was to compute subgraph similarities of those transcripts. One could observe that our graph is a bipartite graph because relationships exists only between a node of type *Gene* and a node of another type.

The third **design choice** was to use the Neo4j inbuilt node similarity algorithm on the Jaccard metric. For two sets U and V, the Jaccard similarity is computed as follows:

$$J(U,V) = \frac{|U \cap V|}{|U \cup V|}$$

In order to work with any algorithm from the Neo4j library, a projection of the graph has to be stored under a user-defined name. The projection specifies which nodes and which relationships from the database have to be included in the projection. Our projection included all nodes from all labels and all relationships between transcripts of genes and the nodes corresponding to their annotations, as it can be depicted in Figure 5.2.

The node similarity algorithm used was the *gds.nodeSimilarity.stream()* function which takes as an arguments the projection of the graph in question. The query is shown in Figure 5.3. The node similarity function compares each node that has outgoing relationships with each another such node. In our case, only nodes describing transcripts of genes are considered for similarity. The algorithm computes pair-wise similarities between a node and all the other nodes, but outputs only a fraction of the results. This is due to memory bounds. Hence, our query had to be modified in order to include a very important parameter - *topK*. This specifies that only the first topK best values for similarities between a transcript of gene and its counterparts will be displayed.

```
CALL gds.graph.project(
    'myProjection',
    ['Gene', 'Pfam','Panther', 'Ko', 'Go', 'Ec', 'Kog', 'Locus'],
    {
        GO: {
            properties: {
                strength: {
                    property: 'strength',
                    defaultValue: 1.0
                }
            }
        },
        .........
   }
);
```

FIGURE 5.2: Neo4j query for building a graph projection considering a subset of nodes and relationships from the graph database

```
CALL gds.nodeSimilarity.stream('myProjection', {topK: 800})
YIELD node1, node2, similarity WHERE
gds.util.asNode(node1).name starts with "S" AND
gds.util.asNode(node2).name starts with "A"
RETURN gds.util.asNode(node1).name AS Gene1,
gds.util.asNode(node2).name AS Gene2, similarity
ORDER BY similarity DESCENDING, Gene1, Gene2
```

FIGURE 5.3: Neo4j query for computing node similarities

As a result, in order to avoid duplicated outputs, our query is designed to compare every Solanum Lycopersicum transcript with any other Arabidopsis Thaliana transcript based on the projection of our graph, looking at the first 800 best results for each transcript.

The value for parameter topK was used after empiric observations: many of the pairs from the ortholog files did not have any computer similarity score outputted, therefore we chose this value because:

- with higher values, the memory resources would not be sufficient on our machine with following specs:

OS	Windows Edition 11 Pro 64-bit
RAM	32.0 GB (31.8 GB usable)
Processor	Intel(R) Core(TM) i7-10750H CPU @ 2.60GHz 2.59 GHz

TABLE 5.1: Table presents specifications for the machine

- using topK: 800, we remain with only 600 transcripts with no outputted values, transcripts with very little scores nonetheless, for which we simply chose to give a computed similarity of zero.

5.4 Text mining

To enhance the richness of our datasets, we turned our attention to a cutting-edge platform that employs a unique approach to construct small-scale knowledge graphs. This platform harnesses the power of annotation to identify and classify various biomedical entities found within the titles and abstracts of biomedical papers published on PubMed. By leveraging this innovative text-mining platform, we aimed to augment our datasets with valuable information extracted from scientific literature that can potentially be related to the genes we were interested in. The platform is meticulously presented by Kroll et al., 2023 their paper.

To expand our exploration and uncover potential new insights, we employed the dataset introduced in Section 4.4. This dataset served as a valuable resource for our quest to identify the most frequently occurring elements within the vast expanse of PubMed.

By mining the entirety of PubMed, we aimed to extract valuable information embedded within the tiny knowledge graphs generated by the text-mining platform: new elements tied to genes from the two plant species of interest can create new bounds between them after being integrated into our initial graph database. For illustrative purposes, we included the figure below:

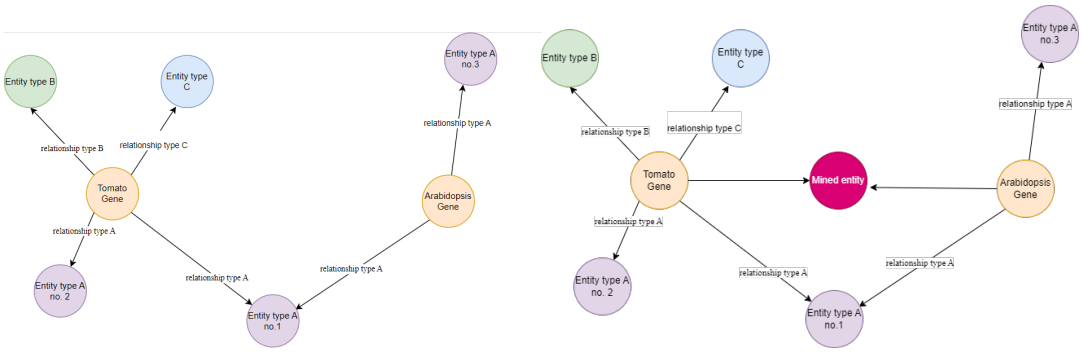


FIGURE 5.4: Figure presents the changed subgraphs of the two different transcripts when a new common entity node is being added

This strategy is employed in order to answer the second research question - whether this form of dataset augmentation can increase the performance of the system when computing the similarity scores based on subgraph similarities or how do they influence the computed similarity scores based only on the annotations established de facto true.

