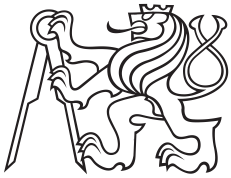
Master's Thesis



Czech Technical University in Prague

F3

Faculty of Electrical Engineering
Department of Computer Science

Automated annotation of non-coding RNAs

Lucie Mühlfeitová

Supervisor: doc. Ing. Jiří Kléma, Ph.D.
Field of study: Medical Electronics and Bioinformatics
Subfield: Bioinformatics
January 2024

MASTER‘S THESIS ASSIGNMENT

I. Personal and study details

Student's name: **Mühlfeitová Lucie** Personal ID number: **483665**

Faculty / Institute: **Faculty of Electrical Engineering**

Department / Institute: **Department of Computer Science**

Study program: **Medical Electronics and Bioinformatics**

Specialisation: **Bioinformatics**

II. Master’s thesis details

Master’s thesis title in English:

Automated annotation of non-coding RNAs

Master’s thesis title in Czech:

Automatická anotace nekódujících RNA

Guidelines:

1. Get acquainted with the automated annotation of non-coding RNAs.
2. Learn about basic methods for building links between molecules.
3. Carry out a literature search for methods ad 1 and 2.
4. Propose a simple tool that will help to find interesting piRNAs and annotate them in the MDS dataset provided by your supervisor. Goals:
a. Find piRNAs that are differentially expressed.
b. Find interacting transpozons and other RNAs.
c. Propose piRNA annotations.
5. Evaluate the tool (the gold standard annotations are not available, but give an overview of the number of findings and verify them in biological literature).

Bibliography / sources:

Oliver, Stephen. "Guilt-by-association goes global." Nature 403.6770 (2000): 601-602.
Cardenas, Jacob, Uthra Balaji, and Jinghua Gu. "Cerina: systematic circRNA functional annotation based on integrative analysis of ceRNA interactions." Scientific reports 10.1 (2020): 1-14.
Liu, Yajun, et al. "Computational methods and online resources for identification of piRNA-related molecules." Interdisciplinary sciences: computational life sciences 13.2 (2021): 176-191.

Name and workplace of master's thesis supervisor:

doc. Ing. Jiří Kléma, Ph.D. Intelligent Data Analysis FEE

Name and workplace of second master's thesis supervisor or consultant:

Date of master's thesis assignment: **08.08.2023** Deadline for master's thesis submission: **09.01.2024**

Assignment valid until: **16.02.2025**

doc. Ing. Jiří Kléma, Ph.D. Supervisor's signature	Head of department's signature	prof. Mgr. Petr Páta, Ph.D. Dean's signature

III. Assignment receipt

The student acknowledges that the master's thesis is an individual work. The student must produce her thesis without the assistance of others, with the exception of provided consultations. Within the master's thesis, the author must state the names of consultants and include a list of references.

.	
Date of assignment receipt	Student's signature

Acknowledgements

I would like to thank my supervisor doc. Ing. Jiří Kléma, Ph.D for his patience, assistence and guidance and help during the work on this diploma thesis. I would also like to thank my family and friends for their support.

Declaration

I declare that the presented work was developed independently and that I have listed all sources of information used within it in accordance with the methodical instructions for observing the ethical principles in the preparation of university thesis.

In Prague, ...

Signature: .....

Abstract

This study focuses on the application of methods for the automated functional annotation of non-coding RNAs (ncRNAs), with particular focus on PIWI-interacting RNA (piRNA). The primary aim of these methods is to assign previously unexplored piRNAs to their correct biological functions. An interaction network between piRNAs, transposons, and genes was constructed using gene expression data and sequence data. Gene annotations are known, and the method of random walks with restart and the general principle of guilt by association has been used to promote the annotation from genes to piRNA. The functionality of the method was validated in the specific domain of myelodysplastic syndrome (MDS). The success of the method was measured by the proportion of MDS-related gene ontology (GO) terms to the number of all assigned GO terms, as there is currently no available database containing the correct assignment of functional annotations to individual piRNAs. It was shown that piRNAs showing more significant differences in the expression levels between groups of patients with MDS and healthy controls were associated with a greater proportion of functional annotations related to MDS.

Keywords: non coding RNA, PIWI-interacting RNA, Myelodysplastic syndrome, Random walks, Random walks with restart, Functional annotations, Gene ontology, Permutation test

Supervisor: doc. Ing. Jiří Kléma, Ph.D.

Abstrakt

Tato práce se zaměřuje na aplikaci metod pro automatizovanou funkční anotaci nekódujících RNA (ncRNA), konkrétně PIWI-interagujících RNA (piRNA). Hlavním cílem těchto metod je přiřadit dosud neprozkoumané piRNA správné biologické funkce. Data genové exprese a sekvenční data byla využita ke konstrukci interakčního grafu mezi piRNA, transpozony a geny. Anotace genů jsou známé, k propagaci anotací od genů k piRNA byla využita metoda náhodných procházek s restartem a obecný princip 'guilt by association'. Přidělení funkční anotace (GO term) bylo posouzeno na základě výsledků permutačních testů. Funkčnost metody byla ověřována v konkrétní doméně myelodysplastického syndromu (MDS). Úspěšnost metody byla vyhodnocována podílem termů genové ontologie (GO) souvisejících s MDS vůči počtu všech přiřazených GO termů, jelikož v současné době neexistuje dostupná databáze obsahující správné přiřazení funkčních anotací k jednotlivým piRNA. Bylo prokázáno, že piRNA vykazující výraznější rozdíly v úrovních exprese mezi skupinami pacientů s MDS a zdravými kontrolami byly spojeny s větším podílem funkčních anotací souvisejících s MDS.

Klíčová slova: nekódující RNA, PiRNA, Myelodysplastický syndrom, Náhodné procházky, Náhodné procházky s restartem, Funkční anotace, Genová ontologie, Permutační test

Překlad názvu: Automatická anotace nekódujících RNA

Contents

Figures

Tables

Chapter 1

Introduction

The fundamental goal of this diploma thesis is to assign Gene Ontology (GO) terms to PIWI-interacting RNAs (piRNA) that could possibly be associated with myelodysplastic syndromes (MDS). The expression data obtained from the Institute of Hematology and Blood Transfusion, which includes measurements from patients suffering from MDS and healthy control subjects, is used for this purpose. We were given expression data for a wide range of piRNAs, transposable elements (TE), and other genes. The concept of differential expression is introduced to select piRNAs of interest to us.

The GO terms assignment of the selected piRNAs will be proposed on the basis of the concept of the guilt by association principle, which states that molecules that interact with one another have functions that are similar and, as a result, share annotations. We benefit from the expanding database of scientifically validated functional annotations for numerous genes to assigned the annotations for our piRNAs. The interactions among the genes in the network will be determined by investigating the correlations of their gene expression data. Furthermore, the connections between piRNAs and TEs will be established by considering their sequence complementarity. Subsequently, the random walks with restart technique will be employed, since it has demonstrated success for other types of non-coding RNA (ncRNA)[3], [4], [5]. The random walks with restart algorithm will be performed on a gene co-expression network consisting of piRNAs, TEs, and other genes. The assignment of GO terms to piRNAs will be determined based on the outcomes of permutation tests.

1.1 Motivation

Non-coding RNAs (ncRNAs) are a class of RNA molecules that do not encode proteins but play crucial roles in various cellular processes. They have gained significant attention in recent years due to their diverse biological functions, including gene regulation, chromatin remodelling, and genome stability. Accurate annotation of ncRNAs is essential for understanding their functional roles and mechanisms of action. However, manual annotation of ncRNAs is labour-intensive and time-consuming. Therefore, the development of automated annotation methods is crucial for efficient analysis and interpretation of ncRNA data [6]. In my diploma thesis, I focus on one class of these ncRNAs, piwi-interacting RNA (piRNA).

PIWI-interacting RNA (piRNA) is a type of small RNA molecules that are not involved in encoding proteins. They play a significant role in regulation of transposable elements (TEs) and ensuring the stability of the genome [7]. The most commonly recognised function of piRNA is to silence transposable elements (TEs). However, studies indicates that numerous piRNA sequences are derived from genomic regions that are unrelated to transposable elements (TEs), implying that piRNAs serve purposes beyond the suppression of TEs [8]. Irregular piRNA expressions have been detected in several diseases, specifically in tumours and disorders of the reproductive system. Therefore, piRNAs offer potential as novel biomarkers for early detection and as targets for precise medicine in therapeutic interventions [1]. This thesis focuses on their possible associaiton with myelodysplastic syndromes (MDS).

1.2 Text structure

Chapter 1 presents the primary aims and approaches that were used in this thesis. It also offers some insight into the significance of investigating this topic. The first part of this thesis centres on the investigation of the subjects related to this study and the methodologies employed to accomplish the objectives. Chapter 2 provides an extended overview of PIWI-interacting RNAs, focusing on their role as transposon silencers and their involvement in different diseases and disorders. Chapter 3 examines techniques that can be employed to automatically assign functional annotations to non-coding RNA. It introduces the concepts of gene co-expression networks, the random walks algorithm, and the principle of guilt by association. Chapter 4 offers information regarding myelodysplastic syndromes.

The second part of this thesis focuses on the actual implementation of the automated annotation method employing random walks. Chapter 5 provides details

about the data set used in our study and the procedures undertaken for adjusting the data to the desired form. Chapter 6 contains all the details regarding the gene co-expression network that was built. This chapter also describes the technique implemented to determine sequence complementarity between piRNAs and TEs. Chapter 7 provides a comprehensive overview of the random walks technique, with a particular focus on random walks with restart. It presents information regarding the application of this method in the given work, as well as introducing the usage of permutation tests. The results of our experiments are presented in chapter 8. Chapter 9 provides a comprehensive overview of the whole project and offers suggestions for future actions.

Part I

Research

Chapter 2

PIWI-interacting RNA

PIWI-interacting RNA (piRNA) is a class of small non-coding RNA molecules that play a crucial role in the regulation of transposable elements (TEs) and the maintenance of genome stability. PiRNAs are primarily expressed in the germline cells of animals. PiRNAs are 26–31 nucleotides in length, which distinguishes them from micro RNAs (miRNAs) that are generally 21–23 nucleotides long or short interfering RNAs (siRNAs) [7]. The piRNA class is the largest and the most diverse of all small non-coding RNAs (sncRNA) [9].

PiRNAs are abundant in animal reproductive organs, where their primary function is to suppress transposons. Transposons are genomic elements that can drive evolutionary change but are also regarded as self-centred DNA parasites. The absence of piRNAs causes transposons to activate, causing genomic damage and complications in the development of reproductive organs, ultimately affecting fertility. As a result, piRNA-mediated transposon silencing is critical for the successful reproduction of sexually reproducing animals. The piRNA pathway has been compared to a genetic immune system, with piRNAs acting as genome guardians against invasive foreign DNA elements [10].

The majority of piRNAs contain complementary sequences to transposon RNAs in the opposite direction. Because of this complementarity, piRNAs can act as oligonucleotides, suppressing and controlling these mobile DNA elements precisely [10].

While our understanding of piRNA biogenesis is far from complete, two summarised mechanisms shed light on the production of mature piRNAs. The first mechanism involves the 'ping-pong' amplification mechanism, in which piRNAs

act as transposon silencers while also enhancing their own presence. The second mechanism is primary synthesis, in which piRNAs produced may play a regulatory role in mRNA expression [9].

The presence of piRNAs was suggested in Drosophila, initially referred to as repeat-associated RNAs (rasiRNAs). However, the true comprehension of piRNA biology only occurred with the advent of next-generation sequencing [11]. In 2006 the existence of piRNA was reported by several independent research groups. At that time, they reported an unknown class of small non-coding RNAs found in fly, mouse and rat germ cells [8]. They revealed that these RNAs were longer (26–31 nt) compared to miRNAs and siRNAs, clustered throughout the genome, mainly matching transposable element (TE) sequences, and specifically present in testes [11]. Since then, significant progress has been achieved in the study of piRNAs, leading to a better understanding of how piRNA clusters are transcribed, the process of piRNA biogenesis, and various aspects of piRNA function [8].

Despite differences in target regulation and formation methods, the three primary categories of non-coding RNAs (ncRNAs) share common functionalities, such as directing Argonaute proteins to nucleic acid targets based on complementary base-pairing rules [8] [12]. Humans have eight Argonaute proteins, four of which are Ago (Argonaute) subfamily proteins and four of which are PIWI (P-element induced wimpy testis) subfamily proteins. AGO proteins are present in various tissues and bind to miRNAs or siRNAs, while PIWI proteins are predominantly expressed in animal germ cells and specifically associate with piRNAs [8]. When Ago proteins are expressed, they bind to siRNAs and miRNAs, transforming double-stranded precursors into mature small RNAs of 20–22 nucleotides (nt) via a dicer-dependent process.

PIWI proteins, on the other hand, form a distinct RNA-induced silencing complex (RISC), known as piRISCs, with a small RNA population known as piRNAs. The biogenesis of primary piRNAs is a dicer-independent process in which different nucleases cut the long single strands into individual piRNA units [12]. Mature piRNAs exhibit a size range of 26–30 nucleotides and possess a distinctive 2'-O-methylation at their 3' ends, characteristics that set them apart from other small RNAs like miRNAs and siRNAs. PiRNA precursors, unlike miRNAs and siRNAs, are single-stranded and lack obvious secondary hairpin structures. These precursors are typically derived from specific genomic regions containing repeating elements, and their synthesis is typically mediated by a dicer-independent pathway. Furthermore, additional posttranscriptional modifications are required for emerging piRNA maturation. PiRNA biogenesis is contributed by two primary pathways: primary synthesis and the secondary amplification cycle, also known as the "ping-pong cycle" [13]. PiRNAs, which are abundant in spermatogenic cells, contribute to stem cell self-renewal and play an important role in preserving germline and genome integrity by concealing insertional mutations from transposons [12].

The main recognised role of PIWI-piRNA complexes is to silence transposable elements (TEs) in animal germ cells at both the transcriptional and post-transcriptional levels. However, it has been observed that many piRNA sequences originate from genomic regions that are not related to TEs, suggesting that piRNAs have functions beyond TE silencing. Growing evidence indicates that the PIWI-piRNA machinery is also involved in regulating protein-coding genes in germ cells [8].

There is a thesis that piRNAs employ a targeting mechanism similar to miRNAs. However, there is a huge difference in the relationship between piRNA and its targets compared to that of miRNA. While miRNA targets develop to be recognised, piRNA targets are genetic elements of a parasitic nature that are pressured to avoid recognition. The collection of piRNAs in animals is typically significantly larger in scale compared to miRNAs, often differing by orders of magnitude. This suggests the need of existence of protective mechanisms to prevent the silencing of the entire germline transcriptome [14].

Irregular piRNA expressions have been observed in various diseases, particularly in tumors and disorders of the reproductive system. PiRNAs show potential as innovative biomarkers for early detection and as targets for precise medicine in therapeutic interventions [1].

2.1 Transposon Silencing

Transposons are important structural components in nearly all eukaryotic genomes, and their mobilisation can cause genetic instability, resulting in harmful mutations. Furthermore, mobile genetic elements contain transcriptional enhancers and insulators, allowing transposition to affect the expression of nearby genes as well as large chromatin domains. This process can cause coordinated changes in gene transcription, which can disrupt development or drive evolutionary changes [15].

