ODYSSEY: A TOOL FOR MICRORNA-MRNA EXPRESSION AND INTERACTION VISUALIZATION

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS
OF
MIDDLE EAST TECHNICAL UNIVERSITY

ALPEREN TACIROGLU

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
BIOINFORMATICS

SEPTEMBER 2018

Approval of the thesis:

ODYSSEY: A TOOL FOR MICRORNA-MRNA EXPRESSION AND INTERACTION VISUALIZATION

submitted by **ALPEREN TACIROGLU** in partial fulfillment of the requirements for the degree of **Master of Science in Bioinformatics, Middle East Technical University** by,

Prof. Dr. Deniz ZEYREK BOZŞAHİN
Dean, **Graduate School of Informatics** _____

Assoc. Prof. Dr. Yeşim AYDIN SON
Head of Department, **Health Informatics, METU** _____

Assist. Prof. Dr. Aybar Can ACAR
Supervisor, **Health Informatics, METU** _____

Assoc. Prof. Dr. Özlen KONU KARAKAYALI
Co-supervisor, **Molecular Biology and Genetics, İ. D. Bilkent University** _____

Examining Committee Members:

Assoc. Prof. Dr. Yeşim AYDIN SON
Health Informatics, METU _____

Assist. Prof. Dr. Aybar Can ACAR
Health Informatics, METU _____

Assoc. Prof. Dr. Özlen KONU KARAKAYALI
Molecular Biology and Genetics, İ. D. Bilkent University _____

Assist. Prof. Dr. Onur ÇİZMECİOĞLU
Molecular Biology and Genetics, İ. D. Bilkent University _____

Assoc. Prof. Dr. Nurcan TUNÇBAĞ
Health Informatics, METU _____

Date: _____

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: ALPEREN TACIROGLU

Signature :

ABSTRACT

ODYSSEY: A TOOL FOR MICRORNA-MRNA EXPRESSION AND INTERACTION VISUALIZATION

TACIROGLU, Alperen

M.S., Department of Bioinformatics

Supervisor : Assist. Prof. Dr. Aybar Can ACAR

Co-Supervisor : Assoc. Prof. Dr. Özlen KONU KARAKAYALI

September 2018, 85 pages

MicroRNAs (miRNAs) are non-coding short RNA molecules that are found in all metazoa studied so far. When distinct metazoa genomes considered up to 200 genes encode for unique miRNAs that show variability between species. Regulatory functions of miRNAs have been studied for 20 years starting after their discovery. The research suggests that they are involved in a wide spectrum of biological activities including apoptosis, tumorigenesis, development, homeostasis and viral infections. miRNAs regulate these cellular processes at the post-transcriptional level by binding to the messenger RNAs (mRNAs), leading to an unstable derivative of the initial biological molecule. miRNA targets are under strict evolutionary pressure which further implicates the importance of underlying biological mechanisms. Although there are several Gene/mRNA-miRNA interaction visualization and analysis tools "Odyssey" was built for improved interactive visualization the interaction network of miRNAs with along with their target expressions for a user uploaded dataset. It is built using Shiny package

of the R programming language leading to seamless online access and modularity. In the end, I aim to provide users a user-friendly web-application which consists of modules that allows: uploading of their own data; performing differential expression (DGEx) analysis; and visualization of the network of which "Odyssey" builds from either experimentally validated or predicted interactions for individual miRNAs queried by the user. Odyssey further enables the user to filter selected nodes of the networks using fold change cut-offs obtained in DGEx step or expand the network using Gene Ontology (GO) terms to act as a strong predictor of the phenotype of interest for the user-specified biological data. Furthermore, the application has been demonstrated using two different public miRNA-mRNA expression datasets.

Keywords: Bioinformatics, microRNA, Gene, Differential Expression, Graph

ÖZ

ODYSSEY: MİKRORNA-GEN EKSPRESYONU VE ETKİLEŞİMLERİNİN GÖRSELLEŞTİRİLMESİ ÜZERİNE BİR YAZILIM

TACIROGLU, Alperen

Yüksek Lisans, Biyoenformatik Programı

Tez Yöneticisi : Doktor Öğretim Üyesi Aybar Can ACAR

Ortak Tez Yöneticisi : Doç. Dr. Özlen KONU KARAKAYALI

Eylül 2018 , 85 sayfa

MikroRNA'lar kodlamayan RNA moleküllerinden olup şimdiye kadar çalışılmış bütün omurgasız hayvanlar dahil pek çok organizmada işlevsel roller alırlar. Farklı omurgasız türleri göz önüne alındığında, 200'e kadar farklı gen, mikroRNA kodlamasında görev alabilir. İlk olarak 20 yıl önceki keşiflerinden bu yana, hücre içerisindeki görevleri, hangi biyolojik yolaklarda regülatör olarak görevler aldıkları çok çeşitli araştırmalara konu olmuştur. Bu araştırmalar sonucunda, apoptoz, homeostaz, gelişimsel biyoloji, kanser olusumu ve virus enfeksiyonları sonucu aktive olan yolaklar dahil farklı birden çok önemli biyolojik mekanizmalarda etken rol aldıkları gözlemlenmiştir. MikroRNA'lar bu mekanizmalara genel olarak post-transkripsiyonel aşamada mesajcı RNA moleküllerine bağlanıp, bu moleküllerin stabilitesini bozarak yön verirler. Evrimsel süreçte incelendiklerinde, nükleotit dizilimlerinin korunmuş olduğu fark edilmiştir ve bu da hücre içerisindeki görevlerinin önemini vurgular niteliktedir. 'Odyssey' mikroRNA'ların bağlandıkları genlerle olan ilişkilerini etkileşimli ağlar yoluyla çizmek için R programlama

dilinin Shiny paketi kullanılarak geliştirildi. Bu web uygulaması geliştirilirken, kullanıcıların kendi biyolojik ekspresyon datasını yükleyebilecekleri ve bu data üzerinden karşılaştırmalı ekspresyon analizi yapabilecekleri, ayrıca bu analiz sonucunda çıkan sonuçları etkileşimli ağ olarak çizebilecekleri kullanımı kolay bir arayüze sahip olmasına özen gösterildi. Ayrıca Odyssey'de bu ağları oluşturan mikroRNA veya genlerin ekspresyon analizi sonucunda çıkan değerler üzerinden filtrelenebilmesi için modüller eklendi. Web uygulamasının nihai amacı olarak kullanıcının analizini yaptığı data fenotipini açıklayıcı nitelikte güçlü bir öngörü aracı olarak kullanılabilmesi hedeflendi.

Anahtar Kelimeler: Biyoenformatik, mikroRNA, Gen, gen regülasyon ağları, gen ekspresyon analizi

To My FAMILY...

Aileme ...

ACKNOWLEDGMENTS

Dr. Aybar Can ACAR and Dr. Özlen KONU KARAKAYALI; for their patient guidance, encouragement and advice

Mr. Hasan MIRZA; for valuable advice on coding

Dr. Andy SIMS; for helping with network building

Boran YAVUZ, MD.; for ideas on how to build a better User Interface

Dr. Zihni Ekim TAŞKIRAN; for his support on the writing of the thesis

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

MTIdb	MicroRNA - Target Interaction database
DGex	Differential Gene Expression Analysis
GRNs	Gene Regulatory Networks
miRNA	microRNA
GO	Gene Ontology
qRT-PCR	quantitative Real Time - Polymerase Chain Reaction
RNA-seq	RNA sequencing
nt	nucleotide
RISC	RNA-induced Silencing Complex
Ago2	Argonaute2
p-bodies	processing bodies
PPI	Protein-Protein interaction
TF	Transcription Factor
NGS	Next Generation Sequencing
CoSE	Compound Spring Embedder
TCA cycle	Tricarboxylic Acid Cycle
HGNC	HUGO Gene Nomenclature Committee
UI	User-Interface
MIAME	Minimum information about a microarray experiment
GEO	Gene Expression Omnibus
EVd	Expression Value difference
SIF	Simple Interaction Format
KEGG	Kyoto Encyclopedia of Genes and Genomes
ANTXR2	Anthrax toxin receptor 2
CAMs	Enriched Cell adhesion molecules

CHAPTER 1

INTRODUCTION

1.1 Motivation

Producing increasing amounts of biological data has created and exposed the need for analyzing these data at a faster pace. We need to find novel ways to uncover the underlying biological meaning of the genome and transcriptome-based datasets. As these data are produced through computational methods, the analysis process also requires sophisticated computational methods. Machine learning in biological data mining [33] is one of the novel ways proposed to extract knowledge from the biological data and the other is using graphs [18]. Generation of graphs from biological data provides a better extract of the system even if only it is to be inspected visually. Researchers are not unfamiliar with observing biological components with a naked eye, as microscopic images had been observed under the microscope for almost 400 years. These cellular images can now be visualized in the digital environment as data in much higher resolutions than before. As there have been developments through computational methods to visualize and extract information from biological images obtained through microscopes, i.e. image processing, there are also newer methods to investigate biological networks laid out using graph generating software, i.e. graph theory-based applications.

There have been many attempts to generate meaningful graphs from biological data and one of the major pieces of software that researchers frequently use is Cytoscape [116] [131] [92]. Other software also exists specifically designed for microRNA-gene networks generation (mentioned in detail in the Software for

Biological Interactions Integration section) however, these software either have a steep learning curve or might lack some important functionality (explained in Scope and Goal section) necessary for efficient transcriptomics analyses. In any case, ongoing research on this area is constantly improving and generating better graphs using applications that are easy to use and have a vast opportunity for further improvement through version updates.

Graphs are widely used by other disciplines as well, which is another reason for encouragement but in this work, I only focused on a new software development to generate biological graphs for mRNA and microRNA interaction as well as annotation of nodes and edges with existing transcriptomics and interactomics data. Being highly conserved through evolution and having experimental documentation in important numerous cellular process, microRNAs take part in different physiological and pathological conditions for multiple complex organisms. microRNA research has improved tremendously during the last two decades, proven by the highly cited articles published recently, suggesting that microRNAs can be candidates to develop therapeutic agents to be used in cancer treatment [59] [96]. Therefore, it is important to better understand their regulatory potential and provide a systems biology approach to microRNA research in different contexts. microRNA – gene interactions are a good candidate to be displayed as graphs. In addition, the expression values of mRNAs and their targeting microRNAs can enhance the information obtained from the graphs.

Furthermore, major tools that are used for graph generation, such as Cytoscape, require installation on local computers. Data to generate graphs, having large storing space needs in the case of next-generation sequencing technologies, also need to be stored within the same computational platform. This creates an issue for many of the researchers with no access to high-performance computational power. This is also one of the reasons why cloud-based software is being widely used in the recent years [110]. My motivation in this study is to reduce the need for local installation of the data and analyses tools and provide a user-friendly and interactive environment for mRNA-miRNA interaction and expression analysis.

1.2 Scope and Goal

The primary objective of this research is to develop a web-based application for microRNA – gene network generation and functional/expression annotation that can be used easily by researchers that are not necessarily experienced online biological tool users. Example data sets are provided as uploaded in the application for new users to test Odyssey in accordance with the documentary provided before moving forward to upload their own data and perform analytic research. As a part of this study, two public mRNA-microRNA expression datasets (meningioma dataset GSE88721 and melanoma dataset GSE35389) have been studied with Odyssey's modules; and networks are drawn to demonstrate findings at a systems biological scale (see Case Study chapter) and discussed in relationship with the existing literature. Emphasis has been placed on a highly automated process of graph generation in Odyssey. This is to make up for shortcomings of the existing relevant online software. User profiles of online bioinformatics tools are diverse. Some users only need to generate a figure or a graph using the tool whereas some users feel the need to learn the tool in detail to correspond to the needs of a project. Bioinformatics tools need to cater to this diverse user profile and make the application user-friendly overall. Odyssey has been developed with a primary aim to provide a user-friendly interface for color-coded visualization mRNA-miRNA interaction-expression networks.

In accordance with above-mentioned microRNA cellular functionalities, another objective of this research through re-analyzing existing datasets is to demonstrate emergent microRNA characteristics for future prospects of the microRNA related research. Apart from restating what has already been reported with published literature, novel network components are anticipated to be revealed by applying Odyssey into a public dataset. In addition, I aimed to allow users to input their own datasets for a systems-biology perspective. Thereby users are encouraged to perform their own analyses using Odyssey and discover novelties on their own. Thus, the final and most important objective of this study is to make a novel impact and contribute to the accumulated knowledge in mRNA-miRNA interaction network software and in relevant biological sciences.

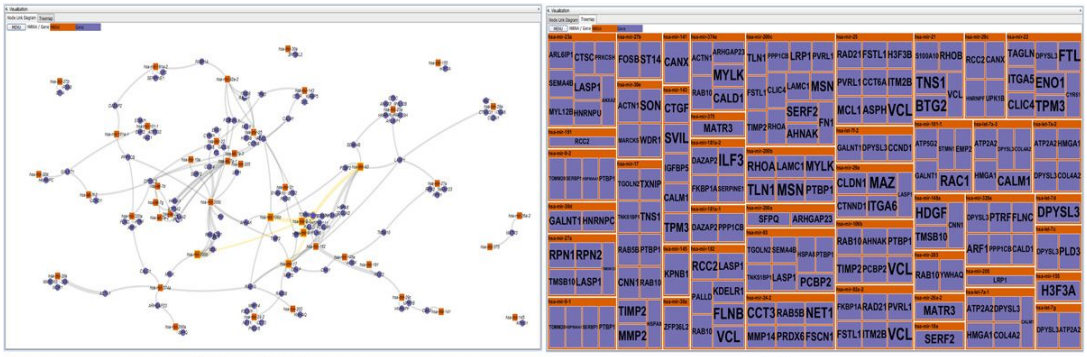
1.3 Existing Alternative Software

As briefly mentioned in the Background section, other tools exist that are similar to Odyssey and/or relevant to the topic of the present study. Alternative and/or improved interpretations of the same dataset by using different tools allow for the identification of different characteristics of the data and the resulting networks. Therefore, why advantages of development of Odyssey have been explained in this section through comparisons with these existing software.

1.3.1 miRTarVis

miRTarVis is a downloadable software for interactive network visualization [58]. Expression data is reflected on the network based on differential analysis of the input data. miRTarVis is capable of filtering network nodes by expression value fold change or p-value cut-offs designated by the user. miRNA - gene pairing is determined by either sequence-based prediction methods or expression profile-based prediction algorithms. Thus, a resulting network is comprised of predicted interactions. miRTarVis is further capable of generating the graph by two different layout algorithms namely; node-link diagram or tree diagram. Finally, in relation with this study, miRTarVis is capable of analysing the networks to generate a report of Gene Ontology Enrichment Analysis. The outcome of this analysis is given as a downloadable report.

However, mirTARVIS is not able to annotate edges differentially as predicted or experimentally verified. Moreover, it is not interactive nor the layout algorithm can be changed dynamically. Log-fold change cut-offs also cannot be changed easily without starting a session from scratch. On the other hand, advantages of MirTarVis includes treemap view of the network which is not provided by other software. MirTarVis is downloaded to local computers and run without an internet connection or uploading user data to remote servers.



A) Node-link diagram B) Treemap

Figure 1.1: An overview of the network layout types through a tutorial session of miRTarVis.

1.3.2 miRNet

Unlike miRTarVis, miRNet is an interactive tool that can be used directly through web browsers without having the need to download and install the tool in a local computer. MiRNet supports expression data from different platforms, i.e. RNA-seq, microarray, qRT-PCR and there are built-in expression data is separated by organisms and tissues for easy differential expression analysis according to the need of researchers researcher's needs. Functional enrichment of the nodes and different layout types are combined with an intuitive user interface which makes miRNet a powerful online tool for network-based visual analysis. [36].

However, although miRNet is interactive and has multiple interchangeable network layouts to incorporate mRNA and microRNA expression levels, example dataset implemented within miRNet does not have a miRNA data counterpart whilst having mRNA data, which increases the steepness of the learning curve of the software. GO term enrichment tool occasionally falls short for recovering statistically significant clusters calculated with p-values mostly due to lacking nodes connected to the query second-degree. In addition, miRNet focuses on experimentally validated targets of microRNAs while there is no integration of predicted target information; it does not allow using intersection or union of validated and predicted microRNA targets.

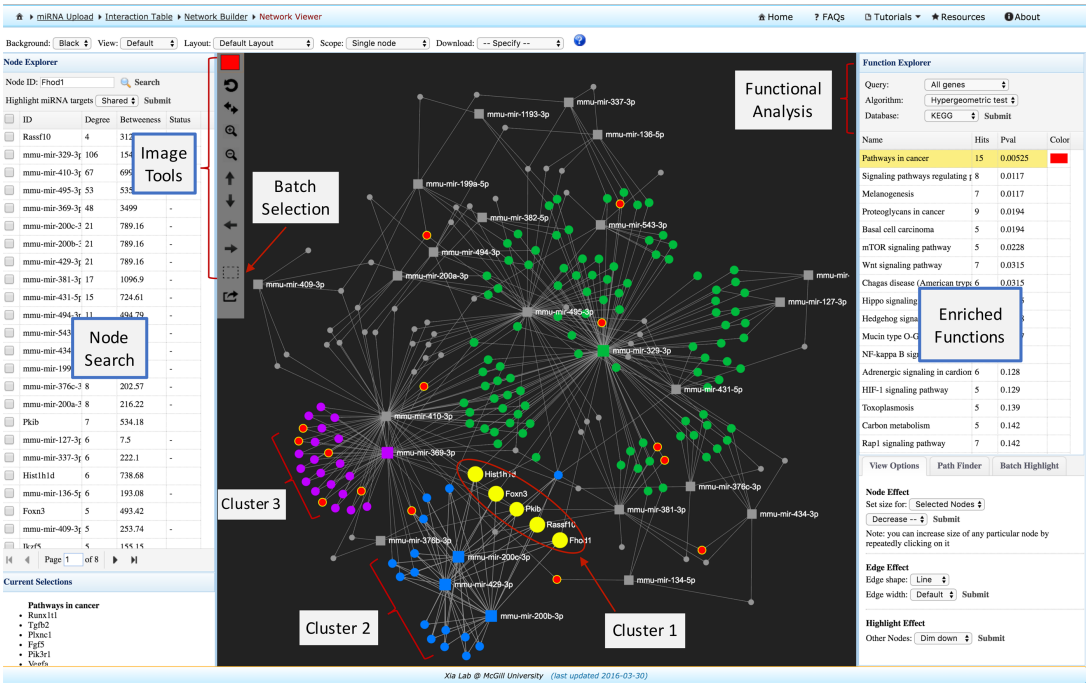


Figure 1.2: An overview of the user interface through an example session of miRNet.

1.4 Outline

The present thesis presents detailed documentation of how Odyssey was developed, what the motivation was behind the research produced Odyssey and how it can be used for generating microRNA - gene networks using exemplary datasets. There are five main chapters comprising the whole document excluding the abstract provided in the first pages and references provided at the end of the document.

A short introduction section is spared for explaining the Motivation, the Scope and Goal (aims under Odyssey development in general) and a comparison of Odyssey with existing similar tools by a brief discussion on its modules. Larger Background chapter is where a literature overview of microRNA related research, explanation of biological networks are given and software that are currently present and open for user experience are briefly debated. Graph layout types are explained in detail, the available type of microRNA - Target Interaction databases and how they are created are explained. Gene Ontology (GO) Terms is a good way to summarize gene functions and also valuable to interpret

the networks in general which is described as a subsection of background part. Finally, the Background section is concluded with describing R programming environment in general and packages to analyze microarray data with and details of the Shiny package in which Odyssey developed in.

In method section, user interface structure of Odyssey is discussed separately for the sidebar and main panels. How Odyssey works is described in detail, together with the underlying reasoning for adding modules like filtering the nodes by expression value difference and expanding the network with second-degree interactions etc. The Methods section is concluded by explaining the data structure processed at the back-end of Odyssey, performance details under why specific data structures are chosen store databases in.

In the case study section, analysis of example data and outcomes of these analyses are explained in detail. The document is finalized with the Conclusion section where results, expectations and shortcomings of Odyssey are debated mainly focusing on the analysis of example data. Future prospects and modules that are under development are explained in the last paragraph of the Conclusion section.

CHAPTER 2

BACKGROUND

2.1 Small inhibitory agents of gene expression: MicroRNA

MicroRNAs(miRNAs) are 22 nucleotides (nt) long, small non-coding RNAs that regulate various cellular and metabolic pathways. Although they were first discovered in Caenorhabditis elegans [2],miRNAs regulate gene expression levels of various other species including Homo sapiens [135].

microRNAs are first transcribed as pri-miRNAs in the nucleus. The Drosha enzyme cleaves pri-miRNA to precursor-miRNA (pre-miRNA) with a length of 70 nucleotides. After being exported from the nucleus by Exportin 5 in the pre-miRNA form, the Dicer enzyme cleaves the pre-microRNAs into a miRNA duplex. Thus, two successive cleavages of the pri-miRNAs through RNAase-III family enzymes Drosha and Dicer play very important roles in the maturation of miRNA molecules [102]. Dicer interacts and couples with double-stranded RNA binding protein TRBP to form a Dicer/TRBP complex which takes part in the next step of maturation of the miRNA duplex. The mature miRNA duplex unwinds and incorporates into RNA-induced silencing complex (RISC). RISC blocks mRNA function by either cleavage or degradation of mRNA through near-perfect complementary binding to 3'UTR region of the mRNAs [146] [126]. Argonaute2 (Ago2, a core component of RISC in human cells) is an important element in microRNA mediated mRNA silencing. Ago2 is responsible for concentrating mRNAs in processing bodies (called P-bodies) where mRNA silencing occurs in animal cells [32].

Many important cellular processes have been documented to be mediated

by microRNA-directed mRNA silencing. Although cell differentiation [94] and developmental pathways [67, 121] are two of the cellular pathways microRNAs function initially have been documented in, pathways regulated by miRNAs include hematopoiesis [5] and also human pathologies like cancer [38].