The initial step was to narrow down the large dataset. This step was required in order to filter out unnecessary documents whose tiny knowledge graphs would not render any valuable information, such as documents lacking any relationships between discovered entities in their titles and abstracts.

Consequently, we extracted from the remaining dataset the entities from the statements which appeared to be in a relationships with transcripts of genes that we were interested in. Therefore, we used two different dictionaries to retain:

- **keys** as entities describing transcripts for our plant species
- **values** as the entities they were recorded to be in a relationships with in the statement sections

The third step consisted of embedding the newly discovered entities and their relationships into the graph database. For this, we extracted the entities which were common for both entities tied to transcripts from Arabidopsis Thaliana and to transcripts from Solanum Lycopersicum. We obtained 51 common entities. However, after another essential filtering we arrived at 47 entities. Retrieving the relationships between mined entities and the transcripts from the database involved several actions:

1. We firstly used the dictionaries to establish relationships between entities and the vocabulary describing our genes of interest. Each item was treated as being written with lowercase letters in order to avoid duplicates.

2. We used the vocabulary dataset to make correspondences between a synonym and genes from both species that were described by it.

3. We established relationships between the transcripts present in the database and the entity it was linked to.

Given the limitations of the synonym dataset, which only contained synonyms for genes (loci) rather than transcripts specific to Arabidopsis Thaliana, it was evident that the available information was insufficient for our intended purposes. In order to overcome this constraint and expand the scope of our analysis, we decided to extrapolate the established relationships between genes and entities to encompass all transcripts associated with each respective gene. In a manner consistent with the original scenario, all newly established relationships were assigned a "strength" property, which defaulted to a value of one. This property served as a measure of significance, reflecting the strength of the relationship between the entities involved.

On the whole, the system was augmented as such: 47 new nodes of type *Entity*, 89 new relationships for transcripts of Solanum Lycopersicum, 15050 new relationships for transcripts of Arabidopsis Thaliana. All relationships were of the same type called *MINED*.

Similarly to the first attempt, the similarity scores were computed using the query presented in Figure 5.3. However, this time the projection of the graph on which the *gds.nodeSimilarity.stream()* function was called had to be modified in order to capture the newly integrated nodes and relationships.

Introducing provenance weight

In order to fully answer the third research question, we thought about introducing into our model considerations about the provenance of information. Once again, the system designed in the basic scenario employed pieces of information which are regarded as the groud truth with respect to the genes of the two species.

In the subsequent scenario, when incorporating text-mined documents from a collection of scientific papers, we confront the challenge of dealing with potentially inconsistent and unreliable new information introduced to the model. This raises an important question: How should we address this issue to accurately depict the disparities between the two knowledge bases?

When evaluating each step from our model that lead to different design choices, a primordial possible tweak is to change accordingly the graph projection. It is required in such a scenario to make a distinction between knowledge bases - this could be easily depicted by choosing different strengths for the different types of relationships in our graph: giving a higher weight for the relationships established to be

of ground truth from the first scenario and lower weights for connections between transcripts and the mined entities. This strategy conveys the goal of reinforcing the idea that the newly added nodes to the graph can inflict unreliable, even misleading influences on the networks.

Therefore, for this case we chose to keep the property *strength* of all relationships linking transcripts to their annotations and to halve the property *strength* for the relationships between transcripts and mined entities. This enabled us to use the Neo4j Weighted Jaccard Similarity:

$$J_W(A, B) = \frac{\sum_i min(a_i, b_i)}{\sum_i max(a_i, b_i)}$$

where $A = (a_i, ..., a_n)$ and $B = (b_i, ..., b_n)$

Chapter 6

Qualitative Research

In this chapter, we aim to examine the impact of dataset quality on the resulting outcomes and identify areas where modifications to the datasets could have been made in order to possibly yield improved results for the computed similarity scores. Accordingly, we have decided two split up the discussion into two parts:

1. discussing the the datasets containing annotations and synonyms presented in 4.1 and 4.2

2. discussing the dataset containing text-mined entities presented in 4.4

When evaluating the quality of datasets, we typically assess several factors, including the degree to which they possess a coherent structure, the level of information contained within the data, and the presence of missing values. While the issue of handling incorrect values is important, it is generally not considered a primary concern in evaluating the quality of biomedical datasets, given that they are typically publicly available and have been curated by the scientific community. As such, the focus tends to be on assessing the overall structure and content of the dataset, as well as identifying any potential gaps or limitations that could impact its utility for research purposes.