Maintaining the genome's integrity is crucial, and effective suppression of transposable element (TE) activity is essential for this purpose. To ensure the production of gametes capable of fertilisation, TEs must undergo silencing. In the male germline, TE silencing takes place prior to meiosis, accomplished through the collaboration of DNMT enzymes and the piRNA pathway [11].

Initially, it was reported that piRNAs played a defensive role against the mobilization of transposons in germline cells of flies. Subsequent validation extended this finding to various organisms, including humans. Transposons, often referred to as jumping genes, resemble endogenous viruses and pose a threat to gene stability

by "copying and pasting" their DNA into the host genome for self-replication. This process has cascading effects: exon insertions disrupt the coding sequence, intron insertions alter splicing patterns potentially leading to the creation of novel and harmful fusion proteins. Transposon insertions also cause DNA nicks and double-strand breaks, and errors in their repair may induce recombination between transposon repeats, resulting in chromosomal duplication, deletion, translocation, or inversion [1].

At the transcriptional level, piRNAs and Piwi proteins directly modify chromatin structure and histone proteins within the nucleus by regulating DNA methyltransferase (DNMT). DNMT methylates CpG islands in promoter regions, suppressing transcription initiation. Piwi proteins guide DNMT to bind with transposable elements or target genes, leading to the silencing of these elements or genes as a consequence [1].

2.2 Disease associations

Recent evidence suggests that piRNAs play important roles in a variety of biological processes, including transposable element slicing in animal germlines, genome defence, and histone modification. Given piRNAs' involvement in gene regulation, there has been a growing interest in understanding their roles in human diseases. Numerous studies indicate that the dysregulation of piRNAs can either encourage or suppress the onset and progression of various diseases, particularly cancers [1]. Studies are increasingly revealing links between abnormal piRNA expression and a variety of diseases, including cancer, neurodegenerative conditions such as Alzheimer's and Parkinson's diseases, and geriatric conditions [16][12].

For instance, in the context of neurodegenerative disorders, differential piRNA expression between the healthy human brain and Alzheimer's disease has been observed. Notably, the Alzheimer's-diagnosed brain exhibits over ten-fold upregulation of five piRNAs piR-hsa-25781, piR-hsa-28467, piR-hsa-1177, piR-hsa-26593, and piR-hsa-29114—among a total of 146 upregulated and 3 downregulated piRNAs. This specific piRNA signature could serve as an effective diagnostic marker for Alzheimer's disease [12].

In the realm of cancers, the expression of piR-651 was found to be elevated in various cancer cell lines, including those associated with gastric, lung, breast, mesothelium, liver, and cervical cancers. Additionally, piR-823 demonstrated significant upregulation in colorectal tumorigenesis, where it interacts with HSF1, enhancing its transcriptional activity and phosphorylation at Ser326, thereby acting

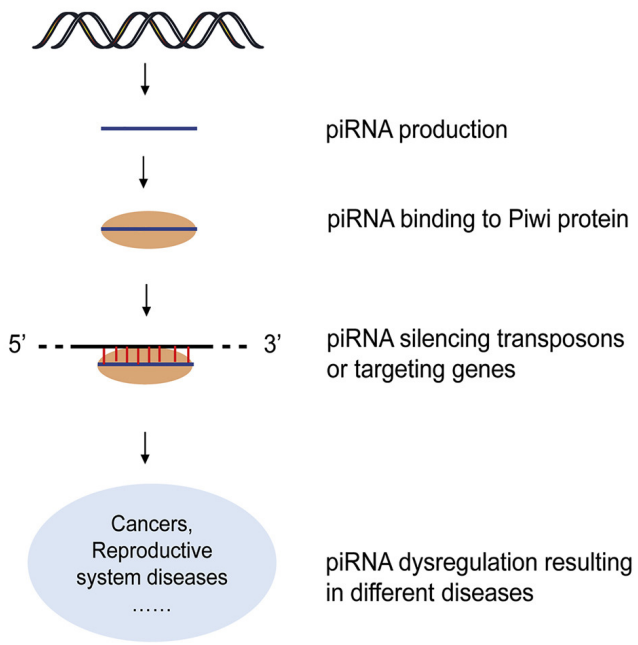


Figure 2.1: Simple graphics showing the role of piRNAs in the development of diseases [1].

as an active promoter of tumour growth [12].

Therefore, biological studies suggest that piRNAs could be used as biomarkers or therapeutic targets for disease diagnosis, prognosis and treatment. As a result, it is critical to identify associations between piRNAs and diseases via the development of computational methods with the goal of unravelling the root causes of these conditions [16][12].

2.2.1 Role of piRNAs in Myelodysplastic Syndromes

Germ cells, stem cells, and cancer cells all exhibit essential biological features like rapid proliferation and self-renewal. Given that the piRNA pathway is important in maintaining the self-renewal mechanism of germline stem cells, it is possible that it also has similar functions in supporting the self-renewal of rapidly dividing hematopoietic stem cells and leukemic cells. Nonetheless, knowledge of piRNA transcription and precise functions in blood cells is still limited [17].

Although the significance of piRNAs in various hematological malignancies (blood-related cancers) like multiple myeloma (MM) and classic Hodgkin lymphoma

has attracted research attention, information regarding myelodysplastic syndromes (MDS) has been limited. The initial study of piRNAs in MDS patients' bone marrow cells found that individuals with low-risk MDS (refractory anaemia) had a higher expression (9%) of piRNAs than those with high-risk MDS (refractory anaemia with excess blasts—2) (2%) and healthy controls (1%). This suggests that piRNAs may play a DNA-protective role in lower-risk MDS. Small non-coding RNAs from plasma and extracellular vesicles were also found to be upregulated in MDS patients compared to controls (hsa-piR-019914 and hsa-piR-020450). Two other piRNAs, hsa-piR-000805 and hsa-piR-019420, were found to be expressed differently in MDS patients with low and high blast counts. The last piRNA was also linked to overall survival in a protective role, but no piRNAs were found to be predictive of azacytidine response in patients. More information is needed on the biological interpretation of these findings and their potential application in routine clinical practice [18].

Although research into transposable elements (TEs) and piRNAs in leukemia is still in its early stages, future advances are expected to contribute to disease classification, monitoring, and therapeutic interventions. This expectation results from our growing understanding of the mechanisms and functions of TE/piRNA processes in both normal and leukemic cells [17].

Chapter 3

Automated annotation methods

As already stated before, understanding the functions of ncRNAs is essential for unravelling their roles in cellular activities. However, annotating and characterising ncRNAs manually can be a tedious and time-consuming process. To overcome this challenge, the tools and methods for the functional automatic annotation of ncRNAs have been developed.

Functional automatic annotation of non-coding RNAs (ncRNAs) is a valuable approach that helps us understand the functions and roles of these RNA molecules. The tools and methodologies developed for functional annotation of earlier discovered microRNAs (miRNAs)[19], long non-coding RNAs (lncRNAs) [20], [21] and circular RNAs (circRNAs) [22], [23] can be adapted for piRNA analysis. The tools can be based on sequence homology [24], structural analysis [25], or machine learning [26]. In this thesis, the main attention will be paid to bioinformatics approaches based on the integration of functional genomics data and the utilisation of biological networks [22], [27], [28].

In systems biology, numerous biological sub-systems can be represented as networks. These networks consist of nodes, which represent entities like genes or proteins, and edges, which represent relationships between pairs of entities. Various biological networks exist, including protein-protein interaction (PPI) networks, gene co-expression networks, metabolic networks, and transcriptional regulatory networks [2]. The method that I use in my thesis for such annotation is based on gene co-expression networks.

3.1 Gene-coexpression networks

Gene-coexpression networks fall under the category of biological correlation networks [28]. A Gene Co-expression Network (GCN) is an undirected network that represents genes as nodes and captures significant co-expression relationships (pairwise correlations) between genes as edges. Gene co-expression networks exhibit a scale-free topology, characterised by a few highly connected nodes called hubs and numerous nodes with only a few connections. Moreover, GCNs display the small-world property, meaning that most genes can be reached from any other gene within a small number of steps. GCNs are commonly employed to identify expression modules, which help extract biological insights by assessing similarity profiles between genes [29].

Correlation networks depict the connections between pairs of nodes based on their correlations. However, constructing a correlation network using simple Pearson's correlations can result in a challenge. This is because the correlations are in most cases non-zero so the network contains all of the edges and results in a complete graph. Therefore, it is necessary to eliminate insignificant relationships in the correlation network, focusing only on the significant edges that demonstrate high correlations between nodes. In order to resolve this issue, either hard thresholds or soft thresholds are applied to Pearson's correlations. This approach enables correlation network methods to identify and capture biologically significant relationships by utilising cutoff values [30].

The commonly used methods involve calculating the similarity between expression profiles of gene pairs and determining a threshold to decide which gene pairs should be connected. However, these methods have some limitations. The main issue with these approaches is that they often rely on arbitrary threshold selection, which can be problematic. Gene CoE exhibits a property known as local scaling, where genes within one cluster may have strong correlations with each other, while genes in another group may have weaker correlations. Consequently, choosing a strict threshold can result in many genes from the weakly correlated group being left unconnected. On the other hand, including more genes in the network would require a lower threshold, leading to a situation where a significant portion of genes are almost completely connected [2]. This is displayed on the graph 3.1 that was constructed using the data I was given. In this case, a correlation network was constructed from nearly 17,000 genes, including piRNA, TE, protein-coding, rRNA, and other genes. In the graph, we can observe that at a threshold of 0.4, the median number of edges per gene remains around 500. However, as the threshold increases, the number of genes without any edges sharply rises. This excessive connectivity makes further analysis challenging .

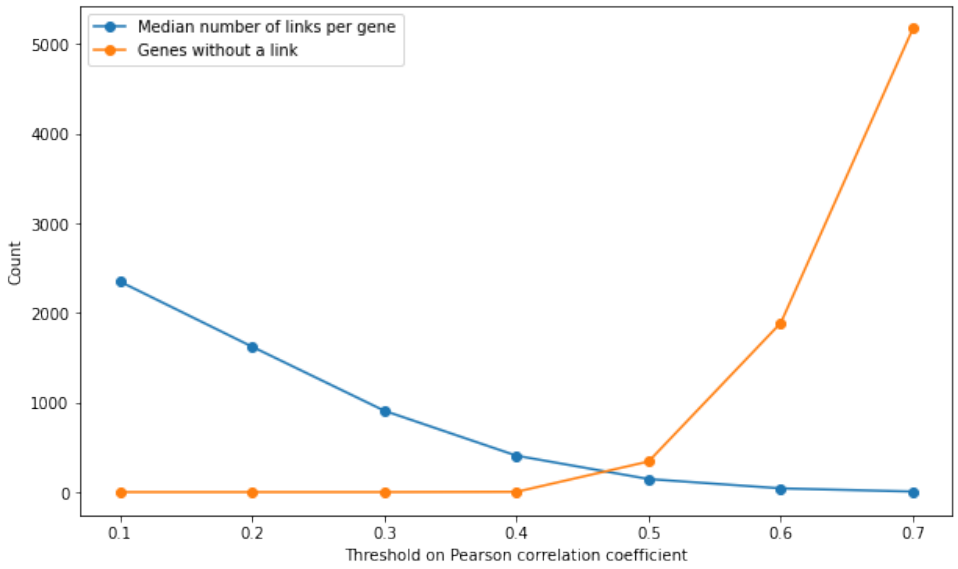


Figure 3.1: The relationship between the threshold set on the Pearson correlation coefficient and two metrics: the median number of links per gene and the count of genes that do not have any links. By varying the threshold, we can observe how these metrics change. Inspired by [2]

To address this problem, there is a solution that involves transforming the similarity matrix based on ranks. Initially, the Pearson correlation coefficient or another suitables similarity measure is calculated for each pair of genes. Next, for every gene, all other genes are ranked based on their correlation coefficients with that gene. Using these ranks, the genes are connected to its top α co-expressed genes. α is a threshold determined by the user and is typically set to a value smaller than five [2].

Genes that belong to the same pathways or functional complexes are often regulated by the same transcription factors (TFs), resulting in similar expression patterns across different conditions. Therefore, an important aspect of gene function analysis is to group genes based on their expression patterns into modules [2]. A co-expression module refers to a cluster of genes that exhibit strong connections within the group while having weaker connections to genes outside the cluster in a network [29]. Additionally, if a significant number of genes in a cluster are known to have specific functions, it is likely that unannotated genes within the same cluster may share similar functions. Common techniques used for clustering gene expression data include hierarchical clustering, k-means clustering, and self-organising maps (SOM) [2].

3.2 Guilt by association principle

An essential objective of gene co-expression networks is to assign function to genes and non-coding RNAs (ncRNAs) that were previously unknown. The guiding principle behind this annotation process is the guilt-by-association (GBA) principle. According to GBA, an unknown gene or ncRNA can be annotated by associating it with terms that have already been linked to protein-coding mRNAs and other ncRNAs whose expression patterns show a strong correlation with the profile being investigated [28]. The GBA approach involves conducting a correlation analysis between the expression patterns of ncRNA and protein-coding mRNA, along with enrichment strategies to associate functional gene sets with the mRNAs that show a correlation with the specific ncRNA of interest [31].

Generally the principle of guilt by association suggests that genes that share functional similarities are often connected as protein interaction partners or exhibit similar expression patterns. GBA is a widely employed principle in biological research and serves as a fundamental approach to analyse and uncover gene function. This principle serves as a fundamental guideline for analysing gene networks, allowing researchers to understand their functional properties and evaluate their ability to encode meaningful biological information [32].

3.3 Random walks

One of the methods of automatic annotations and the first method that I am using in my thesis is the application of random walks. Random walks, rooted in graph theory and stochastic processes, provide a framework for exploring and analysing complex networks. A random walk is a mathematical concept that represents a sequence of random steps taken within a mathematical space. It describes a path formed by a series of unpredictable movements. The concept of random walks was initially introduced by Pearson in 1905 [33].

The general principle of random walks involves a stochastic process of moving through a sequence of steps in a random manner. It begins at a starting point within a defined system (in our case a graph constructed of the gene co-expression network) and progresses through sequential steps by making transitions to neighbouring positions. The selection of the next position is determined randomly, allowing for exploration of the entire system. Random walks are performed iteratively, accumulating information at each step [34].

There are two types of the random walks approach, random walks (RW) and random walks with restart (RWR). A random walk on graphs involves a walker moving from its current node to a randomly chosen neighbouring node in an iterative process, starting from a specified source node, s. Random walks with restart are a variation of random walks where, in addition to the normal transitions, there is also a possibility of restarting the walk at node s at each time step, with a certain probability denoted as r [35].

Random walks were initially devised to explore the overall structure of networks by imitating a particle that moves iteratively from one node to a neighbouring node chosen at random. The concept of restart, which eventually led to the development of the random walk with restart (RWR) algorithm, was first introduced in the context of Internet search engines. The aim was to mimic the behaviour of an internet user who navigates from one webpage to another through hyperlinks but can also restart the browsing process on a new arbitrary webpage. As a result, certain webpages will be visited more frequently than others based on the topological arrangement of the pages and hyperlinks [36].

