


1-1-2017

Integrative Pathway Analysis Pipeline For Mirna And Mrna Data

Diana Mabel Diaz Herrera
Wayne State University,

Follow this and additional works at: https://digitalcommons.wayne.edu/oa_theses

 Part of the [Biostatistics Commons](#), [Computer Engineering Commons](#), and the [Computer Sciences Commons](#)

Recommended Citation

Diaz Herrera, Diana Mabel, "Integrative Pathway Analysis Pipeline For Mirna And Mrna Data" (2017). *Wayne State University Theses*. 559.
https://digitalcommons.wayne.edu/oa_theses/559

This Open Access Thesis is brought to you for free and open access by DigitalCommons@WayneState. It has been accepted for inclusion in Wayne State University Theses by an authorized administrator of DigitalCommons@WayneState.

**INTEGRATIVE PATHWAY ANALYSIS PIPELINE FOR miRNA AND
mRNA DATA**

by

DIANA DIAZ

THESIS

Submitted to the Graduate School

of Wayne State University,

Detroit, Michigan

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

2016

MAJOR: COMPUTER SCIENCE (Bioinformatics)

Approved By:

Advisor

Date

DEDICATION

To my parents and brothers who are my everything.

ACKNOWLEDGEMENTS

I would like to express my gratitude to my advisor, Dr. Sorin Draghici, who generously offered his always wise guidance. I am very thankful for his support and his time. Additionally, I would like to thank Dr. Nathan Fisher and Dr. Alexander Kotov for agreeing to serve on my dissertation committee.

I also want thank to my fellow colleagues and friends in the Intelligent Systems and Bioinformatics Laboratory: Samer Hanoudi, Cristina Mitrea, Sahar Ansari, Azam Peyvandipour, Nafiseh Saberian, Adib Shafi, Becky Tagett, and Behzad Bokanizad for their encouragement. Special thanks to Tin Nguyen and Michele Donato who reviewed my work and provided me useful feedback to make it better.

On a personal level, I would like to thank my friends and family Sumukhi Chandrashekar, Tizita Zewdie, Narimar Ammar, Marcia Arenas, Fabio Navarrete, Natty Sanchez, and Davide Lattanzi for their support and encouragement. I am also thankful to the officers of ACM especially Ridwan Khan, Connor Tukel, and Omer Khan who showed interest in my work and learned how to analyze data using my tool during research club.

Finally, I want to thank the NCBI for the Gene Expression Omnibus (GEO) repository, as well as, the authors and patients who contributed to the GEO database, as well as KEGG and mirTarbase databases. I also want to thank the anonymous reviewers of Bioconductor and PSB 2017 for their valuable comments. I thank the users of my tool that have downloaded it over a thousand times and provided positive feedback.

TABLE OF CONTENTS

Dedication	ii
Acknowledgements	iii
List of Tables	v
List of Figures	v
Chapter 1: Introduction	1
Chapter 2: Related work	5
2.1 Biological background	5
2.2 Current pathway analysis methods	11
2.2.1 Integration of multi-omics data	15
2.3 Pathway analysis	17
2.3.1 Pathway analysis using gene expression	17
2.3.2 Pathway analysis using multi-omics data	21
2.3.3 Topology-aware integrative methods	22
2.4 Omics integration for disease subtyping	27
Chapter 3: Integrating signaling pathways with miRNA target genes: our proposed method	32
3.1 Proposed integrative pathway analysis pipeline	32
3.2 Pathway analysis pipeline	32
3.2.1 Impact analysis using mRNA and miRNA	34
3.3 Proposed algorithm for pathway augmentation	35
Chapter 4: Proposed miRNA augmented pathways database	37
4.1 Databases needed for integration	37
4.1.1 Database of signaling pathways: KEGG pathways	37
4.1.2 Database of miRNA target interactions: mirTarBase	38
4.2 Data representation of the augmented pathways	38
4.2.1 Using the augmented pathways: an example	39

4.3	Proposed graphical visualization of the pathways	40
4.4	Script to augment signaling pathways with miRNA	42
	Chapter 5: Software Implementation	44
5.1	Software features	44
5.2	Pathway analysis in R and Bioconductor	44
5.3	Example of pathway analysis of miRNA and mRNA data	47
	Chapter 6: Method Validation	50
6.1	Validation outline	50
6.2	Descriptive statistics of the augmented pathways	50
6.3	Results	52
	Chapter 7: Reference Manual	59
	Chapter 8: Discussion and Conclusion	76
	References	78
	Abstract	99
	Autobiographical Statement	100

LIST OF TABLES

Table. 2.1	Data types available on TCGA.	16
Table. 6.1	Description of the analyzed datasets	52
Table. 6.2	Results of target pathway identification	55

LIST OF FIGURES

Figure. 2.1	The central dogma of molecular biology.	6
Figure. 2.2	DNA structure and sequence representation. DNA sequence is typically represented in a linear format as a sequence of nucleotides: adenine (A), cytosine (C), guanine (G), or thymine (T).	7
Figure. 2.3	DNA is transcribed to mRNA. a) DNA. b) mRNA structure. c) mRNA nucleotides sequence. In transcription, each adenine (A), cytosine (C), and guanine (G) bases are copied identically. However, thymine (T) bases are copied as uracil (U) bases.	7
Figure. 2.4	Translation from mRNA to proteins.	9
Figure. 2.5	Gene product	10
Figure. 2.6	MicroRNAs bind to mRNA molecules and prevent translation. . .	10
Figure. 2.7	Classification of integrative methods	15
Figure. 2.8	Overview of multi-omics pathway topology techniques	22
Figure. 2.9	An example of factor graphs	25
Figure. 2.10	An example of the pathway model to compute IPA	26
Figure. 2.11	Algorithm to decide if an entity is active or not based on its labels	28
Figure. 3.1	Workflow of pathway analysis using augmented pathways.	33
Figure. 4.1	Model of a miRNA-augmented pathway. Portion of the <i>Colorectal Cancer</i> pathway from KEGG.	39
Figure. 4.2	Visualization of the augmented <i>Sulfur Relay System</i> pathway . .	41
Figure. 5.1	Portion of the miRNA-augmented <i>Colorectal Cancer</i> pathway. . .	46
Figure. 5.2	A screenshot of the mirIntegrator’s graphical user interface . . .	48
Figure. 6.1	Evaluation scheme.	51

Figure. 6.2	Comparison of pathway sizes before and after augmentation . . .	51
Figure. 6.3	Portion of the augmented <i>Amyotrophic Lateral Sclerosis</i> pathway	54
Figure. 6.4	FDR p-values and rankings of the target pathways	57

Chapter 1: Introduction

Cancer is a disease that involves genetic and environmental factors. Knowledge of the roles that genes play in a particular disease is rapidly helping us to understand cancer biology. These functions differ significantly, for example, some genes can contribute to determining the disease state (disease genes) while other genes can interact with particular environmental factors in causing cancer (susceptibility genes). Identifying the roles that genes play in a disease is not an easy task, it requires rigorous biological experiments followed by statistical and computational analyses to interpret the data. High-throughput technologies allow monitoring cell processes at the molecular level.

One of the molecules that are typically measured is ribonucleic acid (RNA), particularly messenger RNA (mRNA). The mRNA is used as a proxy to determine *gene expression*, i.e. the process by which a gene synthesizes to a gene product. These measurements are taken with the purpose of identifying if a gene is over-expressed or under-expressed. Using these technologies, conventional data analysis provides a list of differentially expressed (DE) genes. This analysis is done by comparing the *gene expression* from two groups and statistically identifying the genes that are significantly different between the groups, e.g. one group of healthy individuals versus one group of patients with the disease under study. Lists of DE genes are widely used. However, these lists often fail to elucidate the underlying biological mechanisms.

In the last couple of decades, several approaches have focused on the interactions between genes rather than the study of individual genes. These gene to gene interactions are captured as graphs, named *signaling pathways*, with genes as vertices and the types of interaction on the edges. Each signaling pathway describes a cellular process and contains the genes and interactions that are involved in this process. Researchers have been storing the knowledge about various pathways into

many publicly available databases, such as Kyoto Encyclopedia of Genes and Genomes (KEGG) [74], Biocarta [14], and Nature Pathway Interaction Database (NCI) [127]. Given the availability of such collection of pathways, researchers now could identify the pathways that are significantly impacted by a given condition. Identifying pathways instead of genes increases the explanatory power and gives us a better understanding of the underlying biological phenomenon [77, 79, 102]. Many *pathway analysis* methods have been developed to identify enriched or differentially regulated pathways [9, 12, 44, 47, 137]. These methods can be divided into three different categories: over-representation analysis (ORA), functional class scoring (FCS), and pathway topology methods (PT) [41, 120].

The over-representation analysis (ORA) [141] identifies the pathways with differentially expressed genes significantly greater than expected by chance. This approach ignores all the gene interactions and assumes gene independence, resulting in an incorrect hypothesis testing and thus leading to biased results. Functional class scoring (FCS) methods, such as Gene Set Enrichment Analysis (GSEA) [137] and Gene Set Analysis (GSA) [44], do not assume independence between genes [18, 35]. The hypothesis of FCS methods states that well-coordinated small changes in relevant genes can also have significant effects on pathways besides large changes in individual genes. However, these approaches still do not take into consideration the interactions between genes as described by the pathways, resulting in information loss which in turn leads to both false positives, as well as false negatives [41]. Topology-aware approaches, such as Impact Analysis [41, 140], analyze the pathways as graphs and take into consideration the type and direction of each gene-gene interaction.