6.1 Quality of the annotation datasets

In terms of the annotation datasets, the primary aspects we were interested in were to handle datasets from which we can derive robust, complex graphs, with numerous nodes and edges between them. Therefore, it was chief to inspect how many genes and transcripts were present, how many types of annotations are presented and how connected were transcripts to the other entities. However, upon initial inspection, it was observed that 8687 out of the total number of transcripts were entirely devoid of annotations, representing a significant limitation of our study. This finding underscores the need for caution and thoroughness in data collection and curation efforts, as incomplete or missing data can compromise the validity and utility of the resulting dataset, as stated in 7.3.

Furthermore, Figure 6.1 puts into perspective how connected the transcripts were to other nodes in the graph. It can be understood that the most connected nodes have at most 20 relationships.

6.2 Quality of the synonym datasets

We consider the quality of the synonym datasets to be characterized by the availability of comprehensive, current, and accurate information for each transcript. Unfortunately, we encountered another significant limitation in our implementation due

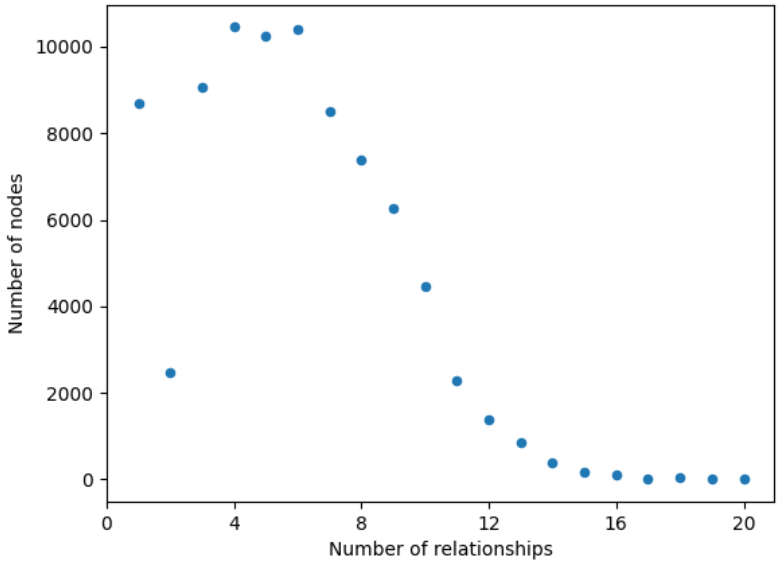


FIGURE 6.1: Number of relationships to number of nodes

to the lack of any synonyms for the entire collection of genes from the synonym dataset for Solanum Lycopersicum.

Additionally, it is noteworthy that the synonym dataset for Arabidopsis Thaliana, as exported from BioMart, had numerous duplicate values for different genes, which had to be aggregated. Nevertheless, this dataset also recorded a substantial number of missing values: out of 27654 transcripts, 21385 were not presenting any synonyms.

Therefore, we assess that the synonym datasets were hardly adequate for our task and we want to emphasize the requirement of having sufficiently large and comprehensive sets of synonyms for efficient text-mining.

6.3 Quality of dataset resulted after text-mining

Given the myriad of JSON lines in the dataset, we encountered the challenge of managing large quantities of data and in order to efficiently achieve our research goals. As such, filtering down our data was an impelling necessity.

A first step was to inspect how frequent the terms "Arabidopsis" and "Solanum" occur in our mined documents, represented each on a new JSON line. This was envisioned as a good indicator for the documents with high probability of clearly mentioning genes and transcripts belonging to the two plant species, and, moreover, for the documents in whose knowledge graphs these genes and transcripts might be in a relation to each other directly.

We were interested, for statistical purposes, to check both for the documents which would provided statements between entities or without any statements at all.

	# of occurrences in documents with statements	# of occurrences in documents without statements
"Arabidopsis"	464148	40099
"Solanum"	1076	32

TABLE 6.1: Table presents the number of occurrences of words "Solanum" and "Arabidopsis" in the dataset's documents

It is evident from the table 6.1, that a higher number of documents mentioned the model plant's name more frequently, which is not surprising due to the extensive research conducted on Arabidopsis Thaliana.

For a better understanding of how the entities are linked, we looked for the most common predicates and for the most common relationships types from the documents with present statements. These are not related to statements regarding the plant species vocabulary, but to all statements. Therefore, Table 6.2 presents all types of recognized relationships in documents' statements and the frequency of their occurrence in the mined dataset. With respect to used predicates, there were almost eleven thousands different predicates extracted.

relationship type	count
"associated"	33915747
"compared	2644928
"induces"	396870
"treats"	291107
"decreases"	177580
"method"	5205151
"inhibits"	622823
"administered"	450442
"metabolises"	127910
"interacts"	2637199

TABLE 6.2: Table presents the relationship types recognized by the mining platform

One potential weakness of this dataset is the large number of papers (over 40k) which did not contain any statements and were therefore useless for constructing knowledge graphs between existing entities. This issue may arise due to the fact that statements are only extracted if at least two entities are mentioned within the same sentences, being therefore connected via a grammatical structure. Upon further inspection, it was discovered that more than 13k papers out of all scraped PubMed documents were missing abstracts.