RWR has emerged as a leading algorithm in the field of network computational biology, particularly in guilt-by-association analysis. It was first applied to identify important disease-related genes. The algorithm ranks all the nodes in a network based on their proximity to known disease-associated nodes, which act as starting points for the analysis. This ranking helps identify nodes that are closely connected to the disease and are likely to play a significant role in its development [36].

Chapter 4

Myelodysplastic syndromes

Myelodysplastic syndrome (MDS) is a heterogeneous group of diseases characterised by an abnormality in the production of blood cells due to genetic mutations. It is a chronic condition characterised by genetic mutations in a type of cells called pluripotent stem cells. These cells have the ability to develop into different types of blood cells. However, in MDS, these mutations disrupt the normal process of cell maturation and differentiation. As a result, the production of healthy blood cells becomes compromised, leading to a condition called dysplastic hematopoiesis. In this condition, the blood cells produced are ineffective and don't function properly. One important concern in MDS is that it has the potential to progress to a more aggressive form of blood cancer known as acute leukemia (AL) [37].

MDS is a group of hematopoietic stem cell (HSC) disorders characterised by impaired hematopoiesis, peripheral blood cytopenia, and a predisposition to progress to leukaemia. In 30-40% of MDS patients, AML with myelodysplasia-related changes (AML-MRC) develops gradually. Several new MDS therapeutic agents have been approved in recent years, with hypomethylating agents like azacitidine or decitabine proving effective in treating both MDS and AML-MRC. In a significant proportion of patients, these agents improve overall survival, clinical outcomes, and quality of life (overall response rate, 40-50%). While the precise mechanism of action is still being researched, it is hypothesised that DNA hypomethylation may reverse tumour-suppressor gene transcription inactivation [17].

MDS is predominantly a disease of older people. Median age of patients is around 70-76 years, although it can rarely affect younger patients as well. The incidence rate of the disease is approximately 5-6 cases per 100,000 people per year in the general population [38], [39]. However, in the population over 70 years of age, the

incidence increases to 26 cases per 100,000 people per year, and among patients over 80 years of age, it rises to 48 cases per 100,000 people per year [38].

The most common symptom of MDS is anemia and anemic syndrome, with 80% of patients having a hemoglobin level below 100 g/L. Patients typically seek medical help with feelings of fatigue, weakness, lack of energy, dizziness, headaches, reduced physical endurance, palpitations, chest pain and difficulty breathing [37].

Disease progression varies greatly among individuals, and treatment approaches must be customised based on the patient's symptom severity, risk of developing acute myeloid leukemia (AML), and any other existing health conditions. The commonly utilised tools for assessing the risk of MDS patients are the International Prognostic Scoring System (IPSS), its revised version (IPSS-R), and the World Health Organisation (WHO)-based prognostic scoring system (WPSS). These systems categorise patients into different risk levels ranging from very low to very high. The classification is determined by evaluating factors such as the cytogenetic features (chromosome abnormalities, recurrent genetic changes), the percentage of bone marrow blasts, and the severity of peripheral blood cytopenias [40].

In low-risk myelodysplastic syndrome (LR-MDS), the primary goal of treatment is to address and improve cytopenias, particularly anemia. On the other hand, in high-risk myelodysplastic syndrome (HR-MDS), the focus of treatment is aimed at slowing down the progression of the disease and extending overall survival [41].

When it is possible, allogeneic stem cell transplantation (allo-SCT) remains the only treatment option that offers the potential for a cure in patients with high-risk myelodysplastic syndrome (HR-MDS). Studies have shown that approximately 40-50% of patients achieve long-term remission without disease recurrence following allo-SCT, and these outcomes have improved over time. Hypomethylating agents, such as azacitidine, are currently considered the initial treatment of choice in the majority of cases involving higher-risk myelodysplastic syndrome (MDS). In a randomised phase 3 clinical trial conducted on patients with higher-risk MDS, azacitidine showed a survival advantage compared to conventional treatment regimens. The study revealed a median survival of 24.4 months with azacitidine, whereas the median survival with conventional care regimens was 15 months. This survival benefit associated with azacitidine was observed regardless of factors such as age, percentage of marrow blasts, and karyotype [41].

Part II

Practical part

Chapter 5

Data set

The main aim of this master's thesis was to create a tool that can identify deferentially expressed piRNAs and assigns them to correct annotations based on the guilt by association principle. To reach this goal this data set was provided.

- Expression data for 556 different piRNAs each measured across a sample of 106 subjects.
- Expression data of 687 transposable elements (TE) across a sample of 114 subjects.
- Expression data of 58216 different kinds of genes, including miRNA, lincRNA, snRNA, protein coding and more. Measured over 86 subjects.
- 15937 different annotations (GO terms) with assigned genes.
- Information about the piRNA and TE sequences.
- Disease information for each subject.
- List of MDS related GO terms obtained from ctd database [42]

The expression data were provided from the Institute of Hematology and Blood Transfusion.

The intersection between the measured subjects from the first three files (TE, piRNA, genes) was 79 subjects. The subjects are divided into four groups based

on the disease or the stage of disease they suffer from. Majority of the subjects suffers from myelodyplastic syndromes (MDS) or Acute myeloid leukemia with myelodysplasia-related changes (AML-MRC). The third group are patients that are currently treated with azacitidine (AZA). And the rest of the subjects are healthy controls. This information is essential for identifying differentially expressed piRNAs.

5.1 Data preparation

The tool was developed in python. The initial phase involved loading and preparing the data. The first step was to find the samples that contained data for all of the genes. This left us with data of 79 subjects. The size of the data set was huge and needed to be reduced by deleting useless data. Therefore the data with zero or very low expression were deleted.

However, even after this step, the data set still contained around ten thousands of genes which would take hours or maybe even days to process. Therefore, for the initial phase of development and testing, the data set was limited to contain only piRNAs and one thousand protein coding genes (the genes that most likely have assigned annotations).

Next step was to normalise the given data. For that purpose the DESeq2 median of ratios normalisation was used.

The annotations provided were not in the form needed for this work. Therefore the table with annotations was transformed into a more suitable form. The original file contained a table that had all the annotations (Go-terms) listed in the first column and all known genes annotated with this annotation were listed in the respective lines. The transformed version is vice versa. The first column contains the list of all the genes and the annotations are listed in the rows.

A subset of all the data had to be made, since our data set was huge and work with such a huge network would not be feasible. The main criterion for selecting suitable piRNAs for the experiment was high differential expression. A high level of differential expression indicates statistically significant differences in the expression levels of two compared groups. In our case, the first group consisted of MDS patients, while the second group consisted of healthy controls. In other words, highly differentially expressed genes are those with high expression in the first group but low expression in the second, or vice versa. The high difference in expression

between these two groups suggests a possible link to myelodysplastic syndrome. As a result, we were looking for piRNAs that were highly differentially expressed.

The same criterion was applied when choosing the subset of the other genes we needed to add to the network. Only this time both highly and lowly differentially expressed genes were chosen, to represent both genes with possible connection to myelodysplastic syndromes (MDS) and genes that are most likely not related to MDS.

Chapter 6

Correlation matrix and adjacency graph

Once the data were successfully normalised, a correlation matrix and an adjacency graph based on this normalised data were constructed. The first constructed graph can be seen in figure 6.1 where the blue nodes represent protein coding genes and the red nodes represent piRNAs. The edges of the graph represent correlation between the genes. In the initial phase of the experiment a non weighted graph was used. The edges were included in the graph if and only if the absolute value of the correlation between the genes was above selected threshold. In this case the threshold was set to 0.3.

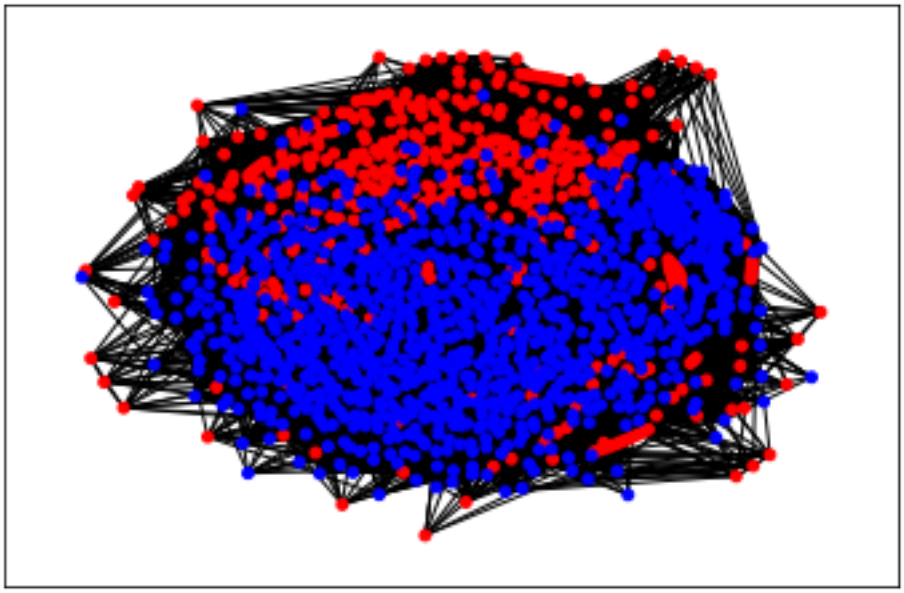


Figure 6.1: Visualisation of the adjacency graph from the initial phase.

However this construction was later proven to be insufficient. The graph contained too many edges and the random walks did not converge. The graph had to be improved and many edges had to be cut. Therefore the method described in Gene-coexpression networks was used. This method is not based on establishing a fixed threshold. Instead, it predefines the number of edges leading from each node. In this way, the number of edges in the graph can be significantly reduced while still maintaining the connectivity of nodes in sparser regions of the graph [2].

It is basically a context of k-nearest neighbour graphs, the objective is to link vertices based on their k-nearest neighbours. However, this definition results in a directed graph due to the asymmetry of the neighbourhood relationship. There are two approaches that can be used to create an undirected graph. The first method involves ignoring edge directions, connecting vertices if either is among the k-nearest neighbours of the other. This results in what is commonly known as the k-nearest neighbour graph. The second option is to connect nodes only if both are among the k-nearest neighbours of each other, forming the mutual k-nearest neighbour graph. In both cases, after establishing connections, the edges are weighted based on the correlations between the connected vertices [43]. In this thesis the first method was used.

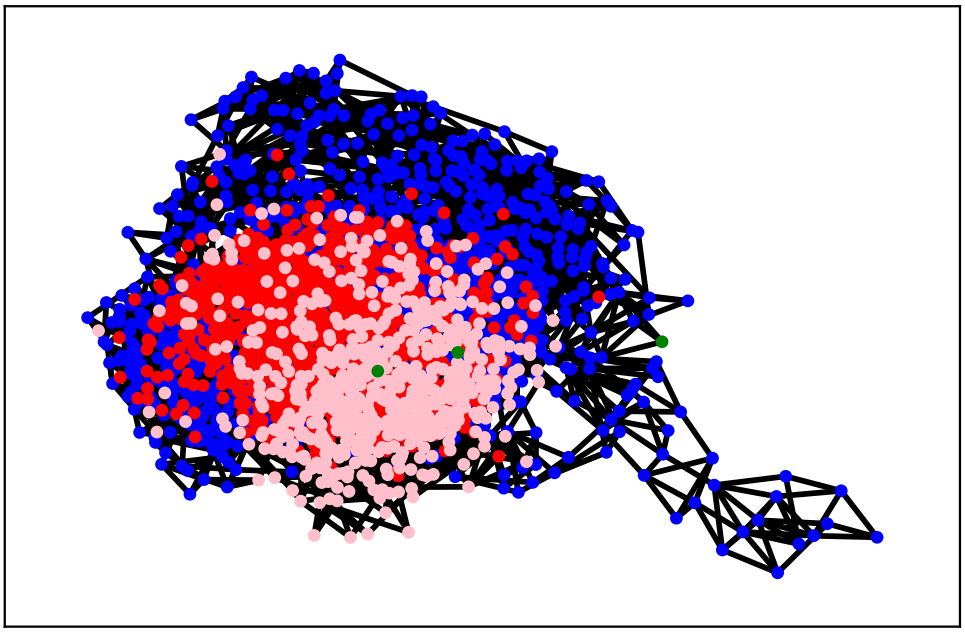


Figure 6.2: Visualisation of the adjacency graph from the advanced phase. (3-nearest neighbour graph) without considering the sequence matches between piRNAs and TEs

In the network visualised in the figure 6.2, there are 3 piRNAs represented as green dots, 600 TEs represented as pink dots, 938 'positive' genes as blue dots and 912 'negative genes shown as red points. Positive genes are those already annotated genes that demonstrate a high value of differential expression between groups of MDS patients and healthy controls. In other words, genes for which we suspect a

link to myelodysplastic syndrome. Negative genes, on the other hand, have lower observed differential expression. The graph is displayed in such a way that nodes with direct connections are displayed close to one another, while nodes with no direct connections are displayed further apart. As a result, mutually interacting genes are shown next to each other.

Two criteria were used to choose the three piRNAs. Sufficiently high differential expression was the initial need. The second requirement was that the chosen piRNA had sequence matches with provided transposons. The method used to identify the sequence matches between piRNAs and TEs is described bellow in section Pirna-TE links. Sequence matches with transposable elements were found for 29 piRNAs. However, none of these piRNAs matched the important differential expression requirement. Therefore some compromises had to be made. We chose two piRNAs with the highest differential expression among the 29 with sequence matches, despite the fact that it was not statistically significant. The third piRNA is a piRNA with the highest and statistically significant differential expression, on the other hand it does not have any sequence match. This third piRNA is shown at the top right of the graph, above the other two closest to the part of the graph with majority of 'positive' genes. As it is the only piRNA with significant differential expression, this could be expected. However the sequence matches are not yet considered in this graph. On the other hand the graph in the figure 6.3 contains additional edges between piRNAs and TEs based on their sequential compatibility. If we compare the graph 6.2 and 6.3, we can see that the piRNAs with the TE links shifted a little, however the shift is not really significant.

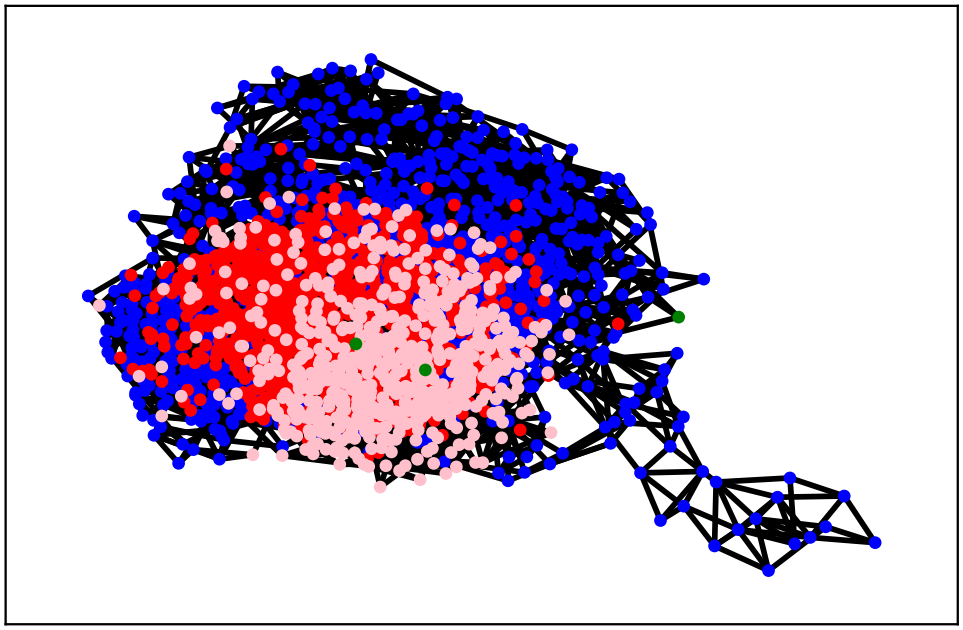


Figure 6.3: Visualisation of the adjacency graph from the advanced phase. (3-nearest neighbour graph) with edges representing sequence compatibility of piRNAs and TEs.