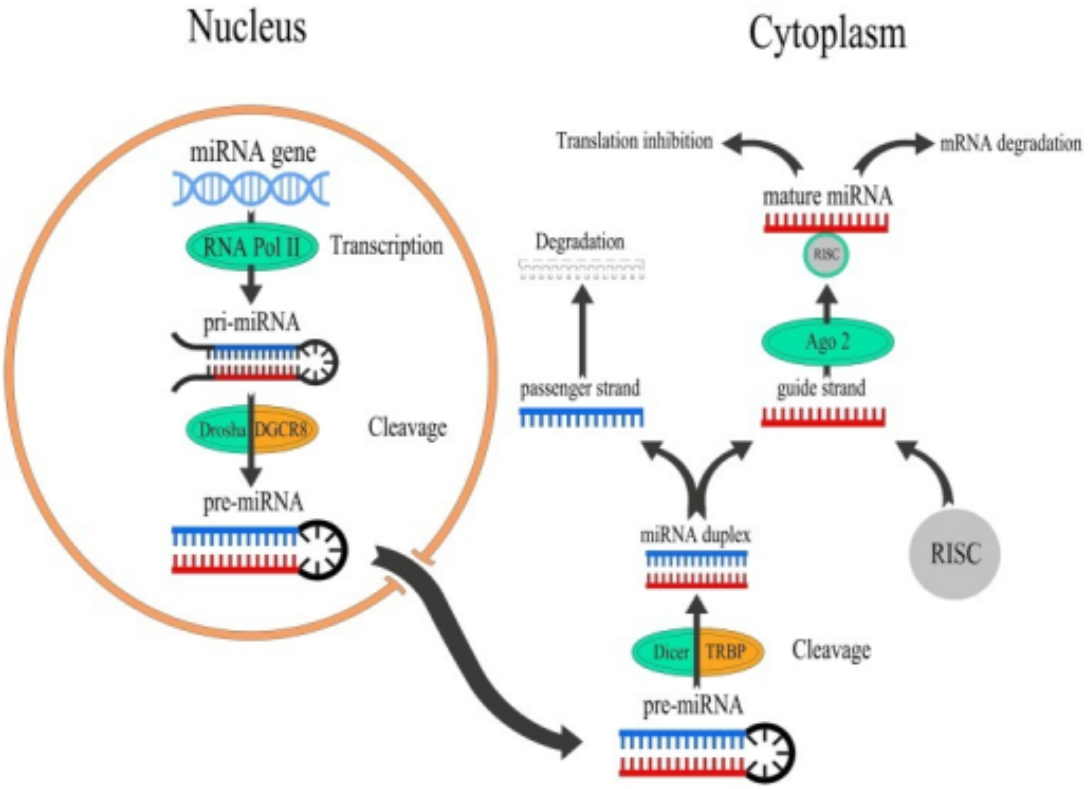


Figure 2.1: Illustration of how microRNAs function within cell.
[Alsharafi et al. [3].]

2.2 Biological Networks

Different sub-disciplines of biological science are represented by different network models because the number of molecules that affect the biological processes is vast (DNA, RNA, protein, drugs, and other chemicals etc.). Nevertheless, these networks are still poorly understood at the systems biology level, although the scientific community putting continuous effort to gain knowledge on these sub-disciplines [105] [68] [69]. In order to infer biological knowledge, different types of biological networks are generated. While generating these networks, researchers aim to capture an assembled summary of the underlying biology from a schematic representation. For that purpose, categories of dif-

ferent biological network types and software that generate them have emerged (See Introduction). Importantly, networks drawn with this purpose are captions that represent the human understanding of underlying biology which in reality cannot be considered separately.

Protein-protein interaction networks, gene co-expression networks, metabolic networks, phylogenetic trees and gene regulatory networks (GRNs) are examples to different categories of biological networks. Gene co-expression networks are generated to be able to capture similar expression patterns of different genes based on stimuli of interest, e.g. drug administration [118].

Metabolic networks are generated to demonstrate chemical processes driven by biological molecules (e.g. ATP) as nodes. Edges in this type of network represent the chemical reactions e.g. dephosphorylation [98].

Protein-protein interaction networks represent interactions between proteins, placed as nodes in the network. Edges in this type of network represent the type of interaction between the nodes they connect, e.g. activation, deactivation [11].

Phylogenetic trees are another type of biological network usually generated to reconstruct the evolutionary history of taxa [125]. Proteins, genes and other similar biological molecular that are shaped by the evolutionary process can be plotted into this type of network as nodes. Edges represent the similarity or dissimilarity between nodes using unit measures derived through mathematical formulation to represent the strength of the evolutionary forces acted upon the nodes connected by it. Timeline in which effect of these forces is being calculated is measured with distant time intervals generally scaling up to millions of years [55].

Gene regulatory networks, apart from co-expression networks, are generated to visualize direct interactions between genes (transcription factor (TF) activation of particular genes) that results in expression difference in the corresponding targeted node [63]. Other internal cellular stimuli (not only TFs) also cause regulation in the gene expression levels, e.g. microRNAs. In the case of gene regulatory networks driven by microRNAs, interaction type is uni-directed microRNAs target genes to deregulate genes that results in a decrease of gene expression levels, usually detected by high-throughput expression profiling as-

says [112] [45] [4]. MicroRNAs are represented in the GRNs as source nodes that an edge emerges directed towards the target nodes that are genes. In this study, I have developed Odyssey to visualize gene regulatory networks integrated with transcriptomics, interactomics and functional data. MicroRNAs are represented in the GRNs as source nodes that an edge emerges directed towards the target nodes that are genes. Such networks are also called "Bipartite Networks" [8]. Nodes in bipartite networks are composed of two different sets of non-interacting entities, miRNAs and mRNAs for the case of Odyssey. Thus, vertices represent interactions between miRNAs to mRNAs and no interaction is observed within the same set of nodes e.g. mRNA to mRNA interaction.

Category	Content	#Resources	Examples
Protein–protein interactions	Mainly pairwise interactions between proteins	105	DIP, BIND, STRING, HPRD
Metabolic pathways	Biochemical reactions in metabolic pathways	60	KEGG, Reactome, ENZYME
Signaling pathways	Molecular interactions and chemical modifications in regulatory pathways	50	STKE, Reactome, TRANSPATH
Transcription factors/Gene regulatory networks	Transcription factors and the genes they regulate	42	JASPAR, TRANSFAC, RegulonDB
Pathway diagrams	Hyperlinked pathway images	30	KEGG, HPRD, SPAD
Protein–compound interactions	Interactions between proteins and compounds	24	ResNet, CLiBE
Protein sequence focused	Diverse pathway information in relation with sequence data	16	REBASE
Genetic interaction networks	Genetic interactions, such as epistasis	6	BIND, BioGRID

Figure 2.2: Software for Different Type Biological Network Generation. [Hecker et al. [49]]

2.3 Software for Biological Interactions Integration and Visual Analysis

Next Generation Sequencing (NGS) technologies have sped up the generation of biological data remarkably [84] [48]. In return, biological data analysis tools are being used more frequently to handle the challenge brought up by NGS technologies, called the "Big Data", according to the report of EMBL-EBI [22]. Rather than analyzing the biological interactions of isolated parts on the molecular level, system level approaches provide valuable and realistic ways to explore cellular biology empowered with high throughput data based on molecular level understanding [47, 65, 111, 119, 143]. Using networks integrated with gene expression data is one such effective way of interpreting the biological data. Cytoscape [124], MiRTarVis [58], miRNet [36] are a few examples of software created to address this issue that generates biological networks, the former being a more general application with potential for integration of plug-ins.

Cytoscape, developed first among these software, is the most cited. With the help of presence of larger amounts of biological data, there has been also a continuous increase in biological databases that house protein and/or chemical interactions over the last two decades [72]. Studying biological phenomena systematically at systems biology level thus has attained greater importance [23]. Researchers continue to develop new software to decipher biological data into knowledge or upgrade frequently used software such as Cytoscape with plug-ins like BiNGO [81]. These software and their comparative relation to this study is explained in the Introduction section.

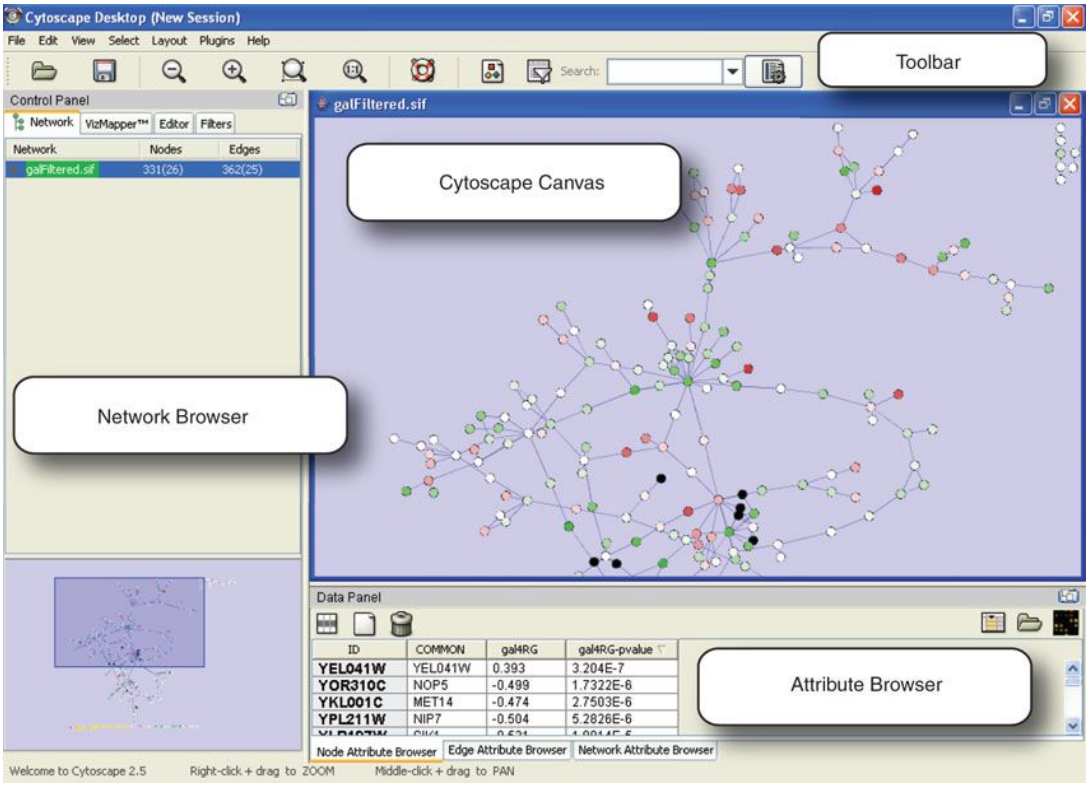


Figure 2.3: Cytoscape Desktop User Interface used for Biological Network Generation.

2.4 Graph Layout Algorithm Types

Types of different graph layout algorithms yield layouts with varying quality in terms of network visualization and performance establishing the node and edge positions with speed. All layout algorithms aim to have a good visual representation of the graph although 'good' is a relative term which may depend on different parameters amongst users. Graph drawing problem can be summarized as the placement of 'n' number of nodes and 'e' number of edges to a surface (generally 2-Dimensional) of 'L' length and 'W' width. Cytoscape js a package is implemented into R language as rcytoscapejs which allows several different types of graph layout algorithms.

2.4.1 Force-directed Graph Layouts

Force-directed graph algorithm, a solution to the problem of placing components on a printed circuit board, has originally been put forward by Quinn et al. (1977) [108]. After their claim of having successfully tackled the issue, the mathematical formulation they came up with has been studied and developed to solve problems from different disciplines including biological sciences. Some force-directed layout algorithms [56] visualize the problem of network generation as a mechanical system where nodes are replaced by constant weights and edges are springs keeping these weights in balance. Nodes generate a repulsive force acting particles trying to get away from the system like through gravitational-like forces whereas edges are attractive forces keeping nodes in place. A successful force-directed layout algorithm tries to have a network where the system is in the state of minimal energy expenditure. Thus, distances between the edges are being adjusted accordingly in this type of layout algorithm [40].
Compound Spring Embedder (CoSE) is a force-directed automated layout algorithm developed to generate the graph in a time efficient manner. The CoSE algorithm can handle non-uniform node sizes, multi-layered nests within the graphs and orphan nodes well [27], and is implemented to be used in R environment. Graph drawn with the CoSE algorithm acts like a singular object located in the dimension where it is drawn because of the balance formed through optimization of the energy within the system.

$$f_a(d) = d^2/k$$
$$f_r(d) = -k^2/d$$

Figure 2.4: Mathematical formulation of energy state calculation of the system in force-directed graph layouts.

The energy of the attractive and repulsive forces are calculated based on the distance between the edges.

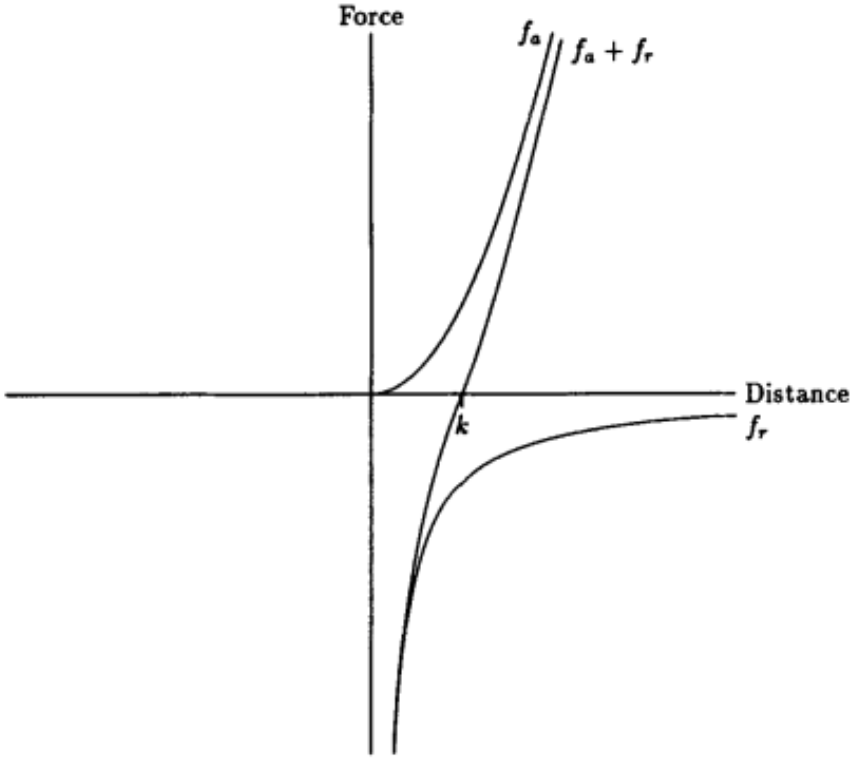


Figure 2. Forces versus distance

Figure 2.5: Figure representation of repulsive and attractive forces strength in force-directed graph layouts.

Point of 'k' represent the distance between the edges where energy is optimised to be theoretically at minimum - Copyright Clearance Centers RightsLink License Number: 4381900146730.

2.4.2 Grid Graph Layout

After force-directed layout approach to the graph drawing problem, other algorithms have been applied to biological networks to decipher the topological structure of numerous biological interactions represented in a graph. In the biological systems, nodes forming the graph are not equal in terms of cellular function. Edges (in other words vertices) connecting these nodes can have various meanings depicting biological processes like activation, repression, complex formation, modification etc. Repulsive forces in between nodes in the force-directed layout algorithms inadequately describe the chemistry between biologi-

cal molecules in most of the biological graph types; because as mentioned, some nodes need to be clumped together forming biologically meaningful clusters [76]. Grid layout algorithms are generally designed to start the drawing the graph by placing the nodes at random places on the graph plane. Cost is calculated between the pairs of nodes in the graph and energy in the system is minimized through a locally minimizing algorithm. The next step is perturbation of the graph and re-plotting as a whole before rerunning the algorithm. Work-flow of the grid layout algorithm is designed to have a more stable output layout as the initial step is random node placement. Other methods to increase stability to the issue of randomization at the beginning of grid layout drawing has been offered in other articles, i.e. introducing node swapping in between locally minimizing the graph [66].

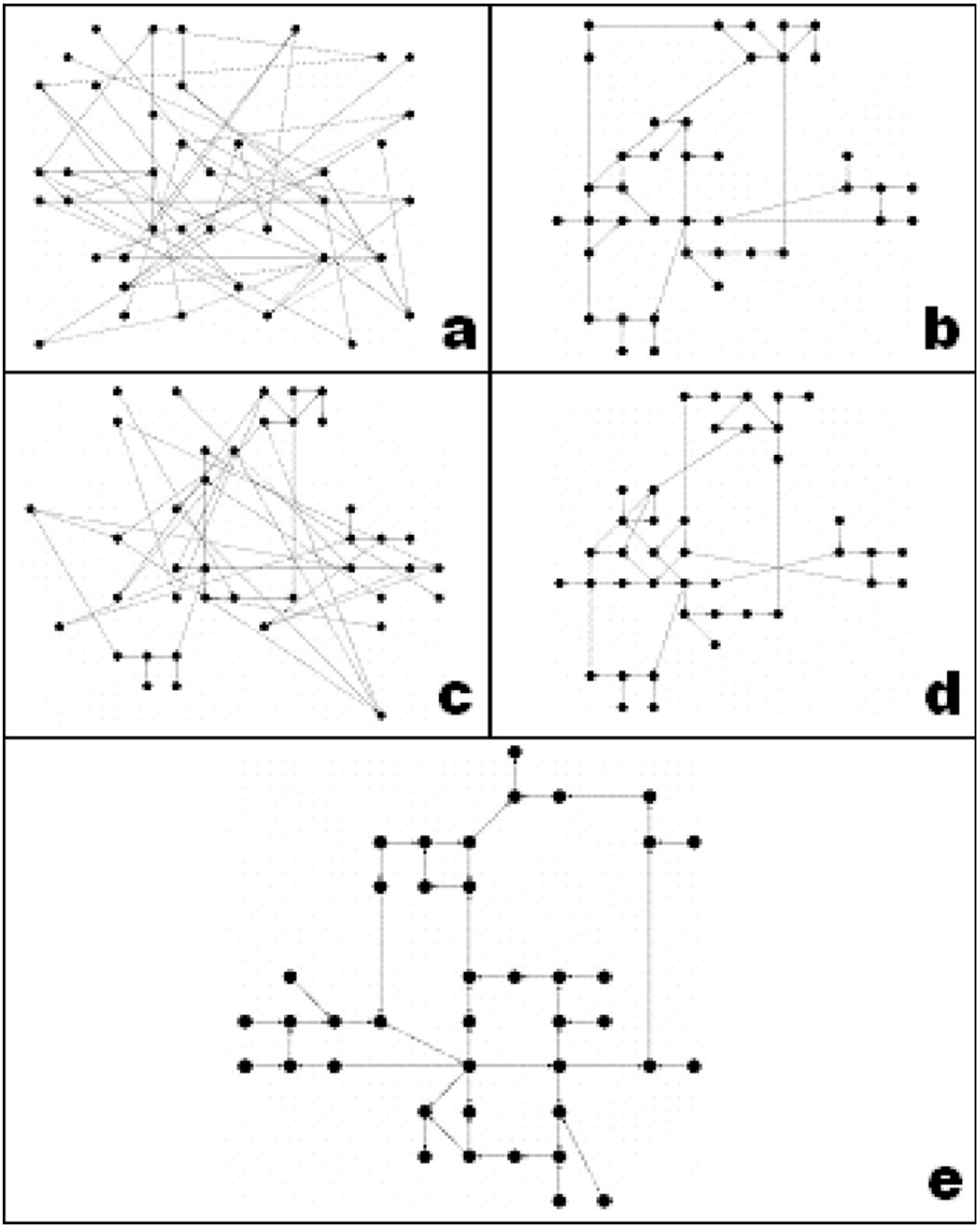


Figure 2.6: A process flow of grid layout of the E.coli heat shock response network.

(a) An initial random layout; (b) a candidate layout obtained through local minimizing the random layout a; (c) a perturbed layout of b; (d) a candidate layout obtained through local minimizing the perturbed layout c; and (e) an output layout. [Li et al. [76]]

2.4.3 Circular Graph Layout

Some biochemical processes form a series of biological molecules that interact with each other in an ordered manner. The Tricarboxylic Acid (TCA) Cycle is an example of the circular graph formed through the evolution of millions of years [82]. To be able to plot the TCA cycle and other similar networks, circular graph layout algorithms have been proposed. These algorithms keep the interconnectedness of the nodes intact and easily visible within the graph and quick to form required clusters is there is such a formation by graph elements [9] [31]. Biological processes are in fact inseparable from another in the sense that they are each, a part of what creates a living organism. However, some series of biological processes that share closely related molecules can be combined to visualize a comprehensive view of the connection existing between their elements. Circular graph layout help researchers to provide such an overview of underlying biology closer to authentic (theoretically or proven in molecular level) state of interactions. All graphs can only represent the focused area of interactions for most complex organisms. Recent publications have reports of being close to having mapped all protein-protein interactions of some living organisms like yeasts [106] [46] however, there has been constant upgrades on these reports over the years.

Figure 2.7: Combination of TCA cycle, Glycolysis and Urea cycle.

2.4.4 Concentric (Centrifugal) Graph Layout

Nodes in the graph can be placed in an infinite number of ways but only some of the placements grant a visual aid in the understanding of the network drawn. One of the popular ways of node placement is called concentric layout. In this type of layout, the nodes are placed in concentric circles [61]. Densely connected nodes are placed towards more central locations of the graph and scarcely connected nodes i.e. nodes connected with one vertice only are placed at the outer circles. Therefore, generation of a weighted type of graph according to the interconnectedness of the nodes is accomplished.