Pathway analysis methods using gene expression (mRNA) have achieved remarkable results [9, 12, 44, 47, 77, 79, 102, 137]. However, mRNA alone is unable to capture the complete picture of cell processes, as other entities also play important roles. For instance, microRNAs (miRNAs) are newly discovered gene regulators that

play a crucial role in diagnosis and prognosis for different types of cancer [88]. miRNAs are small RNA molecules capable of suppressing protein production by binding to gene transcripts. In fact, more than 30% of the protein-coding genes in humans are miRNA-regulated [89]. Given all the evidence of the miRNA's relevance, hundreds of thousands of miRNA targeting genes interactions have been experimentally validated and collected in public databases such as mirTarBase [65], miRWalk 2.0 [42], miRecords [156], and TarBase 7.0 [129]. There are also several algorithms used to predict miRNA targets [72, 85, 89] and databases with predicted interactions such as miRanda [72], TargetScan [89], PicTar [85], and TargetRank [110].

In addition, relevant work has been done to elucidate the important interplay between miRNAs and biological pathways [4, 19, 64, 65, 107, 148]. These studies focus on different directions, some methods search for pathways that are targeted by a particular miRNA [4], and others perform pathway analysis using just miRNA expression, such as mirTar [64, 65] and DIANA-miRPath [148]. Other methods incorporate both mRNA and miRNA for pathway analysis [19, 107]. The earliest tool that implements mRNA-miRNA integration is the miRNA and mRNA integrated analysis (MMIA) [107] which performs Gene Set Analysis (GSA) of the down-regulated genes that are targeted by up-regulated miRNAs. However, as mentioned before, GSA does not take advantage of the knowledge captured by the pathway topology. The state-of-the-art approach for miRNA-mRNA pathway analysis method is microGraphite [19] which uses an empirical gene set approach. microGraphite's primary goal is the identification of signal transduction paths that are mostly correlated with the condition under study [97]. Functional analysis methods that include miRNA are still needed to enhance the knowledge on disease gene regulation [28].

The major drawback of current approaches is that most of them do not take into consideration the knowledge about the interactions between the genes, as well as between genes and miRNAs. In this thesis, we present mirIntegrator, a topology-aware

approach that systematically integrates miRNA and mRNA expressions to identify pathways that are significantly impacted by the studied condition. Our framework is flexible and allows users to integrate signaling pathway databases with miRNA-mRNA interaction databases to produce *miRNA-augmented pathways*. Here we show that pathway analysis performed on these *augmented pathways* offer more statistical power than performing analysis on gene-gene pathways. Our augmented pathways offer a more comprehensive view and a deeper understanding of complex diseases.

This thesis encloses three contributions: a tool for integrating miRNA into signaling pathways (mirIntegrator) [33], a publicly available miRNA-augmented pathway database (mirAP), and examples of applying such augmented pathways to pathway analysis [34, 32, 109]. Our pathway analysis pipeline uses mirAP and adapts Impact Analysis (IA) [41, 140], a topology-aware pathway analysis method previously developed by our group. To demonstrate the advantage of our method, we analyze nine real datasets studying seven different diseases with mRNA and miRNA expression. We show that the proposed approach is able to identify the pathways that describe the underlying conditions as significant. We compare our integrative method with the traditional Impact Analysis and the state-of-the-art approach microGraphite [19]. The proposed method produces p-values and rankings of the disease pathways significantly smaller than those obtained without data integration as well as those obtained using microGraphite.

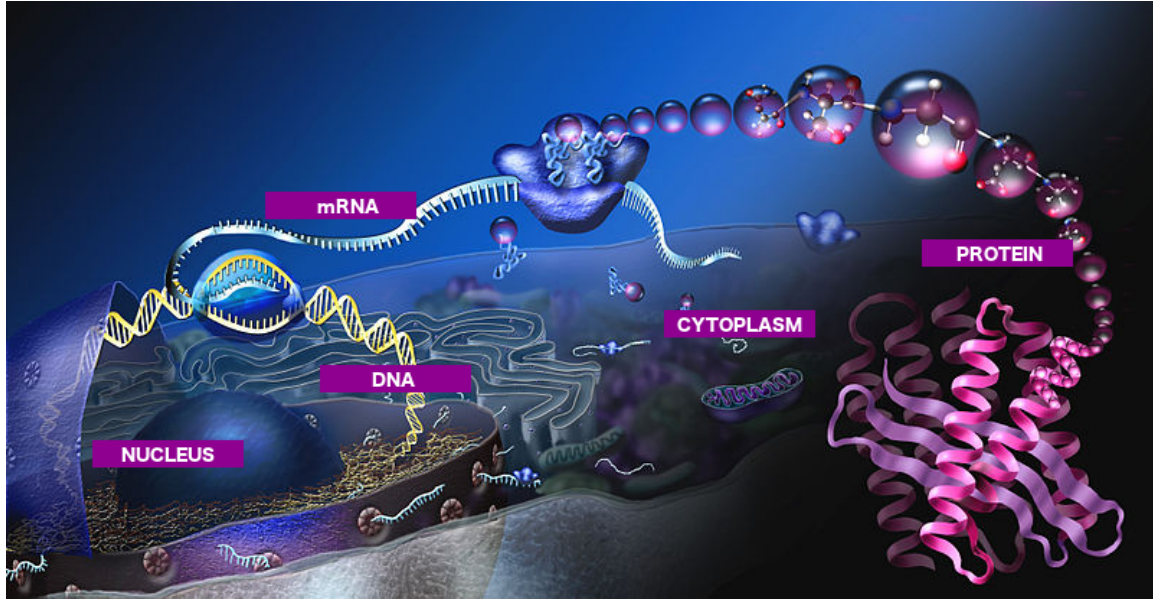
Chapter 2: Related work

2.1 Biological background

This section provides a basic background in molecular biology which is key to understand our proposed method. In particular, we present the central dogma of molecular biology and introduce the function of microRNAs (miRNAs).

Molecular biology can be defined as the study of life at the molecular level. It is an interdisciplinary approach that combines genetics (i.e. the study of heredity) and biochemistry (i.e. the study of the chemistry of living things). Here, we are interested in the molecular biology of the genes rather than other components of the cell, i.e. the study of genes, how they translate to proteins, and its clinical significance. For example, if we were interested in studying human skin cells we would consider both disciplines: genetics and biochemistry. The genetics of human skin would focus on identifying genes that have an influence on skin traits, such as the human TYRP1 gene and the mutations in this gene that are associated with oculocutaneous albinism type III [126] for example. The biochemistry of skin describes the chemical compounds found in the skin, such as melanosomal proteins: tyrosinase, tyrosinase-related protein 1, and DOPAchrome tautomerase [131]. Finally, a comprehensive study would include the TYRP1 gene, the regulatory mechanisms for which the gene translate into the tyrosinase-related protein 1, and how these gene and protein relate to the occurrence of oculocutaneous albinism type III [83].

Figure 2.1 illustrates a simple representation of the flow of genetic information from genes to proteins. This process is known as *the central dogma of molecular biology* which has two main steps: *translation* and *transcription*. First, a piece of information in the DNA (a gene) is *transcribed* into messenger-RNA (mRNA) in the cell nucleus. Then, the mRNA is transported to the cytoplasm to be *translated* into a polypeptide chain (protein) by the action of a ribosome and multiple transfer-RNAs.

Figure 2.1: The central dogma of molecular biology.

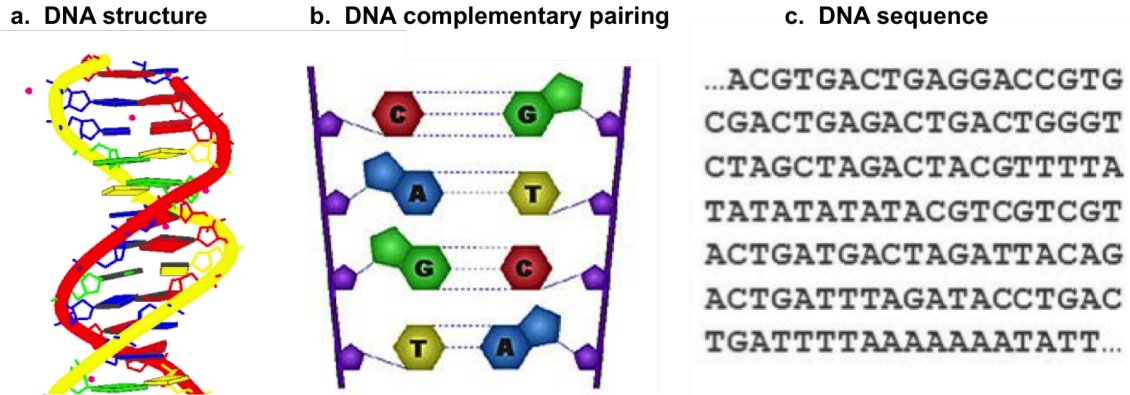
Credit: Nicolle Rager, National Science Foundation

The input material of transcription is deoxyribonucleic acid (DNA). The genetic information necessary for cell functioning is carried in the form of DNA which is made up of *nucleotides*. Each DNA nucleotide contains one of these four bases: adenine (A), cytosine (C), guanine (G), or thymine (T). These bases bind nucleic acids together by complementary pairing. Adenine base pairs with thymine and cytosine with guanine (see Figure 2.2.b). The DNA structure contains two strands of complementary nucleotide chains forming a double helix [153] as shown in Figure 2.2.a. Typically, DNA is represented in a linear format as a sequence of nucleotides (see Figure 2.2.c).

The output of transcription is ribonucleic acid (RNA). DNA is transcribed to messenger-RNA (mRNA) which is transported out of the nucleus. RNA is a single strand of nucleotides (Figure 2.3.b), where each RNA nucleotide contains one of these four nitrogen bases: adenine (A), cytosine (C), guanine (G), or uracil (U). In transcription, each thymine base is copied as an uracil base. Typically, mRNA is described in a linear format as a sequence of nucleotides (Figure 2.3.c).