As already stated before, the edges of the graph were limited to have an predefined number of edges coming from each node. The numbers of edges we were working with were 3, 5 and 10. The random walks method and the annotations assignment were made for each node degree setting and the results were then compared. To understand the gene co-expression network better few histograms showing the distributions of the correlation values are presented.

In the graph 6.4 there are three correlation distributions, one for each node degree setting. This graph looks at every edge in the adjacency graph and shows us the strengths of the correlations in the network. It is possible to see that increasing the degree of the vertexes does not significantly alter the distribution of the correlation. In other words, adding more edges to the network does not significantly lower the mean and median of the correlation values. The median of the values of correlations for each node degree is also displayed in the graph 6.4. With more edges in the network, the median moves slightly towards lower values which was expected since the the network always contains the x strongest correlations for each gene, where x is the degree of the nodes in the network. That means that with increasing node degree the links added are expected to be weaker than the ones already in the network. However the difference in the median value is not large. For these three settings of the network the median ranges between 0.59 for node degree 10 and 0.64 for node degree 3. The exact values of all the mean and median values are presented in table 6.1.

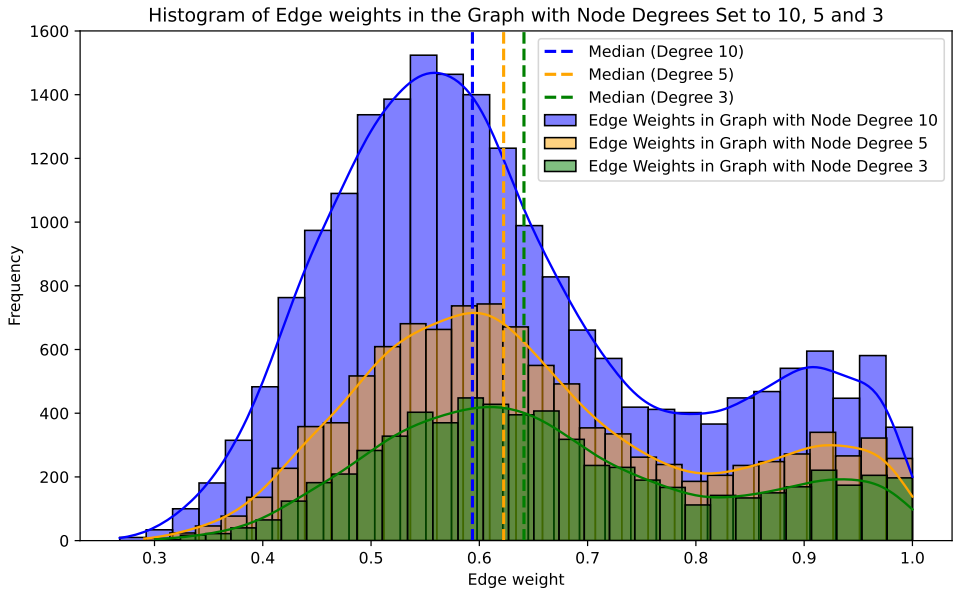


Figure 6.4: Histogram showing correlations values that were used for creating the three adjacency graphs with node degrees 3, 5 and 10.

Node Degree	Mean	Median
3	0.67	0.64
5	0.65	0.62
10	0.62	0.59

Table 6.1: The mean and median values of the correlation values distribution for each setting of the degree of the nodes in the network.

The histogram in graph 6.4 gives us some understanding of the distribution of the correlation values in the gene co-expression network. However it does not completely tell us how the values are distributed. It does not tell us whether there are genes that are highly correlated and other genes from lower correlated areas or if the difference between the strongest correlation and other correlations for that gene are significant. In other words if there are genes in the network that are only strongly correlated with one or two other genes with much lower correlations to others. Therefore to understand the network even better, additional graphs were created.

The graph 6.5 shows the distribution of the highest correlation values throughout all genes. This time there is only one histogram for all the node degree settings since the strongest correlation for each gene is included in every setting therefore all settings can be displayed together. From this graph it is possible to see that there are in fact genes with really high strongest correlation and also genes with the strongest correlation only somewhere around 0.4. However the vast majority of the genes has its strongest correlation value 0.5 or higher. The mean value of the strongest correlations is 0.71 and the median is 0.69.

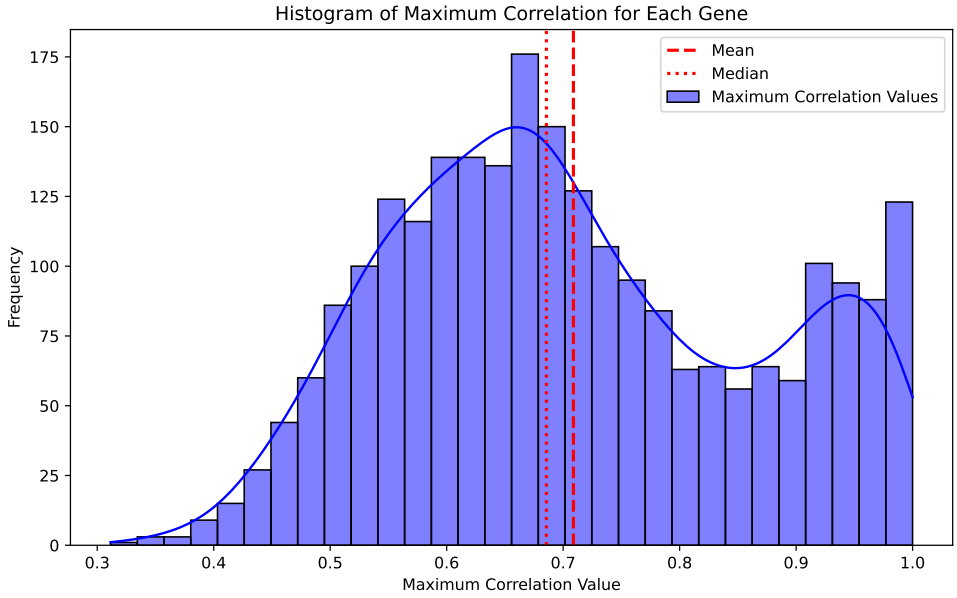


Figure 6.5: Histogram showing the strongest correlation value for each gene.

Both graphs 6.6 and 6.7 show the differences between the the strongest and the weakest correlation used in the network throughout all genes. However there is a difference at what was considered the weakest correlation for each gene. The graph 6.6 simply shows the distribution of differences between the strongest correlation for each gene and its 3rd/5th/10th strongest correlation based on the node degree setting. Therefore, we can see how significant the differences are between the strongest correlations of each gene and infer if there are areas of highly correlated genes or not. As we can see in the graph 6.6 the differences between the highest and lowest correlation value throughout the genes are very small which suggests that majority of the genes either have more correlation of an similar values and therefore we can suggest that they form some kind of clusters that are more or less correlated together. The differences increase slightly with the increasing degrees of the nodes which is expected since weaker edges are being added. However the mean difference does not go over 0.1. The exact mean and median values are displayed in table 6.2.

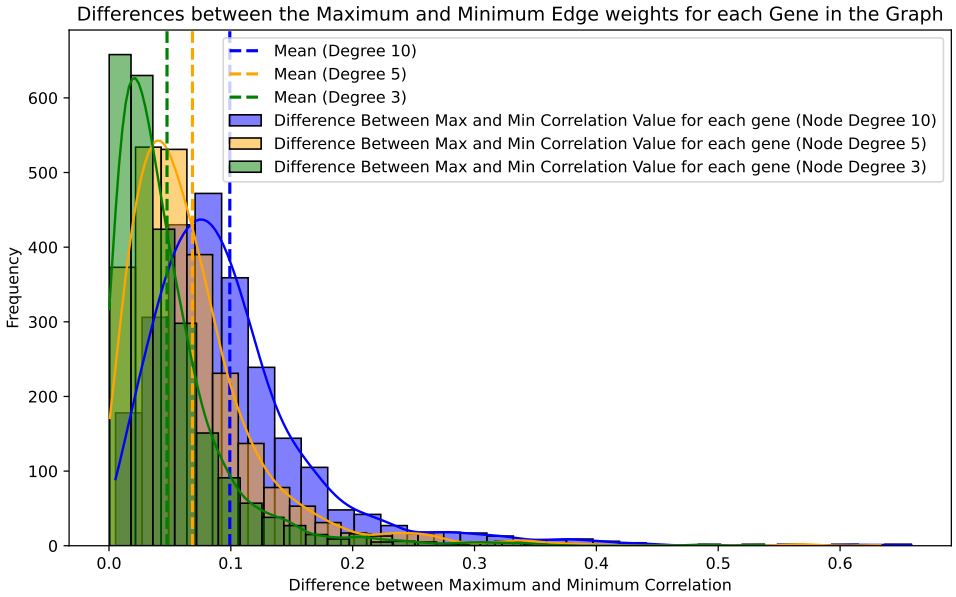


Figure 6.6: Histogram presenting the difference between the strongest correlation value and the weakest correlation value that was used for the adjacency graph (3rd, 5th or 10th strongest correlation value based on the node degree parameter) for each gene.

Node Degree	Mean	Median
3	0.048	0.034
5	0.069	0.054
10	0.99	0.085

Table 6.2: The mean and median values of the differences between the strongest correlation values and the 3rd/5th/10th strongest correlation value based on the node degree.

However as already stated before, adding only x strongest correlations for each gene, where x is the desired degree of the nodes, results in a directed graph since the fact that a specific correlation value is in the top x correlation for a specific gene does

not mean that the same correlation value is also in the top x correlation for the gene that it leads to and vice versa. To solve this problem and avoid the oriented graph, an approach was used and therefore an edge is added to the network if it is in the top x correlations of any of the two genes. The graph 6.7 takes this fact into account and shows the distribution of the differences between the highest and lowest correlation value for each gene considering all the edges. The difference between the graphs 6.6 and 6.7 is that this time the lowest correlation does not have to come from the top correlation values of the gene. As we can see in the graph 6.7 the differences raised quite significantly which suggests that there are a lot of links between the generally lower correlated genes and the higher correlated genes in the network. The mean and median are again displayed in the table 6.3.

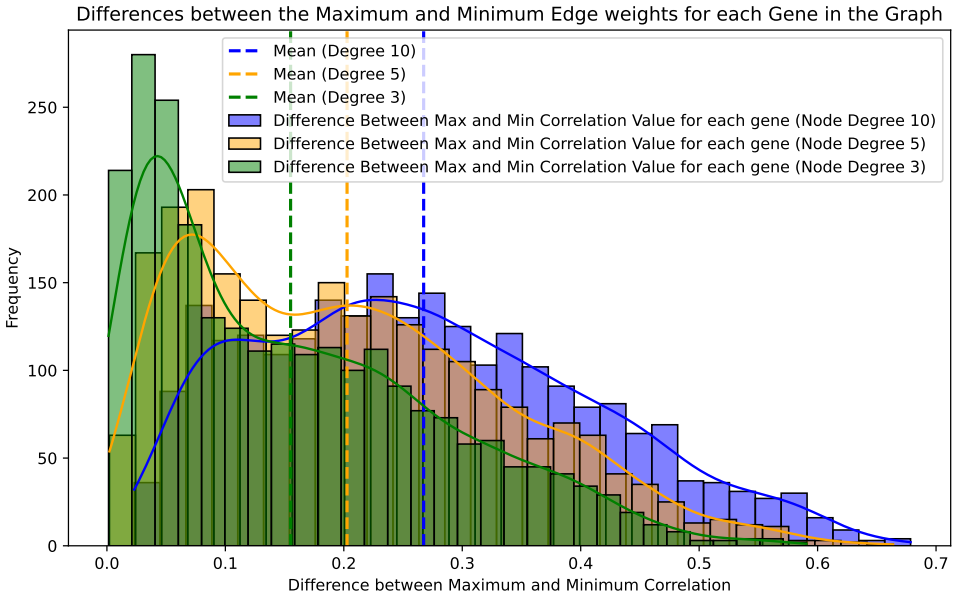


Figure 6.7: Histogram displaying the difference between the strongest and the weakest correlation used in the adjacency graph for each gene. The weakest correlation value might not necessary be the 3rd/5th/10th strongest correlation for each gene since the graph is not oriented and the edge might have been created based on the gene on the other side of the edge.

Node Degree	Mean	Median
3	0.16	0.13
5	0.20	0.19
10	0.27	0.25

Table 6.3: The mean and median values of the differences between the strongest correlation values and the weakest correlation value used in the network for each gene.

6.1 Pirna-TE links

Since one of the roles of piRNAs in silencing transposons, it was expected that there would be connections between piRNAs and transposable elements (TEs) in the gene co-expression networks. Nevertheless, the expression levels of piRNAs and TEs generally showed no significant correlations. As a result, the concept of forming links based on their compatibility in terms of sequence formed.

The implemented method was based on the pirScan web software [44], [45]. PirScan is a specialised tool designed to predict the locations where piRNAs are most likely to bind in a given sequence of C. elegans. It proposes potential silent mutations that could be introduced to prevent the silencing of transgenes by piRNAs. Users can input either a mature mRNA or spliced DNA sequence and select either the default or personalised piRNA targeting rules for the search. The results are displayed in a simple and structurally organised manner, presenting all expected piRNA target sites within the provided sequence through graphs and tables. PirScan applies established targeting rules to predict the endogenous piRNA targeting sites in an input sequence. PirScan visually presents the precise locations of piRNA targeting sites within an input sequence, along with the corresponding pairing information at each site. The outcomes of every piRNA target prediction can be additionally downloaded [44].

Although we attempted to work with pirScan to detect piRNAs and TEs that are sequentially compatible, we encountered a difference between our piRNA database and the one in pirScan. Additionally, using pirScan for all TEs in our database would require an extensive amount of time and effort. Consequently, it was necessary to develop a similar tool based on pirScan. For our work, exact location of piRNA targeting sites was not necessary. Therefore, a significantly simplified version of pirScan was created. Our tool attempts to align the piRNA sequence to every possible position of the TE sequence, evaluates the number of mismatches, calculates a 'match score', and provides the match score of the best alignment. The algorithm aligns the piRNA sequence with the reversed TE sequence and examines the presence of complementary base pairs. Matches are identified whenever 'AT' and 'GC' pairs are found. Other combinations are treated as mismatches, with one exception, the 'GT' pair is considered to be a mismatch, however, the penalty for this pair is less severe than the penalty for other mismatches.