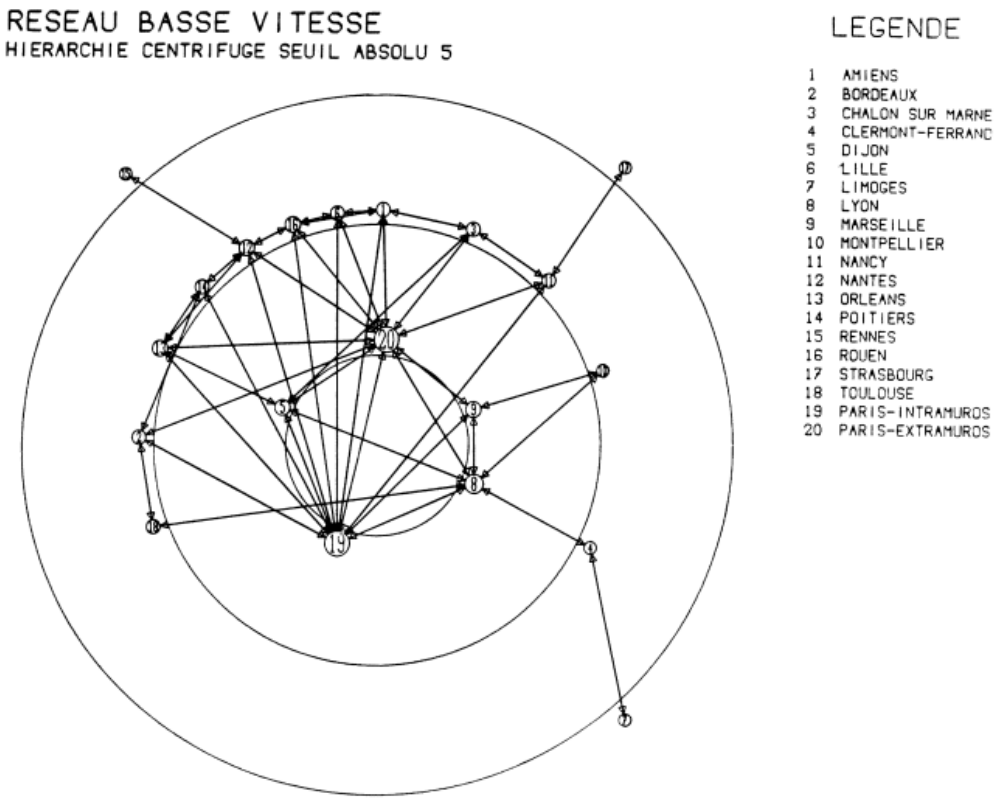


Figure 2.8: French data transmission system displayed in concentric (centrifugal) graph layout format.

[Carpano et al. [16]]

2.4.5 Breadth-first Graph Layout

Graphs can be analyzed using a search algorithm based on the number of vertices connecting the elements of the network. This is particularly important

for large graphs because these algorithms are used in the identification of clusters or sudden changes in the network which otherwise could not be detected due to the massiveness of the network [7] [21]. Breadth-first search algorithms developed to search through graphs depends on the memory of the system rather than CPU power because search relies on finding the shortest paths, connected components in the network. Research has been performed to increase the efficiency of these kinds of algorithm [147].

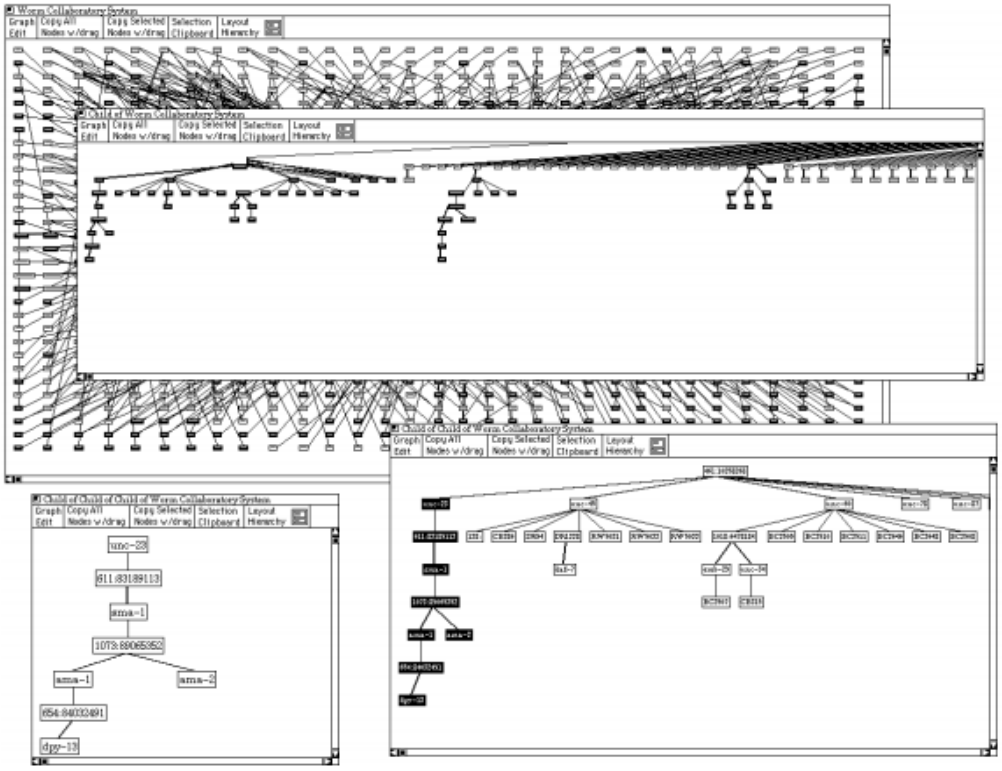


Figure 2.9: Multiple hierarchical view of a large graph.

Search algorithms are used to extract the nodes of known hierarchical status. [Henry et al. [50]]

2.5 MicroRNA - Target Interaction (MTI) Databases

Large number of interaction databases have emerged in the past years corresponding to the need of researchers to search among numerously documented microRNAs each of which has multiple targets [73, 74, 95, 113, 123, 139]. There are two types of interaction databases classified according to how MTI are doc-

umented. While some databases use only computational approaches to predict the interactions as in the case of TargetScan [1], other databases prefer to have only experimentally validated MTI, miRTarBase [51] or combination of both experimentally validated and computationally predicted MTI, miRNet [36].

2.5.1 Predicted Interactions

Computational approaches predict hundreds of interactions based on seed pairing of microRNAs with their targets, as seed pairing is a key factor of microRNA regulated gene expression inhibition [13, 14]. Contradictory research that questions the potency of using seed pairing for prediction of MTI also has been published [25]. Algorithms created for MTI prediction use diverse methodologies including Pattern-Based prediction methods [90] and Hidden Markov Model based prediction methods [60]. However, most of them take advantage of the evolutionary conservation of target sites for predictions [109].

2.5.2 Validated Interactions

Experimental methods are a less error-prone method of interaction validation which is also used to test the prediction strength of computational approaches [44]. Different methods to validate the authenticity of computationally predicted or manually curated microRNA - Gene interactions have been established in the literature [62, 78, 132, 133].

2.6 Gene Ontology (GO) Terms

Understanding the biology on the cellular level requires classification of its fundamental parts into comprehensible categories. Gene Ontology consortium [6] aims to classify genes using established vocabulary to clarify their functions. Whilst establishing a biological network with multiple nodes and edges, GO terms can be used to add new nodes to (expand) the network, i.e. adding new genes that share common GO terms with existing genes, or filter the nodes out

for visualizing the network on a specified level.

2.7 Gene and MicroRNA Nomenclature

After the initial discovery of the microRNAs and realization of their biological significance, identifying microRNAs with a robust scheme became necessary. miRBase [43] [104] is a source for such nomenclature that is widely accepted. Designated scheme eases the burden of having a common microRNA language to be used in peer reviews, flexible to changes and updates that come with novel research and can be used across different species regarding the homology in between.

Gene nomenclature, as in microRNA nomenclature but raised as an issue long before microRNA discovery, was essential to be established as a scheme for the same reasons why microRNA nomenclature was needed. Having a designated gene naming system for researchers from different backgrounds to use collaboratively indeed speeds up the pace of research. HUGO Gene Nomenclature Committee (HGNC) [103] have provided a widely accepted scheme which enables assigning unique names to genes to be discovered.

2.8 R Programming and Shiny Package

The R programming language is derived from the S language which was used by statisticians and data analysts frequently in the past. Similarly, R is an effective tool for data processing and analysis and in the field of Bioinformatics, it is used frequently [42] [85] [99]. The community of R users expresses the need to accomplish a task, bioinformatics related or otherwise, in the form of R packages. Bioconductor [53] is an open source software project for developers to contribute package development for R programming. At the time of writing this thesis, 934 packages can be found in the Bioconductor libraries to be used freely.

The Shiny package, although not a Bioconductor based package, has grown amongst users to create a user-friendly interface for bioinformaticians to deliver

their scripts to biological research community without having the need to learn or write R scripts on their own [20]. R community has expanded their efforts to correspond to the need of having a UI for using the bioinformatics tools developed for the rest of the community [10] [86] [70] [28]. However, Shiny has not been used yet to develop a miRNA-mRNA interaction/expression network tool.

2.9 Microarray Data Processing and Analysis

Microarray platforms provide a high-throughput way of measuring gene expression levels although next-generation sequencing technologies have been predicted to surpass the need for microarrays in the future [54]. As there are multiple platforms offered by different microarray provider companies, having a standard post-processing methodology of the microarray data and revealing the biological significance with the technical noise of the platform filtered out [97]. Minimum information about a microarray experiment (MIAME) [12] created an opportunity for standardized microarray data. Researchers can use different platforms [144] by themselves and/or take public microarray datasets to merge them into bigger datasets of different platform compositions [114] for deriving larger biological meaning out, by paying a lesser cost.

R programming, supported by a large research community, has numerous packages offered to users for normalization, background correction and differential expression analysis of microarray data [41] [122] [115].

CHAPTER 3

METHODOLOGY

3.1 ODYSSEY

Odyssey - a web application based on the Shiny package of R, is developed to create a network representation of MTI. The name "Odyssey" comes from the classic Greek epic poems attributed to Homer. Poems accommodate numerous characters and complex social interactions between these characters that are visualized as a network on literary research for reaching a better understanding of the epic poem [91].

CSS and HTML components were added to the design of the UI for creating a friendly, easy to use UI and to avoid a steep learning curve. Resulting networks can be enriched for specific Gene Ontology terms and filtered out based on expression data uploaded prior to network generation. For easy visualization and transfer between peers, functionality to download the generated network in the HTML format has been added as a feature. Odyssey is capable of analyzing mRNA microarray data of GPL17692, GPL570, GPL96, GPL97 platforms and miRNA microarray data of GPL21572, GPL8786 platforms from Affymetrix. In the future versions, the capability to analyze data from Agilent microarray platforms and RNA-seq data will be added.

To establish a user-friendly interface, modularity of the application has been regarded as essential. Buttons are colored with selected black color whereas slider bars are colored blue for the interval selected. Help tips are placed at the bottom of each module to aid users what to expect by changing the settings from default. Some functionality like choosing between different graph layouts

are dynamically created as the user chooses between different layouts aimed to give a smoother user experience.

Odyssey, apart from listed other tools, can build microRNA - gene/mRNA networks to second-degree and thus aims to generate a more comprehensive network of the underlying biology. The process of expanding the network is supplied with one check-box which is checked as default for user convenience. Besides this functionality, GO term based network expansion is also a novelty comes with Odyssey. miRNet and miRTarVis generate reports for GO term enrichment while Odyssey has the capability to provide a network expansion based on the GO terms enriched in the network. This enrichment is guided by the userś selection of GO terms found already enriched in the network in varying degrees. Network expansion with GO terms that are not enriched in the network (de novo GO term expansion) is not currently implemented in Odyssey however, it is noted as a future goal of this research.

Initially filtering of the nodes has been performed using a default (-1,1) expression value interval for both microRNA and gene nodes. Since this results in less flexibility, I have updated this filtering feature such that newer versions perform filtering by automatically calculated 5% and 95% quantile values based on the expression value distributions. This change in the filtering process is committed to having better automated graph generation, thus aiming to further increase the quality of user experience. Apart from other tools, Odyssey determines filtering criteria automatically, calculated specifically for each data.

Future upgrades of the software aim to generate graphs with non-uniform node sizes. Node sizes are judged to be better adjusted with regards to the degree of node interconnectivity to grant visual aid for better understanding the knowledge present in the graph.

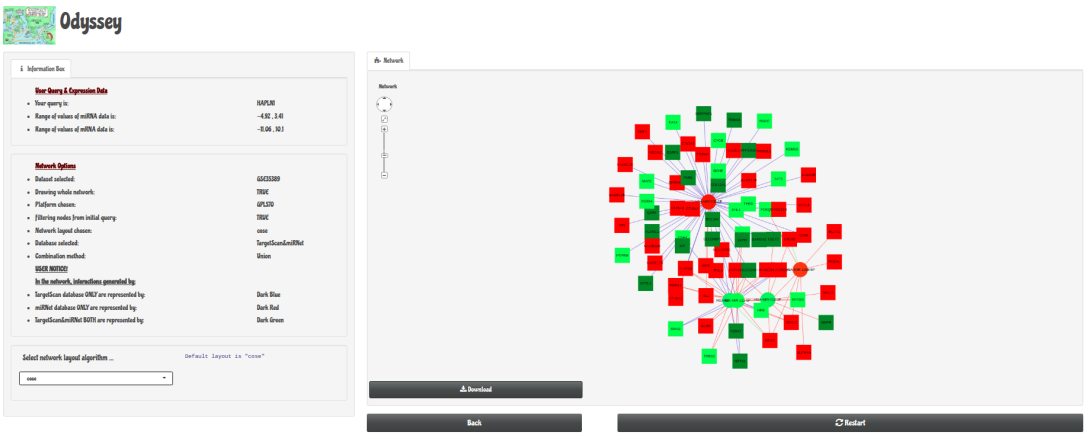


Figure 3.1: An overview of network drawn using Odyssey.
GSE35389 data and HAPLN1 gene have been used to generate the graphics.

3.2 Designing the User Interface (UI)

UI of Odyssey is designed to modularize the functionality provided. Whenever the user proceeds to a module from another module in the application, parts of UI that are relevant to the previous module should be replaced with related parts of UI. For example, replacing *Cytoscape tab* in the main-panel with Network tab in the same panel when the user visualizes the Cytoscape file and presses *'Generate Network'* button is an example of the desired behaviour Odyssey executes. This is achieved through mentioned JavaScript libraries implemented in R.

Warning texts and explanatory texts are displayed with red, small font size lettering directly below the interactive parts of UI they are related to, i.e. *"Check the box to extend the network to second degree interactions"* is displayed below the *"Show comprehensive network"* check-box.

Home button added onto the logo that is displayed at the very top left of the page. When clicked, a refreshed page of Odyssey opens up in the web environment, discarding the selections that were made in the previous session and allowing a new session of Odyssey. At the bottom of the *Network tab*, a *refresh button* is placed to provide a similar functionality to *Home button* mentioned. One difference between "Home Button" and "Refresh Button" is, the user might need to re-authenticate herself after the "Home Button" is clicked whereas "Refresh Button" clears the selections of the current session while keeping the

session intact.

Odyssey UI is divided into *Sidebar Panel* and *Main Panel* to modularize the application into two main parts. *Sidebar Panel* is reserved for obtaining input whereas *Main Panel* displays the output of application depending on the input supplied in *Sidebar Panel*. The functionality of these two panels is explained in detail below.

3.3 Sidebar Panel

3.3.1 Authentication Tab

Authentication tab in the side panel of the interface is created to hold a Google Authentication button. Via Google authentication tokens, users who are white-listed by the server can access Odyssey to proceed with data upload or example data selection. By using a Google authentication interface, besides giving permission to only registered and approved users to access the app, the responsibility of securing the credentials of users - a password that would be required with the registration - and also the maintenance of the access account is transferred to Google. Google authentication provides a more reliable environment for users to sign in. Blocking of the rest of the application for users with no valid authentication token is implemented via JavaScript-based libraries [83] of R. Hide and disable functionalities that performed this task is supplied by another JavaScript-based package shinyjs [80] developed for shiny R.

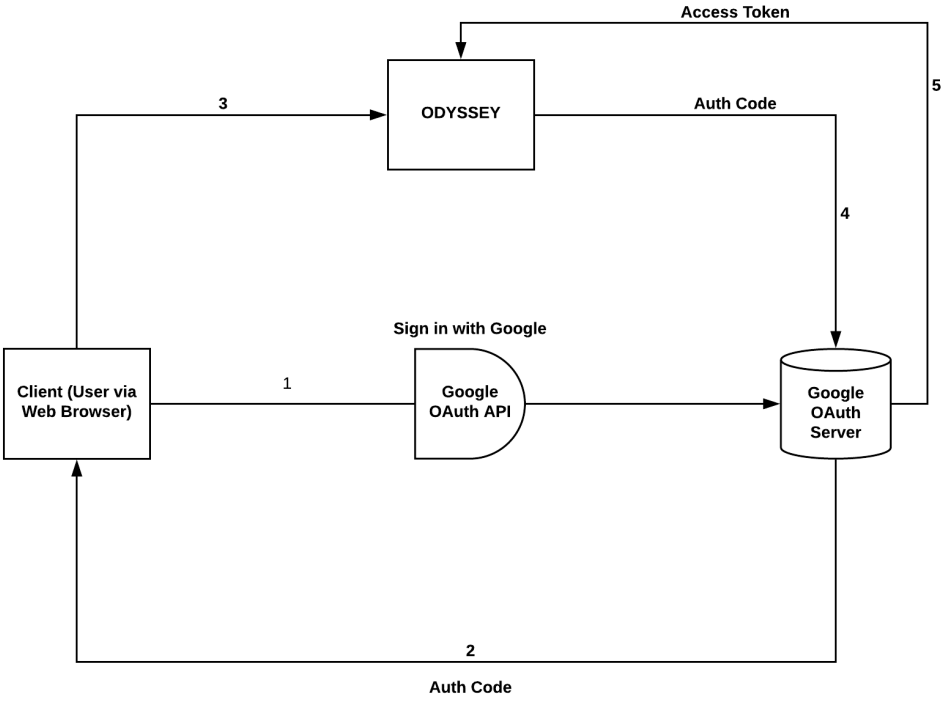


Figure 3.2: Schematic representing Google OAuth Authentication.

3.3.2 Data Selection Tab

Data Selection tab is composed of three user input dependent sections. Most fundamental part of the Odyssey is where the user enters a miRNA or a gene query in a text-box to be processed in that session of the web application. Query input via text-box is where the user is allowed to enter many possibilities without being restricted to the gene or microRNA related strings. Unrelated queries had to be processed with proper user guidance for a better user experience. For that purpose, this module is highly integrated with error handling functionalities to make up for countless text that can be input. Apart from unrelated text, some genes or microRNAs are not found within miRNet, TargetScan or other databases or within the expression data to be used. Moreover, miRNAs and genes have many different identifications (aliases as well as versions/sense/antisense) i.e. HSA-MIR-29C molecule is separated into two different identifiers; HSA-MIR-29C-3P and HSA-MIR-29C-5P. When textbox is input with HSA-

MIR-29C, Odyssey is programmed to return these two possible choices to select from, with a proper warning that query had multiple matches with the integrated databases. Some gene queries might also return multiple matches. To be able to explain that issue, 'PTX' gene query and how Odyssey handles such a query is explained below. 'PTX' query does not have an exact match within the databases, however, PTX is sub-strings of PTX3, PTX4, and APTX which all are present in the databases to be processed in a session. In that scenario, Odyssey creates a temporary user interface based on the initial query to be displayed directly below, with a multiple matches warning and expects the user to select the most appropriate choice from the drop-down menu created for handling the issue or change the initial query altogether.

Second part is where the user will select the MTI database. *TargetScan*, *miRNet*, or *TargetScan& miRNet* options are given in a drop-down menu format. If the user selects TargetScan& miRNet, another drop-down menu will appear asking the user if intersection or union (Intersect, Union as they appear in the drop-down menu) of both databases should be taken. This will most likely to change the number of targets a miRNA has in the final network. The version of miRNet updated in 11/16/2017 and TargetScan 7.2 Release on March 2018 are currently implemented with Odyssey at the time of writing this document.

The third user dependent input section is designed to provide users two options to select for expression value representations that to be reflected on a final created network. Users can either select to upload their own expression data or use one of the two example dataset already have been loaded and ready to use. This is arranged with a check-box placed within the tab. If the check-box for *'Use Example Data'* is clicked, a drop-down menu will be revealed giving users a selection of example dataset implemented with Odyssey. If this check-box remains unclicked, a data upload panel will remain visible.

In the Data Selection tab, there are two separate upload buttons. One of the upload buttons is for mRNA data while the other is reserved for miRNA data upload. Uploading the data section is not mandatory. Choosing not to upload gives users the option to draw networks without expression data. In the case of not using expression data for the network, filtering with expression data functionality of Odyssey will become unavailable. Uploaded data need to be

in the format of series matrix file downloaded from Gene Expression Omnibus (GEO) [34]. Series matrix file format found in GEO consists of pre-processed microarray experiments which generally does not require more processing. By using this structure of data upload, I aimed to have a standardized flow of network generation which is essential for obtaining consistent results. However, uploading a custom pre-processed expression dataset is still permitted with this system. In custom data uploads, tab, comma or semicolon delimited files are accepted. After the upload of the data completed, Odyssey checks the format of the data for handling errors that might be caused from incorrect or unsupported data uploads. First measure is reading the uploaded file with read.table() function in R wrapped in try() error handling functionality. If the data format cannot be read through this system, user will be notified with an error message. Series matrix file format is read with these functions with an additional parameter to designate lines that start with symbol of '!' as the comment lines which leaves only the expression data to further analyze with Odyssey. Second measure is checking whether the data is log2 transformed. Dynamic range of the microarray data is verified with max() function within R applied on the data.frame. Maximum expression value of +20 is chosen for an upper limit for any log2 transformed data point. This selected upper limit is safely within the boundaries of range of expression values reported in literature [77] [107]. Lastly, the example data integrated in Odyssey can be downloaded via appointed *'Download'* buttons in the *Data Visualization Tab* for learning more about the format of expression data that is allowed in the system.