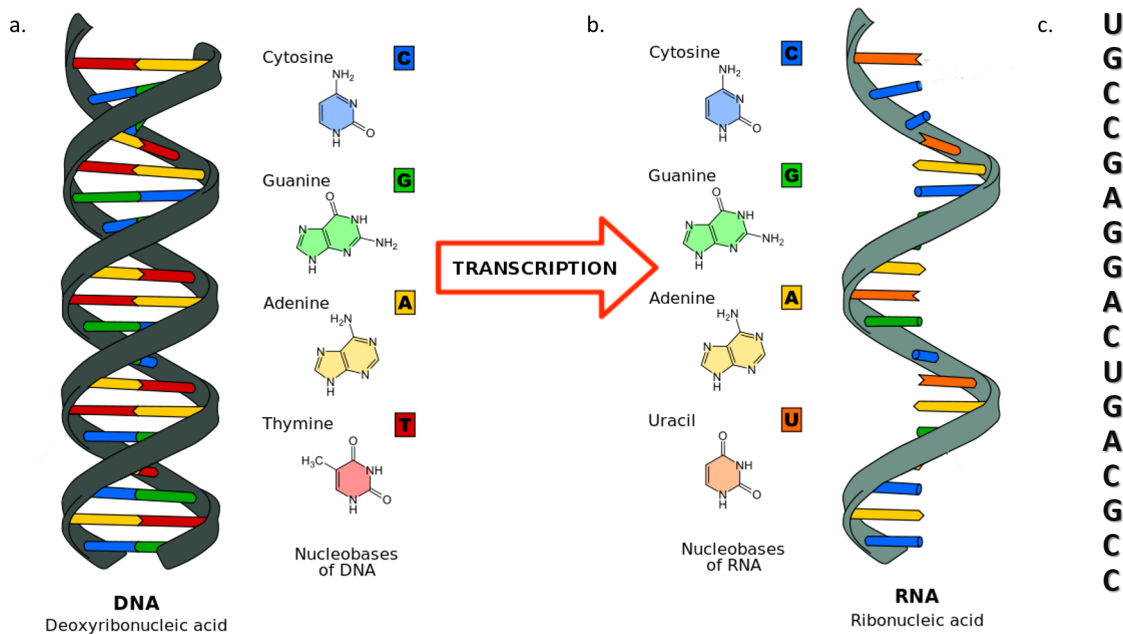
Each triplet of mRNA nucleotides, named *codon*, is translated to an amino acid

Figure 2.2: DNA structure and sequence representation. DNA sequence is typically represented in a linear format as a sequence of nucleotides: adenine (A), cytosine (C), guanine (G), or thymine (T).



Credit: The protein data bank (PDB:3BSE) and wiki commons.

Figure 2.3: DNA is transcribed to mRNA. a) DNA. b) mRNA structure. c) mRNA nucleotides sequence. In transcription, each adenine (A), cytosine (C), and guanine (G) bases are copied identically. However, thymine (T) bases are copied as uracil (U) bases.



Credit: Wikimedia commons.

(see Figure 2.4.a.). In humans, there are 20 types of amino acids, and each amino acid is mapped from more than one codon. Figure 2.4.b. displays a codon wheel that shows which codon encodes which amino acid. The inner circle is the first nucleotide in the codon, the second ring the second nucleotide and the third ring the third nucleotide. Amino acids are shown around the wheel. The amino acids translated from an mRNA strand bond together to form proteins, i.e. polypeptide chains. Proteins are involved in almost all functions in a cell.

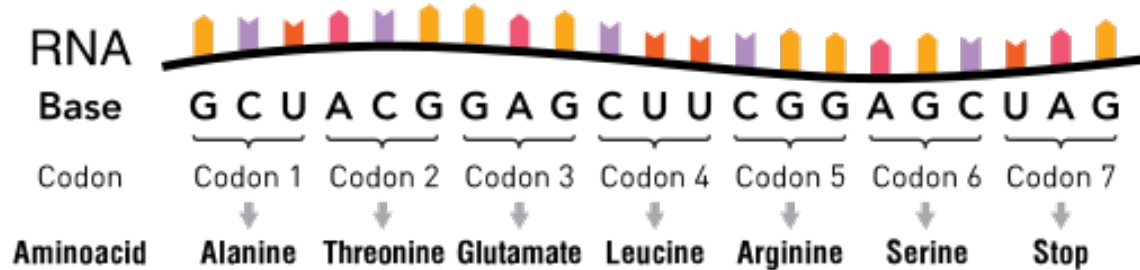
There are two main categories of genes: protein-coding genes and non-protein-coding genes (see Figure 2.5) Protein-coding genes are transcribed and then translated into protein. Non-protein-coding genes are transcribed but never translated; their final product is non-coding RNA (ncRNA). Gene expression is the process by which a particular gene information (DNA) is transformed to a gene product, i.e. either ncRNA or protein. The basic central dogma model does not include crucial ncRNAs, such as microRNAs (miRNAs).

microRNAs (miRNAs) are small RNA molecules of approximately 22 nucleotides capable of suppressing protein production by binding to gene transcripts (see Figure 2.6). In fact, more than 30% of the protein-coding genes in humans are miRNA regulated [90]. Additionally, miRNAs have been shown to play a significant role in diagnosis and prognosis for different types of diseases [88]. Several efforts have been done to identify mRNA-miRNA target interactions, i.e. which miRNAs regulate which genes. Most microRNA-target interactions are statistically predicted, and some are experimentally validated.

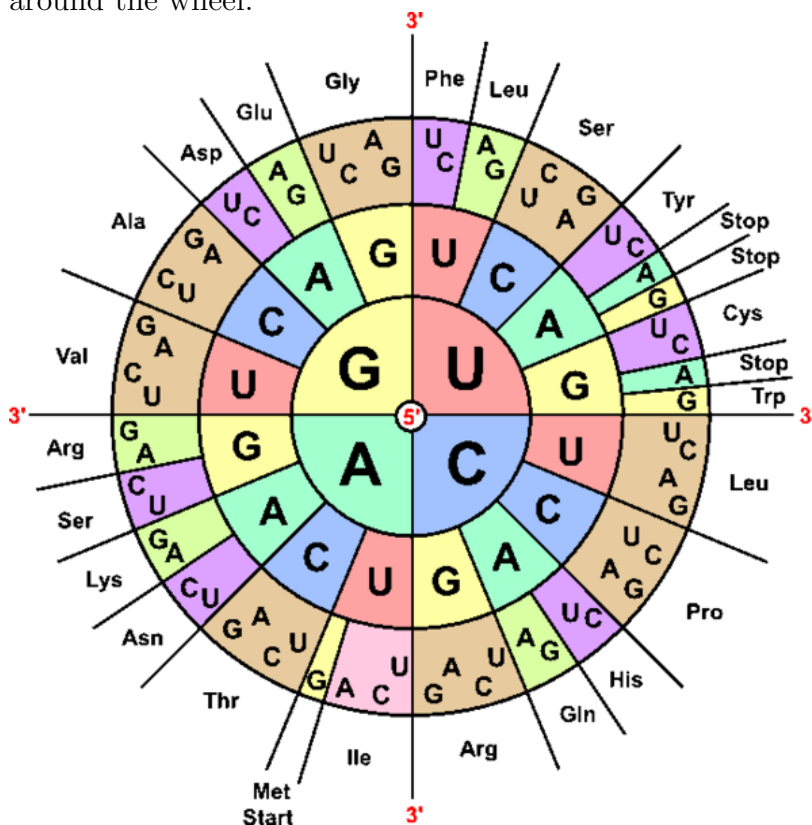
Given the importance of miRNAs, hundreds of thousands of miRNA-targeting-genes interactions have been experimentally validated and collected in public databases such as mirTarBase [65], miRWalk 2.0 [42], miRecords [156], and TarBase 7.0 [129]. There are also several algorithms used to predict miRNA targets [72, 90, 85] and databases with predicted interactions such as miRanda [72],

Figure 2.4: Translation from mRNA to proteins.

a. mRNA codons are translated to amino acids to form proteins.