Chapter 7

Random Walks and Permutation Tests

The method used for the piRNA annotation was the random walks method. The algorithm itself is described in function 1. The random walk starts in a piRNA node and in each step it randomly progresses to one of the neighbours of the current nodes. The next step is dependent only on the current state and it is not influenced by the preceding sequence. The random walk takes a pre-selected number of steps.

Function 1: Random walks

```
Result: Path
Input: graph, startNode, numSteps
current = startNode;
path = [current];
for i = 1 to numSteps do
    neighbors = graph.getNeighbors(current);
    if length(neighbors) equal to 0 then
        break;
    nextNode = randomlySelect(neighbors);
    Add nextNode to path;
    current = nextNode;
```

In this diploma thesis a special version of random walks called random walks with restart was used. In a random walk with restart, there is an added probability that the walker might "restart" its journey from the initial node rather than moving to a neighbouring node at each step. This restarting probability introduces a bias, favouring paths that begin at the first node. The random walks with restart algorithm is described in function 2.

Function 2: Random walks with restart

```
Result: Path
Input: graph, startNode, numSteps, restartProbability
current = startNode;
path = [current];
for i = 1 to numSteps do
    neighbors = graph.getNeighbors(current);
    if length(neighbors) equal to 0 then
        break;
    nextNode = randomlySelect(neighbors) or startNode (with probability ==
     restartProbability);
    Add nextNode to path;
    current = nextNode;
```

At the end of each random walk, after the number of steps we chose, we look at the node that the random walk ended in and the annotations assigned to it. Each of the GO terms assigned to the gene at the end of the random walk is considered as a possible GO term for the piRNA. The annotations (GO terms) with significantly larger than random occurrence at the end gene of the random walks are assigned to the piRNA.

The decision whether the GO term occurrence is statistically significant is made based on permutation tests. The key idea behind permutation tests is to generate a null distribution by randomly permuting or reordering the data, simulating a scenario in which the observed effect is entirely due to random. The observed test statistic is then compared to the statistical distribution resulting from many such random permutations. If the observed statistic is in the extreme tail of the null distribution, it indicates that the observed effect is unlikely to have occurred by chance alone, implying that the null hypothesis must be rejected. In other words, the permutation test determines the likelihood that the observed accuracy happened by chance. The p-value denotes the proportion of randomised data sets in which the classifier performed as well as or better than it did in the real data, assuming a specific null hypothesis [46].

Randomised data sets are generated using the random walks method applied to the graph, where GO terms are shuffled randomly throughout the network. The process is visually illustrated in Figure 7.1 on a small, simple network. This involves taking the correct gene co-expression network, which includes information about nodes, edges, and assigned GO terms. The method requires temporarily removing the correct GO term assignments followed by reassigning them to a random node. It removes the correct assignments of the GO terms and then put them back to a randomly selected node. As a result, the network's structure remains unchanged, while the GO terms are shuffled across the network and assigned to different genes

than before. Throughout this process, the density of individual GO terms within the network is preserved.

For each randomisation of the network, a predetermined number of random walks are performed. This procedure is repeated several times in order to ensure proper distribution. The method improves and becomes more precise as the number of random walks per randomisation and the number of randomisations of the network increases. On the other hand the more random walks per randomisation and the more randomisations of the network are performed, the longer the execution times becomes. Therefore some compromise between precision and execution time had to be made.

Throughout the majority of the experiment, 100 random walks per randomisation and 500 differently shuffled networks were conducted. The same number of random walks was applied to the correctly annotated gene co-expression network. Consequently, we obtained a value between zero and one hundred for each GO term in the network, representing how frequently a random walk finished in a gene annotated with that particular GO term in the correctly annotated network. With a total number of 11,892 GO terms in the network, we then generated a distribution of 1000 numbers (dependent on the number of performed shuffles) for each GO term from the randomised networks. Subsequently, we compared the obtained number and distribution, calculating a p-value. GO terms exhibiting a significant p-value are then assigned to the piRNA.

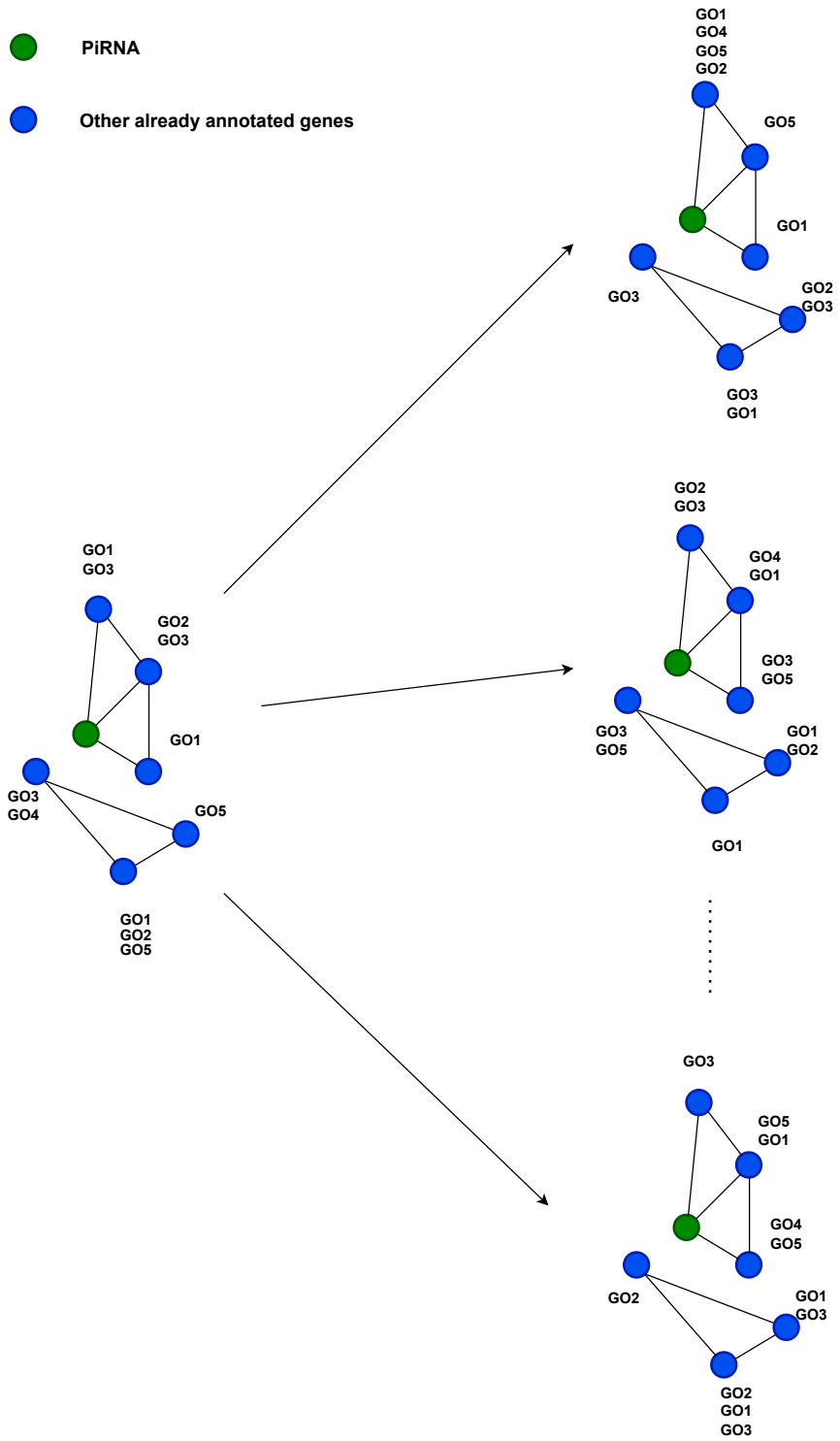


Figure 7.1: Graphical explanation of the random shuffling of the GO terms in the network. GO terms, represented as GO1-5 are randomly shuffled among the network. The density of each GO term remains the same.

The figure 7.2 shows the distribution of the results of the random walks that started from piRNA 'hsa-piR-018849' performed on the correctly annotated gene co-expression network. We can see that the vast majority of the GO terms was not visited in any of the 100 random walks. These are mostly the GO terms that belong to the genes that are not highly correlated with examined piRNA and therefore are located further away in the network and cannot be reached with the random walks. They can also be GO terms with really low density in the graph so they are harder to reach with only 100 random walks. Since the number of the GO terms with no visits is so huge, the rest of the distribution cannot be properly seen. For this reason, the zoom version of this histogram was created and is displayed in figure 7.3.

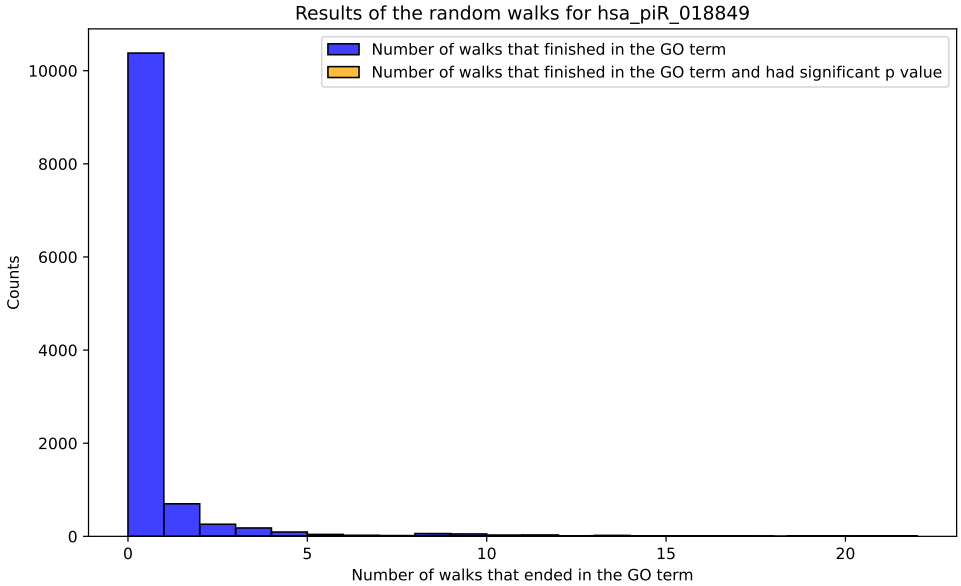


Figure 7.2: Distribution of Random Walk Endpoints on Gene Ontology Terms: A histogram illustrating the frequency of random walks reaching different Gene Ontology (GO) terms. The x-axis represents the number of random walks terminating in each GO term, while the y-axis indicates the count of GO terms for each corresponding endpoint count. This analysis provides insights into the distribution of random walk outcomes across the GO hierarchy. Blue distribution shows all GO terms, while orange distribution only shows GO terms with statistically significant occurrence compared to the results of random walks on the randomly shuffled network.

Figure 7.3 shows a more detailed look at the lower count values in the distribution of results from random walks initiated from the piRNA 'hsa-piR-018849' on a correctly annotated gene co-expression network. This visualisation presents the distribution of occurrences of Gene Ontology (GO) terms at the final point of random walks, satisfying the statistical significance threshold.

Notably, a relatively high frequency does not guarantee statistical significance. A GO term with a low occurrence, on the other hand, may still be considered significant. This variability occurs because the p-value is influenced not only by the results of

random walks in the correctly annotated network, but also by those in the shuffled network. That is because, the density of the GO term in the network plays an important role in determining significance.

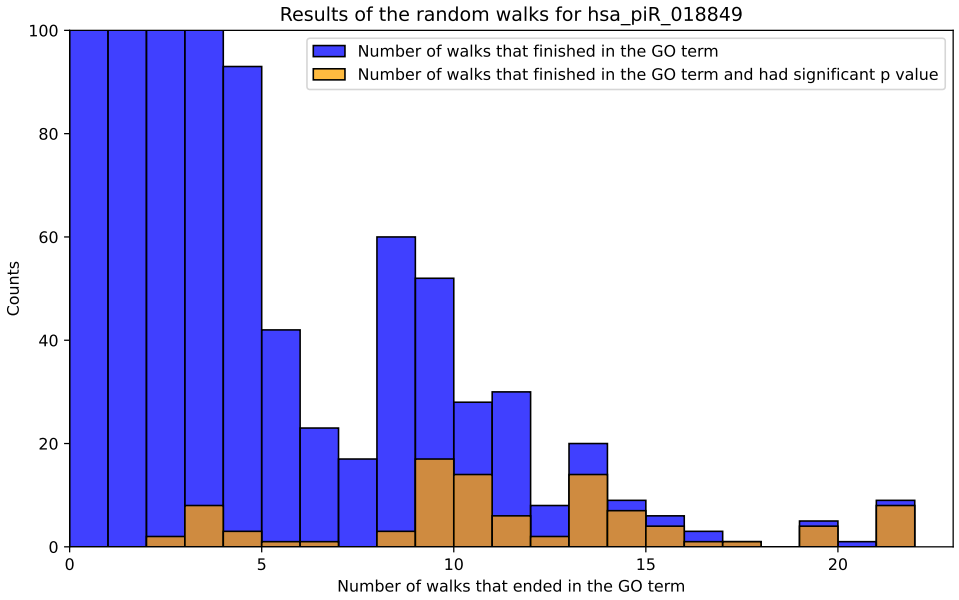


Figure 7.3: Zoom on the distribution of Random Walk Endpoints on Gene Ontology Terms: A histogram illustrating the frequency of random walks reaching different Gene Ontology (GO) terms. The x-axis represents the number of random walks terminating in each GO term, while the y-axis indicates the count of GO terms for each corresponding endpoint count. This analysis provides insights into the distribution of random walk outcomes across the GO hierarchy. Blue distribution shows all GO terms, while orange distribution only shows GO terms with statistically significant occurrence compared to the results of random walks on the randomly shuffled network.

Figure 7.4 displays a distribution of the numbers of occurrences of individual GO terms in our gene co-expression network. The blue distribution represents all of the 11925 GO terms assigned to genes in our gene co-expression network. The orange distribution includes only the 1243 MDS related GO terms assigned to genes in our network. According to the graph majority of the GO terms are only assigned to only few genes. In other words, most of the GO terms have low density in the network. This distribution looks fairly similar with distribution portraying the GO terms at the finish states of the random walks displayed in figure 7.2. This similarity is not coincidental as these two distributions are dependent on each other as the random walks are generally more likely to end up in the more abundant GO term.