However, the most salient weakness is the introduction of ambiguous connections, which is rather a consequence resulted from the use of a weak vocabulary, as presented in 6.2. Ambivalent terms can be a cause of confusion in different scenarios, either representing different concepts or not being representative at all for the transcripts appearing in different abstracts or titles. To be more specific, synonyms such as "mutant", "polar", "mania", "grounded", "circadian rhythms", "circadian rhythm",

"woody", "ball", "scream", "family 77" and many more are misleading, while synonyms such as "thiamine", "superoxide dismutase" may point to other biomedical entities such as enzymes, vitamins etcetera.

Chapter 7

Evaluation

The evaluation chapters reflects our results and our considerations with respect to the methodology and to the final system. We address two cases:

1. **the first study case** - experiment 1 - which involved computing the similarity scores between genes which are considered orthologous by the scientific community using only the annotation datasets presented in 4.1

2. **the follow-up case** - experiment 2 - which entailed using PubMed text-mined information, presented in 4.4, related to our genes and transcripts in order to extend the previously used knowledge base with novel pieces of data, therefore changing the topology of the graph on which the similarity scores are computed and introducing potentially unreliable information

The objective of the evaluation process is not limited to providing a straightforward responses in terms of the accuracy of annotations in representing orthology relationships, but rather we aim to quantify the degree to which the selected methodology, design decisions, and overall framework are capable of replicating the given orthology relationships. The evaluation process seeks to provide numerical values that indicate the likelihood of success in reproducing the orthology relationships, considering the specific approaches we adopted.

7.1 Results for experiment 1

The results for the original experiment are emphasized in table 7.1. With respect to figure 7.1, it is clearly that the most pairs of ortholog genes tend to have a high computed similarity score based only on the annotations.

ranges of similarity scores	count
0	623
0.1 - 0.19	261
0.2 - 0.29	565
0.3 - 0.39	588
0.4 - 0.49	608
0.5 - 0.59	2037
0.6 - 0.69	2767
0.7 - 0.79	2170
0.8 - 0.89	2815
0.9 - 1.0	13007

TABLE 7.1: Table summarizes the number of transcripts' pairs having a computed similarity score in a particular range of values

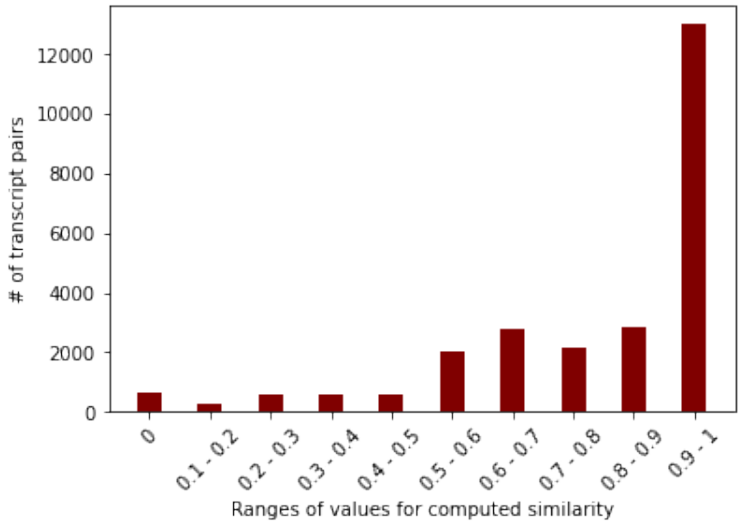


FIGURE 7.1: The distribution of transcript pairs having a computed similarity score between a range of values

The analysis revealed that 12,652 pairs of transcripts obtained a similarity score of 1.0, which means that the subgraphs comprising their nodes and their annotations, along with edges between them, were identical. Nearly half of the entire dataset of ortholog pairs were correctly depicted using the subgraph similarities with perfect accuracy.

Nevertheless, is it evident that the number of orthologous pairs with low computed similarities is not negligible: 10% of the total number of orthologous pairs obtained similarity scores below 50%.

7.2 Results for experiments 2

Figure 7.2 depicts the number of transcript pairs which had the similarity scores between different ranges.

Our analysis involved comparing the computed similarities derived from the initial experiment with those obtained in the subsequent experiment, utilizing the augmented graph and employing the same query. The results of our investigation reveal unfortunate findings. Specifically, approximately 87% of the orthologous pairs (amounting to 22,090 pairs) yielded identical outcomes, suggesting that the augmented graph did not significantly impact their computed similarity scores. However, we observed that 13% of the pairs, an equivalent to 3,351 pairs, exhibited lower computed similarity scores subsequent to the incorporation of the new scraped entities into the graph database. Remarkably, none of the pairs demonstrated an improvement in their similarity scores, highlighting the absence of positive impact resulting from the augmentation.