3.3.2.1 Example Data Collection

I integrated Odyssey with two example datasets to select from. These data aimed to be used for test runs of the application. Both of the datasets are downloaded from publicly available GEO with series id GSE35389 [138] and GSE88721 [24].

Data with id GSE35389 has 15 samples in GEO, 7 miRNA samples from the GPL8786 platform, 8 mRNA samples from GPL570 platform Affymetrix microarrays. Data with id GSE88721 has 30 samples in GEO, 15 miRNA samples

from the GPL21572 platform and 15 mRNA samples from GPL17692 platform Affymetrix microarrays.

In the process of example data integration, datasets with matching miRNA-mRNA samples were taken into consideration to be able to reflect the matching microarray data on the network. Thus, I aimed to lower technical variation and increase significant biological signal derivation out of the data integrated.

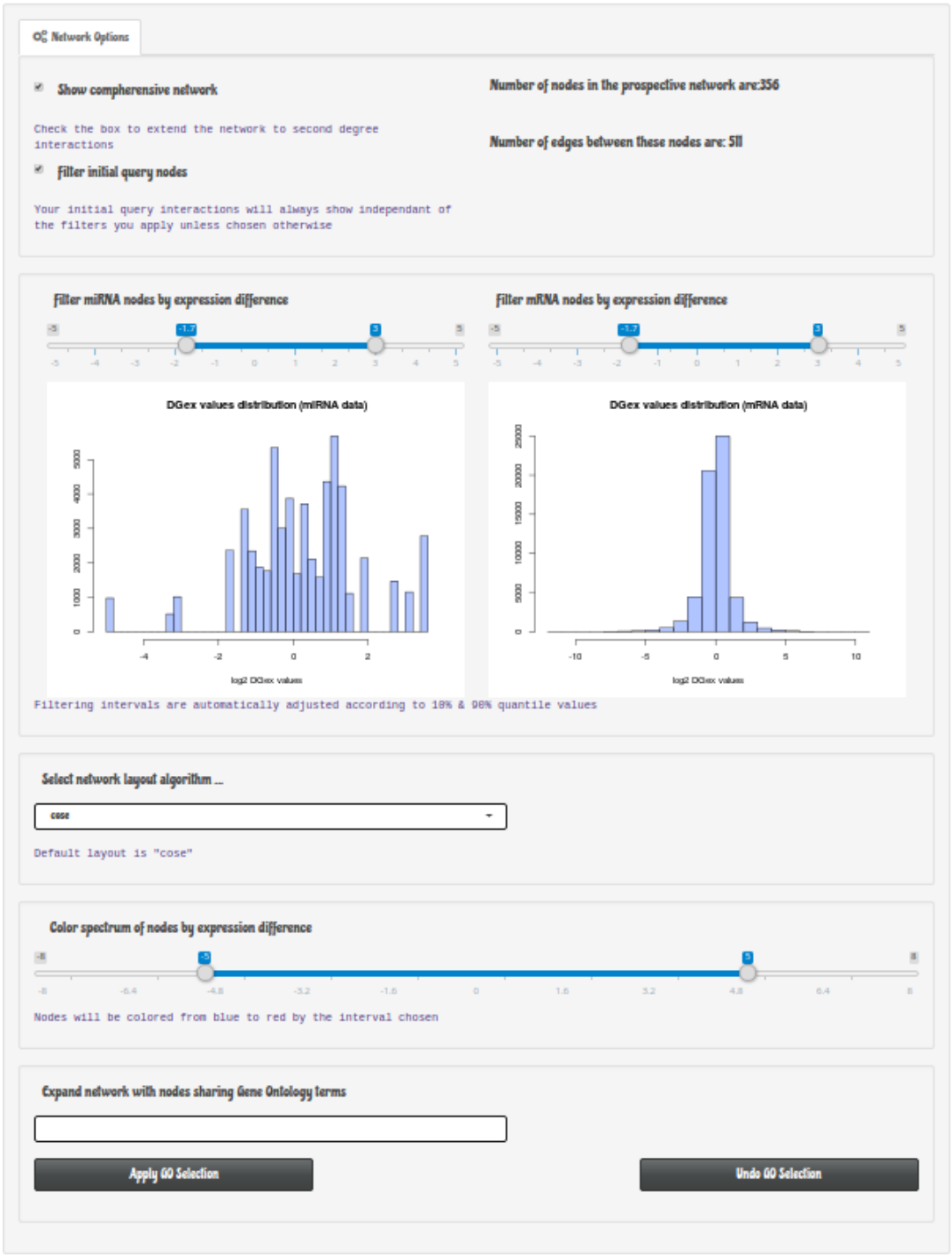


Figure 3.3: Network Options tab located in the Sidebar Panel.

3.3.3 Network Options Tab

After data selection part is completed by the user, access to *Cytoscape tab* and *Network Options tab* are provided. If any data are uploaded during the

previous step, filtering parameters of the network nodes based on expression values will be determined in this window. By default, Odyssey will omit the nodes with logarithmically transformed expression values that are beyond the automatically calculated 5% and 95% quantile values based on expression values lists, separately for both miRNA and mRNA nodes. Expression value distributions for miRNA and mRNA data are plotted using a histogram and displayed under the slider bars provided to users for filtering optionality (Figure 3.3.). Allowed filtering interval for expression values can be widened up to minus 5 and plus 5 (-5, 5). Widening the filtering interval will result in a more strict filtering and yielding networks that may have fewer nodes consequently (explained in detail in Filter nodes by expression difference section). The stringency of the filtering should be adjusted according to the density of the network which can be understood by observing the Cytoscape tab. (See Cytoscape Tab in the Main Panel section).

Using the 5% quantile and 95% quantile values generated networks of varying sizes depending on the query and the data selected for the analysis. Some of the networks drawn this way could still be over-crowded and consume more than available resources on the server. Moreover, this issue reduces the quality of user experience on Odyssey because underlying biology behind large networks is not easy to understand. For that purpose, a reactive network size calculation has been added on Network Options panel. Because most of the resources are spent on the visualization of the network, dynamic calculation of network size based on parameters that change network size, e.g. filtering and expanding the network to second-degree interactions, does not significantly reduce the performance of Odyssey. Through this reactive network size calculation, Odyssey aims to provide users with a network with controlled size and complexity prior to network generation without changing pages online, which is not provided by other similar software.

3.3.3.1 Display of the Comprehensive Network

Odyssey has the option of enlarging the network to the second-degree nodes from the initial query targets, which I call 'Comprehensive Network'. When

'Show Comprehensive Network' check-box is clicked on, resulting network will consequently be larger by including nodes of miRNAs and mRNAs that have second-degree relations with the node of interest and its first neighbours. First degree networks are useful for visualizing the interactions of the biological query, however, increasing the network density to the second-degree connections increases the degree of nodes. As the degree of nodes increases in a network, *expected force* of those nodes also increase, a term used for measuring the overall influence of a node within a network [71]. Odyssey aims to provide a more complete way of representation of the underlying biology through this module. Increasing the network degree depending on user input might be implemented in the following versions of the Odyssey, however, at this point, it is not included. The reason for that decision lies with the inclination of keeping relevant nodes and discarding information that might not be relevant to the initial query. Even after this option is selected, node compositions will depend on filtering options that are selected below at the same panel which again aiming to filter irrelevant biological elements from final network based on educated input from the user. Node degree and network degree are two similar terminologies used here that should not be confused for one another. Node degree stands for a number of connections a node makes with other nodes while in this thesis network degree is used for expansion of network from *initial query node* by addition of more interactions to the MTIdb declared connections of the initial query.

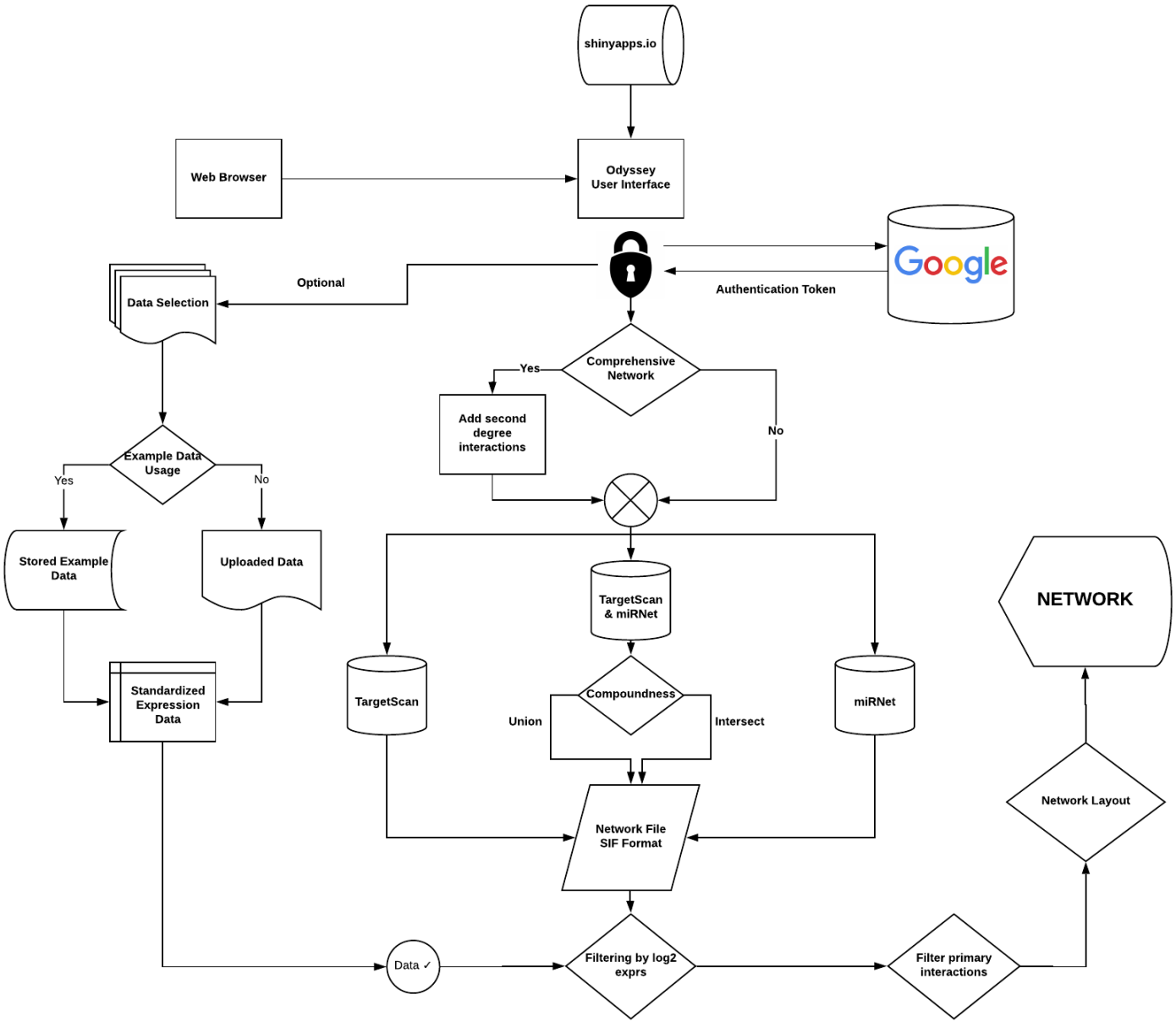


Figure 3.4: Odyssey Flowchart.

3.3.3.2 Filtering of initial query nodes

In of the previous development versions of Odyssey, the user has ended up with a network entirely unrelated to the query she/he initially searched for. This issue has been caused by expanding and then filtering the network in the process of creating it on the back-end of the application. In other words, adding nodes to the initial query and then filtering out the whole network based on expression data logFC selections could remove the initial query node and the singleton nodes remaining in the network after the initial query node has been removed. While this situation might result in an interesting network to be created, it could also be misleading or undesirable depending on the user's question of interest. To prevent that, an optionality to keep the initially created network out of the filtering process has been added. If the initial query is a microRNA, gene products targeted by this microRNA recorded in a parallel process and the newly added nodes are built upon on this parallel process. Resulting network does not comply with the user selected filtering options for expression value consequently for the initial network nodes that are kept regardless.

3.3.3.3 Filter nodes by expression difference

Odyssey has the option to reduce network size by removing microRNA or gene nodes by the expression difference (logFC) calculated from expression data. Whenever expression dataset allows the user to do any of these "tests vs. control", "healthy vs. diseased", "normal vs. cancer" sample comparisons, an expression difference based on the log2 expression value of the selected samples is calculated. Expression difference is calculated for each row in both miRNA and gene expression data and stored as two columns to be visualized in *Cytoscape tab* of Odyssey. One row of the file visualized in *Cytoscape tab* is in the following format: 'ID'(row number of entries)- 'miRNA' - 'gene' - 'mRNA expression difference' (abbreviated as Diff_ mRNA) - 'miRNA expression difference' (abbreviated as Diff_ miRNA). Column names are abbreviated to prevent data corruption because, in some packages of R programming language, multi-worded column names are not handled well. Filtering of the nodes by the expression

difference complies with the 'initial query nodes' issue explained in the section above. Filtering module is set to remove nodes with log2 expression value difference (EVd) within the boundaries. The filtering boundaries can be adjusted using the slider bar provided, within (-5, 5) interval.

In order to aid users to generate standardized networks and give general information about the expression data, differential expression gene distribution plots are added below slider bars responsible for filtering the nodes by expression value. Also, the filtering intervals are calculated data specifically to exclude miRNA and gene nodes with differential expression value 5% quantile - 95% quantile rounded to first decimal place. Default filtering interval can be changed by the user to cover a larger or smaller interval.

3.3.3.4 Choose Graph Layout Algorithm

Six different layout algorithms are offered by Cytoscape.js [39] implemented into R programming. As the default, networks are drawn with 'CoSE' layout. However, users have 'CoSE', 'grid', 'circle', 'concentric', breadthfirst' or 'random' algorithms to choose from. Depending on the algorithm selected, final networks may have rigidly different layouts. The final graph layout depends on the size of the network and interconnectedness of the nodes within the network as well. Odyssey reactively changes graph layout as user drifts between different layout options from the drop-down menu provided in the sidebar panel.

All of the five listed layout algorithms are explained in the Graph Layout Algorithm Types Section in the Background part of this work apart from random type layout. Random layout type is partially explained in Grid Graph Layout. In random layout type, nodes are randomly distributed to the places on the graph dimensional space and connecting vertices are added to finalize the output graph.

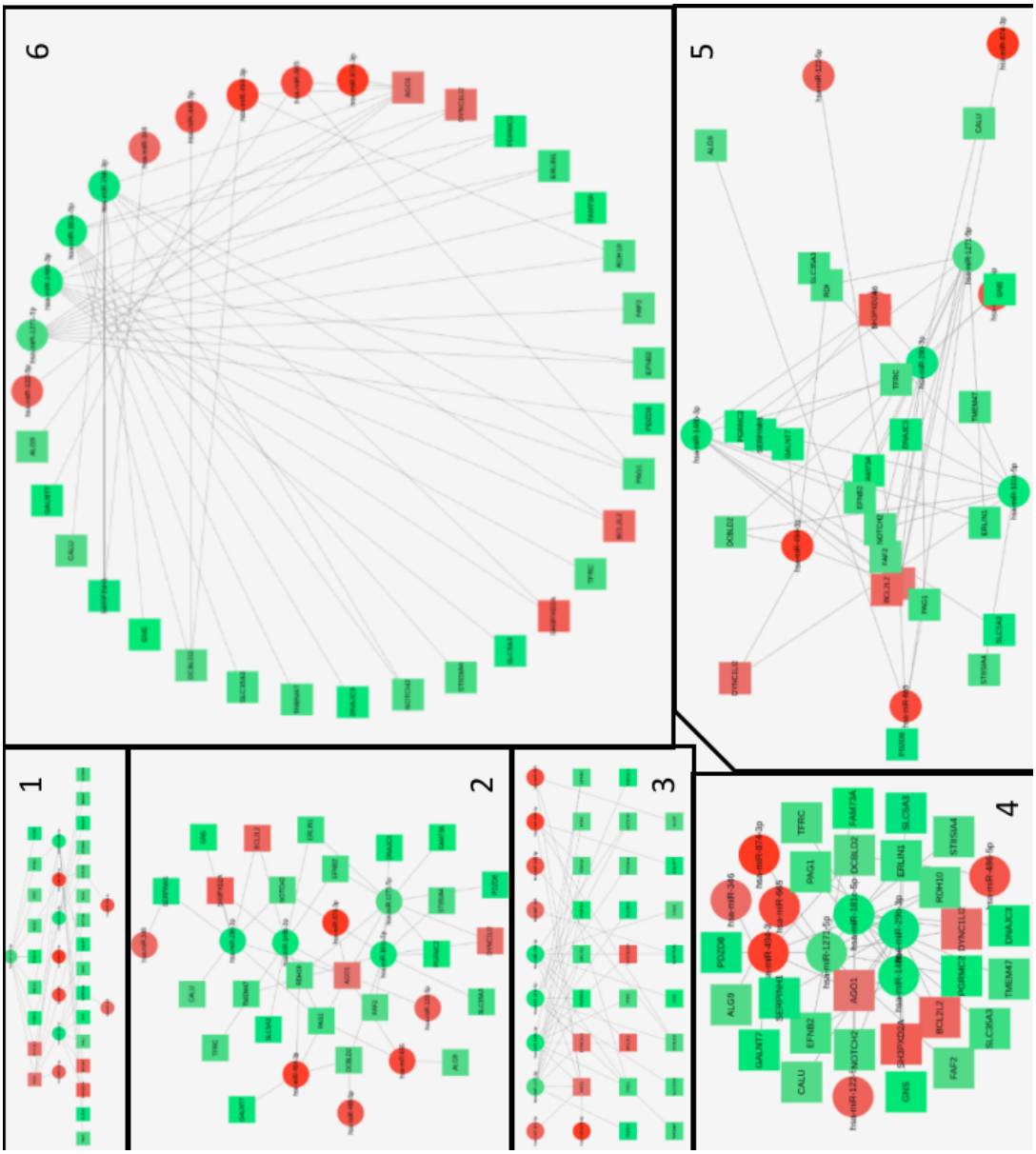


Figure 3.5: Graph Layout Algorithm Options provided by Odyssey.

1-) breadthfirst, 2-) cosy, 3-) grid, 4-) concentric, 5-) random, 6-) circle. In the above graphic, GSE35389 has been used with a query miRNA hsa-miR-29b-3p.

3.3.3.5 Color spectrum of nodes by expression difference

In this part of User Interface, a slider bar is provided to users to change the EVd interval for node colour adjustment. By default, nodes are coloured according to the interval (-5, 5). Nodes on the negative side of the interval are coloured green and nodes on the positive side of the interval coloured red (nodes with -5 EVd coloured green only, no red colour in the mix and nodes with +5 EVd coloured red only, no green colour in the mix). This optionality is provided for flexible network generation with datasets composed of skewed expression value distribution amongst samples. Positive skewness (samples with low expression value overall) might require using a narrower interval on the positive side while negative skewness might require using a narrower interval on the negative side of initially defined (-5, 5) interval. Adjusting the settings for datasets with expression value distribution skewness results in increased visibility for the nodes closer to mode EVd in the distribution. However, it is advised to note that keeping the default setting through multiple datasets yields standardized networks.

3.3.3.6 GO Term Enrichment and Graph Expansion

After the output graph is drawn, users are expected to visualize the network and note results for their session. Gene Ontology terms enriched for output network is also supplied in the sidebar panel determined by the gene-based nodes and Panther database [88] implemented within Odyssey using official gene symbols. Sometimes network nodes have significantly enriched Gene Ontology terms for important biological terms. In these cases, it can be beneficial to expand the network by adding more genes that share a common GO term with existing network genes. Through this module, I aimed to grant a better visual to the user for a GO term of interest, which can be a biological process or cellular component or molecular function.

Addition of new nodes to the current network is important and can be used to help the analysis for several reasons. New nodes that are not present before node expansion through GO term are either nodes that have not been filtered out during filtering step or the nodes that gene expression data supplied have

not contained. In either case, these newly added nodes will be coloured grey (or shade of grey if the filtering criteria for expression value difference were strict) for their lack of having a significant expression value difference. Therefore, these new nodes are distinguishable from the nodes that were present before graph expansion. The degree of expansion and biological meaning of newly added nodes by this process is an opportunity for deduction of biological signature present in the experiment by analyses executed through Odyssey.

3.3.4 Information Box Tab

Information Box is designed to provide a summary of the network options selected, altered from or remained as default settings. This tab also displays expression data information for the user to be able to handle EVd based filtering with ease, i.e. for a network with EVd (-3.4, 1.2) for miRNA nodes, using a filtering for the interval (-3.5, 1.5) will not produce a network. This is because the above-mentioned filtering interval will filter out all the miRNA nodes that might have connected to gene targets.

Apart from the EVd value interval for miRNA and gene nodes separately, a) selected dataset, b) show comprehensive network option, c) the platform of the selected microarray experiment, d) filtering initial query nodes, e) chosen graph layout algorithm and f) MTIdb selected is also displayed in the information box.

3.4 Main Panel

3.4.1 Data Visualization Tab

Data selection step in the sidebar panel might be concluded either with data upload from user or selection of an example data stored in the server. If the user decides to run Odyssey with example data provided, selected data are displayed in the Data Visualization Tab in two separate panels. MiRNA microarray data are displayed at the top panel and mRNA microarray data are displayed at the panel below. Expression data on the hand are stored in data.table [29]

data structure defined in R. This format allows users to limit the number of rows that are displayed in the table. A search box is also implemented with data.table structure that allows users to do an exhaustive search through the table, which can be used to extract expression values through samples for specific probe id of interest.