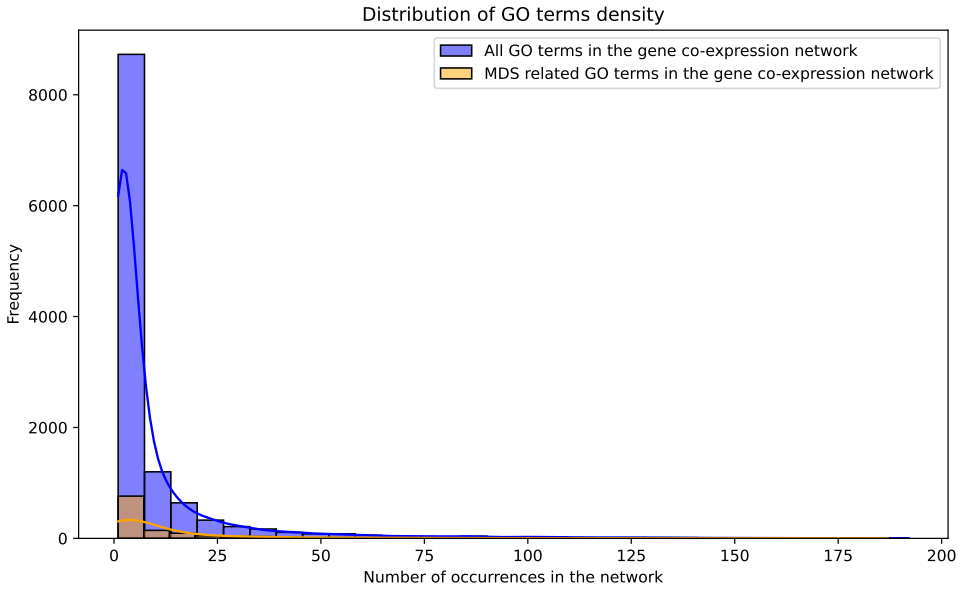


Figure 7.4: Histogram depicting the distribution of Gene Ontology (GO) term occurrences in the gene co-expression network. The x-axis represents the number of occurrences, while the y-axis indicates the frequency of GO terms corresponding to each occurrence level. Two distributions are provided, blue one represents all GO terms in the network, orange represents MDS related terms.

For a better comparison of the distributions of all GO terms and MDS related GO terms, histogram displayed in figure 7.5 was constructed. It presents the same distributions as the histogram in figure 7.4, only this time not in absolute numbers on the y axis. Graph in figure 7.5 has density value on the y axis which is more suitable for comparison between two distributions of different sizes. It is possible to see that both distributions are similar, meaning that many GO terms in both populations have only few occurrences in the network. However the MDS related terms seem to be slightly more abundant than the population with all GO terms. The mean value of MDS related GO term abundance is slightly higher than the mean value of all GO terms abundance. The mean value for MDS related GO terms is 14.79, for all GO terms it is 9.52.

The graph in figure 7.6 illustrates the successful assignment of a GO term based on the outcome of a permutation test. The assigned Gene Ontology term was accessed 15 times in the correctly annotated network. The results of the shuffled network are displayed as the blue histogram. It is evident that the number of visits in the accurately annotated network exceeds that of any of the shuffled networks, which in this particular case equals to 1000.

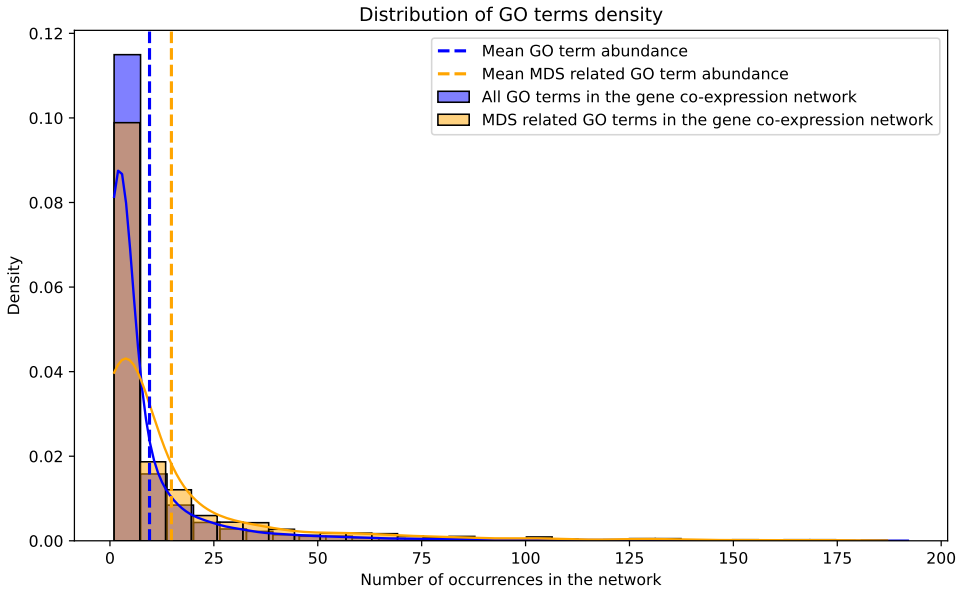


Figure 7.5: Histogram depicting the distribution of Gene Ontology (GO) term occurrences in the gene co-expression network. The x-axis represents the number of occurrences, while the y-axis the density of the number of GO terms for each occurrence level. Two distributions are provided, blue one represents all GO terms in the network, orange represents MDS related terms.

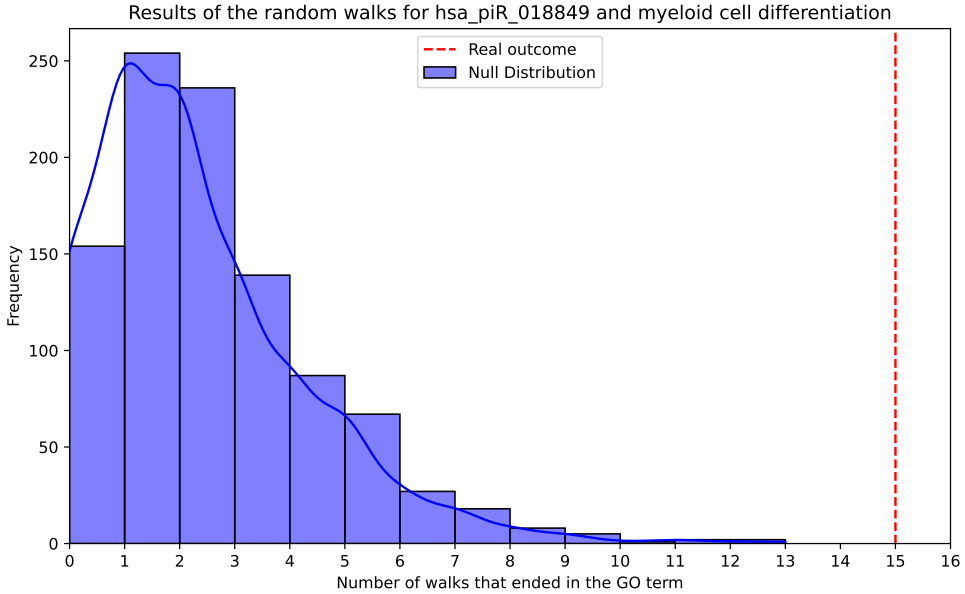


Figure 7.6: Comparison of the distribution obtained from the random walks with restart performed on the shuffled network and the true outcome of the random walks with restart on the correct network for the piRNA hsa-piR-018849 and GO term myeloid cell differentiation.

Chapter 8

Results

During the initial stages of testing and experimentation, the network pictured in figure 6.1 was used. The network consisted of 1500 protein coding genes, displayed as blue nodes, and 423 piRNAs, displayed as red nodes. Protein coding genes were selected due to the fact that the majority of them possess officially assigned Gene Ontology (GO) terms. The piRNAs were selected from the data-set, including all of those with a non-zero expression. At this point, differential expression was not considered. The decision whether a link between two genes was included in the network was determined by applying a predetermined threshold of 0.3. That indicates that the genes were linked in the network only if the correlation between them had an absolute value greater than 0.3. At this time, the random walk method is working properly and successfully assigning GO terms.

Nevertheless, the network was insufficient due to a large amount of edges, resulting in a complete lack of convergence in the random walks. Consequently, the results in this stage appeared to be comparable to a random generator of GO terms. Furthermore, the genes that were used in the network were not precisely selected. Due to the absence of differential gene expression analysis, we were unable to tell the difference between interesting piRNAs and others.

For all these reasons major changes in the network had to be made. Some of the changes were already discussed in chapter Correlation matrix and adjacency graph. The number of edges in the gene co-expression network had to be reduced, to do that the method already described in chapters Gene-coexpression networks and Correlation matrix and adjacency graph was used. This way each gene was left with a predetermined number of edges. An edge between two genes was added into the network if it was in the top x correlations for at least one of the two genes,

where x is the desired degree of the nodes in the network. Next the network had to consist of completely different subset of available genes. This time considering the values of differential expression to focus on piRNAs that are potentially linked to myelodysplastic syndromes (MDS). In this network transposable elements (TE) were also included since piRNAs are known for interacting with them. TEs in general do not have GO terms assigned to them but they could serve as links between piRNAs and other genes with previously assigned GO terms. However the correlations between piRNAs and TEs in most cases were not significant. From that the idea to connect piRNAs and TEs according to their sequential compatibility arose. The method used to add those piRNA-TE links is described in section Pirna-TE links.

In order to obtain the final results of our work, different network than the one in the initial experiment was used. The final network, previously discussed in chapter Correlation matrix and adjacency graph, includes three piwi interacting RNAs (piRNA) that have been identified as worth mentioning. Additionally, it includes 600 transposable elements (TE), with the elimination of TEs that were expressed at very low levels. Finally, the network comprises 1850 genes that should already possess verified annotations. The genes were selected out of two categories: genes exhibiting significant differential expression between patients with myelodysplastic syndromes (MDS) and healthy individuals, indicating a potential association with MDS; and genes exhibiting little differential expression, suggesting no connection to MDS.

The selection of the noteworthy piRNAs was based on two criteria: statistically significant differential expression between the groups of interest and sequence compatibility with the available transposable elements (TEs). However, none of the piRNAs met both conditions satisfactorily. Consequently, a compromise had to be formed. We identified 29 piRNAs that exhibited sequential compatibility with TEs. Among these, we selected the two piRNAs with the greatest differential expressions, although the statistical significance of those differences was not confirmed. The last newly selected piRNA demonstrated statistically significant differential expression, but did not possess any sequentially compatible transposable elements (TEs). The chosen piRNAs are presented in table 8.1.

piRNA	matched TE sequences	log2FoldChange	Adjusted pValue
hsa-piR-014626	6	-1.38	0.94
hsa-piR-021121	3	1.25	0.94
hsa-piR-018849	0	5.97	0.02

Table 8.1: Table presenting information about piRNAs selected for the experiment. It displays number of sequence compatible transposable elements (TE) and results of the differential expression analysis, specifically log2 fold change and adjusted p-value.

In order to evaluate the efficiency of our approach, it is essential to have access to officially recognised annotations that can serve as a benchmark for evaluating our findings. Unfortunately, we were unable to find a database offering piRNA annotations that included the piRNAs in our data-set. As a result, it is not possible to

accurately determine the success rate of the method we implemented. Therefore, the only approach for evaluating the success is by examining the quantity of assigned Gene Ontology (GO) terms and analysing the proportion of those terms that are associated with Myelodysplastic syndromes (MDS). The p-values were calculated based on the cumulative distribution function corresponding to the hypergeometric distribution given the above values. The p-values provide the probability of observing a number of positive GO terms (related to MDS) or higher in a sample of randomly drawn GO terms (the total number of assigned GO terms), taking into account the sizes of the two populations (the number of different GO terms in the network and the number of MDS-related GO terms in the network).

The mathematical calculation of the mentioned above probability is provided in the following equation.

$$P(X \geq x) = 1 - \sum_{k=0}^{x-1} \frac{\binom{K}{k} \cdot \binom{N-K}{n-k}}{\binom{N}{n}} \tag{8.1}$$

where

- N is the population size (Total number of different GO terms in the network)
- K is the number of success elements in the population (Number of different MDS related GO terms in the network)
- n is the number of draws (Number of assigned GO terms)
- x is the number of desired successes (Number of assigned MDS related GO terms)

8.1 Determining the best random walks parameters

The initial phase of the final experiment involved conducting random walks with restart on this network using various parameter configurations in order to potentially achieve optimal parameters. The parameters had an impact on both the network structure and the random walks algorithm. The parameters that were analysed were the node degree in the network, the length of the random walks, and the probability of returning to the initial position of the random walk. In this phase of the experiment,

100 random walks were conducted for each configuration, along with 500 random shuffles of GO terms.

The tables 8.2, 8.4 and 8.6 display the amounts of assigned Gene Ontology (GO) terms, along with the assigned GO terms related to myelodysplastic syndromes (MDS). Each table corresponds to one of the three piRNAs. Tables 8.3, 8.5 and 8.7 present the p-values derived from the cumulative distribution function of the hypergeometric distribution. The p-values that are statistically significant at a significance level of 0.001 are highlighted in green.

It is noticeable that hsa-piR-018849 has a greater amount of assigned GO terms compared to the other two piRNAs. Furthermore, it exhibits a higher proportion of GO terms related to myelodysplastic syndromes (MDS). The explanation for this may be found in figure 6, since the piRNA hsa-piR-018849 exhibits significantly higher differential expression values compared to other piRNAs in the data-set. Consequently, it is mainly linked with other genes that also show significant differential expression and already have assigned Gene Ontology (GO) terms. As a result, most of the random walks end in the annotated genes, leading to a greater number of assigned GO terms in the final analysis. Furthermore, given that the genes in close contact to this piRNA exhibit significant differences in expression between MDS patients and healthy controls, the likelihood of obtaining a GO term associated with MDS increases.

On the contrary, the remaining two piRNAs (hsa-piR-014626 and hsa-piR-021121) are mainly situated near genes and transposable elements (TE) that show minimal differential expression. Transposable elements (TEs) lack assigned annotations, therefore an abundance of TEs results in numerous random walks terminating at genes without annotations, consequently reducing the number of assigned Gene Ontology (GO) terms. Furthermore, due to the lack of a lot of highly differentially expressed genes in their surroundings, the probability of assigning MDS-related Gene Ontology terms is also reduced.

	Probability of restart	0.05	0.1	0.25
Degree	**Random walks length**			
10	**5**	136 (28)	130 (28)	88 (17)
	10	121 (23)	84 (18)	60 (17)
	50	29 (7)	33 (9)	102 (20)
5	**5**	57 (16)	100 (17)	104 (16)
	10	95 (21)	149 (24)	86 (17)
	50	58 (17)	98 (21)	68 (15)
3	**5**	97 (19)	94 (17)	85 (19)
	10	120 (21)	108 (19)	65 (15)
	50	72 (14)	172 (29)	49 (10)

Table 8.2: Table presenting the results of the random walks with restart method for the piRNA 'hsa-piR-018849' across 27 parameter combinations. Each cell displays the count of assigned Gene Ontology (GO) terms alongside the number of assigned terms related to Myelodysplastic syndromes (MDS), separated by a comma. Parameters include random walk length, restart probability, and the desired degree of the nodes in the gene co-expression network.

	Probability of restart	0.05	0.1	0.25
Degree	**Random walks length**			
10	**5**	3.2×10^{-4}	1.4×10^{-4}	8.7×10^{-3}
	10	3.1×10^{-3}	2.2×10^{-3}	9.0×10^{-5}
	50	2.6×10^{-2}	5.4×10^{-3}	3.9×10^{-3}
5	**5**	1.6×10^{-4}	2.9×10^{-2}	7.2×10^{-2}
	10	6.5×10^{-4}	2.0×10^{-2}	6.9×10^{-3}
	50	5.6×10^{-5}	9.9×10^{-4}	3.7×10^{-3}
3	**5**	4.9×10^{-3}	1.7×10^{-2}	1.0×10^{-3}
	10	1.2×10^{-2}	1.6×10^{-2}	2.4×10^{-3}
	50	1.5×10^{-2}	6.2×10^{-3}	2.8×10^{-2}

Table 8.3: P-Values obtained from the Cumulative distribution function of hypergeometric distribution for results of the random walks with restart method performed for piRNA 'hsa-piR-018849'. The table corresponds with Table 8.2. P-values significant on level 0.001 are highlighted in green.