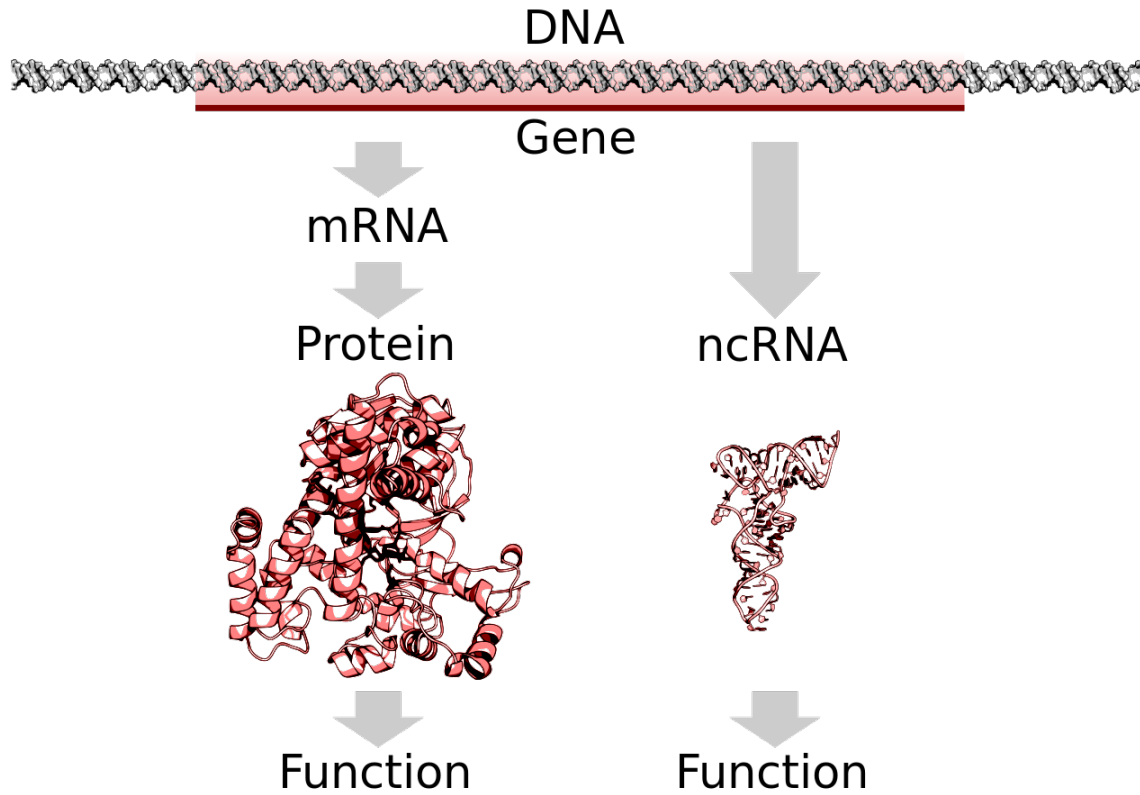


b. The standard genetic code (or translation table) shows which codon encodes which amino acid. The inner circle is the first nucleotide in the codon, the second ring the second nucleotide and the third ring the third nucleotide. Amino acids are shown around the wheel.



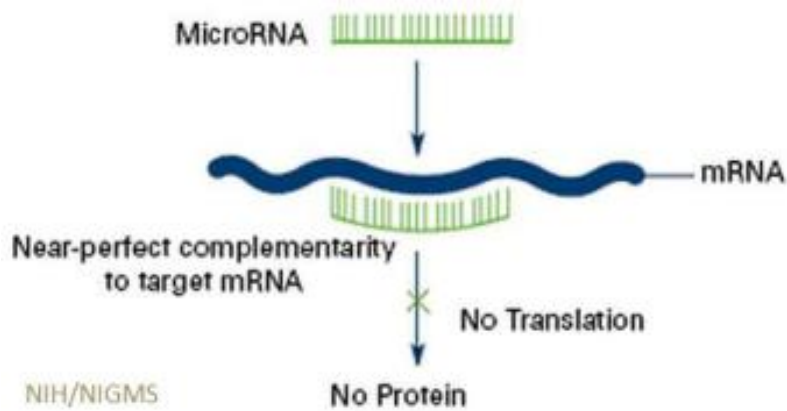
Credit: Wikimedia commons.

Figure 2.5: Gene product



Credit: Wikimedia commons.

Figure 2.6: MicroRNAs bind to mRNA molecules and prevent translation.



Credit: U.S. National Institute of General Medical Sciences

TargetScan [90], PicTar [85], and TargetRank [110]. There are also find miRNA-disease interaction databases [65, 70, 91] which are growing rapidly.

2.2 Current pathway analysis methods

High-throughput molecular biological methods perform thousands of simultaneous measurements of biological molecules to observe a particular state of cells. Recent technologies have extended the breadth of available high-throughput molecular biological data. Nowadays, most of the molecular data types are analyzed separately which has provided important discoveries, such as biomarker identification. However, analyzing various data types together can lead to a more consistent understanding of cell processes [50].

The term *high-throughput data* is used here as large measures of genetic data taken in a short time. These data are generated by different technologies commonly referred as “omics technologies” which are the foundation for systems biology [117]. Omics seek to quantify, describe, and identify all of the components of cellular systems with spatial and temporal dimensions [123]. There are several data types of high-throughput measurements from which four categories are the most important: proteomics, transcriptomics, metabolomics, and genomics [157]. Proteomics is the study of proteins present in cells. Transcriptomics measures all gene expression values. Metabolomics aims for the quantification and identification of metabolites. Genomics includes the large-scale genotyping of SNPs (single nucleotide polymorphisms). Each of these data types is unique and provide different perspectives on the cellular processes.

There are several computational solutions for analyzing omics data in isolated fashion [11]. However, single data type analyses have not given enough understanding for successfully perform disease diagnosis and treatment. Some of the ultimate goals for integrating multiple-omics are the identification of pathways relevant to a

condition and disease subtyping.

The identification of pathways that are involved in a particular phenotype is typically referred as *pathway analysis*. Identifying pathways that are relevant to a condition is important because it gives insights that can be used to further disease treatment or diagnosis. The standard input of pathway analysis techniques is the log-fold change of a large set of genes (around 25,000). Fold change is computed as the ratio of gene expression between two different groups, commonly one group of control subjects and another group with patients. The output of pathway analysis is a ranked list of statistically significant biological pathways. These pathways are considered to be related to the condition under study. Biological pathways are graphical representations of common knowledge about genes and their interaction with biological processes. In particular, signaling pathways are represented as graphs with a set of genes as nodes and the biochemical and physical interactions as edges. These pathways are typically made by mining the literature and then manually curating the retrieved information [74]. Signaling processes of the cell are captured in pathways that describe the interactions between protein-coding genes and DNA [79].

Disease subtyping is another important goal for omics integration. Generating clinical meaningful disease subtyping is critical for prognosis and further treatment determination. Based on statistical information and the patient's profile, the objective is to identify the subtype of disease that the patient more likely belongs to. The input for disease subtyping is molecular and clinical data from several patients with the same condition but have different outcomes. The expected output is well-identified groups that highly correlate with the observed outcomes (e.g. a group of long-term survival patients and another group of short-term survival patients). It is also important to identify possible patterns that are shared among members of each subtype and differences with other subtypes. This is commonly expressed as a clustering problem where the main goal is to search for similarities among the data points.

All these applications show how important is integrating various biomolecular data types. There are more applications of data integration, such as signaling networks reconstruction [30, 53, 81, 105] and biological networks visualization [133], but here we focus on pathway analysis.

From the computer science perspective, the term data integration refers to the integration of fragmented information from different physical databases or data warehouses and different representations. Several authors have proposed platforms and languages to integrate databases (typically using XML) [1]. Even though data fragmentation is a significant problem, we are not studying here that type of data integration. In bioinformatics, the terms data integration and data fusion are synonymous. In computer science, data fusion is referred as the process of integrating information acquired from various heterogeneous types into a single compound knowledge. Here, we define data integration and data fusion as the integration of knowledge without focusing on the representation. Additionally, data fusion is valuable for acquiring more reliable information than the raw measurements from a single type of source. The primary issue in data fusion (DF) is to provide fused data with increased correctness, conciseness, and completeness when compared with the original disjoint data. Correctness measures whether the fused data conform to the reality of the object under study. This occurs when more than one data source can confirm the same hypothesis which increases the confidence of the data. Conciseness refers to the reduction of ambiguity which means that the fused data from multiple sources have decreased the set of hypotheses about the object of study. Finally, completeness measures the amount of information from the fused data which increases the robustness because one measurement can contribute information where others measurements are incapable. To make this process successful, we need to define an outline for resolving conflicts. Data conflicts can occur when there is uncertainty or when there are contradictions. Un-

certainty occurs when there is missing information, such as gene levels not included in the measured platform, or in a particular sample. A contradiction occurs when there is not missing information, but the information that we can extract from one source is completely different to the one that can be obtained from another source.

Data fusion techniques have been applied mainly to the graphical computation context. Numerous data fusion algorithms have been developed giving users different levels of detail. Some technologies have been categorized by the USA Joint Directors of Laboratories [92] into three basic levels according to the amount of information that they provide. The first level corresponds to raw data uncorrelated. The second level, or feature level, provides a greater degree of inference and some interpretative meaning. Finally, the third level, or decision level, delivers additional explanatory meaning. It is designed to provide recommendations to users.

In the high-throughput biomolecular data context, data integration is typically performed in four different manners. One is to analyze each data type separately first and then integrate the final findings. Another manner is to pre-process each type of data independently, then perform cross-platform normalization across the data types, then combine the normalized figures and finally perform an overall analysis. The third type of integration consists of performing a statistical integration. The fourth approach is to integrate the data by modeling the data types based on the biological meaning of the molecules and their interactions (see Figure 2.7).