Second option is for the user to upload her own data which requires the user to manually select "control" and "treatment" samples (based on the experiment setup, sample annotations might be different, i.e. 'cancer' and 'healthy' etc.). Selection of the samples is accomplished through the data.table structure which allows selection of multiple columns at once. Two from "Gene Controls", "Gene Treatments", "miRNA Controls" and "miRNA Treatments" buttons appear simultaneously as data upload completion at the bottom of each data.table corresponding to the data uploaded. These buttons allow the user to save selected column(s) as control or treatment samples to be processed. After one of these four buttons is clicked, column selection on the table will reset, allowing the user to proceed to the next selection without requiring to 'unselect' the samples selected before the button has been clicked. A differential gene expression analysis will be conducted based on the sample selection conducted in *Data Visualization Tab* should the user chooses to upload her own data. If example data is chosen to run Odyssey with, however, manual annotation of samples step will be skipped. This is due to example data stored in the server has predefined control and treatment samples.

3.4.2 Cytoscape Tab

After the data visualization part in the previous tab, pressing the *forward button* in the bottom right of the *Data Visualization Tab* triggers the creation of Cytoscape compatible 'Simple Interaction Format' (SIF) which is displayed in *Cytoscape Tab*. SIF file is also built using data.table structure, which means studying the file is viable through the same methodology described. Generally, a row of the SIF file will be composed of four cells, "miRNA", "target", "Diff_ mRNA", "Diff_ miRNA". If the user had run Odyssey without expression data selection, the last two cells will be omitted and one row of SIF will be composed

of either two or three cells. Variability in the data structure of the SIF file is due to optional data selection part and also MTIdb selection undertaken in the previous steps, i.e. if the union of TargetScan& miRNet MTIdbs is chosen, another column describing where the corresponding miRNA-target interaction information for each row of SIF file is taken from added to the table. Namely, values in the added column can be: 'TargetScan', 'miRNet' or 'both' indicating that the same interaction information is present in both MTIdbs.

Another important information to take out of data.table in Cytoscape Tab is the number of rows in the non-filtered SIF file. If user proceeds with default expression value filtering settings number of rows in SIF file will decrease due to filtering. However, the number of rows indicated at the bottom of the panel displaying the data can be put to use for an educated guess as to how many nodes will be in the non-filtered network. I predict a relatively flat learning curve of Odyssey to allow users to use filtering options with competence after a short period of usage. It should be noted networks larger than 50 nodes will take time to build and consume server resources depending on the graph layout algorithm selected.

3.4.3 Network Tab

Network Tab is where the output network is displayed. Networks are created using cytoscapejs implementation package in R. This package allows interactive network creation in HTML format. The user can zoom in to the network generated, move through the network and reset the zoom percentage to default setting using the gadgets provided in the upper left corner of the network. Also, a download button and a refresh button are provided at the bottom of the network. Download button allows network download in HTML file format to be transferred among peers. Refresh button allows the user to clear the parameters entered in the text boxes or selected through drop-down menus without terminating the session. Previous session information is saved for one run of Odyssey to prevent recalculation of the same input data entered, should the user choose to run Odyssey to visualize the previous session once more.

3.5 Data Handling

In this section, I will discuss how example expression data and MTI databases are processed and stored suitably for time efficient and convenient runs of the application.

3.5.1 Example Data Handling

Selected public expression data are downloaded from GEO webpage in cel file format. Cel files are processed in R using affy package as mentioned previously and assigned to an R object in data.table format mentioned. R programming has varying efficiency scores to handle large delimited data. Microbenchmark package [87] of R offers information as to the efficiency of these data handling functions. In this study, I used ".RData" for efficient saving and loading of example data selected which is the most efficient flow of large data handling except read_ feather function. Efficiency of load(*.RData) function is currently sufficient for handling the data stored in Odyssey, however in the later versions of Odyssey converting to the most efficient method is considered given the collection of example data provided in Odyssey becomes larger.

3.5.2 MTI Database Handling

In Odyssey, there are four different type of MTI databases (TargetScan, miRNet, TargetScan& miRNet - Intersection, TargetScan& miRNet - Union) that are being stored. The total number of MTI databases (MTIdb) stored is eight, however, this is the updated version of Odyssey which differs from development versions which stored four MTIdb's total. The increment in the MTI databases is done considering user input type, which can be either a gene or miRNA. Instead of converting the MTIdb according to user input dynamically, storing a gene and a miRNA input specific correspondence of each MTIdb is considerably time efficient and eases the work-load on the server processor. Handling of these databases robustly for different runs of the application is con-

sidered a significant part of time-efficient runs of the application. MTIdbs are converted into R specific "list of vectors" format and stored as ".RData" for performance considerations mentioned above. In the list of vectors format, each gene or miRNA is an object that returns a vector of targeted genes (for miRNA queries) or vector of targeting miRNAs (for gene queries). Through this implementation method, the issue of looping through the entire MTIdb is relieved which reduces performance significantly for R programming scripts.

CHAPTER 4

CASE STUDIES

4.1 Introduction to Case Studies

In order to demonstrate how Odyssey works on microarray expression data, public datasets from GEO database are downloaded and integrated onto Odyssey. Melanoma and Meningioma datasets were selected for the demonstration, primarily because these datasets are paired expression modulation datasets containing both miRNA expression data and mRNA expression data. In the generated networks, consequently, both miRNA and mRNA nodes are coloured by the expression values obtained from differential expression analysis run on the datasets. Thus, I aim to observe correlation on expression values of miRNA nodes and targeted mRNA nodes and reach a biological insight on the cases.

There are further reasons for the selection of Melanoma and Meningioma datasets. Both cases have published articles linked to the GEO database display webpage. After investigating the articles linked, I realized the opportunity to explain the underlying biological network because comprehensive network analysis was not conducted by the authors for either of the cases. In addition, novel biomarkers both reported by the authors and also established melanoma and meningioma driver genes in the literature presented a good opportunity for reproducing the published results and thus to verify that Odyssey brings results parallel to the established literature. Finally, the endeavour of reproducing established results may also conceive to novel biomarker discovery through following Odyssey's novel methodology.

4.2 Melanoma Case, Identifying mRNA, MicroRNA and Protein Profiles of Melanoma Exosomes

4.2.1 Introduction to Case

In this study, authors have compared the expression profiles of melanoma cells and normal melanocytes focusing on the profile of exosomes. Exosomes are bio-compartments that are secreted out of the cell through a process called exocytosis. Authors declared that by studying the expression profile of exosomes they are aiming for novel biomarker discovery that might successfully mark melanoma-related bioprocesses in the organism (human for the case of particular research). These biomarkers may be used to detect melanoma formation and presence. Depending on the signature comes as biomarkers, diagnosis can be made in an early stage and correct therapeutic agent administration can be done in an appropriate dosage.

For their analysis expression data have been obtained from two different commercially available cell lines. Human malignant melanoma cell line A375 and human normal epidermal melanocyte cell line HEMa-LP have been used for that purpose.

Table 4.1: GSE35389 Sample Overview.

Case	miRNA Samples	mRNA Samples
1	GSM867220, melanoma cell A375	GSM867227, melanoma cell A375
2	GSM867221, melanoma cell A375	GSM867228, melanoma cell A375
3	GSM867222, melanoma exosome A375	GSM867229, melanoma exosome A375
4	GSM867223, melanoma exosome A375	GSM867230, melanoma exosome A375
5	GSM867224, melanocyte HEMa-LP	GSM867231, melanocyte HEMa-LP
6	GSM867225, melanocyte HEMa-LP	GSM867232, melanocyte HEMa-LP
7	GSM867226, melanocyte HEMa-LP	GSM867233, melanocyte HEMa-LP exosome
8	- - -	GSM867234, melanocyte HEMa-LP exosome

4.2.2 Odyssey Analysis

GSE35389 dataset is comprised of exosome samples and melanocytes both for miRNA platform GPL8786 array and mRNA platform GPL570 array (Table 4.1). Differential expression analysis can be made through several comparisons. Selection of samples to perform differential expression analysis with is important for results to be unbiased and biologically meaningful. In this work, after samples were selected and DGex analysis was performed, corresponding networks were drawn and resulting networks were investigated taking important notes from literature. The aim of this section to propose an example method on how to perform an efficient analysis using Odyssey for mRNA-miRNA interactions in the context of the GSE35389 dataset.

4.2.2.1 Analysis on Melanomas vs Normal Human Melanocytes

Article published with the corresponding dataset makes important remarks for the comparison of melanocyte and melanoma cell exosome expression profiling. In this section of the document, a comparison has been done on cell miRNA and mRNA coexpression/interaction rather than exosome profiling. Thus, I aimed to conduct a novel analysis by selecting cellular expression data instead of exosome samples which are also present in the dataset. For that purpose, I selected GSM867224, GSM867225 normal melanocyte miRNA array samples obtained from HEMa-LP human melanocyte cell line as controls and GSM867220, GSM867221 melanoma cell miRNA array samples obtained from A375 human malignant melanoma cell line as treatment (See Table 4.1). Corresponding mRNA array GSM867231, GSM867232 normal melanocyte samples obtained from HEMa-LP human melanocyte cell line and GSM867227, GSM867228 melanoma cell samples obtained from A375 human malignant melanoma cell line are selected as the treatment to be compared via differential expression analysis. The aim of the study is to extract mRNA-miRNA pairs that interact with each other and at the same time are transcriptionally regulated and combine them into a concise network to further help researchers to validate their findings.

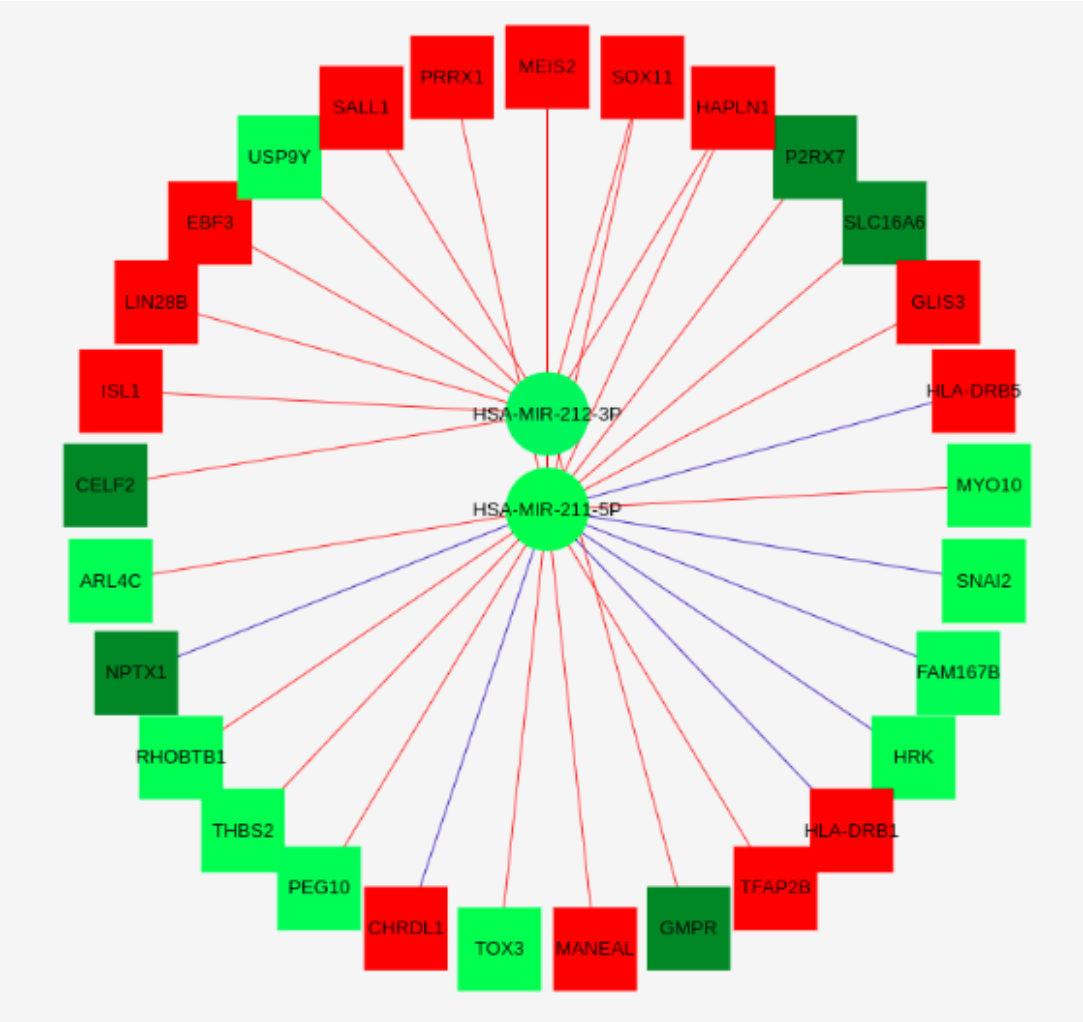


Figure 4.1: Network generated by Odyssey through query search "HAPLN1" on Melanoma Case.

TargetScan and miRNet Union database is used for the session. Dark red edges represent interactions reported by miRNet only and dark blue edges represent interactions reported by TargetScan only. Green edges are the microRNA - gene interactions reported both by TargetScan and miRNet. Rectangular nodes are genes whereas circular nodes are microRNAs. microRNAs are filtered according to (-3, 3.5) interval to leave out microRNAs that have log fold change within the interval. mRNAs in the network are filtered by the interval (-3.5, 3.5). The concentric layout was used for the generating of the network

HAPLN1 gene was reported as a finding by Xiao D. et al. authors of Melanoma Case Study. In the published article, HAPLN1 stated to be a differentially expressed melanoma exosomal protein which has been reported by other sources to be a tumorigenic protein with strong evidence playing roles in tumour invasion and metastasis [57] [142].

In the network generated by Odyssey (Figure 4.1), hsa-miR-212-3p is found in the middle of a cluster of highly upregulated genes like HAPLN1 (confirming the findings by Xiao D. et al.) and other reported tumour growth related genes e.g. ISB1 [127]. microRNAs play roles in proliferation and metastasis of different cancer types. Hsa-miR-212-3p (converted to all capital letters for ease of query search by the user), which is found in the centre of the cluster by Odyssey, is reported to have such a function, reported being carried by exosomes of tumour cells to other cells in the tumour microenvironment to promote tumour invasion and metastasis [26]. There seems to be a negative correlation in the gene expression between hsa-miR-212-3p and itś targets. This specific microRNA is down-regulated in the melanoma cells while the targets are mostly upregulated. Another microRNA in the centre of the network is hsa-miR-211-5p. Hsa-miR-211-5p is reported to suppress tumour invasion and proliferation in Triple Negative Breast Cancer [19]. Hsa-miR-211-5p is found to be down-regulated in the melanoma cells after differential expression analysis I made in Odyssey compared against normal human melanocytes. The findings of Odyssey help confirm the literature findings as well as providing other genes that might be modulated by these two downregulated miRNAs. Since miRNAs target multiple genes, their inhibitory activity is partitioned into their targets. One can assume a decrease in miRNA expression might lead to an increase in its target mRNA expression. In this respect, has-miR-213-3p has more negatively correlated mRNA targets than the 211-3p. In addition Odyssey show two different color-coded edges for the network in Fig. 4.1 Red edges represent interactions indicated within miRNet database while blue represent interactions within the TargetScan database. Yet no green nodes indicating existence in both miRNet and TargetScan which demonstrates that these two target prediction/validation databases diverge between themselves significantly. The number of red edges is not clearly different from blue edges in the count. The reason for this could

be the disjunct way MTI prediction algorithms work compared to experimental methods.

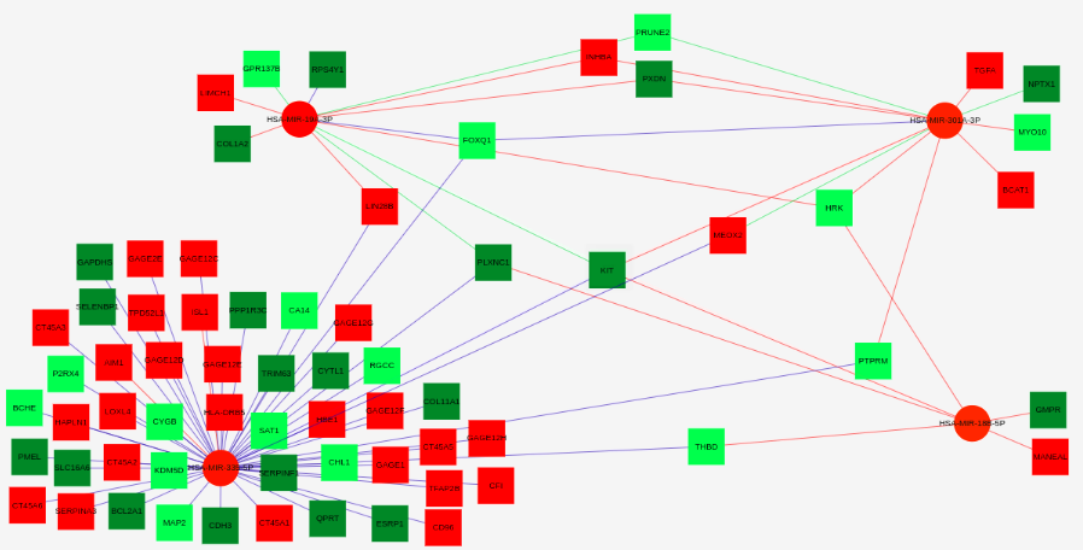


Figure 4.2: Network generated by Odyssey through query search "KIT". TargetScan and miRNet Union database is used for the session. Dark red edges represent interactions reported by miRNet only and dark blue edges represent interactions reported by TargetScan only. Green edges are the microRNA - gene interactions reported both by TargetScan and miRNet. Rectangular nodes are genes whereas circular nodes are microRNAs. microRNAs are filtered according to (-3, 3) interval to leave out microRNAs that have log fold change within the interval. mRNAs in the network are filtered by the interval (-4, 4). Cose layout was used for the generating of the network and then nodes were manually adjusted for a better visualization.

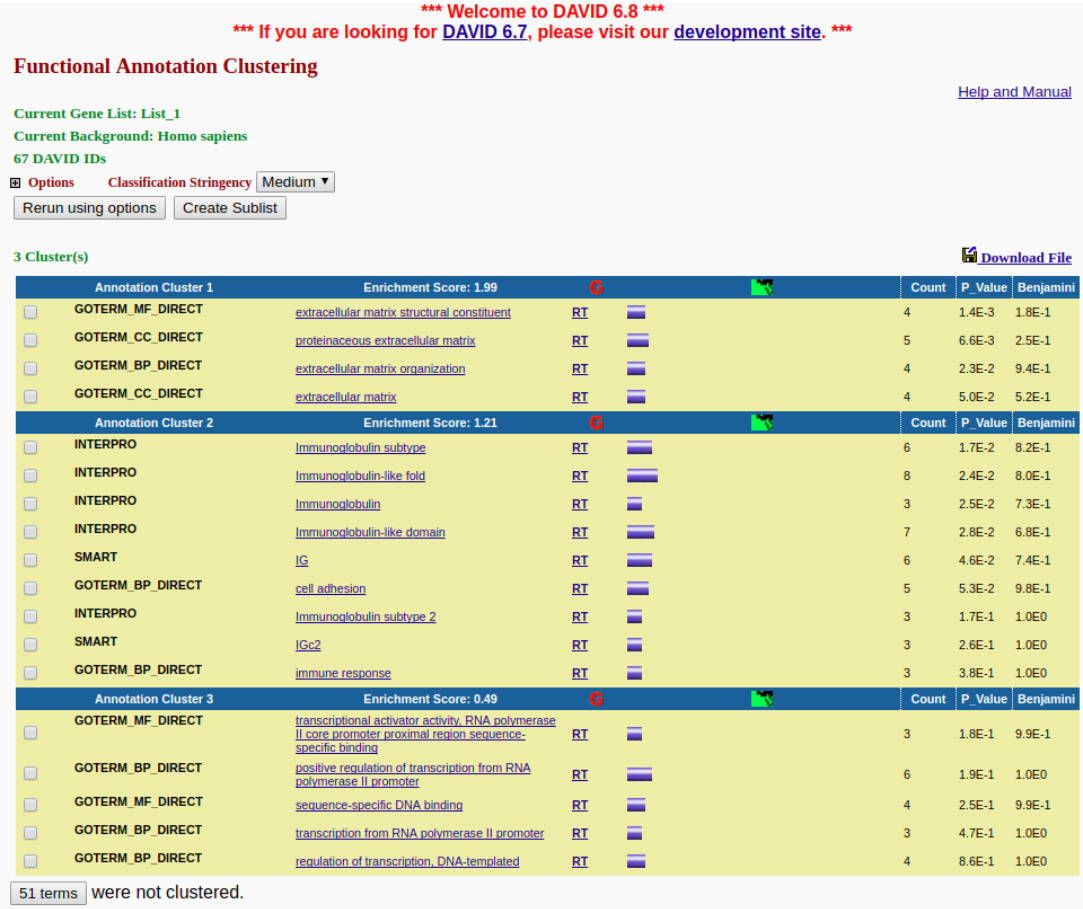
KIT is a well-known melanoma driver gene. There are studies on mutations of the KIT gene resulting in different forms of dermatosis, therefore, mutations on KIT is considered important for patients suffering from any of these diseases [129]. Interestingly, both under-expression and over-expression of KIT has been detected to cause abnormalities depending on the underlying pathway disruptions consequence of the specific mutation. Therapeutic agents are developed to target KIT to help battle melanoma [145]. In the network generated through Odyssey (Figure 4.2), the KIT gene is seen to be at the centre of the network. Four different microRNAs target KIT, namely; HSA-MIR-301A-3P, HSA-MIR-335-5P, HSA-MIR-19A-3P and HSA-MIR-18B-5P. Published literature confirms these microRNAs to belong to affected microRNAs in tumorigenesis and most of them have evidently played roles in studied melanoma cases. HSA-MIR-

301A-3P is found to be significantly upregulated in bladder cancer [35] and early progression of melanoma [93]. Research indicates HSA-MIR-335-5P can be used for the prognosis of gastric cancer [141]. Other studies report HSA-MIR-335-5P plays role in metastasis and invasion of gastric cancer [120]. Although these listed microRNAs (HSA-MIR-335-5P, HSA-MIR-301A-3P) are already reported in melanoma cases with strong evidence, HSA-MIR-19A-3P and HSA-MIR-18B-5P still need research on their involvement on carcinogenesis of melanoma. HSA-MIR-19A-3P targets 5-Lipoxygenase which is a key player in immunologic responses including macrophage, monocyte recruitment, and dendritic cell maturation [15]. Differential expression of HSA-MIR-19A-3P might cause direct regulation of 5-Lipoxygenase which in turn affects immunological response to the tumour mass, tumour microenvironment or overall immune response of the organism. HSA-MIR-19A-3P is found to be differentially expressed in plasma-derived exosomes of metastatic melanoma patients when compared to healthy melanocytes [101]. However, the way HSA-MIR-19A-3P operates in melanoma patients, the degree of malignancy it might cause and whether these findings can be used in the diagnosis or prognosis of patients still need work. HSA-MIR-18B-5P is another biological molecule effects in melanoma patients need to investigated like HSA-MIR-19A-3P. Functions of HSA-MIR-18B-5P is studied in relation with hepatocellular carcinoma patients and it has been reported that high expression of this particular microRNA significantly correlates with shorter relapse-free survival time period of the patients. I strongly believe that the network (Figure 4.2) put forward through analysis of Odyssey provides multiple candidate genes/miRNAs for prospective research for functional studies. In addition, network elements can be used as signatures for melanoma patients for both prognostic and diagnostic purposes.