When looking at the numerical difference between the two computed similarities, we concluded that the declines in similarity scores fell mostly within specific ranges:

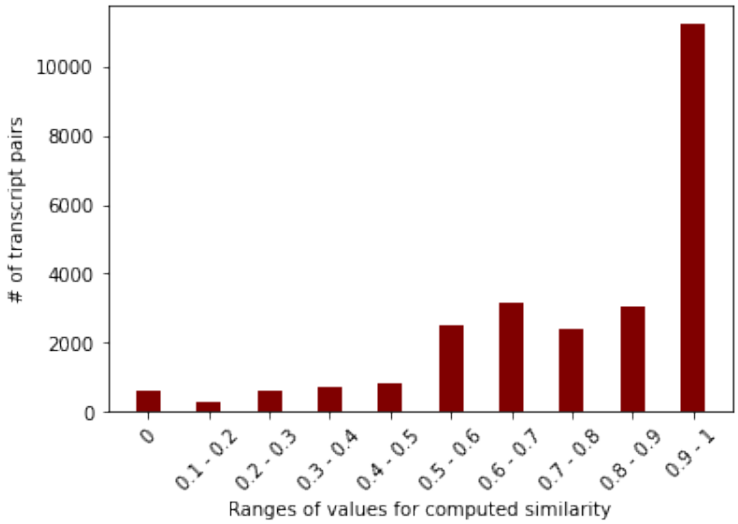


FIGURE 7.2: The distribution of transcript pairs having a computed similarity score after introducing mined entities in the graph database between a range of values

ranges of similarity scores	count
0 - 0.1	690
0.1 - 0.2	1162
0.2 - 0.3	682
0.3 - 0.4	438
0.4 - 0.5	249
0.5 - 0.6	51
0.6 - 0.7	51
0.7 - 0.8	18
0.8 - 0.9	3
0.9 - 1	0
1 - 2	7

TABLE 7.2: Table summarizes the number of transcript pairs having the difference between the computed similarity scores in a particular range of values

The most striking observation is that three pairs of transcripts recorded a difference between the computed similarity scores equal to 2. These instances are of particular interest, as they indicate a transition from a perfect similarity score of 1 in the initial scenario to being completely unrelated, as evidenced by an assigned similarity score of -1."

When inspecting the pairs in the Neo4j graph, it was easily recognizable that the transcript belonging to the Arabidopsis Thaliana species was now linked to many more nodes of type *Entity*, therefore downgrading the similarity score obtained using the Jaccard metric.

The fruit of this discussion germinates from the biomedical entities mined which were common for both species. It is essential to look which were these entities, as far as they are represented in the vast vocabulary of the mining tool. The curated items are presented in Figure 7.3 . As mentioned in Section 5.4, everything entity was treated as lowercase, narrowing down from 51 entities to 47.

'protein', 'kinase', 'superoxide dismutase', 'peroxidase', 'tobacco', 'parents', 'death', 'ethylene', 'shock', arabidopsis', 'Arabidopsis', 'cloning', 'beta-glucuronidase', 'rice', 'hydrogen peroxide', 'shape', 'tomato', 'magnesium', 'gus', 'fusarium', 'pod', 'acyltransferase', 'gene expression analysis', 'leucine', 'EIN2', 'transmembrane protein', 'Fusarium', 'exome sequencing', 'dioxygenases', 'circadian rhythms', 'heat-shock proteins', 'qpcr', 'sequence analyses', 'solanum', 'mamps', 'NAC', 'sterility', 'male sterility', 'solanum lycopersicum', 'NBS', 'Tomato', 'lyase', 'engase', 'sHSP', 'ERECTA', 'Solanum habrochaites', 'P. aegyptiaca', 'S. lycopersicum', 'shsp', 'solyc09g075080', 'solyc01g068560'

FIGURE 7.3: Common entities for transcripts from Arabidopsis Thaliana and Solanum Lycopersicum

It is obvious that terms such as "arabidopsis", "tomato", "solanum lycoperiscum", "solanum", "S. lycopersicum" are terms which can inflict a lot of confusion in our experiments as they

7.2.1 Tackling provenance

The results obtained while keeping the strength property of the relationships of type *MINED* at half of the initial value are presented as following:

1. only 3,345 pairs recorded different values between the two cases
2. overall, 3,269 pairs had better similarity scores when all relationships were treated equally in contrast to when the the newly embedded relationships' strength property was halved.

7.3 Limitations

In this section we address the limitations of our study and whether we are capable to suggest, summarizing all our observations, any possible solutions or steps for improvement for future researches.