For example, researchers have integrated mRNA and microRNA paired data by analyzing each data type independently and then interpreting the results manually [25]. Sometimes the results of these experiments can lead to conflicting and unexplained outcomes. A second scenario is given when researchers having sample-paired data decide to merge the two data tables into a single table and analyze this new merged table. This practice requires cross-normalization, and it is very dangerous

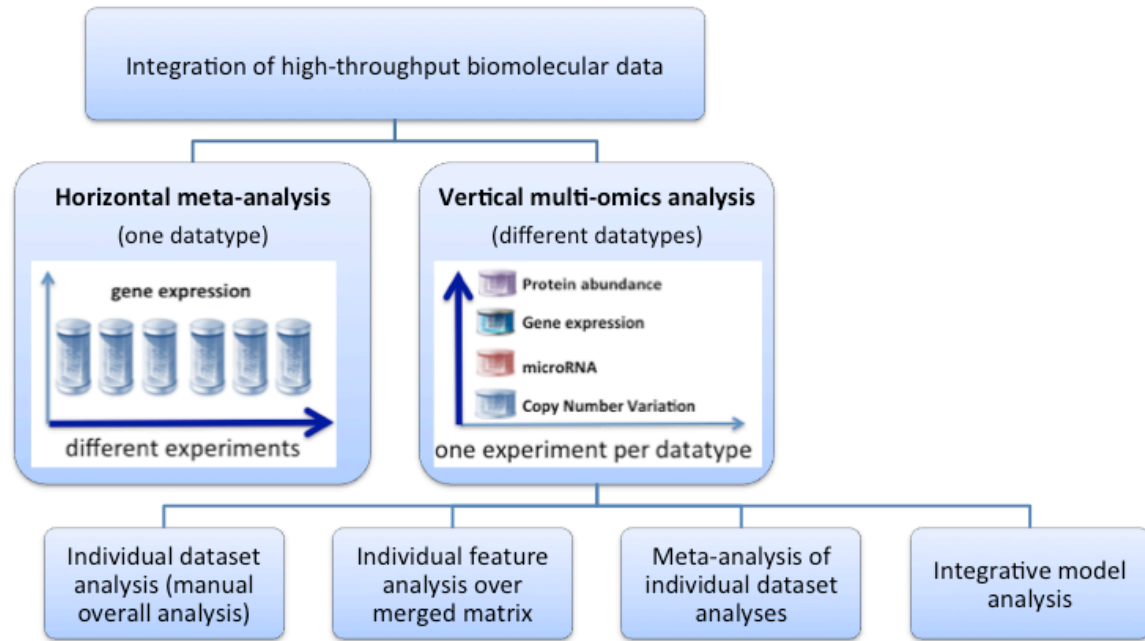


Figure 2.7: Classification of integrative methods for high-throughput biomolecular data because each data type has different scales, volumes, and properties.

2.2.1 Integration of multi-omics data

During the last two decades, immense progress has been made toward understanding the molecular processes that are altered in cancer patients. Traditional approaches compare gene expression levels between samples of cancer patients and normal individuals. Integrating gene expression with other data types has become the new challenge in our age. Integrative approaches have shown to be successful in finding cohesive perspectives of complex cellular systems [15, 87, 112]. Yet, analyzing multiple data types is extremely difficult due to data heterogeneity and high-dimensionality. To give an example of the magnitude of this problem, The Cancer Genome Atlas¹ (TCGA) [142] portal contains datasets from nine data levels (excluding clinical data and images) for a total of 26 different data types (see Table 2.1). Life scientists that intend to an-

¹TCGA is an effort of the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI).

Table 2.1: Data types available on TCGA.

Data Level	Data Type
Microsatellite Instability (MSI)	Microsatellite Instability (MSI)
DNA Sequencing	Whole exome sequence
DNA Sequencing	Whole genome sequence
DNA Sequencing	Trace-Sample Relationship
DNA Sequencing	Mutations
miRNA Sequencing	miRNA sequence
miRNA Sequencing	Isoform
Protein expression	Protein expression
mRNA Sequencing	mRNA sequence
mRNA Sequencing	Exon
mRNA Sequencing	Gene sequence from mRNA
mRNA Sequencing	Splice
mRNA Sequencing	Isoform
Total RNA sequence	Total RNA sequence
Total RNA sequence	Exon
Total RNA sequence	Gene sequence
Total RNA sequence	Splice Junction
Total RNA sequence	Isoform
Array based expression	Gene expression
Array based expression	Exon
Array based expression	miRNA
DNA Methylation	Bisulfite sequencing
DNA Methylation	Array based
Copy number variation	SNP array
Copy number variation	CN array
Copy number variation	Low-pass DNA sequencing

analyze these datasets in pairs would have to perform 324 different analyses to compare every possible pair of data types. To help biologists to analyze this complex data flood, bioinformaticians have been developing computational methods that facilitate the integration of multiple omics.

A vast amount of high-throughput data has been accumulated in many publicly available repositories, such as Gene Expression Omnibus [7, 43], The Cancer Genome Atlas [142], and ArrayExpress [17, 125]. To take advantage of this information, researchers are trying to integrate data from multiple datasets and multiple measurements of the same set of patients from different sources. There are two general

directions to integrate data: i) horizontal meta-analysis and ii) vertical multi-omics analysis [146]. Horizontal meta-analysis is also known as cross-cohort data integration. Its purpose is to integrate the same type of data from independent but related studies. A vertical multi-omics analysis integrates multiple types of data from the same set of patients. Both of these can also incorporate information from biological pathways or other knowledge databases. These studies require interdisciplinary expertise, such as molecular biology, statistics, and computer science. In this thesis, we focus on vertical multi-omics analysis and investigate integrative approaches in the domain of pathway analysis.

2.3 Pathway analysis

This section is organized as follows. First, we briefly introduce a typical comparative analysis, we discuss the importance of pathway analysis using only gene expression, and we describe the existing knowledge-based pathway analysis methods. Third, we explain the need for multi-omics data integration to identify the impacted pathways for a better understanding of the biological mechanisms that are relevant to the disease under study. Finally, we summarize the main strategies used to integrate multiple data types for the purpose of pathway analysis.

2.3.1 Pathway analysis using gene expression

High-throughput technologies for gene and protein profiling, such as DNA microarray or RNA-Seq, have transformed biomedical research by allowing for comprehensive monitoring of biological processes. A typical data analysis often yields a set of genes that are differentially expressed (DE) when comparing patients versus healthy samples. The lists of DE genes helps to identify genes that take part in the underlying phenomenon. However, there are two drawbacks. First, they often fail to reveal the

underlying mechanisms [79, 145]. Second, independent experiments often yield completely different lists of DE genes, making the interpretation extremely difficult [138].

High-throughput technologies for gene and protein profiling, such as DNA microarray or RNA-Seq, have enhanced biomedical research by allowing for comprehensive monitoring of biological processes. Typical data analysis often yields a set of genes that are differentially expressed (DE) when comparing two groups, e.g. patients with a given condition versus normal samples. The selection of DE genes is made by comparing the distributions of the two groups using statistical tests, such as t-test. The obtained p-value is then compared with a chosen significance level (usually, $p \leq 0.05$) and the fold change is compared with a predefined threshold (< 1.5 or 2) [66]. The lists of DE genes help to identify genes that take part in the underlying phenomenon. However, there are two drawbacks. First, they often fail to reveal the underlying mechanisms [79, 145]. Second, independent experiments of the disease often yield completely different lists of DE genes, making the interpretation extremely difficult [45, 46, 138].

To address this challenge, researchers have developed a large number of knowledge bases. Biological processes, in which genes are known to interact with each other, are described in pathway databases, such as Kyoto Encyclopedia of Genes and Genomes (KEGG) [74, 111], or Biocarta [14]. Pathway analysis methods [47, 66, 79, 84, 102] have been developed to identify pathways that are related to the condition under study.

There are three main strategies for pathway analysis using gene expression data: over-representation analysis (ORA), functional class scoring (FCS), and pathway topology (PT) based methods. The input of pathway analysis in general consists of two parts: i) the molecular measurements using a high-throughput technology, e.g. gene expression, and ii) functional annotations of the corresponding genome, e.g. a pathway database. Gene expression data is often presented as a matrix where

the columns represent the samples and the rows represent genes. For example, a DNA microarray assay [38, 52] of 20 diabetes patients and 10 healthy patients can be presented as a matrix of 30 columns and about 20,000 rows. Each column represents a patient while each row represents the expression of a gene across all patients. The second input, the pathway database, is a list of known functional gene modules. A functional module can simply be a set of genes [3, 6, 20, 23, 101, 100] that are known to be involved in a biological process, or can be a complicated network or graph where the nodes represent genes and the edges represent interactions between genes [26, 73, 74, 75, 111, 98].

The earliest pathway analysis methods use the over-representation analysis (ORA) [9, 12, 37, 40, 78, 96] to identify the gene sets that have more differentially expressed genes than expected by chance. This approach starts by identifying genes that are differentially expressed between the two phenotypes, e.g. disease versus control. Statistical methods for identifying DE genes include t-test [58, 116], regularized t-test [5, 62], and linear models [135]. For each pathway, ORA calculates the probability of obtaining the same number of DE genes or more, using hypergeometric or Fisher's exact test [48].

The ORA approach is available in a large number of tools and has widespread usage [79]. However, ORA has a number of limitations. First, this approach only takes into consideration the number of DE genes and completely ignores the change in expression, i.e. it ignores gene expression values. However, gene expression and fold-change can be useful in assigning different weights to the DE genes. Second, ORA typically uses the most significant genes and completely ignore other genes. For example, genes that are marginally less significant, e.g. $p\text{-value} = 0.011$, are not considered resulting in information loss. Finally, ORA assumes that the difference in expression of a gene is independent of the other genes. However, this assumption is invalid since biological systems are complex systems of interactions between genes

and their products. This assumption ignores the structural correlation between genes, resulting in incorrect hypothesis testing and thus leads to biased results.

The second class of methods in pathway analysis is the functional class scoring (FCS). Methods in this class include the gene set enrichment analysis (GSEA) [137, 104], gene set analysis (GSA) [44], sigPathway [144], Category [71], SAFE [8], GlobalTest [57], PCOT2 [82], SAM-GS [35], Catmap [18], FunCluster [63], and PADOG [139]. The hypothesis of this approach is that not only large changes in individual genes can have significant effects, but well-coordinated small changes in functionally related genes can also have significant effects on pathways. FCS methods mainly consist of three steps. First, they calculate the gene-level statistics, i.e. differential expression of individual genes between two phenotypes. Examples include correlation [115], Q-statistic [57], t-test [2], or Z-score [80]. Second, they aggregate the gene-level statistics into pathway-level statistics, one for each pathway. Existing pathway-level statistics include Kolmogorov-Smirnov (used in GSEA) [104, 137], sum, mean, or median of gene statistics for all genes in the pathway (used in Category) [71], or the maxmean statistic (used in GSA) [44].