	Probability of restart	0.05	0.1	0.25
Degree	**Random walks length**			
10	**5**	32 (1)	43 (3)	19 (0)
	10	32 (6)	0 (0)	22 (4)
	50	24 (2)	10 (0)	21 (2)
5	**5**	14 (0)	23 (3)	12 (0)
	10	24 (5)	12 (2)	18 (0)
	50	13 (2)	27 (6)	18 (0)
3	**5**	45 (4)	41 (1)	29 (2)
	10	47 (3)	6 (0)	7 (0)
	50	55 (4)	4 (0)	161 (2)

Table 8.4: Table presenting the results of the random walks with restart method for the piRNA 'hsa-piR-014626' across 27 parameter combinations. Each cell displays the count of assigned Gene Ontology (GO) terms alongside the number of assigned terms related to Myelodysplastic syndromes (MDS), separated by a comma. Parameters include random walk length, restart probability, and the desired degree of the nodes in the gene co-expression network.

	Probability of restart	0.05	0.1	0.25
Degree	**Random walks length**			
10	**5**	0.97	0.84	1.00
	10	0.11	1.00	0.19
	50	0.73	1.00	0.66
5	**5**	1.00	0.44	1.00
	10	0.10	0.36	1.00
	50	0.40	0.06	1.00
3	**5**	0.70	0.99	0.82
	10	0.88	1.00	1.00
	50	0.84	1.00	1.00

Table 8.5: P-Values obtained from the Cumulative distribution function of hypergeometric distribution for results of the random walks with restart method performed for piRNA 'hsa-piR-014626'. The table co-responds with table 8.4.

	Probability of restart	0.05	0.1	0.25
Degree	**Random walks length**			
10	**5**	12 (0)	27 (3)	92 (7)
	10	10 (1)	115 (13)	73 (8)
	50	8 (0)	59 (7)	55 (4)
5	**5**	17 (0)	34 (3)	45 (2)
	10	21 (4)	28 (2)	80 (9)
	50	69 (10)	32 (3)	93 (17)
3	**5**	26 (3)	29 (5)	42 (6)
	10	28 (3)	42 (8)	105 (18)
	50	91 (21)	35 (4)	46 (7)

Table 8.6: Table presenting the results of the random walks with restart method for the piRNA 'hsa-piR-021121' across 27 parameter combinations. Each cell displays the count of assigned Gene Ontology (GO) terms alongside the number of assigned terms related to Myelodysplastic syndromes (MDS), separated by a comma. Parameters include random walk length, restart probability, and the desired degree of the nodes in the gene co-expression network.

	Probability of restart	0.05	0.1	0.25
Degree	**Random walks length**			
10	**5**	1.00	0.55	0.86
	10	0.67	0.42	0.50
	50	1.00	0.42	0.84
5	**5**	1.00	0.70	0.96
	10	0.17	0.81	0.46
	50	0.18	0.66	0.01
3	**5**	0.52	0.18	0.27
	10	0.57	0.07	0.02
	50	3.52×10^{-4}	0.50	0.20

Table 8.7: P-Values obtained from the Cumulative distribution function of hypergeometric distribution for results of the random walks with restart method performed for piRNA 'hsa-piR-021121'. The table co-responds with table 8.6. P-values significant on level 0.001 are highlighted in green.

8.1.1 Network with added piRNA-TE links based on sequence compatibility

The results of analysing the sequence compatibility were not taken into account in the above-mentioned tests. Furthermore, two additional experiments were conducted on the piRNAs, which targeted compatible transposable elements (TEs). The links between the piRNAs and those transposable elements (TEs) were additionally included in the network. Figure 6.3 displays the newly established network. Subsequently, random walks with restart were executed on this revised network. The purpose of doing so was to determine whether the addition of the new link would have any influence on the outcomes. Nevertheless, the impact was minimal and predominantly unfavourable, leading us to walk away from any additional involvement with this particular version. The outcomes of these experiments are displayed in tables 8.8 and 8.10. The p-values obtained from the cumulative distribution function of the hypergeometric distribution are displayed in tables 8.9 and 8.11.

	Probability of restart	**0.05**	**0.1**	**0.25**
Degree	**Random walks length**			
10	**5**	20 (0)	0 (0)	10 (0)
	10	3 (0)	41 (11)	15 (1)
	50	9 (2)	69 (3)	44 (0)
5	**5**	27 (1)	15 (2)	12 (2)
	10	13 (0)	10 (0)	9 (0)
	50	4 (0)	4 (0)	2 (0)
3	**5**	51 (1)	25 (0)	17 (1)
	10	10 (1)	5 (0)	35 (1)
	50	20 (1)	6 (1)	49 (3)

Table 8.8: Table presenting the results of the random walks with restart method for the piRNA 'hsa-piR-014626' on the updated network with added piRNA-TE links across 27 parameter combinations. Each cell displays the count of assigned Gene Ontology (GO) terms alongside the number of assigned terms related to Myelodysplastic syndromes (MDS), separated by a comma. Parameters include random walk length, restart probability, and the desired degree of the nodes in the gene co-expression network.

	Probability of restart	0.05	0.1	0.25
Degree	**Random walks length**			
10	**5**	1.00	1.00	1.00
	10	1.00	0.00	0.81
	50	0.24	0.98	1.00
5	**5**	0.95	0.47	0.36
	10	1.00	1.00	1.00
	50	1.00	1.00	1.00
3	**5**	1.00	1.00	0.85
	10	0.67	1.00	0.98
	50	0.89	0.48	0.90

Table 8.9: P-Values obtained from the Cumulative distribution function of hypergeometric distribution for results of the random walks with restart method performed on the updated network with added piRNA-TE links for piRNA 'hsa-piR-014626'. The table co-responds with table 8.8.

	Probability of restart	0.05	0.1	0.25
Degree	**Random walks length**			
10	**5**	2 (0)	58 (7)	7 (0)
	10	7 (1)	9 (1)	90 (14)
	50	10 (0)	29 (2)	23 (3)
5	**5**	17 (2)	21 (0)	27 (4)
	10	30 (8)	38 (2)	22 (3)
	50	80 (8)	70 (14)	72 (15)
3	**5**	12 (0)	66 (13)	36 (6)
	10	19 (2)	15 (0)	44 (4)
	50	4 (0)	18 (0)	52 (8)

Table 8.10: Table presenting the results of the random walks with restart method for the piRNA 'hsa-piR-021121' on the updated network with added piRNA-TE links across 27 parameter combinations. Each cell displays the count of assigned Gene Ontology (GO) terms alongside the number of assigned terms related to Myelodysplastic syndromes (MDS), separated by a comma. Parameters include random walk length, restart probability, and the desired degree of the nodes in the gene co-expression network.

	Probability of restart	0.05	0.1	0.25
Degree	**Random walks length**			
10	**5**	1.00	0.40	1.00
	10	0.54	0.63	0.08
	50	1.00	0.82	0.44
5	**5**	0.54	1.00	0.31
	10	0.01	0.92	0.41
	50	0.60	0.01	0.01
3	**5**	1.00	0.02	0.17
	10	0.60	1.00	0.69
	50	1.00	1.00	0.17

Table 8.11: P-Values obtained from the Cumulative distribution function of hypergeometric distribution for results of the random walks with restart method performed on the updated network with added piRNA-TE links for piRNA 'hsa-piR-021121'. The table co-responds with table 8.10.

8.2 Final Experiment

For the final phase of the experiment, an entirely new set of piRNAs was selected for examination, and a specific combination of parameters was chosen. The selection of the parameter combination was determined by the outcomes obtained for piRNA hsa-piR-018849, as the results for the other piRNAs were not sufficiently satisfying. Selecting the optimal parameter combination proved challenging due to the experiment's limited size caused by limitations on time and high computational requirements. The random walks with restart did not achieve complete convergence in this small experiment, suggesting the possibility that a more optimal parameter combination may exist but was not discovered. We selected the combination of a node degree set to 5 and a restart probability of 0.05, as this combination consistently provided statistically significant p-values. The random walk length was set to 50 by selecting the one with the lowest p-value among the three lengths.

The piRNAs for this final experiment were selected after finalising the decision on parameter combination. We decided to select three piRNAs from our data-set that were the only ones showing statistically significant differential expression between patients with myelodysplastic syndromes (MDS) and healthy controls. Additionally, we randomly chose one piRNA that exhibited only minimal differential expression. Among the three piRNAs that show differential expression, only one is up-regulated in the group of MDS patients. This particular piRNA also exhibits the largest log2 fold change, indicating the greatest difference in expression between the two groups under examination. It is also the piRNA used in the previous experiment was also utilised, and the parameter combination was determined based on the results obtained from it. Two additional piRNAs that are differentially expressed are down-regulated in the group of patients with MDS. Table 8.12 contains all the piRNAs that have been chosen. The expectation is to achieve a higher ratio of MDS-related GO terms for the three differentially expressed piRNAs compared to the last piRNA, which does not exhibit significant differential expression.

piRNA	baseMean	log2FoldChange	pvalue	padj
hsa-piR-018849	89.58	5.97	9.35×10^{-05}	0.020
hsa-piR-020828	64.68	-2.17	6.22×10^{-4}	0.045
hsa-piR-009051	92.91	-1.41	6.09×10^{-4}	0.045
hsa-piR-021032	183.51	0.25	4.02×10^{-1}	0.996

Table 8.12: Table presenting piRNAs selected for the final experiment. Information obtained from the differential expression analysis are displayed.

The largest feasible experiment was conducted for the given piRNAs and parameter combination, taking into account our limitations, such as limitation of time and the insufficient computational power. We conducted 500 random walks on the accurate gene co-expression network, as well as 500 random walks with restart on 5000 gene co-expression networks that had GO terms shuffled in different ways.

The outcomes of this experiment are displayed in table 8.13. The table provides the piRNA identifiers, the total count of assigned Gene Ontology (GO) terms to each piRNA, the count of assigned GO terms related to myelodysplastic syndromes (MDS), and the p-value obtained from the cumulative distribution function of the hypergeometric distribution. The piRNA hsa-piR-018849 exhibited the most favourable outcomes, which was demonstrated by a p-value of 0.0016. This was expected, given that it possesses the highest differential expression values among all piRNAs. The piRNA hsa-piR-009051 also demonstrated a significant p-value at a significance level of 0.05. However, the third piRNA (hsa-piR-020828) exhibited the most unfavourable outcomes among the four piRNAs, even surpassing the non-interesting piRNA that, as expected, failed to produce a significant p-value. Following this discovery, we analysed the location of the piRNA hsa-piR-020828 within the network. The network is depicted in figure 8.1. The piRNA in debate is portrayed as a yellow node, situated at the boundary between differentially expressed genes and those without significant differential expression. This position of the node may explain the unsatisfactory outcome.

piRNA	# assigned GO terms (MDS)	p-value
hsa-piR-018849	122 (24)	1.6×10^{-3}
hsa-piR-020828	36 (4)	0.53
hsa-piR-009051	118 (22)	4.9×10^{-3}
hsa-piR-021032	34 (5)	0.28

Table 8.13: Results of the final experiment. Numbers of assigned annotations and assigned MDS related annotations are presented as well as the p-value obtained from the cumulative distribution function of hypergeometric distribution.

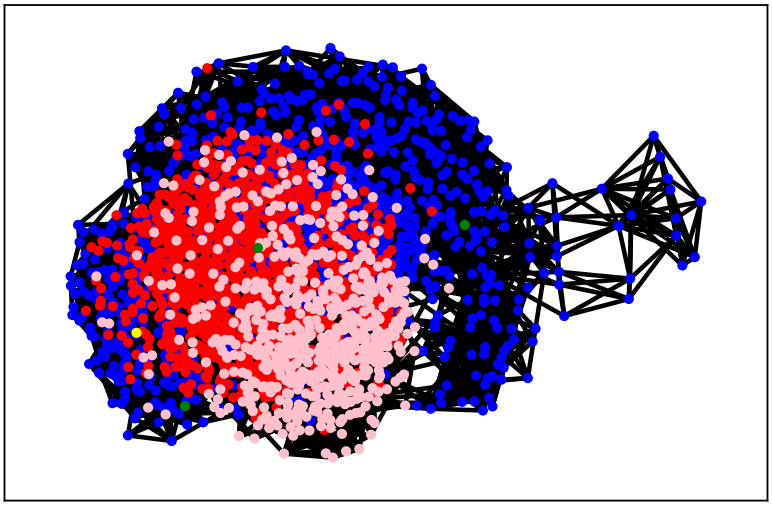


Figure 8.1: Visualisation of the gene co-expression network used in the final experiment. Blue nodes represent genes with significant differential expression between MDS patients and healthy controls, red nodes are genes with minimal differential expression between previously mentioned groups, pink nodes represent transposable elements (TE). Yellow node represents piRNA hsa-piR-020828 that exhibited statistically significant differential expression but did not presented statistically significant results from random walks. Other three piRNAs are displayed as green nodes.

The GO terms assigned to piRNA hsa-piR-018849 at a significance level of 0.005 are displayed in Figure 8.2. The graph was generated using the enrichment map plugin within the cytoscape software. The MDS related GO terms are represented by red nodes, while the remaining assigned GO terms are represented by blue nodes. It is evident that the majority of the assigned GO terms are concentrated in clusters, indicating strong associations between the GO terms. This suggests that the findings were not just coincidental.

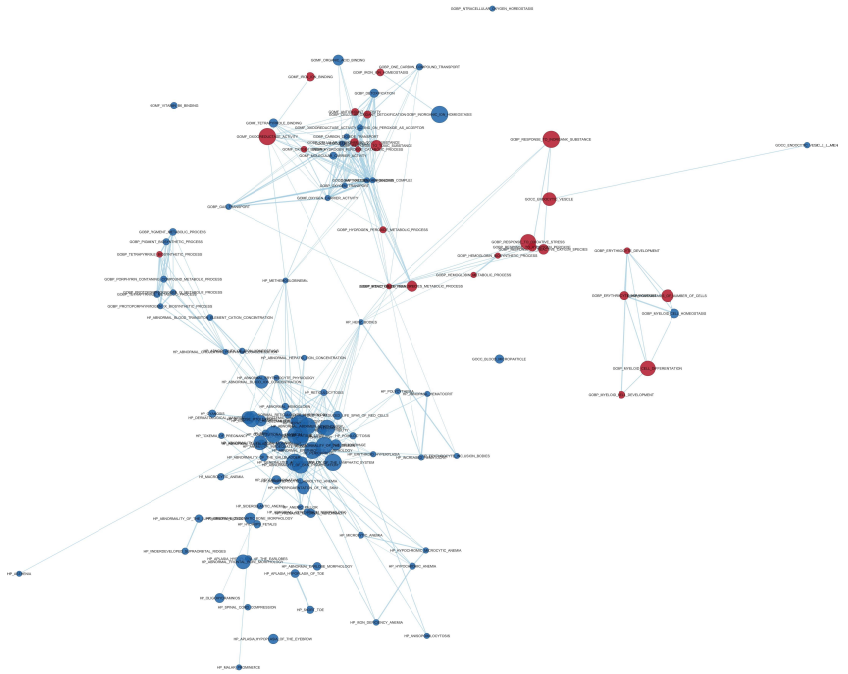


Figure 8.2: Graphical visualisation of GO terms assigned to piRNA hsa-piR-018849 created with the enrichment map plugin of the cytoscape software. The myelodysplastic syndromes related GO terms are visualised as red circles, blue circles represent GO terms that are not related to MDS.