7.3.1 Quality of datasets

Chapter 6 depicted an encompassing view of how the quality of the datasets influenced the results and our design decisions. In general, our study could have potentially generated further insights and achieved higher scores for a greater number of pairs between transcripts of the two plant species if the following conditions were met:

- The annotation datasets contained fewer instances of missing values, ensuring more complete and reliable information.
- The dataset outlined in Section 4.4 possessed a more robust underlying vocabulary, thereby minimizing the presence of ambiguous terms and enhancing the accuracy of the results.
- A Solanum Lycopersicum synonym dataset could have been available, enabling more comprehensive analysis and improving the accuracy of cross-species comparisons.

7.3.2 Neo4j inbuilt functions

Using pre-existing inbuilt functions by Neo4j can offer an immediate and convenient solution for calculating the similarity scores. However, this can also account for a source of limitation in terms of available similarity metrics and their potential to capture unique characteristics of the datasets and the relationships among nodes. As introduces in 5.3.2, we relied on the Jaccard metric for the node similarity function, even though there was another option delivered by Neo4j to use the Overlap metric.

Moreover, for similar projects, it might be required that custom made similarity functions have to be developed, taking into considerations the particularities of the problem itself and of the underlying datasets it is modelled after. Even though Neo4j offers flexibility with respect to which are the parts of the database considered for a projection in-memory, respectively, to the properties allocated for relationships, one has to considered whether stepping away from the platform's functions can better capture the problem.

An alternative approach would be the use of **node embeddings** or **subgraph embeddings**. Node embeddings are low-dimensional vector representations of nodes in a graph, frequently employed in machine learning problems. Graph embeddings have the considerable advantage of retaining rich information about the network and the property of nodes and relationships within a graph.

Neo4j GDS library has several node embedding algorithms: the FastRP technique is used especially to preserve the similarity between nodes and their neighbours. Embeddings are calculated in iterations by using random walks in the graph, and the number of iteration is a tunable hyperparameter which can highly influence the final embeddings assigned to each node. Only one iteration will consider the direct neighbouring nodes (i.e. in our case - the annotations for each transcripts), while more iterations will include information containing neighbours which are further away. Additionally, FastRP is supposed to work with undirected graphs, but choosing orientation for relationships (outgoing vs. incoming) can also affect the embeddings tremendously and yield unexpected results.

7.3.3 Insufficient Memory

As discussed in 5.3.2, memory bounds can be a dire limitation once the graph database increases in size considerably. For our particular experiments, the number of pairwise computed similarity scores which were outputted by Neo4j were constrained by the memory available for it. Notwithstanding the memory constraints, there could be four straightforward solutions for alike experiments where one could:

- employ machines with increased memory capacities
- use alternative graph projections
- compute similarities for a smaller subset of nodes
- employing a threshold for displaying similarity scores

7.3.4 Graph properties

We saw in Section 5.4 that the newly established relationships between transcripts from both plants species and the mined, common entities had all been embedded

into the graph database under the same label: *MINED*. However, with better understanding of the biological underlines of each term, different items could be aggregated under different types of nodes, instead of a general type *Entity*, and with different types of relationships to the corresponding transcript nodes. This could benefit the system by allowing multiple and disjoint properties per type of relationship.

7.3.5 Provenance management

Lastly, an evident area for improvement lies in the design choice of dealing with provenance management. We conceived that creating different types of relationships and with a better understanding of the connections between subsets of transcripts to subsets of entities, different weights can be attributed to them. For instance, one modality of tackling provenance could be using the metadata contained per document from the dataset deployed by the text-mining platform service presented in 4.4 in order to track the authority of the journals and/or conferences where the papers amassed in PubMed were presented. Needless to say, one could consider that top A conferences and journal could instantiate more accurate and complete information. The statements extracted from papers with higher recognition would be assigned a higher weight as well in our model.

Chapter 8

Conclusions

At the beginning of our this study, we set out to learn more about Arabidopsis Thaliana, about gene similarity and about key concepts in the real of comparative genomics. The main goal was to design a framework which could encapsulate large datasets containing annotations of two different plant species, which could yield valuable information about orthologous genes.

However, throughout the course of our research, we inevitably encountered a common challenge prevalent in various research domains: the issue of having an insufficient number of observations in our data, thereby hindering the generation of meaningful estimates. Notwithstanding this fact, we firmly consider that such problems account for worthwhile lessons.

We believe that this research presents a valuable methodology for studying gene similarity using graph-based techniques, offering several advantages. Firstly, it provides contextual information by leveraging subgraph similarities, which take into account the relationships and interdependencies between nodes, thereby capturing the complexity inherent in biological networks. Secondly, this methodology allows for the integration of multiple data sources, enabling a more comprehensive understanding of gene relationships and functions. Lastly, it addresses the issue of annotation incompleteness by identifying similar subgraph structures, bridging gaps in knowledge regarding gene functions, even when annotations for two genes may be incomplete or divergent.

On the other hand, despite our efforts, the results of this study were not entirely satisfactory, primarily due to several factors. The prevalent conclusion we draw was that **data quality** is a major factor when determining the outcomes of such experiments. We cannot overstate the significance of data integrity in research of this nature. When computing the similarity scores between Arabidopsis Thaliana and Solanum Lycopersicum, we encountered ambiguity regarding whether the results unveiled intriguing cases of orthology between genes or if we were simply dealing with incomplete data, as both annotation datasets contained a considerable number of unannotated transcripts.