The strategy used in FCS methods offers a great improvement over ORA methods. However, it also has several limitations [79]. First, although FCS methods do not assume the independence between genes, they still assume the independence between pathways. However, this is not true because a gene can function in more than one pathway. Therefore, FCS methods fail to address the crosstalk between pathways and thus lead to biased analysis and increase in false positives. Second, they do not take into consideration the interaction between genes. For example, consider a gene that is known to interact with many other genes in a pathway. A significant change in expression of this gene would result in a large perturbation in the pathway. This gene should be weighted much more than a gene that does not interact with any other genes.

The third class of pathway analysis methods is pathway topology-based approaches (PT) [41, 55, 56, 67, 68, 120, 134, 140, 154]. Methods in this class include ScorePAGE [120], impact analysis (IA) [41], signaling pathway impact analysis (SPIA) [140], NetGSA [134], TopoGSA [56], DEGraph [68], MetPA [154], BPA [67], and EnrichNet [55]. These methods take advantage of the interaction between genes provided in the pathway databases. Some PT methods, such as IA [41] and SPIA [140], model each pathway as a directed graph, where the nodes are genes or gene products and the edges are the known interactions between the nodes. These methods perform two statistical tests. The first test focuses on the differential expression of genes falling on a given pathway. The p-value of this first test can be obtained from ORA or FCS methods described above. The second test focuses on the number of perturbation factors accumulated on the given pathway. This test is concerned with the topological position, magnitude and sign of changes in expression of genes in the given pathway. The null distribution of the pathway perturbation is obtained by permuting the genes at different locations in the pathway graph. The two p-values obtained from the two independent tests are then combined using Fisher’s method.

2.3.2 Pathway analysis using multi-omics data

Although pathway analysis using gene expression has achieved applaudable results [79], recent research has proven that integrating heterogeneous types of data offers a more comprehensive view of complex cellular systems [106], resulted in a wave of methods for the purpose of data integration [21, 86, 130, 152]. We divide multi-omics pathway analysis methods into two categories: topology-aware methods and non-topology aware methods. Topology-aware approaches incorporate gene topology and interactions between entities into the analysis, i.e. methods that make use of nodes and edges of the pathways. Non-topology aware methods are methods that treat a pathway as a set of genes or entities without considering their topology

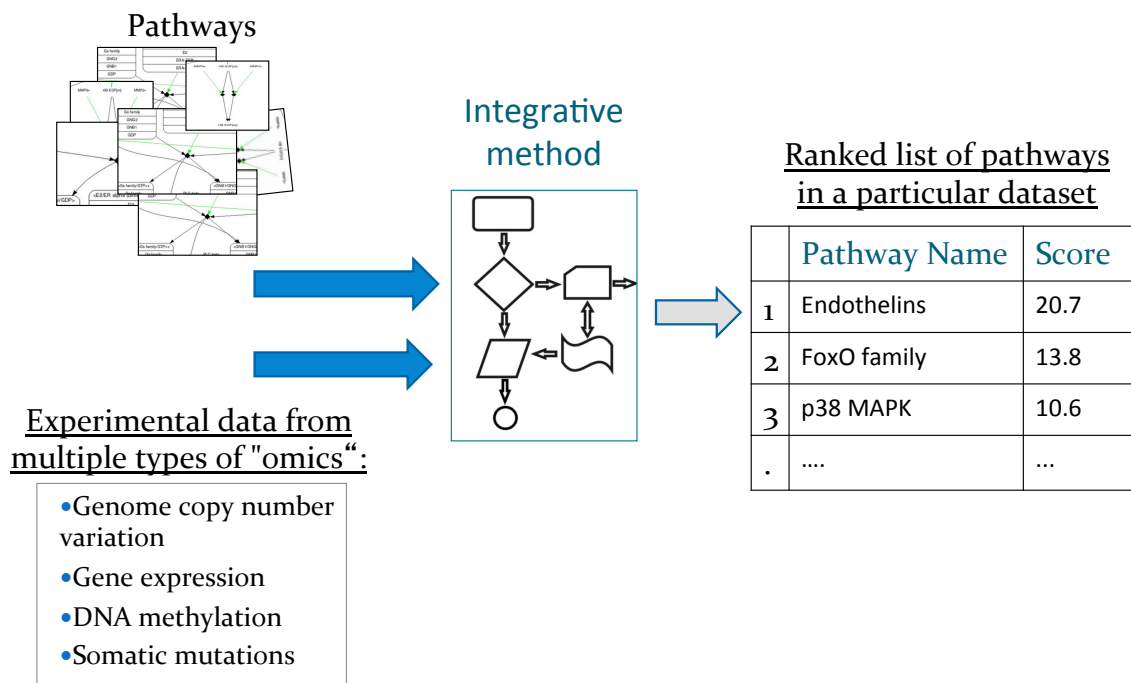


Figure 2.8: General overview of multi-omics pathway topology techniques. The input of these techniques includes different types of molecular measurements for the same set of patients, and pathway knowledge from the databases. The output is a list of pathways ranked according to their statistical significance, e.g. p-values or scores.

or interactions.

Figure 2.8 shows the overall pipeline of integrative pathway analysis methods. The input includes a set of signaling pathways and experimental data from multiple data types coming from the same set of patients. Integrative methods output a list of pathways ranked by statistical significance, i.e. p-value or score.

2.3.3 Topology-aware integrative methods

Topology-aware integrative methods are based on the hypothesis that incorporating the structure of biological processes on the analysis will provide better results. We have reviewed several methods of this type [21, 86, 130, 152] and identified two main categories, graphical extensions and probabilistic graphical models. Methods belonging to the first category extend the existing signaling pathways with molecules

or nodes that were not included in the original pathways. Methods belonging to the second category transform pathways to probabilistic graphical models and include additional relations among multiple types of data.

Methods in the first category expand the existing graphs by adding new nodes, relations, and interactions to use traditional pathway analysis methods on the expanded graphs. The added nodes and relations represent the new molecules or new data types. After the pathways are expanded, the integrative pathway analysis problem can be mapped to the classical pathway analysis, where the original pathways are replaced with the expanded pathways, and the gene expression data is replaced with multi-omics data. The main advantage of this approach is that traditional methods have been evaluated and accepted by the scientific community; therefore, can be adapted rapidly to expanded networks. The main disadvantage of this approach is that some data types cannot be mapped directly to gene interactions because their effect on gene expression is not completely understood, for example, DNA methylation data [59]. Given that current signaling pathway databases contain information about gene interactions and ignore remaining data types, enhancing them is crucial [113, 119].

One data type that can be easily integrated to current pathways is microRNAs (miRNAs). miRNAs are gene regulators that have shown to play important roles in the development of cancers and many complex diseases [88, 93]. Relevant work has been done to elucidate the important interplay between miRNAs and biological pathways [4, 19, 64, 65, 107, 148]. The state of the art approach for miRNA-mRNA pathway analysis is microGraphite [19] which uses an empirical gene set approach. microGraphite’s main goal is the identification of signal transduction paths correlated with the condition under study [97]. microGraphite integrates miRNA and mRNA expressions by wiring the miRNA-mRNA interactions into the formal pathway representations. After expanding the pathway, microGraphite performs pathway analysis

using a method previously proposed by their authors named CliPPER [97].

The pipeline of microGraphite consists of five steps. In the first step, microGraphite integrates microRNAs into existing pathways downloaded from pathway databases, such as Kyoto Encyclopedia of Genes and Genomes (KEGG) [74, 111], Nature Pathway Interaction Database (NCI) [127], Reactome [73], or Biocarta [14]. There are two types of microRNA-target interactions that are integrated to the pathways: in-silico predicted interactions and validated interactions. Validated interactions are obtained from miRecords[156] and mirTarbase [65]. In the second step, microGraphite performs pathway analysis to obtain an initial set of significant pathways. In the third step, it carries an analysis across the significant pathways to score the coherent paths inside the pathways. In the fourth step, microGraphite selects the paths with the highest score and then join these paths to form a connected network called *meta-pathway*. Finally, microGraphite performs pathway analysis among the paths to identify the most significant paths. The authors validated their pipeline on an ovarian cancer dataset, obtaining a meta-pathway that guided biological experiments further performed by them.

Methods in the second category use graphical models, such as Bayesian networks and factor graphs, to represent the interaction between data types and gene expression. These models are more versatile because they can describe more complex type of interactions. These methods rely on the fact that each type of genomic data contains valuable information, so integrating them in an equivalent variable makes the analysis more complete.

An example of these approaches is PARADIGM [147]. This method integrates and analyzes different types of genomic data by producing a single measurement called Inferred Pathway Activity (IPA). Having an IPA per patient allows us to perform pathway analysis for an individual while current approaches need a group of

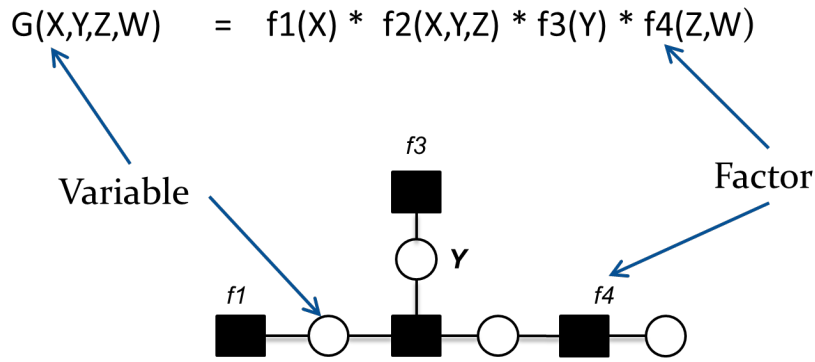


Figure 2.9: An example of factor graphs. This factor graph represents a global function G as the product of the local functions f_1 , f_2 , f_3 , and f_4 . Black squares of the graph represent local functions or factors and circles represent variables. Each factor is a function of its neighbor variables.

samples. In order to compute the IPA, PARADIGM connects the different types of measurements by adding causal-effect relations and the interaction between genes in a factor graph model (see Figure 2.9). Then, the likelihood of having a gene activated or not in each particular cancer patient and the IPA per gene is computed by performing a Bayesian inference algorithm. The method was evaluated by performing pathway analysis in two different diseases, breast cancer and glioblastoma multiform (GBM), and comparing the results with those obtained with SPIA [140]. The authors claimed that PARADIGM analysis generates fewer false positives, and they were able to identify different groups of GBM with significantly different survival profiles [147]. However, in spite of numerous efforts, we were not able to reproduce these results. This method has been included as an official tool into The Cancer Genome Atlas (TCGA) [142].