The novelty Odyssey brings is a systems biological view on KIT expression and coregulation by microRNAs in using paired mRNA/miRNA data on the same subjects. Although miRtarvis and MAGIA and miRNet can do similar extractions have their own specific advantages/disadvantages; the advantages of using Odyssey in identification of KIT in melanoma is easy UI, clear visualization of the elements of the network, edge annotation indicating both predicted and/or validated targets, and ability to expand the network to second degree

starting with a single gene/miRNA.

GO terms and protein domains displayed in Figure 4.3 are significantly enriched for the gene set provided. Extracellular matrix and extracellular matrix structural constituent GO terms are in parallel with melanoma exosomes studied in the dataset. Cluster 2 is mostly separated from protein domains related to immune system components. Cluster 3 is composed of general GO terms and this cluster is not discussed here because of itś unspecificity.



*** Welcome to DAVID 6.8 ***
*** If you are looking for DAVID 6.7, please visit our development site. ***

Functional Annotation Clustering

Help and Manual

Current Gene List: List_1
Current Background: Homo sapiens
67 DAVID IDs

Options Classification Stringency Medium

Rerun using options Create Sublist

3 Cluster(s) Download File

Annotation Cluster 1	Enrichment Score: 1.99				Count	P_Value	Benjamini
GOTERM_MF_DIRECT	extracellular matrix structural constituent	RT			4	1.4E-3	1.8E-1
GOTERM_CC_DIRECT	proteinaceous extracellular matrix	RT			5	6.6E-3	2.5E-1
GOTERM_BP_DIRECT	extracellular matrix organization	RT			4	2.3E-2	9.4E-1
GOTERM_CC_DIRECT	extracellular matrix	RT			4	5.0E-2	5.2E-1
Annotation Cluster 2	Enrichment Score: 1.21				Count	P_Value	Benjamini
INTERPRO	Immunoglobulin subtype	RT			6	1.7E-2	8.2E-1
INTERPRO	Immunoglobulin-like fold	RT			8	2.4E-2	8.0E-1
INTERPRO	Immunoglobulin	RT			3	2.5E-2	7.3E-1
INTERPRO	Immunoglobulin-like domain	RT			7	2.8E-2	6.8E-1
SMART	IG	RT			6	4.6E-2	7.4E-1
GOTERM_BP_DIRECT	cell adhesion	RT			5	5.3E-2	9.8E-1
INTERPRO	Immunoglobulin subtype 2	RT			3	1.7E-1	1.0E0
SMART	IGc2	RT			3	2.6E-1	1.0E0
GOTERM_BP_DIRECT	immune response	RT			3	3.8E-1	1.0E0
Annotation Cluster 3	Enrichment Score: 0.49				Count	P_Value	Benjamini
GOTERM_MF_DIRECT	transcriptional activator activity, RNA polymerase II core promoter proximal region sequence-specific binding	RT			3	1.8E-1	9.9E-1
GOTERM_BP_DIRECT	positive regulation of transcription from RNA polymerase II promoter	RT			6	1.9E-1	1.0E0
GOTERM_MF_DIRECT	sequence-specific DNA binding	RT			4	2.5E-1	9.9E-1
GOTERM_BP_DIRECT	transcription from RNA polymerase II promoter	RT			3	4.7E-1	1.0E0
GOTERM_BP_DIRECT	regulation of transcription, DNA-templated	RT			4	8.6E-1	1.0E0

51 terms were not clustered.

Figure 4.3: DAVID Bioinformatics Tool outcome based on Odyssey results for query search "KIT" on Melanoma Case.

Results are extracted and the resulting gene list is input to DAVID Bioinformatics Tool. [Huang et al. [52]] Enriched GO terms and disease related terms and protein domains are displayed in the figure.

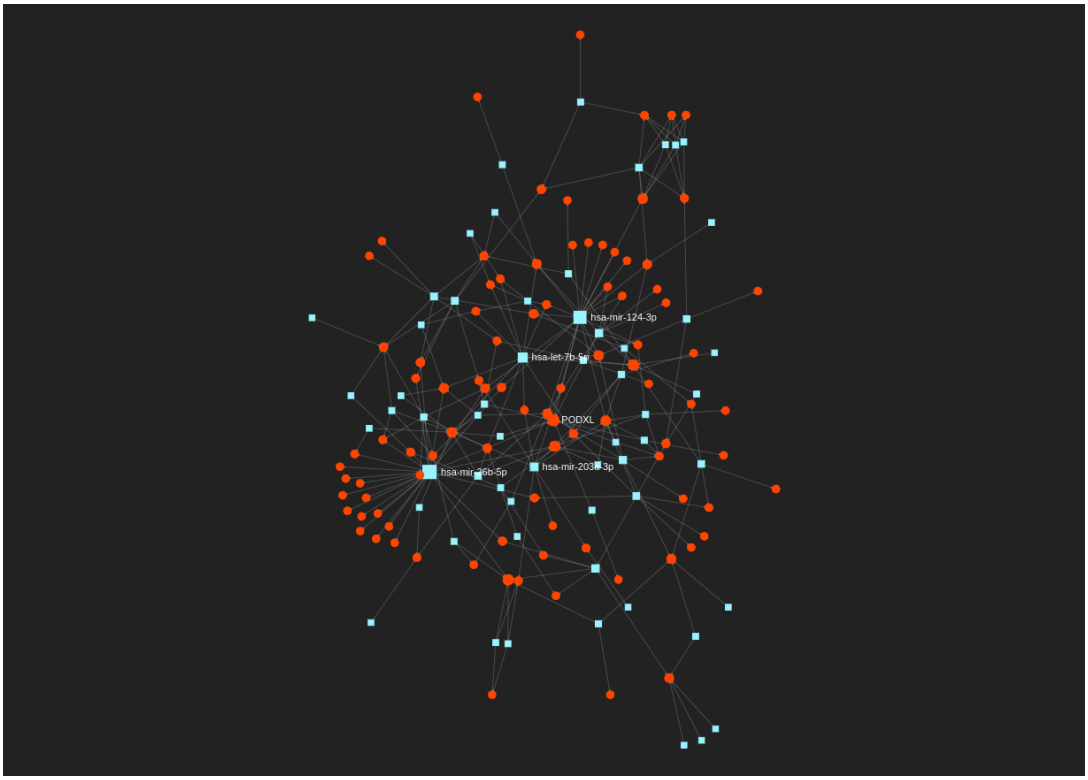


Figure 4.4: Network generated with miRNet on Melanoma Case.

Differentially expressed genes and microRNAs are extracted as list and microRNAs are filtered according to (-3, 3) interval to leave out microRNAs that have log fold change within the interval. mRNAs in the network are filtered by the interval (-4, 4). Resulting network is subjected Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment for the nodes in the network.

4.2.3 Analysis of Melanoma vs Normal Human Melanocytes via miRNet

Because miRNet is a similar software in terms of functionalities of Odyssey, an analysis was performed using the same samples for GSE35389 and the results were compared with those of Odyssey. For that purpose, differentially expressed filtered gene list used to generate Figure 4.2 is extracted and the miRNet session is run with "1.0" degree cut-off parameter to obtain an interactive network displaying the interactions between microRNAs and genes. Replicating the procedure followed for analysis with Odyssey, GSE35389 dataset GSM867231, GSM867232 normal melanocyte samples were selected as control and GSM867227, GSM867228 melanoma cell samples were selected as the treat-

ment to be compared via differential expression analysis using miRNet user interface (See Figure 4.4). Two major nodes of the network appeared as HSA-MIR-355-5P and KIT. These nodes were significant parts of the network generated with Odyssey as well. GO term enrichment test applied through miRNet User Interface by selecting "GO Term: Molecular Function" enrichment. Identical protein binding, protein homodimerization activity, and cysteine-type endopeptidase regulator activity involved in the apoptotic process were the most enriched GO terms with an insignificant p-value > 0.2 each. Some of the highly connected nodes in this network were, documented genes or microRNAs that have functions in cancer i.e. PODXL, reported to involved in Willms tumorigenesis via p53-mediated transcriptional repression [130]: hsa-miR-let7b, a known inhibitor of melanoma cell proliferation [140]: and hsa-miR-124-3p, a well-studied actor of proliferation inhibitor in multiple types of cancer including melanoma [128] and lung adenocarcinoma [136]. Highlighted nodes in the miRNet analysis were different from the findings of Odyssey but not unrelated to case study explored in this section. GO term enrichment analysis via miRNet was inconclusive for this network as none of the terms was significantly enriched for the p values obtained. One apparent shortcoming of miRNet was, UI did not allow uploading microRNA expression data which rendered it unable to filter via microRNA based log expression fold change between the control and treatment samples. In addition coloring of the nodes based on log fold changes was not possible neither the annotation of the edges as predicted or validated or both.

4.2.4 Discussion on the Results

Multiple queries based on melanoma data have been processed and analysed with Odyssey in this section. Based on the results, there are important remarks supported by existing literature which indicates efficient ways of using Odyssey, and also some results might shed light on future melanoma research. Further, I aimed to have a clear comparison of the results between Odyssey and the competitive existing software, miRNet.

Based on the literature research of the results given by both software, highly functional and previously documented melanoma-related genes and miRNAs

have been revealed by the networks. Some promising nodes, HSA-MIR-19A-3P and HSA-MIR-18B-5P for Odyssey results and PODXL for miRNet results might be beneficial for future research as knowledge associating these elements to melanoma cases is still poorly understood. There are multiple ways proposed to explore the functionality of these elements and attain the findings of melanoma cases. One of which is uploading more microRNA - gene matching melanoma expression data to Odyssey or miRNet and analyse the results. Another way (but not necessarily the last) is to explore the functionality of these molecules on the molecular level and try to correlate the findings to melanoma cases without using software discussion in this document. Both of the experimental ways listed would be interesting for the purposes of scientific discovery however, they are out of context with this thesis. I hope that motivation for this proposed research comes from results obtained with Odyssey that are in parallel with literature.

MicroRNA expression data can be uploaded to Odyssey and filtering can both be done on microRNAs and genes which is an advantage compared to miRNet. Without this functionality, microRNA nodes in the network have no attained expression correlation with targeting nodes which causes only significantly expressed genes to be predictive on the phenotype of data studied. On the other hand, miRNet allows more filtering options on the network like degree filtering or shortest path filtering which is useful for extracting highly regulated components of the network and eliminating singularities which might be irrelevant to the study. GO term analysis of miRNet yielded no significant clusters whereas gene list obtained from Odyssey revealed interesting GO terms clusters that are parallel to data. This coordination between the literature search and GO term enrichment analysis of results are esteemed to be an indication of promising analyses will be conducted using Odyssey.

At this point, Odyssey does not allow users to enter multiple queries to be analysed which is a fundamental difference between miRNet and Odyssey. Odyssey rather builds a second-degree network based on the initial query which is the main reason why the results are different between the two software. This single query based way of work results in more isolated networks to be produced. Capability to expand the network to second-degree interactions is beneficial for a more comprehensive network view which is in parallel with high-throughput

data analysis. Comprehensiveness of the network is limited by the initial query, however, which in the end is able to produce a more focused and detailed network from a single nodeś perspective. This phenomenon is also highly dependant on the user because filtering parameters are subjected to change in every session for Odyssey but also for miRNet as well. Odyssey might have an advantage for calculation of initial filtering parameters that are automated with quantile values mentioned in the methods section.

As a final remark, it should be noted miRNet has other functionalities apart from drawing microRNA - gene networks for having documented lncRNA - miRNA interactions or RT-qPCR - gene interactions etc.. These functionalities at this point are, beyond the purpose of why Odyssey was developed.

4.3 Meningioma Case for simultaneous analysis of miRNA-mRNA in human meningiomas by integrating transcriptome: A relationship between PTX3 and miR-29c

4.3.1 Introduction to Case

Meningioma is a type of brain cancer that causes serious morbidity [100]. Around 30 percent of all brain tumours are classified as meningioma and the underlying biology behind this type of cancer is poorly understood [137]. In this dataset, authors aimed to explore molecular mechanisms behind meningioma cases by studying microRNA, mRNA relations through comparative transcriptome analysis. Dataset comprises of 30 samples equally divided to miRNA and mRNA assays in numbers. One control (normal meningial cell sample) is used each against 14 meningioma samples for both miRNA and mRNA subset of the whole dataset to be able to extract relative expression profiles. Authors published an article (Dalan et al., cited in Example Data Collection section) that focuses on the relation between PTX3 and miR-29c. In this article, PTX3 and miR-29c are claimed to have a negative correlation in expression levels which is interpreted as an indication of regulation of PTX3 by miR-29c. Same samples have been chosen to be used in comparison to the analysis conducted in my own

research. Odyssey is used to query PTX3 and miR-29c separately. However, observing that PTX3 query has resulted in an inconclusive network with one miRNA node interacting with multiple genes with varying expression levels I have chosen to omit this particular network from the Case Study section of this thesis. The same dataset also has been used for an analysis in miRTarVis which is a similar software to Odyssey. Results obtained from Odyssey were planned to be compared both with miRTarVis results and also the results published in the article by Dalan et al, however, after the analysis with miRTarVis is abandoned (explained in miRTarVis Analysis section below), miRNet software is used for comparison for this dataset also.

Table 4.2: GSE88721 Sample Overview.

Case	miRNA Samples	mRNA Samples
1	GSM2344678 Meningioma 1 [miRNA]	GSM2344693, Meningioma 1 [gene]
2	GSM2344679 Meningioma 2 [miRNA]	GSM2344694, Meningioma 2 [gene]
3	GSM2344680 Meningioma 3 [miRNA]	GSM2344695, Meningioma 3 [gene]
4	GSM2344681 Meningioma 4 [miRNA]	GSM2344696, Meningioma 4 [gene]
5	GSM2344682 Meningioma 5 [miRNA]	GSM2344697, Meningioma 5 [gene]
6	GSM2344683 Meningioma 6 [miRNA]	GSM2344698, Meningioma 6 [gene]
7	GSM2344684 Meningioma 7 [miRNA]	GSM2344699, Meningioma 7 [gene]
8	GSM2344685 Meningioma 8 [miRNA]	GSM2344700, Meningioma 8 [gene]
9	GSM2344686 Meningioma 9 [miRNA]	GSM2344701, Meningioma 9 [gene]
10	GSM2344687 Meningioma 10 [miRNA]	GSM2344702, Meningioma 10 [gene]
11	GSM2344688 Meningioma 11 [miRNA]	GSM2344703, Meningioma 11 [gene]
12	GSM2344689 Meningioma 12 [miRNA]	GSM2344704, Meningioma 12 [gene]
13	GSM2344690 Meningioma 13 [miRNA]	GSM2344705, Meningioma 13 [gene]
14	GSM2344691 Meningioma 14 [miRNA]	GSM2344706, Meningioma 14 [gene]
15	GSM2344692 Meningial Cells [miRNA]	GSM2344707, Meningial Cells [gene]

4.3.2 Odyssey Analysis

HSA-MIR-29C query returned two choices to select from which were both present in the TargetScan& miRNet - Union database; HSA-MIR-29C-3P and HSA-MIR-29C-5P. As reported by Dalan et al., the analysis has been carried on by HSA-MIR-29C-3P selection which had lower log fold change value of -3.627 when compared to HSA-MIR-29C-5P which had -1.570 log fold change. Result-

ing network had 34 nodes and 37 edges (See Figure 4.5). Importantly, the PTX3 gene was not amongst the listed targets of HSA-MIR-29C-3P for both miRNet and TargetScan databases. Because of that, claims of the authors cannot be verified using Odyssey. However, targets of HSA-MIR-29C-3P were displayed in the network and some of these genes have indications in the literature for the roles they might play in meningioma patients. Anthrax toxin receptor 2 (ANTXR2) is such a gene having proof of relation to meningioma although indirectly documented in Li et al. [75]. YAP is an oncogene when knocked down, causes an up-regulation in a list of genes including ANTXR2. ANTXR2 is highly upregulated in the network drawn with Odyssey and it has been targeted by several other microRNAs that are highly down-regulated. HSA-MIR-27B-3P was another modulator of ANTXR2 which has been shown to be down-regulated in meningiomas along with HSA-MIR-29C-3P in Ludwig et al. [79]. Thus, ANTXR2 appears to be a promising biological molecule to investigate in prospective research dealing with meningioma.

LOX gene was shown as another target of HSA-MIR-29C-3P and targeted by multiple down-regulated microRNAs. LOX has been shown to be upregulated in some meningioma cases and reported to take a role in cell adhesion in Fèvre-Montange et al. [37] by its functional importance in collagen and elastin cross-linking.

There has been only one gene, PRELP, emerged presently out of applied filtering. PRELP is a candidate predictor of meningioma and chronic lymphoid leukaemia (CLL) according to two studies where PRELP has been documented in several meningioma and CLL cases to be differentially expressed [17] [89].

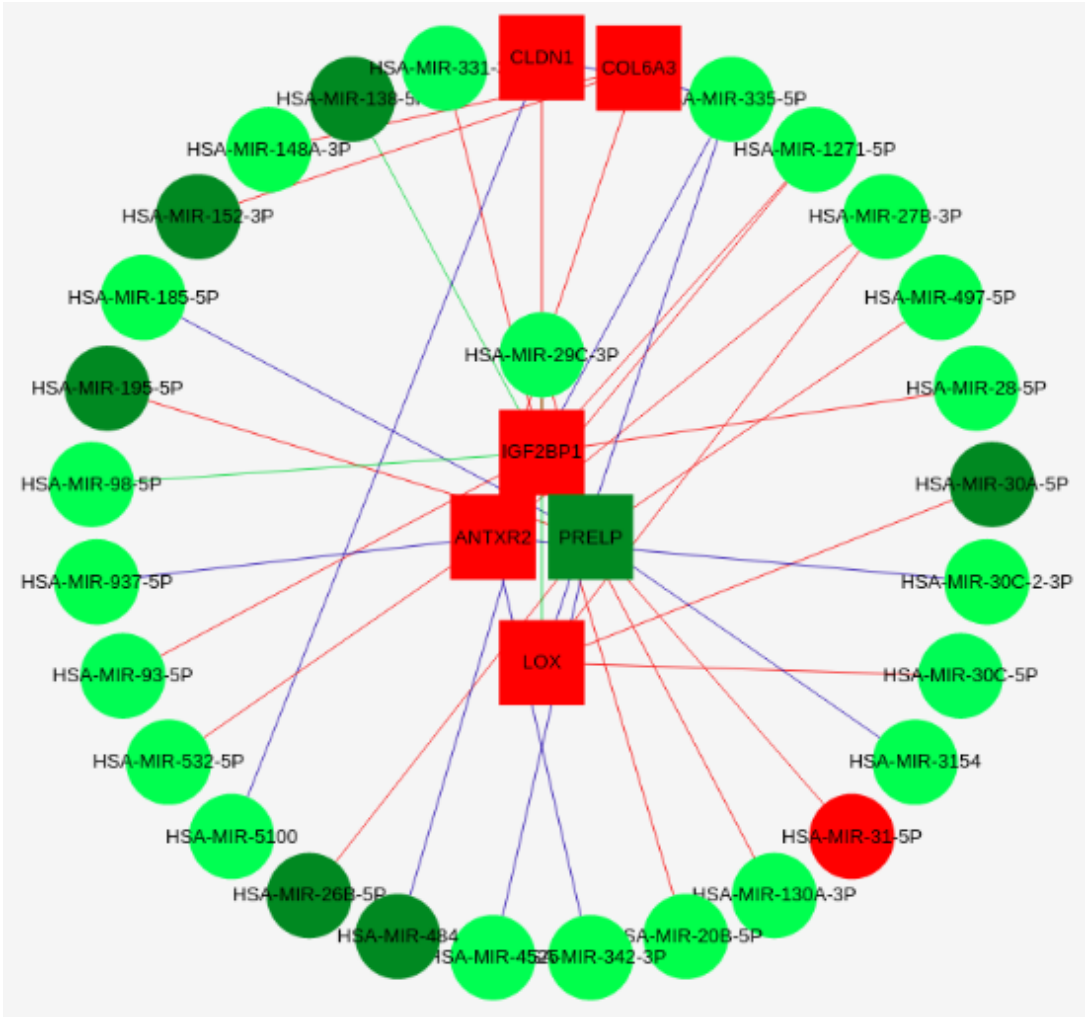


Figure 4.5: Network generated by Odyssey through query search "HSA-MIR-29C-3P" on Meningioma Case.

TargetScan and miRNet Union database is used for the session. Dark red edges represent interactions reported by miRNet only and dark blue edges represent interactions reported by TargetScan only. Green edges are the microRNA - gene interactions reported both by TargetScan and miRNet. Rectangular nodes are genes whereas circular nodes are microRNAs. microRNAs are filtered according to (-3.5, 3.5) interval to leave out microRNAs that have log fold change within the interval. mRNAs in the network are filtered by the interval (-4.0, 4.0). The concentric layout was used for the generating of the network.