Additionally, our exploration of using Natural Language Processing (NLP) techniques to enrich the annotation datasets yielded uninformative results. Once again, we attribute this outcome primarily to data quality issues. It is imperative to emphasize the importance of employing clean vocabulary datasets specifically tailored for extracting biomedical entities from scientific papers. Furthermore, the integration of mined entities should only occur after a comprehensive and thorough analysis to ensure that the additional information genuinely pertains to other biomedical entities, as opposed to mere general words such as "shape" or "parent."

In summary, this study provides valuable lessons into gene similarity through

the application of graph-based methods. It highlights the significance of data quality, underlining the need for comprehensive and accurate annotation datasets. Furthermore, it outlines challenges associated with integrating text-mined information, emphasizing the requirement for meticulous analysis and the utilization of reliable sources. By addressing these considerations, future research endeavors in this field can enhance the reliability and effectiveness of similar methodologies.

Bibliography

Altenhoff, Adrian M and Christophe Dessimoz (2012). "Inferring orthology and paralogy". In: *Evolutionary Genomics: Statistical and Computational Methods, Volume 1*, pp. 259–279.

Altenhoff, Adrian M et al. (2010). "OMA 2011: orthology inference among 1000 complete genomes". In: *Nucleic acids research* 39.suppl_1, pp. D289–D294.

Altschul, Stephen F et al. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs". In: *Nucleic acids research* 25.17, pp. 3389–3402.

Bayat, Ardeshir (2002). "Science, medicine, and the future: Bioinformatics". In: *BMJ: British Medical Journal* 324.7344, p. 1018.

Chen, Feng et al. (2006). "OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups". In: *Nucleic acids research* 34.suppl_1, pp. D363–D368.

Conceição, Sofia IR and Francisco M Couto (2021). "Text mining for building biomedical networks using cancer as a case study". In: *Biomolecules* 11.10, p. 1430.

Consortium, Gene Ontology (2004). "The Gene Ontology (GO) database and informatics resource". In: *Nucleic acids research* 32.suppl_1, pp. D258–D261.

DeLuca, Todd F et al. (2006). "Roundup: a multi-genome repository of orthologs and evolutionary distances". In: *Bioinformatics* 22.16, pp. 2044–2046.

Dennis, Elizabeth S et al. (2008). "Genetic contributions to agricultural sustainability". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 363.1491, pp. 591–609.

Dessimoz, Christophe et al. (2005). "OMA, a comprehensive, automated project for the identification of orthologs from complete genome data: introduction and first achievements". In: *Comparative Genomics: RECOMB 2005 International Workshop, RCG 2005, Dublin, Ireland, September 18-20, 2005. Proceedings 3*. Springer, pp. 61–72.

Ensembl (n.d.). Accessed: 2022-10-15. URL: https://plants.ensembl.org/index.html.

Genetwister (n.d.). Accessed: 2023-02-27. URL: https://www.genetwister.nl/.

Harbola, Aditya et al. (2022). "Bioinformatics and biological data mining". In: *Bioinformatics*. Elsevier, pp. 457–471.

Hufford, Matthew B et al. (2012). "Comparative population genomics of maize domestication and improvement". In: *Nature genetics* 44.7, pp. 808–811.

InParanoiDB (n.d.). Accessed: 2022-04-27. URL: https://inparanoidb.sbc.su.se/.

Jensen, Lars Juhl et al. (2007). "eggNOG: automated construction and annotation of orthologous groups of genes". In: *Nucleic acids research* 36.suppl_1, pp. D250–D254.

JGI Phytozome (n.d.). Accessed: 2023-02-27. URL: https://phytozome-next.jgi.doe.gov/biomart/martview/fc1d7f71ddc3c5174d4b3076fb19bbf8.

Jothi, Raja et al. (2006). "COCO-CL: hierarchical clustering of homology relations based on evolutionary correlations". In: *Bioinformatics* 22.7, pp. 779–788.

Jumper, John et al. (2021). "Highly accurate protein structure prediction with AlphaFold". In: *Nature* 596.7873, pp. 583–589.

Jurca, Gabriela et al. (2016). "Integrating text mining, data mining, and network analysis for identifying genetic breast cancer trends". In: *BMC research notes* 9, pp. 1–35.

Kawashima, Koya, Wenjun Bai, and Changqin Quan (2017). "Text mining and pattern clustering for relation extraction of breast cancer and related genes". In: *2017 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*. IEEE, pp. 59–63.

Kriventseva, Evgenia V et al. (2007). "OrthoDB: the hierarchical catalog of eukaryotic orthologs". In: *Nucleic acids research* 36.suppl_1, pp. D271–D275.