Figure 2.10 shows a pathway example with three nodes: MDM2, TP53, and Apoptosis; and two relationships: MDM2 represses TP52 and TP52 activate Apoptosis. The first step of PARADIGM is to represent the pathway as a probabilistic model using factor graphs (see Figure 2.9). A factor graph is a bipartite graph that represents a global function as the product of local functions [128]. Each of the local functions

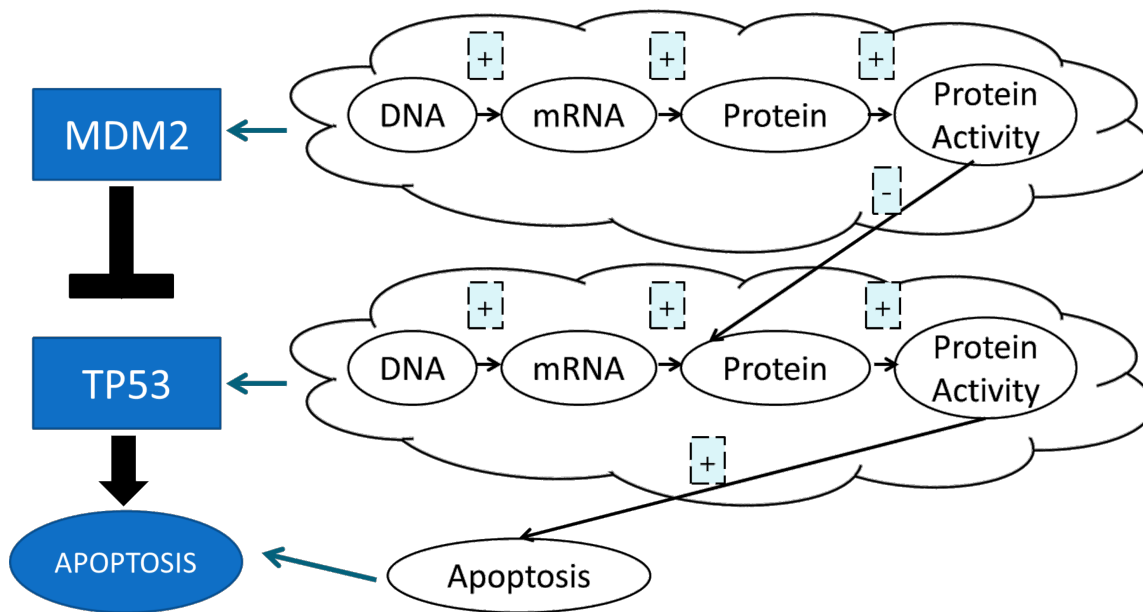


Figure 2.10: An example of the pathway model to compute IPA [147]. The biological pathway contains three entities MDM2, TP53, and Apoptosis. The omics data of every entity is translated to variables in a logic representation DNA, mRNA, Protein, Protein activity, and Apoptosis. The links among variables are labeled according to the logical interpretation of the biological pathway.

is a factor and is represented in the graph as a black square. Each factor has its own variables and the global function will be represented in terms of the overall sets of variables. The authors represent every data type as a variable in their model. The molecules included in the model are protein-coding genes, small compounds, protein complexes, gene families, and abstract processes. Protein-coding genes are measured with four data-types: copy number variation (CNV), mRNA expression (mRNA), protein level (protein), and protein activity status (activity). In this toy example (Figure 2.10), we have two protein-coding genes (MDM2 and TP53) and one abstract process (apoptosis). Therefore, we have four variables per each gene and one variable for the process. The authors also defined a set of labels to apply depending on the type of relationships among components. In this example, the relationships are “activation” and “repression” and the labels “positive” and “negative” are respec-

tively applied. Now that the variables and relations are identified, we can construct the factor graph. Each variable will represent a differential state of each entity in comparison with a control level. The possible states can be *activated*, *nominal*, or *deactivated*. A factor is attached to each variable and represents the expected state of such a variable. To decide the state of the variables, a voting system will be applied as described on Figure 2.11. For each link of a variable on the factor graph, the algorithm starts checking the type of label assigned: *minimum*, *maximum*, *positive*, or *negative*. If the label is *minimum*, the vote is computed as the minimum state among the parent variables connected to such a link. If the label is *maximum*, the vote is computed as the maximum state among the parent variables connected to such a link. If the label is *positive*, the vote is the same state of the parent connected to such a link. If the label is *negative*, the vote is computed as the inverse state of the parent connected to such a link. Once the votes of all the edges are counted, the next expected state is equal to the state with more votes. If there is a tie, the expected state is *-1*.

The second step is to define the prior probability. For this, the authors use the expectation maximization algorithm [31, 61, 99] to estimate if a particular hidden variable is likely to be in a particular state. Once the prior probabilities are estimated, they use a belief propagation algorithm [49] to find the maximum likelihood that a variable is in a particular state along with all the other observations made for the patient.

2.4 Omics integration for disease subtyping

A vast majority of the diseases develop differently making them heterogeneous. Precise classification of patients into subtypes has important indications for medicine. Moreover, identifying subtypes that are relevant to survival profiles or related to bi-

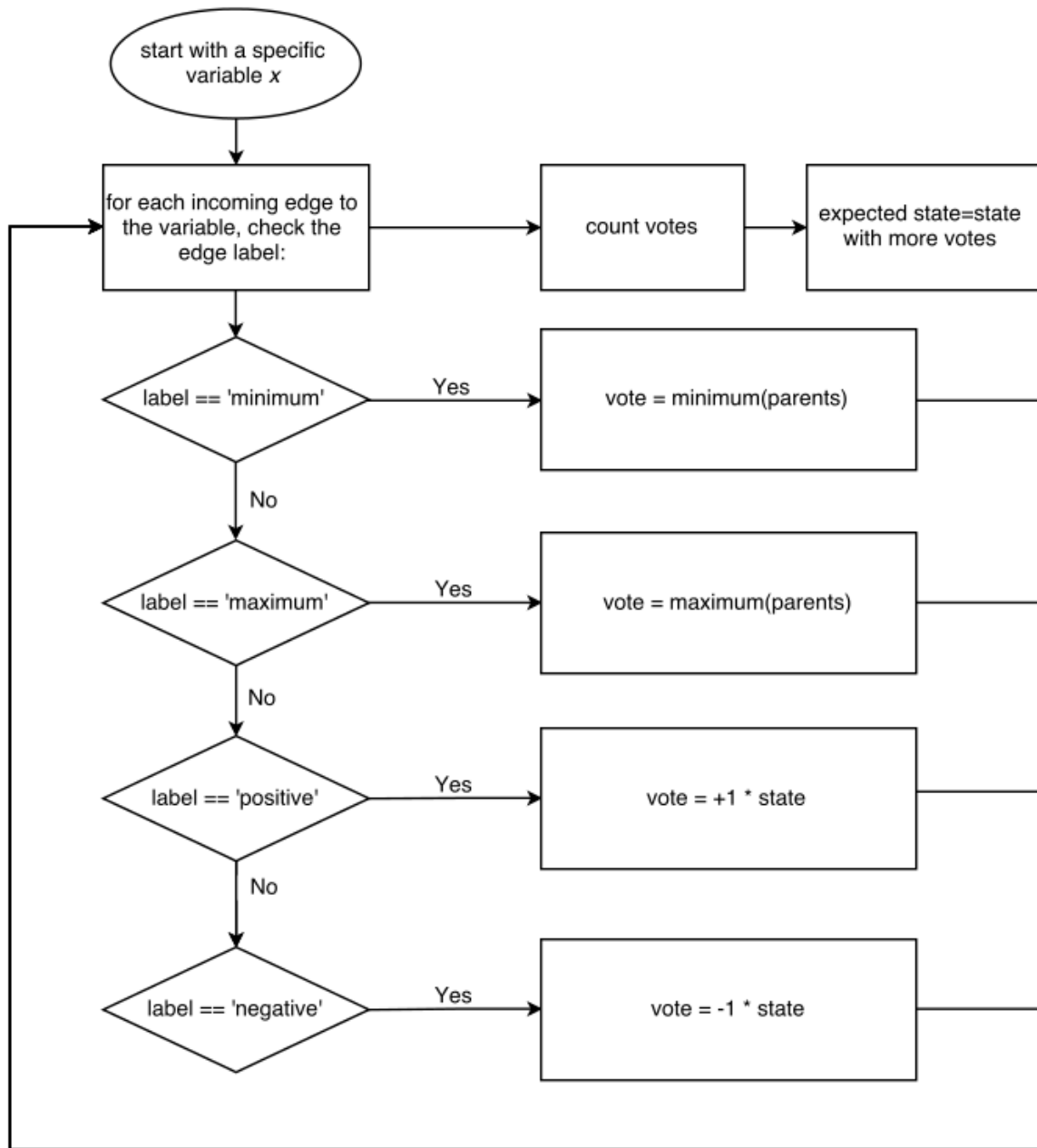


Figure 2.11: Algorithm to decide if an entity is active or not based on its labels. For each link of a variable on the factor graph, we start by checking the type of label assigned: minimum, maximum, positive, or negative. If the label is ‘minimum’, the vote is computed as the minimum state among the parent variables connected to such a link. If the label is ‘maximum’, the vote is computed as the maximum state among the parent variables connected to such a link. If the label is ‘positive’, the vote is the same state of the parent connected to such a link. If the label is ‘negative’, the vote is computed as the inverse state of the parent connected to such a link. Once the votes of all the edges are counted, the next expected state is equal to the state with more votes. If there is a tie, the expected state is -1.

ological patterns is crucial. This is to identify more homogeneous disease subtypes and their corresponding genetic signatures. Subtype distinction can advance diagnosis classification which can improve clinical decision and treatment matching. Most methods for disease subtyping perform clustering using clinical data from patients. These methods do not use molecular measurement and the outcome subtypes are prone to be suboptimal [136].