4.3.3 miRTarVis Analysis

GSE88721 data has been uploaded miRTarVis which runs on local computers, unlike Odyssey. Data format and sample names are important for data uploaded to be completed successfully. The first row of the data holds the sample names. This row was modified for miRTarVis to recognize which sample columns were "Treatment" and "Control" samples following the data sample demonstration within miRTarVis folder downloaded; in the case of this dataset, they were meningioma samples and normal meningial cells, respectively. MiRTarVis can run one or two-tailed t-test on the samples for both miRNA and mRNA data and asks whether to apply log transformation before running these tests. The two-tailed t-test was chosen because of the identification of candidate genes and miRNAs on both sides of the data distribution considered important. However, the log transformation option was not applied because data were log transformed before analysing it with Odyssey. MiRTarVis enables filtering on both p-value and log fold change obtained from the two-tailed t-test. After the data upload, statistical analysis conducted by miRTarVis yielded all miRNAs and genes to have 0 fold change and 0 p-value. Thus, filtering was not possible to apply because any filtering applied would cause node count within the network reduced to 0. After many unsuccessful attempts to fix the issue by going back and forth to documentation supplied in the website of miRTarVis, the whole miRTarVis analysis is abandoned and miRNet is used again for comparison with Odyssey.

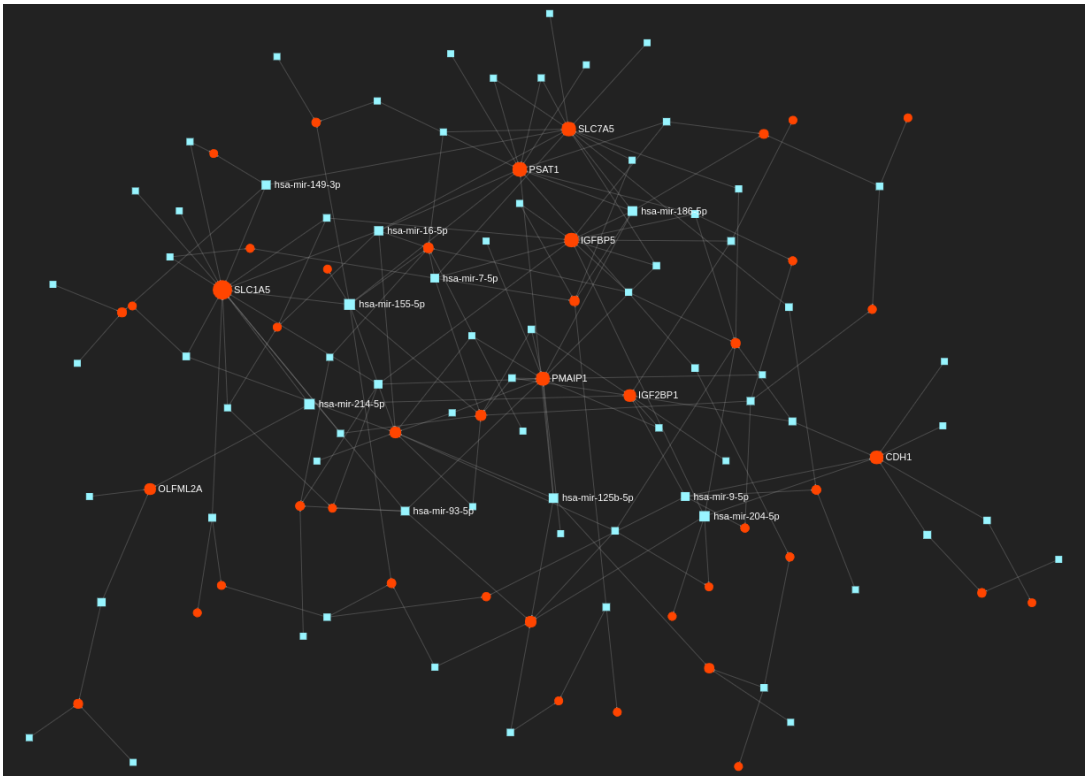


Figure 4.6: Network generated by miRNet through on Meningioma case.

Brain samples were chosen in miRNet and limma based differential expression analysis is performed on the samples. mRNAs in the network are filtered by the interval (-4.0, 4.0) and, 1.0-degree cut-off is applied on the graph. The force atlas layout was used for the generating of the network.

4.3.4 miRNet Analysis

Compared to miRTarVis, miRNet analysis goes fairly straight-forward. MiRNet provides better documentation on how to run an analysis with working examples of example data. User-interface of miRNet designed as compact and intuitive for smoother user experience. After mRNA data upload, (-4.0, 4.0) log fold change interval has been selected to be removed from the network for mRNA nodes and, 1.0 degree cut off is also used to remove singleton nodes. Thus, only nodes with higher inter-connectivity remain in the network which I find is a better way of extracting knowledge out. Degree cut-off parameter remains optional, however, for users to experiment and find the best suiting criteria for themselves.

After the graph has been generated several nodes took more space in the

network observed from their node diameters. These nodes represented higher connectivity within the network and as mentioned above, more suitable candidates for further investigation. One such node was gene PMAIP1. PMAIP1 have some evidence in meningioma cases although not thoroughly investigated, to be playing roles in recurrence and progression of meningioma cases [64] and, is shown to be taking part in apoptosis-related pathways mediated by p53 expression [30]. Hsa-miR-155-5p is another such candidate and have stronger evidence on brain neoplasia including but not limited to meningioma cases reported to cause increased proliferation of cancer cells when over-expressed [134]. Moreover, hsa-miR-155-5p expression can be used as a predictor of numerous other types of cancer including breast cancer and head and neck squamous cell carcinoma for its property of being an oncogenic microRNA [117].

Finally, network nodes were enriched with functional and utility annotations using KEGG database. Enriched Cell adhesion molecules (CAMs) and Complement and coagulation cascades were the first two terms on the list with subsequently 4 and 3 hits with insignificant p-value 0.291 for both terms. GO terms: Molecular Function database is used next for enrichment. Amino acid transmembrane transporter activity (4 hits, p-value 0.0144), neutral amino acid transmembrane transporter activity (3 hits, p-value 0.0144) and L-amino acid transmembrane transporter activity (3 hits, p-value 0.0268) were the most enriched terms found at the very top of the list amongst some other insignificantly enriched terms. Solute Carrier Family proteins SLC7A5, SLC1A5 and SLC7A1 were the highlighted nodes resulted in the enrichment of transporter activity related GO terms.

4.3.5 Discussion on the Results

Progress of whole analysis needs to be visualized with proper error reports that guide the user to be able to complete the analysis. Also, comprehensive documentation on demo usage of the tool is very important.

Both miRNet and Odyssey results have come up with links in the literature published on cancer studies including meningioma and other brain neoplasia. These candidates can be used for further analyses, however, at this stage, they

alone provide insufficient information to be able to reveal the phenotypical background of the samples. Additional comprehensive literature search seems fundamental for evaluation of the findings and supporting the claims. Functional enrichment analysis done in miRNet software does not appear to be useful in meningioma dataset also. Enriched terms either have insignificant p-value or they are very general terms like amino acid transmembrane transporter activity which can be interpreted as a natural cellular activity for normal and diseases samples.

Findings of Odyssey conflicts with the literature on PRELP because in Figure 4.5, PRELP is observed to be highly down-regulated compared to normal meningial cells whereas according to literature up-regulation of the gene was expected. There could be many explanations for this issue. One possible explanation is that case-specific studies do not always reveal the underlying biology behind the sample analysed. By increasing the sample size or number of experiments and evaluating the findings altogether, expression profiling of more genes can be brought to light with higher confidence. The second possible explanation is that microRNA - gene interactions do not always have the dramatic effect to regulate the pathways by itself, as there are other biological agents in effect to suppress or further elevate the observed condition. Increasing the experiment size as mentioned above as the first explanation might be helpful in revealing the true biology underneath. Integration of multiple OMICS data alongside with expression value data could be beneficial as well to help see how other biological molecules affect the system.

One other explanation is that the network is too narrow for identifying how each system element behaves under various conditions. Increasing the size of the network might be helpful to understand the reasons why some nodes do not follow the expected patterns in expression levels. However, even after all of the listed methods are applied conflicting results between experiments can occur which makes the collective evaluation of results more important. To the prospective users of Odyssey, I advise not to jump into conclusions and use Odyssey as a guiding tool rather than a definitive one, which I humbly believe should be the way for any biological software present.

CHAPTER 5

CONCLUSIONS AND FINAL REMARKS

Odyssey was developed for paired and unpaired miRNA and mRNA data analysis and visualization of the analyses on an interactive graph. It is not the first software to be developed for this purpose, miRNet, MAGIA and miRTarVis were developed for the similar rationale. There are pros and cons of these tools when functionalities are matched against each other. MiRTarVis and Odyssey allow both paired and unpaired data analysis whereas miRNet does not have this functionality. Paired data analysis is important for making correlations between microRNAs and their targets in the biological system of interest. Using this module, clusters within the network can be detected as candidate components that are differentially regulated in the samples uploaded. MiRNet and Odyssey allows analysis without any expression data upload, which makes them practical for observing the sole interactions of any gene or miRNA query.

One of the most important conclusions of the present thesis is that microRNA - mRNA interactions might not explain all the variability on the samples investigated. Up-scaling the experiment set up to include more samples, running additional experiments and backing the results with existing literature is crucial for reaching conclusions with higher confidence. As discussed, integration of multi-OMICs data might reveal more information that is on systems biology level which is an outlook for the future of Odyssey.

Odyssey falls behind of other similar software in statistical back up of the session run. Fold change is an important filtering parameter to be used, however, the addition of p values is also important for clarifying insignificant results and confidently reporting candidates in the network. Besides, some of the network nodes that are being filtered out due to low fold change values might also be

important candidates for explaining the phenotype of the samples. Additional statistical parameters like p-value or degree filtering might be helpful in terms of more meaningful graph generation and reducing these graphs to most meaningful sections. Both Odyssey and existing other software lack correlation analysis of miRNAs with its targets based on the network studied. Implementing correlation scores might be important to understand expression profiles of studies coming from various sources. I intend to implement a module to calculate correlation scores on various experiments on Odyssey. This prospective module will be keeping this information to be reflected upon future Odyssey sessions. Storing additional data on each user session should strictly be dependant on user permissions.

New bioinformatics tools emerge at an increasing pace because of the need to analyse massive amounts of biological data generated. These tools need to be flexible to adapt ever-changing biological sciences and shape their functionalities accordingly. Online based tools have the advantage at this era of biological sciences because local computers lack sources to work with large amounts of data. Moreover, error handling, updates, and related documentation are easier to establish with online software as these software continuously collect session data in log files based on real-time usage.

Software mentioned in this document provide a good guide for understanding the biological features of an experiment run in a session. However, in my personal opinion bioinformatics tools need to be improved towards using the Systems Biology features; combining multiple aspects of biological data e.g. RNA-seq, ChIP-seq. Exploring the underlying biology of a system from a singular perspective i.e. transcriptomics exclusively, is not enough for reaching a comprehensive understanding.

REFERENCES

[1] V. Agarwal, G. W. Bell, J. W. Nam, and D. P. Bartel. Predicting effective microRNA target sites in mammalian mRNAs. *eLife*, 4(AUGUST2015), 2015.

[2] M. I. Almeida, R. M. Reis, and G. A. Calin. MicroRNA history: Discovery, recent applications, and next frontiers, 2011.

[3] W. A. Alsharafi, B. Xiao, M. M. Abuhamed, and Z. Luo. miRNAs: biological and clinical determinants in epilepsy. *Frontiers in Molecular Neuroscience*, 8, 2015.

[4] F. Amati, G. Chillemi, and G. Novelli. *Gene expression analysis during development by high-throughput methods*. 2009.

[5] V. Ambros. The functions of animal microRNAs, 2004.

[6] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: Tool for the unification of biology, 2000.

[7] O. G. Attia, T. Johnson, K. Townsend, P. Jones, and J. Zambreno. CyGraph: A reconfigurable architecture for parallel breadth-first search. In *Proceedings of the International Parallel and Distributed Processing Symposium, IPDPS*, pages 228–235, 2014.

[8] S. Banerjee, M. Jenamani, and D. K. Pratihar. Properties of a projected network of a bipartite network. In *Proceedings of the 2017 IEEE International Conference on Communication and Signal Processing, ICCSP 2017*, volume 2018-January, pages 143–147, 2018.

[9] M. Y. Becker and I. Rojas. A graph layout algorithm for drawing metabolic pathways. *Bioinformatics*, 17(5):461–467, 2001.

[10] C. Beeley. *Web Application Development with R using Shiny*, volume 2. 2013.

[11] G. Blin, F. Sikora, and S. Vialette. Querying graphs in protein-protein interactions networks using feedback vertex set. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 7(4):628–635, 2010.

[12] A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C. A. Ball, H. C. Causton, T. Gaasterland, P. Glenisson, F. C. Holstege, I. F. Kim, V. Markowitz, J. C. Matese, H. Parkinson, A. Robinson, U. Sarkans, S. Schulze-Kremer, J. Stewart, R. Taylor, J. Vilo, and M. Vingron. Access : Minimum information about a microarray experiment (MIAME)|[mdash]|toward standards for microarray data : Nature Genetics. *Nature genetics*, 29(4):365–71, 2001.

[13] J. Brennecke, A. Stark, R. B. Russell, and S. M. Cohen. Principles of microRNA-target recognition. In *PLoS Biology*, volume 3, pages 0404–0418, 2005.

[14] J. P. Broughton, M. T. Lovci, J. L. Huang, G. W. Yeo, and A. E. Pasquinelli. Pairing beyond the Seed Supports MicroRNA Targeting Specificity. *Molecular Cell*, 64(2):320–333, 2016.

[15] S. Busch, E. Auth, F. Scholl, S. Huenecke, U. Koehl, B. Suess, and D. Steinhilber. 5-Lipoxygenase Is a Direct Target of miR-19a-3p and miR-125b-5p. *The Journal of Immunology*, 194(4):1646–1653, 2015.

[16] M. J. Carpano. Automatic Display of Hierarchized Graphs for Computer-Aided Decision Analysis. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-10(11):705–715, 1980.

[17] X. Castells, J. J. Acebes, S. Boluda, A. Moreno-Torres, J. Pujol, M. Julià-Sapé, A. P. Candiota, J. Ariño, A. Barceló, and C. Arús. Development of a predictor for human brain tumors based on gene expression values obtained from two types of microarray technologies. *Omics : a journal of integrative biology*, 14(2):157–64, 2010.

[18] S.-Y. Chao. Graph Theory and Analysis of Biological Data in Computational Biology. In *Advanced Technologies*. InTech, oct 2009.

[19] L. L. Chen, Z. J. Zhang, Z. B. Yi, and J. J. Li. MicroRNA-211-5p suppresses tumour cell proliferation, invasion, migration and metastasis in triple-negative breast cancer by directly targeting SETBP1. *British Journal of Cancer*, 117(1):78–88, 2017.

[20] J. Cheng. Modularizing Shiny app code, 2015.

[21] E. H.-H. Chi. A taxonomy of visualization techniques using the data state reference model. *Information Visualization, 2000. InfoVis 2000. IEEE Symposium on*, (Table 2):69–75, 2000.

[22] S. Chojnacki, A. Cowley, J. Lee, A. Foix, and R. Lopez. Programmatic access to bioinformatics tools from EMBL-EBI update: 2017. *Nucleic Acids Research*, 45(W1):W550–W553, 2017.

[23] M. E. Cusick, N. Klitgord, M. Vidal, and D. E. Hill. Interactome: Gateway into systems biology. *Human Molecular Genetics*, 14(SUPPL. 2), 2005.

[24] A. B. Dalan, S. Gulluoglu, E. C. Tuysuz, A. Kuskucu, C. K. Yaltirik, O. Ozturk, U. Ture, and O. F. Bayrak. Simultaneous analysis of miRNA-mRNA in human meningiomas by integrating transcriptome: A relationship between PTX3 and miR-29c. *BMC Cancer*, 17(1), 2017.

[25] D. Didiano and O. Hobert. Perfect seed pairing is not a generally reliable predictor for miRNA-target interactions. *Nature Structural and Molecular Biology*, 13(9):849–851, 2006.

[26] G. Ding, L. Zhou, Y. Qian, M. Fu, J. Chen, J. Chen, J. Xiang, Z. Wu, G. Jiang, and L. Cao. Pancreatic cancer-derived exosomes transfer miRNAs to dendritic cells and inhibit RFXAP expression via miR-212-3p. *Oncotarget*, 6(30):29877–29888, 2015.

[27] U. Dogrusoz, E. Giral, A. Cetintas, A. Civril, and E. Demir. A layout algorithm for undirected compound graphs. *Information Sciences*, 179(7):980–994, 2009.

[28] J. DOI, G. POTTER, J. WONG, I. ALCARAZ, and P. CHI. Web Application Teaching Tools for Statistics Using R and Shiny Journal. *Technolgy Innovations in Statstics Education*, 9(1):1–10, 2016.

[29] M. Dowle. Package 'data.table'. *Cran*, 2016.

[30] M. G. Dozmorov, Q. Yang, W. Wu, J. Wren, M. M. Suhail, C. L. Woolley, D. G. Young, K. M. Fung, and H. K. Lin. Differential effects of selective frankincense (Ru Xiang) essential oil versus non-selective sandalwood (Tan Xiang) essential oil on cultured bladder cancer cells: A microarray and bioinformatics study. *Chinese Medicine (United Kingdom)*, 9(1), 2014.

[31] U. Doğrusōz, B. Madden, and P. Madden. Circular layout in the graph layout toolkit. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 1190, pages 92–100, 1997.

[32] T. Du. microPrimer: the biogenesis and function of microRNA. *Development*, 132(21):4645–4652, 2005.

[33] G. Dumancas, I. Adrianto, G. Bello, and M. Dozmorov. Current Developments in Machine Learning Techniques in Biological Data Mining. *Bioinformatics and Biology Insights*, 11(0), 2017.

[34] R. Edgar. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1):207–210, 2002.

[35] H. Egawa, K. Jingushi, T. Hirono, Y. Ueda, K. Kitae, W. Nakata, K. Fujita, M. Uemura, N. Nonomura, and K. Tsujikawa. The miR-130 family promotes cell migration and invasion in bladder cancer through FAK and Akt phosphorylation by regulating PTEN. *Scientific Reports*, 6, 2016.

[36] Y. Fan, K. Siklenka, S. K. Arora, P. Ribeiro, S. Kimmins, and J. Xia. miRNet - dissecting miRNA-target interactions and functional associations through network-based visual analysis. *Nucleic acids research*, 44(W1):W135–W141, 2016.

[37] M. Fèvre-Montange, J. Champier, A. Durand, A. Wierinckx, J. Honnorat, J. Guyotat, and A. Jouvet. Microarray gene expression profiling in meningiomas: Differential expression according to grade or histopathological subtype. *International Journal of Oncology*, 35(6):1395–1407, 2009.

[38] W. Filipowicz, S. N. Bhattacharyya, and N. Sonenberg. Mechanisms of post-transcriptional regulation by microRNAs: Are the answers in sight?, 2008.

[39] M. Franz, C. T. Lopes, G. Huck, Y. Dong, O. Sumer, and G. D. Bader. Cytoscape.js: A graph theory library for visualisation and analysis. *Bioinformatics*, 32(2):309–311, 2015.

[40] T. M. Fruchterman and E. M. Reingold. Graph drawing by forcedirected placement. *Software: Practice and Experience*, 21(11):1129–1164, 1991.

[41] L. Gautier, L. Cope, B. M. Bolstad, and R. A. Irizarry. Affy - Analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, 20(3):307–315, 2004.

[42] R. Gentleman. *R Programming for Bioinformatics*. 2009.

[43] S. Griffiths-Jones, R. J. Grocock, S. van Dongen, A. Bateman, and A. J. Enright. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Research*, 34(suppl_1):D140–D144, 2006.

[44] A. Grimson, K. K. H. Farh, W. K. Johnston, P. Garrett-Engele, L. P. Lim, and D. P. Bartel. MicroRNA Targeting Specificity in Mammals: Determinants beyond Seed Pairing. *Molecular Cell*, 27(1):91–105, 2007.

[45] M. Hackenberg. MicroRNA expression profiling and discovery. In *Bioinformatics for High Throughput Sequencing*, pages 191–208. 2012.

[46] G. T. Hart, A. K. Ramani, and E. M. Marcotte. How complete are current yeast and human protein-interaction networks? *Genome Biology*, 7(11), 2006.

[47] L. H. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray. From molecular to modular cell biology. *Nature*, 402(6761supp):C47–C52, 1999.

[48] H. Hauswedell, J. Singer, and K. Reinert. Lambda: The local aligner for massive biological data. In *Bioinformatics*, volume 30, 2014.

[49] M. Hecker, S. Lambeck, S. Toepfer, E. van Someren, and R. Guthke. Gene regulatory network inference: Data integration in dynamic models-A review. *BioSystems*, 96(1):86–103, 2009.

[50] T. R. Henry and S. E. Hudson. End User Controlled Visualization of Large Graphs.

[51] S. D. Hsu, F. M. Lin, W. Y. Wu, C. Liang, W. C. Huang, W. L. Chan, W. T. Tsai, G. Z. Chen, C. J. Lee, C. M. Chiu, C. H. Chien, M. C. Wu, C. Y. Huang, A. P. Tsou, and H. D. Huang. MiRTarBase: A database curates experimentally validated microRNA-target interactions. *Nucleic Acids Research*, 39(SUPPL. 1), 2011.

[52] D. W. Huang, B. T. Sherman, and R. A. Lempicki. Systematic and integrative analysis of large gene lists using\nDAVID bioinformatics resources. *Nat. Protoc.*, 4(1):44–57, 2009.

[53] W. Huber, V. J. Carey, R. Gentleman, S. Anders, M. Carlson, B. S. Carvalho, H. C. Bravo, S. Davis, L. Gatto, T. Girke, R. Gottardo, F. Hahne, K. D. Hansen, R. A. Irizarry, M. Lawrence, M. I. Love, J. MaCdonald, V. Obenchain, A. K. Oles, H. Pagès, A. Reyes, P. Shannon, G. K. Smyth, D. Tenenbaum, L. Waldron, and M. Morgan. Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods*, 12(2):115–121, 2015.