Kroll, Hermann et al. (2023). "A discovery system for narrative query graphs: entity-interaction-aware document retrieval". In: *International Journal on Digital Libraries*, pp. 1–22.

Li, Li, Christian J Stoeckert, and David S Roos (2003). "OrthoMCL: identification of ortholog groups for eukaryotic genomes". In: *Genome research* 13.9, pp. 2178–2189.

Lin, Zeming et al. (2023). "Evolutionary-scale prediction of atomic-level protein structure with a language model". In: *Science* 379.6637, pp. 1123–1130.

Linard, Benjamin et al. (2011). "OrthoInspector: comprehensive orthology analysis and visual exploration". In: *BMC bioinformatics* 12.1, pp. 1–13.

Meinke, David W et al. (1998). "Arabidopsis thaliana: a model plant for genome analysis". In: *Science* 282.5389, pp. 662–682.

Mouse Genome Informatics (n.d.). Accessed: 2022-04-27. URL: https://www.informatics.jax.org/.

Muller, Jean et al. (2010). "eggNOG v2. 0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations". In: *Nucleic acids research* 38.suppl_1, pp. D190–D195.

National Human Genome Institute (n.d.[a]). Accessed: 2023-02-27. URL: https://www.genome.gov/genetics-glossary/DNA-Sequencing#:~:text=DNA%20sequencing%20refers%20to%20the,use%20to%20develop%20and%20operate.

National Human Genome Institute (n.d.[b]). Accessed: 2023-02-27. URL: https://www.genome.gov/genetics-glossary/Transcription#:~:text=Transcription%2C%20as%20related%20to%20genomics,protein%20information%20encoded%20in%20DNA..

National Human Genome Institute (n.d.[c]). Accessed: 2023-03-19. URL: https://www.genome.gov/genetics-glossary/Duplication#:~:text=Duplication%2C%20as%20related%20to%20genomics,Duplications%20occur%20in%20all%20organisms..

National Library of Medicine (n.d.). Accessed: 2023-02-27. URL: https://pubmed.ncbi.nlm.nih.gov/.

Östlund, Gabriel et al. (2010). "InParanoid 7: new algorithms and tools for eukaryotic orthology analysis". In: *Nucleic acids research* 38.suppl_1, pp. D196–D203.

Overbeek, Ross et al. (1999). "The use of gene clusters to infer functional coupling". In: *Proceedings of the National Academy of Sciences* 96.6, pp. 2896–2901.

Pesquita, Catia et al. (2009). "Semantic similarity in biomedical ontologies". In: *PLoS computational biology* 5.7, e1000443.

Remm, Maido, Christian EV Storm, and Erik LL Sonnhammer (2001). "Automatic clustering of orthologs and in-paralogs from pairwise species comparisons". In: *Journal of molecular biology* 314.5, pp. 1041–1052.

Rindflesch, Thomas C et al. (1999). "EDGAR: extraction of drugs, genes and relations from the biomedical literature". In: *Biocomputing 2000*. World Scientific, pp. 517–528.

Schlicker, Andreas et al. (2006). "A new measure for functional similarity of gene products based on Gene Ontology". In: *BMC bioinformatics* 7, pp. 1–16.

Sevilla, Jose L et al. (2005). "Correlation between gene expression and GO semantic similarity". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 2.4, pp. 330–338.

Shackelton, Laura A and Edward C Holmes (2004). "The evolution of large DNA viruses: combining genomic information of viruses and their hosts". In: *Trends in microbiology* 12.10, pp. 458–465.

Singhal, Ayush et al. (2016). "Pressing needs of biomedical text mining in biocuration and beyond: opportunities and challenges". In: *Database* 2016.

Sivashankari, Selvarajan and Piramanayagam Shanmughavel (2007). "Comparative genomics-a perspective". In: *Bioinformation* 1.9, p. 376.

Smith, Temple F, Michael S Waterman, et al. (1981). "Identification of common molecular subsequences". In: *Journal of molecular biology* 147.1, pp. 195–197.

Sonnhammer, Erik LL and Eugene V Koonin (2002). "Orthology, paralogy and proposed classification for paralog subtypes". In: *TRENDS in Genetics* 18.12, pp. 619–620.

Spreafico, Roberto et al. (2020). "Advances in genomics for drug development". In: *Genes* 11.8, p. 942.

Tair - About Arabidopsis (n.d.). Accessed: 2023-03-19. URL: https://www.arabidopsis.org/portals/education/aboutarabidopsis.jsp.

Tatusov, Roman L, Eugene V Koonin, and David J Lipman (1997). "A genomic perspective on protein families". In: *Science* 278.5338, pp. 631–637.

Wall, DP, HB Fraser, and AE Hirsh (2003). "Detecting putative orthologs". In: *Bioinformatics* 19.13, pp. 1710–1711.

Wang, Haiying et al. (2004). "Gene expression correlation and gene ontology-based similarity: an assessment of quantitative relationships". In: *2004 Symposium on Computational Intelligence in Bioinformatics and Computational Biology*. IEEE, pp. 25–31.