However, in order to obtain more reliable results, a significantly larger experiment would need to be conducted, specifically involving a greater number of random walks with restart per shuffle. Unfortunately, the execution of such a massive experiment was unfeasible due to the limited amount of time and the insufficient computational power of my computer. Increasing the number of random walks with restart (RWR) per shuffle generally leads to improved convergence of the RWR and strengthens the reliability of the results. The resolution of the p-value and the reliability of the results are influenced by the number of GO term shuffles.

Chapter 9

Conclusions

The fundamental goal of this diploma thesis was to develop a tool capable of assigning functional annotations (GO terms) to PIWI-interacting RNAs (piRNA), with a particular focus on potential associations with myelodysplastic syndrome (MDS). The Institute of Haematology and Blood Transfusion provided expression data for this purpose. The data were measured across multiple groups of subjects. We focused primarily on two groups: patients suffering with myelodysplastic syndromes and healthy individuals, serving as a reference for comparison. The data-set contained expression data for PIWI-interacting RNAs, transposable elements, and various other genes.

The assignment of GO terms was conducted using a random walks with restart (RWR) algorithm, accompanied by the performance of permutation tests. The implementation of the RWR algorithm was motivated by the concept of the guild by association principle, which claims that molecules that interact with each other frequently possess similar functions. Due to the large size of the provided data-set and our limited computational capacity, the initial task was to determine the selection of genes to be included in the gene co-expression network, which functioned as the framework for the RWR algorithm. The concept of differential expression was introduced for this purpose. As a result, we had the ability to identify and choose genes that are particularly relevant to our research. The noteworthy genes were those that displayed significant differences in expression levels between our groups of interest, namely MDS patients and healthy controls. After obtaining the information about differential expression of genes, the genes could be selected and the gene co-expression network could be constructed.

Throughout our experiments, we encountered an issue caused by an excessive

amount of links in the network. Therefore, an approach had been used to decrease the quantity by establishing a fixed degree for the nodes. Following this, a decision had to be made concerning the optimal number of edges for the network. The RWR algorithm required the establishment of certain variable parameters, such as the length of the walks and the probability of restart. As a result, the initial phase of the experiment was focused on discovering the most suitable combination of parameters. After selecting the combination, a more extensive experiment was conducted. Four piRNAs were picked for the annotation assignment in this experiment. Three piRNAs exhibited significant differential expression, indicating a potential association with myelodysplastic syndromes. The last piRNA exhibited minimal differential expression and served to support the belief that piRNAs with lower differential expression would be associated with fewer MDS-related GO terms.

Due to the absence of an available database containing the correct GO term assignments for our piRNAs, the only possible way to evaluate the results was to examine the number of assigned GO terms along with the proportion that related to MDS. For this purpose, the cumulative distribution function of the hypergeometric distribution was introduced to help us in determining whether the proportion of MDS related GO terms is statistically significant given the sizes of the populations of GO terms/MDS GO terms in the gene co-expression network.

The p-values for two out of the three selected piRNAs with significant differential expression (hsa-piR-018849, hsa-piR-009051) were determined to be statistically significant at a significance level of 0.05. This suggests a potential association with myelodysplastic syndromes. The p-value for the selected piRNA with non-significant differential expression was as expected, demonstrating a lack of statistical significance. However, the p-value for the last piRNA also turned out to be non-significant, despite the fact that this piRNA showed statistically significant differential expression between the groups of myelodysplastic syndrome patients and healthy controls.

9.1 Future plans

The database containing verified annotations for the provided piRNAs currently remains unavailable. It would be beneficial to revisit this method once the annotations that have been verified through biological means are accessible. This might be helpful in accurately assessing the efficiency of our implemented method.

Until then, it would be beneficial to conduct an experiment with a significantly higher number of random walks per shuffle and an increased number of shuffles in order to obtain more reliable outcomes.

Bibliography

[1] X. Wu, Y. Pan, Y. Fang, J. Zhang, M. Xie, F. Yang, T. Yu, P. Ma, W. Li, and Y. Shu, "The biogenesis and functions of pirnas in human diseases," *Molecular Therapy - Nucleic Acids*, vol. 21, pp. 108–120, 2020.

[2] J. Ruan and W. Zhang, "Identification and evaluation of functional modules in gene co-expression networks," in *Systems Biology and Computational Proteomics* (T. Ideker and V. Bafna, eds.), (Berlin, Heidelberg), pp. 57–76, Springer Berlin Heidelberg, 2007.

[3] X. Lei and C. Bian, "Integrating random walk with restart and k-nearest neighbor to identify novel circrna-disease association," *Scientific Reports*, vol. 10, no. 1, 2020.

[4] J. Li, X. Li, X. Feng, B. Wang, B. Zhao, and L. Wang, "A novel target convergence set based random walk with restart for prediction of potential lncrna-disease associations," *BMC Bioinformatics*, vol. 20, no. 1, 2019.

[5] L. Wang, M. Shang, Q. Dai, and P.-a. He, "Prediction of lncrna-disease association based on a laplace normalized random walk with restart algorithm on heterogeneous networks," *BMC Bioinformatics*, vol. 23, no. 1, 2022.

[6] A. T. Vivek and S. Kumar, "Computational methods for annotation of plant regulatory non-coding rnas using rna-seq," *Briefings in Bioinformatics*, vol. 22, pp. 1–24, Jul. 2021.

[7] A. Aravin, D. Gaidatzis, S. Pfeffer, M. Lagos-Quintana, P. Landgraf, N. Iovino, P. Morris, M. J. Brownstein, S. Kuramochi-Miyagawa, T. Nakano, M. Chien, J. J. Russo, J. Ju, R. Sheridan, C. Sander, M. Zavolan, and T. Tuschl, "A novel class of small RNAs bind to MILI protein in mouse testes," *Nature*, vol. 442, pp. 203–207, June 2006.

[8] X. Wang, A. Ramat, M. Simonelig, and M.-F. Liu, "Emerging roles and functional mechanisms of piwi-interacting rnas," *Nature Reviews Molecular Cell Biology*, vol. 24, no. 2, pp. 123–141, 2023.

[9] Y. Liu, A. Li, G. Xie, G. Liu, and X. Hei, "Computational methods and online resources for identification of pirna-related molecules," *Interdisciplinary Sciences: Computational Life Sciences*, vol. 13, no. 2, pp. 176–191, 2021.

[10] S. Hirakata and M. C. Siomi, "pirna biogenesis in the germline: From transcription of pirna genomic sources to pirna maturation," *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, vol. 1859, no. 1, pp. 82–92, 2016. SI: Clues to long noncoding RNA taxonomy.

[11] J. Castañeda, P. Genzor, and A. Bortvin, "pirnas, transposon silencing, and germline genome integrity," *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, vol. 714, no. 1, pp. 95–104, 2011.

[12] S. D. Ali, H. Tayara, and K. T. Chong, "Identification of pirna disease associations using deep learning," *Computational and Structural Biotechnology Journal*, vol. 20, pp. 1208–1217, 2022.

[13] D.-D. Jia, H. Jiang, Y.-F. Zhang, Y. Zhang, L.-L. Qian, and Y.-F. Zhang, "The regulatory function of pirna/piwi complex in cancer and other human diseases: The role of dna methylation," *International Journal of Biological Sciences*, vol. 18, no. 8, pp. 3358–3373, 2022.

[14] T. A. Anzelon, S. Chowdhury, S. M. Hughes, Y. Xiao, G. C. Lander, and I. J. MacRae, "Structural basis for pirna targeting," *Nature*, vol. 597, pp. 285–289, Sep. 2021.

[15] J. S. Khurana and W. Theurkauf, "piRNAs, transposon silencing, and Drosophila germline development," *Journal of Cell Biology*, vol. 191, pp. 905–913, 11 2010.

[16] J. Hou, H. Wei, B. Liu, and S. Bozdag, "Ipida-gcn: Identification of pirna-disease associations based on graph convolutional network," *PLOS Computational Biology*, vol. 18, Oct. 2022.

[17] M. Merkerova and Z. Krejcik, "Transposable elements and piwi-interacting rnas in hemato-oncology with a focus on myelodysplastic syndrome (review)," *International Journal of Oncology*, vol. 59, Nov. 2021.

[18] V. Georgoulis, E. Koumpis, and E. Hatzimichael, "The role of non-coding rnas in myelodysplastic neoplasms," *Cancers*, vol. 15, no. 19, 2023.

[19] S. Nam, B. Kim, S. Shin, and S. Lee, "Mirgator: an integrated system for functional annotation of micrornas," *Nucleic Acids Research*, vol. 36, pp. D159–D164, Dec. 2007.

[20] M. Dostalová Merkerová, J. Kléma, D. Kundrát, K. Szikszai, Z. Krejčík, A. Hruštincová, I. Trsová, A. V. Le, J. Čermák, A. Jonášová, and M. Beličková, "Noncoding rnas and their response predictive value in azacitidine-treated patients with myelodysplastic syndrome and acute myeloid leukemia with myelodysplasia-related changes," *Cancer Genomics - Proteomics*, vol. 19, pp. 205–228, Feb. 2022.

[21] Q. Liao, H. Xiao, D. Bu, C. Xie, R. Miao, H. Luo, G. Zhao, K. Yu, H. Zhao, G. Skogerbo, R. Chen, Z. Wu, C. Liu, and Y. Zhao, "Ncfans: a web server for functional annotation of long non-coding rnas," *Nucleic Acids Research*, vol. 39, pp. W118–W124, Jun. 2011.

[22] P. Ryšavý, J. Kléma, and M. D. Merkerová, "Circgpa: circrna functional annotation based on probability-generating functions," *BMC Bioinformatics*, vol. 23, no. 1, 2022.

[23] J. Cardenas, U. Balaji, and J. Gu, "Cerina: systematic circrna functional annotation based on integrative analysis of cerna interactions," *Scientific Reports*, vol. 10, no. 1, 2020.

[24] C. S. Copeland, M. Marz, D. Rose, J. Hertel, P. J. Brindley, C. B. Santana, S. Kehr, C. S.-O. Attolini, and P. F. Stadler, "Homology-based annotation of non-coding rnas in the genomes of schistosoma mansoni and schistosoma japonicum," *BMC Genomics*, vol. 10, no. 1, 2009.

[25] A. Zampetaki, A. Albrecht, and K. Steinhofel, "Corrigendum: Long non-coding rna structure and function," *Frontiers in Physiology*, vol. 10, Sep. 2019.

[26] N. Amin, A. McGrath, and Y.-P. P. Chen, "Evaluation of deep learning in non-coding rna classification," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 246–256, 2019.

[27] E. A. R. Serin, H. Nijveen, H. W. M. Hilhorst, and W. Ligterink, "Learning from co-expression networks: Possibilities and challenges," *Frontiers in Plant Science*, vol. 7, Apr. 2016.

[28] V. Kunc and J. Kléma, "On functional annotation with gene co-expression networks," in *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 3055–3062, 2022.

[29] A. Emamjomeh, E. Saboori Robat, J. Zahiri, M. Solouki, and P. Khosravi, "Gene co-expression network reconstruction: a review on computational methods for inferring functional information from plant-based expression data," *Plant Biotechnology Reports*, vol. 11, pp. 71–86, Apr 2017.

[30] D. Yu, M. Kim, G. Xiao, and T. H. Hwang, "Review of biological network data and its applications," vol. 11, no. 4, 2013.

[31] S. Lefever, J. Anckaert, P.-J. Volders, M. Luypaert, J. Vandesompele, and P. Mestdagh, "Decoderna— predicting non-coding rna functions using guilt-by-association," *Database*, vol. 2017, Jan. 2017.

[32] J. Gillis, P. Pavlidis, and A. Rzhetsky, ""guilt by association" is the exception rather than the rule in gene networks," *PLoS Computational Biology*, vol. 8, Mar. 2012.

[33] F. Xia, J. Liu, H. Nie, Y. Fu, L. Wan, and X. Kong, "Random walks: A review of algorithms and applications," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 4, no. 2, pp. 95–107, 2020.

[34] L. Lovász, "Random walks on graphs," *Combinatorics, Paul erdos is eighty*, vol. 2, no. 1-46, p. 4, 1993.

[35] S. Köhler, S. Bauer, D. Horn, and P. N. Robinson, "Walking the interactome for prioritization of candidate disease genes," *The American Journal of Human Genetics*, vol. 82, no. 4, pp. 949–958, 2008.

[36] A. Valdeolivas, L. Tichit, C. Navarro, S. Perrin, G. Odelin, N. Levy, P. Cau, E. Remy, and A. Baudot, "Random walk with restart on multiplex and heterogeneous biological networks," *Bioinformatics*, vol. 35, pp. 497–505, 07 2018.

[37] M. J. Vondráková, "Myelodysplastickỳ syndrom, diagnostika a léčba," *Interní medicína pro praxi*, vol. 12, no. 11, pp. 535–539, 2010.

[38] L. Červinek *et al.*, "Myelodysplastickỳ syndrom," *Onkologie*, vol. 9, no. 3, pp. 112–118, 2015.

[39] T. B. Lauritsen, J. M. Nørgaard, S. O. Dalton, K. Grønbæk, T. C. El-Galaly, and L. S. G. Østgård, "10-year nationwide trends in incidence, treatment patterns, and mortality of patients with myelodysplastic syndromes in denmark," *Leukemia Research*, vol. 128, 2023.

[40] J. P. Bewersdorf, H. Carraway, and T. Prebet, "Emerging treatment options for patients with high-risk myelodysplastic syndrome," *Therapeutic Advances in Hematology*, vol. 11, 2020.

[41] P. Fenaux, U. Platzbecker, and L. Ades, "How we manage adults with myelodysplastic syndrome," *British Journal of Haematology*, vol. 189, no. 6, pp. 1016–1027, 2020.

[42] "Phenotype information for myelodysplastic syndromes." `https://ctdbase.org/detail.go?type=disease&acc=MESH%3AD009190&view=phenotype`. Accessed: November 2023.

[43] U. von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, p. 395–416, Aug. 2007.

[44] W.-S. Wu, W.-C. Huang, J. S. Brown, D. Zhang, X. Song, H. Chen, S. Tu, Z. Weng, and H.-C. Lee, "Pirscan: a webserver to predict pirna targeting sites and to avoid transgene silencing in c. elegans," *Nucleic Acids Research*, vol. 46, pp. W43–W48, Jul. 2018.

[45] D. Zhang, S. Tu, M. Stubna, W.-S. Wu, W.-C. Huang, Z. Weng, and H.-C. Lee, "The pirna targeting rules and the resistance to pirna silencing in endogenous genes," *Science*, vol. 359, pp. 587–592, Feb. 2018.

[46] M. Ojala and G. C. Garriga, "Permutation tests for studying classifier performance.," *Journal of machine learning research*, vol. 11, no. 6, 2010.