[54] P. J. Hurd and C. J. Nelson. Advantages of next-generation sequencing versus the microarray in epigenetic research. *Briefings in Functional Genomics and Proteomics*, 8(3):174–183, 2009.

[55] D. H. Huson and D. Bryant. Application of phylogenetic networks in evolutionary studies, 2006.

[56] K. Inoue, S. Shimozono, H. Yoshida, and H. Kurata. Application of approximate pattern matching in two dimensional spaces to grid layout for biochemical network maps. *PLoS ONE*, 7(6), 2012.

[57] A. V. Ivanova, C. M. Goparaju, S. V. Ivanov, D. Nonaka, C. Cruz, A. Beck, F. Lonardo, A. Wali, and H. I. Pass. Protumorigenic role of HAPLN1 and its IgV domain in malignant pleural mesothelioma. *Clinical Cancer Research*, 15(8):2602–2611, 2009.

[58] D. Jung, B. Kim, R. J. Freishtat, M. Giri, E. Hoffman, and J. Seo. MiRTarVis: An interactive visual analysis tool for microRNA-mRNA expression profile data. *BMC Proceedings*, 9, 2015.

[59] P. J. Kaboli, A. Rahmat, P. Ismail, and K.-H. Ling. MicroRNA-based therapy and breast cancer: A comprehensive review of novel therapeutic strategies from diagnosis to treatment. *Pharmacological Research*, 97:104–121, 2015.

[60] S. Kadri, V. Hinman, and P. V. Benos. HHMMiR: Efficient de novo prediction of microRNAs using hierarchical hidden Markov models. In *BMC Bioinformatics*, volume 10, 2009.

[61] T. Kamada and S. Kawai. An algorithm for drawing general undirected graphs. *Information Processing Letters*, 31(1):7–15, 1989.

[62] K. D. Kasschau, Z. Xie, E. Allen, C. Llave, E. J. Chapman, K. A. Krizan, and J. C. Carrington. P1/HC-Pro, a viral suppressor of RNA silencing, interferes with Arabidopsis development and miRNA function, 2003.

[63] J. J. B. Keurentjes, J. Fu, I. R. Terpstra, J. M. Garcia, G. van den Ackerveken, L. B. Snoek, A. J. M. Peeters, D. Vreugdenhil, M. Koornneef, and R. C. Jansen. Regulatory network construction in Arabidopsis by using genome-wide gene expression quantitative trait loci. *Proceedings of the National Academy of Sciences of the United States of America*, 104(5):1708–13, 2007.

[64] Y. Kishida, A. Natsume, Y. Kondo, I. Takeuchi, B. An, Y. Okamoto, K. Shinjo, K. Saito, H. Ando, F. Ohka, Y. Sekido, and T. Wakabayashi. Epigenetic subclassification of meningiomas based on genome-wide DNA methylation analyses. *Carcinogenesis*, 33(2):436–441, 2012.

[65] H. Kitano. Systems biology: a brief overview. *Science (New York, N.Y.)*, 295(5560):1662–4, 2002.

[66] K. Kojima, M. Nagasaki, E. Jeong, M. Kato, and S. Miyano. An efficient grid layout algorithm for biological networks utilizing various biological attributes. *BMC Bioinformatics*, 8, 2007.

[67] A. M. KRICHEVSKY. A microRNA array reveals extensive regulation of microRNAs during brain development. *RNA*, 9(10):1274–1281, 2003.

[68] R. Kumar, P.-J. Lahtvee, and J. Nielsen. *Systems biology: Developments and applications*. 2014.

[69] V. Kumar, M. Baweja, P. K. Singh, and P. Shukla. Recent Developments in Systems Biology and Metabolic Engineering of Plant–Microbe Interactions. *Frontiers In Plant Science*, 7(1421):1–12, 2016.

[70] K. Lakshmanan, A. Prakasham Peter, S. Mohandass, S. Varadharaj, U. Lakshmanan, and P. Dharmar. SynRio: R and Shiny based application platform for cyanobacterial genome analysis. *Bioinformation*, 11(119):973–2063, 2015.

[71] G. Lawyer. Understanding the influence of all nodes in a network. *Scientific Reports*, 5, 2015.

[72] B. Lehne and T. Schlitt. Protein-protein interaction databases: keeping up with growing interactomes. *Human genomics*, 3(3):291–297, 2009.

[73] B. P. Lewis, C. B. Burge, and D. P. Bartel. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, 120(1):15–20, 2005.

[74] B. P. Lewis, I. H. Shih, M. W. Jones-Rhoades, D. P. Bartel, and C. B. Burge. Prediction of Mammalian MicroRNA Targets. *Cell*, 115(7):787–798, 2003.

[75] W. Li, J. Cooper, L. Zhou, C. Yang, H. Erdjument-Bromage, D. Zagzag, M. Snuderl, M. Ladanyi, C. O. Hanemann, P. Zhou, M. A. Karajannis, and F. G. Giancotti. Merlin/NF2 Loss-Driven Tumorigenesis Linked to CRL4DCAF1-Mediated Inhibition of the Hippo Pathway Kinases Lats1 and 2 in the Nucleus. *Cancer Cell*, 26(1):48–60, 2014.

[76] W. Li and H. Kurata. A grid layout algorithm for automatic drawing of biochemical networks. *Bioinformatics*, 21(9):2036–2042, 2005.

[77] W. Li, Y. J. Suh, and J. Zhang. Does logarithm transformation of microarray data affect ranking order of differentially expressed genes? In *Annual International Conference of the IEEE Engineering in Medicine and Biology - Proceedings*, pages 6593–6596, 2006.

[78] C. Llave, Z. Xie, K. D. Kasschau, and J. C. Carrington. Cleavage of Scarecrow-like mRNA targets directed by a class of Arabidopsis miRNA. *Science*, 297(5589):2053–2056, 2002.

[79] N. Ludwig, Y. J. Kim, S. C. Mueller, C. Backes, T. V. Werner, V. Galata, E. Sartorius, R. M. Bohle, A. Keller, and E. Meese. Posttranscriptional deregulation of signaling pathways in meningioma subtypes by differential expression of miRNAs. *Neuro-Oncology*, 17(9):1250–1260, 2015.

[80] W. Luo, C. Brouwer, M. S. Friedman, K. Shedden, K. D. Hankenson, P. J. Woolf, D. Attali, E. Bailey, M. Lawrence, W. Huber, H. Pagès, P. Aboyoun, M. Carlson, R. Gentleman, M. Morgan, V. Carey, L. Trestle Technology, and M. Futschik. shinyjs: Perform Common JavaScript Operations in Shiny Apps using Plain R Code. *{PLoS} Computational Biology*, 9(8):1830–1831, 2013.

[81] S. Maere, K. Heymans, and M. Kuiper. BiNGO: A Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks. *Bioinformatics*, 21(16):3448–3449, 2005.

[82] R. J. Mailloux, R. Bériault, J. Lemire, R. Singh, D. R. Chénier, R. D. Hamel, and V. D. Appanna. The tricarboxylic acid cycle, an ancient metabolic network with a novel twist. *PLoS ONE*, 2(8), 2007.

[83] Mark Edmondson. googleAuthR: Authenticate and Create Google APIs.

[84] V. Marx. The big challenges of big data. *Nature*, 498(7453):255–260, 2013.

[85] N. Matloff. The Art of R Programming. *Book*, page 373, 2011.

[86] P. J. McMurdie and S. Holmes. Shiny-phyloseq: Web application for interactive microbiome analysis with provenance tracking. *Bioinformatics*, 31(2):282–283, 2015.

[87] O. Mersmann. R: microbenchmark, 2014.

[88] H. Mi, X. Huang, A. Muruganujan, H. Tang, C. Mills, D. Kang, and P. D. Thomas. PANTHER version 11: Expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Research*, 45(D1):D183–D189, 2017.

[89] E. Mikaelsson, A. Österborg, M. Jeddi-Tehrani, P. Kokhaei, M. Ostadkarampour, R. Hadavi, M. Gholamin, M. Akhondi, F. Shokri, H. Rabbani, and H. Mellstedt. A proline/arginine-rich end leucine-rich repeat protein (PRELP) variant is uniquely expressed in chronic lymphocytic leukemia cells. *PloS one*, 8(6):e67601, 2013.

[90] K. C. Miranda, T. Huynh, Y. Tay, Y. S. Ang, W. L. Tam, A. M. Thomson, B. Lim, and I. Rigoutsos. A Pattern-Based Method for the Identification of MicroRNA Binding Sites and Their Corresponding Heteroduplexes. *Cell*, 126(6):1203–1217, 2006.

[91] P. Miranda, M. Baptista, and S. Pinto. Analysis of communities in a mythological social network. *arXiv preprint arXiv:1306.2537*, 2013.

[92] J. H. Morris, L. Apeltsin, A. M. Newman, J. Baumbach, T. Wittkop, G. Su, G. D. Bader, and T. E. Ferrin. ClusterMaker: A multi-algorithm clustering plugin for Cytoscape. *BMC Bioinformatics*, 12, 2011.

[93] D. W. Mueller, M. Rehli, and A. K. Bosserhoff. MiRNA expression profiling in melanocytes and melanoma cell lines reveals miRNAs associated with formation and progression of malignant melanoma. *Journal of Investigative Dermatology*, 129(7):1740–1751, 2009.

[94] I. Naguibneva, M. Ameyar-Zazoua, A. Polesskaya, S. Ait-Si-Ali, R. Groisman, M. Souidi, S. Cuvellier, and A. Harel-Bellan. The microRNA miR-181 targets the homeobox protein Hox-A11 during mammalian myoblast differentiation. *Nature Cell Biology*, 8(3):278–284, 2006.

[95] S. Nam, B. Kim, S. Shin, and S. Lee. miRGator: An integrated system for functional annotation of microRNAs. *Nucleic Acids Research*, 36(SUPPL. 1), 2008.

[96] S. P. Nana-Sinkam and C. M. Croce. MicroRNA dysregulation in cancer: opportunities for the development of microRNA-based drugs. *IDrugs*, 13(12):843–846, 2010.

[97] M. Newton, C. Kendziorski, C. Richmond, F. Blattner, and K. Tsui. On Differential Variability of Expression Ratios: Improving Statistical Inference about Gene Expression Changes from Microarray Data. *Journal of Computational Biology*, 8(1):37–52, 2001.

[98] G. A. Pavlopoulos, M. Secrier, C. N. Moschopoulos, T. G. Soldatos, S. Kossida, J. Aerts, R. Schneider, and P. G. Bagos. Using graph theory to analyze biological networks, 2011.

[99] R. D. Peng. R Programming for Data Science. *Aging*, 7(11):132, 2015.

[100] A. Perry, S. L. Stafford, B. W. Scheithauer, V. J. Suman, and C. M. Lohse. Meningioma grading: An analysis of histologic parameters. *American Journal of Surgical Pathology*, 21(12):1455–1465, 1997.

[101] S. Pfeffer, K. Grossmann, P. Cassidy, C. Yang, M. Fan, L. Kopelovich, S. Leachman, and L. Pfeffer. Detection of Exosomal miRNAs in the Plasma of Melanoma Patients. *Journal of Clinical Medicine*, 4(12):2012–2027, 2015.

[102] R. S. PILLAI. MicroRNA function: Multiple mechanisms for a tiny RNA? *RNA*, 11(12):1753–1761, 2005.

[103] S. Povey, R. Lovering, E. Bruford, M. Wright, M. Lush, and H. Wain. The HUGO Gene Nomenclature Committee (HGNC), 2001.

[104] C. C. Pritchard, H. H. Cheng, and M. Tewari. MicroRNA profiling: Approaches and considerations, 2012.

[105] M. Pröschel, R. Detsch, A. R. Boccaccini, and U. Sonnewald. Engineering of Metabolic Pathways by Artificial Enzyme Channels. *Frontiers in Bioengineering and Biotechnology*, 3, 2015.

[106] S. Pu, J. Wong, B. Turner, E. Cho, and S. J. Wodak. Up-to-date catalogues of yeast protein complexes. *Nucleic Acids Research*, 37(3):825–831, 2009.

[107] J. Quackenbush. Microarray data normalization and transformation, 2002.

[108] N. R. Quinn and M. A. Breuer. A Forced Directed Component Placement Procedure for Printed Circuit Boards. *IEEE Transactions on Circuits and Systems*, 26(6):377–388, 1979.

[109] N. Rajewsky. Microrna target predictions in animals. *Nature Genetics*, 38(6S):S8–S13, 2006.

[110] R. Rauscher. Cloud computing considerations for biomedical applications. In *Proceedings - 2012 IEEE 2nd Conference on Healthcare Informatics, Imaging and Systems Biology, HISB 2012*, page 142, 2012.

[111] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A. L. Barabási. Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586):1551–1555, 2002.

[112] R. R. S. Reddy and M. V. Ramanujam. High throughput sequencing-based approaches for gene expression analysis. In *Methods in Molecular Biology*, volume 1783, pages 299–323. 2018.

[113] M. REHMSMEIER. Fast and effective prediction of microRNA/target duplexes. *RNA*, 10(10):1507–1517, 2004.

[114] D. R. Rhodes, J. Yu, K. Shanker, N. Deshpande, R. Varambally, D. Ghosh, T. Barrette, A. Pander, and A. M. Chinnaiyan. ONCOMINE: A Cancer Microarray Database and Integrated Data-Mining Platform. *Neoplasia*, 6(1):1–6, 2004.

[115] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research*, 43(7):e47, 2015.

[116] C. G. Rivera, R. Vakil, and J. S. Bader. NeMo: Network module identification in cytoscape. *BMC Bioinformatics*, 11(SUPPLL.1), 2010.

[117] E. D. Robertson, C. Wasylyk, T. Ye, A. C. Jung, and B. Wasylyk. The oncogenic microRNA hsa-miR-155-5p targets the transcription factor ELK3 and links it to the hypoxia response. *PLoS ONE*, 9(11), 2014.

[118] C. Sabatti. Co-expression pattern from DNA microarray experiments as a tool for operon prediction. *Nucleic Acids Research*, 30(13):2886–2893, 2002.

[119] L. Salwinski and D. Eisenberg. Computational methods of analysis of protein-protein interactions, 2003.

[120] A. Sandoval-Bórquez, I. Polakovicova, N. Carrasco-Véliz, L. Lobos-González, I. Riquelme, G. Carrasco-Avino, C. Bizama, E. Norero, G. I. Owen, J. C. Roa, and A. H. Corvalán. MicroRNA-335-5p is a potential

suppressor of metastasis and invasion in gastric cancer. *Clinical Epigenetics*, 9(1), 2017.

[121] L. F. Sempere, S. Freemantle, I. Pitha-Rowe, E. Moss, E. Dmitrovsky, and V. Ambros. Expression profiling of mammalian microRNAs uncovers a subset of brain-expressed microRNAs with possible roles in murine and human neuronal differentiation. *Genome Biology*, 5(3):R13, 2004.

[122] N. Servant, E. Gravier, P. Gestraud, C. Laurent, C. Paccard, A. Biton, I. Brito, J. Mandel, B. Asselain, E. Barillot, and P. Hupé. EMA - A R package for Easy Microarray data analysis. *BMC Research Notes*, 3, 2010.

[123] P. Shahi, S. Loukianiouk, A. Bohne-Lang, M. Kenzelmann, S. Küffer, S. Maertens, R. Eils, H.-J. Gröne, N. Gretz, and B. Brors. Argonaute–a database for gene regulation by mammalian microRNAs. *Nucleic acids research*, 34(Database issue):D115–8, 2006.

[124] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: A software Environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11):2498–2504, 2003.

[125] S. I. Sheikh, T. Kahveci, S. Ranka, and J. Gordon Burleigh. Stability analysis of phylogenetic trees. *Bioinformatics*, 29(2):166–174, 2013.

[126] S. K. Shenouda and S. K. Alahari. MicroRNA function in cancer: Oncogene or a tumor suppressor?, 2009.

[127] Q. Shi, W. Wang, Z. Jia, P. Chen, K. Ma, and C. Zhou. ISL1, a novel regulator of CCNB1, CCNB2 and c-MYC genes, promotes gastric cancer cell proliferation and tumor growth. *Oncotarget*, 7(24):36489–36500, 2016.

[128] X.-B. Shi, L. Xue, a H Ma, C. G. Tepper, R. Gandour-Edwards, H.-J. Kung, and R. W. deVere White. Tumor suppressive miR-124 targets androgen receptor and inhibits proliferation of prostate cancer cells. *Oncogene*, (October 2012):4130–4138, 2012.

[129] A. Slipicevic and M. Herlyn. KIT in melanoma: Many shades of gray, 2015.

[130] P. Stanhope-Baker, P. M. Kessler, W. Li, M. L. Agarwal, and B. R. G. Williams. The Wilms tumor suppressor-1 target gene podocalyxin is transcriptionally repressed by p53. *Journal of Biological Chemistry*, 279(32):33575–33585, 2004.

[131] G. Su, J. H. Morris, B. Demchak, and G. D. Bader. Biological Network Exploration with Cytoscape 3. *Current Protocols in Bioinformatics*, 2014:8.13.1–8.13.24, 2014.

[132] G. Tang, B. J. Reinhart, D. P. Bartel, and P. D. Zamore. A biochemical framework for RNA silencing in plants. *Genes and Development*, 17(1):49–63, 2003.

[133] D. W. Thomson, C. P. Bracken, and G. J. Goodall. Experimental strategies for microRNA target identification, 2011.

[134] M. Visani, G. Acquaviva, G. Marucci, M. Ragazzi, E. Fraces-Chi, A. A. Brandes, G. Tallini, A. Pession, and D. De Biase. International Journal of Brain Disorders and Treatment MicroRNA in Brain Neoplasia: A Review. *Int J Brain Disord Treat*, 1:1, 2015.

[135] T. Wang, X. Zhang, L. Obijuru, J. Laser, V. Aris, P. Lee, K. Mittal, P. Soteropoulos, and J. J. Wei. A micro-RNA signature associated with race, tumor size, and target gene activity in human uterine leiomyomas. *Genes Chromosomes Cancer*, 46(4):336–347, 2007.

[136] X. Wang, Y. Liu, X. Liu, J. Yang, G. Teng, L. Zhang, and C. Zhou. MIR-124 inhibits cell proliferation, migration and invasion by directly targeting SOX9 in lung adenocarcinoma. *Oncology Reports*, 35(5):3115–3121, 2016.

[137] J. Wiemels, M. Wrensch, and E. B. Claus. Epidemiology and etiology of meningioma, 2010.

[138] D. Xiao, J. Ohlendorf, Y. Chen, D. D. Taylor, S. N. Rai, S. Waigel, W. Zacharias, H. Hao, and K. M. McMasters. Identifying mRNA, MicroRNA and Protein Profiles of Melanoma Exosomes. *PLoS ONE*, 7(10), 2012.

[139] F. Xiao, Z. Zuo, G. Cai, S. Kang, X. Gao, and T. Li. miRecords: An integrated resource for microRNA-target interactions. *Nucleic Acids Research*, 37(SUPPL. 1), 2009.

[140] D. Xu, J. Tan, M. Zhou, B. Jiang, H. Xie, X. Nie, K. Xia, and J. Zhou. Let-7b and microRNA-199a inhibit the proliferation of B16F10 melanoma cells. *Oncology Letters*, 4(5):941–946, 2012.

[141] Z. Yan, Y. Xiong, W. Xu, J. Gao, Y. Cheng, Z. Wang, F. Chen, and G. Zheng. Identification of hsa-miR-335 as a prognostic signature in gastric cancer. *PLoS ONE*, 7(7), 2012.

[142] C. Yau, L. Esserman, D. H. Moore, F. Waldman, J. Sninsky, and C. C. Benz. A multigene predictor of metastatic outcome in early stage hormone receptor-negative and triple-negative breast cancer. *Breast Cancer Research*, 12(5), 2010.

[143] S.-H. Yook, Z. N. Oltvai, and A.-L. Barabási. Functional and topological characterization of protein interaction networks. *PROTEOMICS*, 4(4):928–942, 2004.

[144] M. Zahurak, G. Parmigiani, W. Yu, R. B. Scharpf, D. Berman, E. Schaeffer, S. Shabbeer, and L. Cope. Pre-processing Agilent microarray data. *BMC Bioinformatics*, 8, 2007.

[145] D. Zhang, R. Zhu, H. Zhang, C. H. Zheng, and J. Xia. MGDB: A comprehensive database of genes involved in melanoma. *Database*, 2015, 2015.

[146] Y. Zhao and D. Srivastava. A developmental view of microRNA function, 2007.

[147] R. Zhou and E. A. Hansen. Breadth-first heuristic search. *Artificial Intelligence*, 170(4-5):385–408, 2006.

APPENDIX A

EK A

A.1 Copyright License Numbers

Some figures used are subjected to copyright permission. Usage of these figures for this document is licensed. Copyright license numbers for permission obtained are written below.

Figure 1.1 - Thesis / Dissertation Reuse. Open Access. Does not require a formal reuse license

Figure 1.2 - Copyright Clearance Centers RightsLink License Number : 4385281085963

Figure 2.2 - RightsLink / Springer Nature Copyright Clearance Center - Order Number: 4365920061294

Figure 2.3 - RightsLink / Springer Nature Copyright Clearance Center - Order Number: 4365870085947

Figure 2.4 - Copyright Clearance Centers RightsLink License Number: 4381900146730

Figure 2.5 - Copyright Clearance Centers RightsLink License Number: 4381900146730

Figure 2.6 - Copyright Clearance Centers RightsLink License Number: 4381951279056

Figure 2.7 - Copyright Clearance Centers RightsLink License Number : 4381961511695

Figure 2.8 - Thesis / Dissertation Reuse. The IEEE does not require individuals working on a thesis to obtain a formal reuse license

Figure 2.9 - Thesis / Dissertation Reuse. Does not require individuals working on a thesis to obtain a formal reuse license