A contemporary method that integrates genetic data for disease subtyping is SNF (Similarity Network Fusion) [152]. The input of SNF includes multiple matrices for the same set of patients, each matrix is the molecular measurement of a data type. SNF first constructs a patient similarity matrix (PSM) for each data type based on Euclidean distance. It then constructs a patient similarity network for each data type where the nodes are patients and the edges are the similarity between them. It then iteratively fuses these networks into one network that represents the overall similarity between patients for the multi-omics data. In each iteration, the fused network discards the weak similarities to eliminate contradictions. After each iteration, the networks from multiple data types are more similar to each. The algorithm stops when the networks are identical. Finally, a similarity-based clustering, such as spectral clustering [108], is performed on the fused network to identify subtypes of the disease.

The authors validated the discovered subtypes using Kaplan-Meier survival curves, Cox regression [24, 143], and Silhouette score [124]. The method is compared with existing methods, such as iCluster [132] and Consensus Clustering [103]. The data analysis was done using five different cancer datasets downloaded from TCGA: glioblastoma multiform data (GBM), breast invasive carcinoma (BIC), kidney renal clear cell carcinoma (KRCCC), lung squamous cell carcinoma (LSCC), and colon adenocarcinoma (COAD). For all the five datasets and all the metrics used, SNF achieved the best result.

Cox log-rank test is one of the methods to decide if certain groups have clearly different survival behavior or not. This method is used in survival analysis. Survival analysis is performed in studies that aim to investigate the change of state of a variable. In this context, we refer to studies that investigate life duration of the certain sample. The basic setup for the analysis is that certain subjects (e.g. cancer patients) are tracked over time until the event happens (e.g. death) or the subject is lost from the sample and cannot be track anymore (e.g. patients do not return to the center that is performing the study and the center does not know if the patient is alive or not). Survival in this context means how long people stay in the sample. At the beginning of the study all the subjects are in the sample; therefore, the survival is 100%. Over time, events start happening and the survival start decreasing until the study is over. Typically, the analysis will include a survival curve to visualize the behavior of survival over time. It also includes hazard rates, which represent the risk of failure or what is the chance that the event will happen before a certain time period. In this case, the hazard is the probability of dying at a particular moment. The dependent variable is the duration of measurement wich is a combination of three variables the time variable (the length of time until the event happened or being in the study), the event variable (1 if the event happened or 0 if the event has not happened yet), and the censored variable (indicating if the measurement was taken or not). These survival studies can have several extensions, one of these is the use of more that one group of participants in the same study. Survival analysis can be made for nonparametric models, parametric models, or semi-parametric models. Nonparametric models are useful for descriptive purposes and to visualize the shape of the survival and hazard functions before using a parametric model. Hazard curves are nonmonotonic and survival curves are strictly non-increasing curves. There are two estimators commonly used for non-parametric models: Nelson-Aalen estimator of the cumulative hazard function and The Kaplan-Meier estimator for the survival function.

Once the nonparametric model has been run, one can include some independent variables that may affect those functions by using a parametric or semi-parametric model. These models depend on the form of the functions. The Cox proportional hazard model is used for estimation of semiparametric models. The cluster number is included on the survival analysis as an independent variable. A comparison of two survival curves can be performed using the statistical hypothesis test named log-rank test. It tests the null hypothesis that there is zero difference between the survival curves. The dependent variable for this model is the hazard [13]. The p-value of the test indicates if the difference between groups borders on the statistical significance or not, to determine if there is strong evidence that a variable is associated with length of survival. The Cox regression test is useful to find which variables are significant for the complete set of samples. The Log-rank Mantel-Cox test is a test for two or more groups and aims to tell if these groups have different survival curves. When comparing more than two groups, correction for multiple comparisons is needed.

In SNF, the validation of label prediction was achieved by comparing SNF with two other approaches, PAM50 and iCluster. They used breast cancer data from the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) [27] and computed the Cox log-rank test p-values for each of the methods. They also computed the concordance index (CI) [60] for discovery and validation cohort for risk of death prediction. The authors also presented a general validation of the fused network by comparing the Cox log-rank test p-values obtained when using individual data types versus the fused data. For this experiment, they used the five cancer data sets from TCGA mentioned before: GBM, BIC, KRCCC, LSCC, and COAD.

Chapter 3: Integrating signaling pathways with miRNA target genes: our proposed method

In this section, we propose an algorithm for integrating miRNA into signaling pathways. The integration of miRNA into signaling pathways have multiple applications, such as pathway analysis and disease subtyping. We also describe a pipeline to use the miRNA-augmented pathways (mirAP) in the context of pathway analysis (PA). This analysis is used in biological studies comparing genetic samples from two different phenotypes (e.g. disease vs. control samples). Our PA pipeline integrates miRNA and mRNA expression data and identifies pathways that are related to the disease under study. The standard input of pathway analysis includes gene expression data from two different phenotypes and a set of signaling pathways. The input of our pathway analysis includes miRNA and gene expression data comparing two different phenotypes (e.g. condition vs. control) and a set of signaling pathways and a list of miRNA-gene interactions.

3.1 Proposed integrative pathway analysis pipeline

3.2 Pathway analysis pipeline

The identification of pathways that are significantly perturbed in a given phenotype helps us understand the underlying biological processes. Traditional pathway analysis techniques aim to infer the impact on individual pathways using only gene expression measurements (mRNA). However, gene expression does not capture the complete picture of the biological mechanisms involved, as many other entities play important roles in the processes. By ignoring them, we ignore potentially crucial information. One type of these entities is microRNA (miRNA), newly discovered gene regulators that have been shown to play a significant role in diagnosis and prognosis

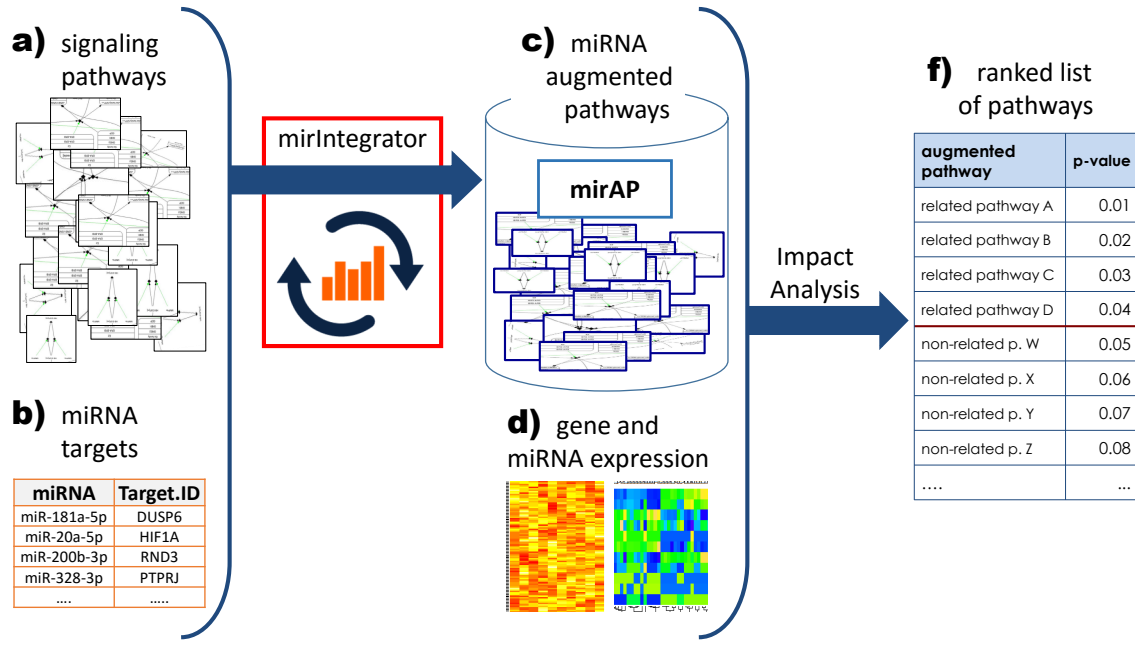


Figure 3.1: Workflow of pathway analysis using augmented pathways.

for different types of diseases [88].

Researchers investigating miRNA cannot perform pathway analysis using traditional techniques because current pathways datasets do not contain miRNA-target interactions. Our pipeline fills this gap by integrating miRNAs into signaling pathways.

Our pathway analysis pipeline consists of three main steps (shown in Figure 3.1):

1. In the first step, we find the miRNAs that target the genes of each signaling pathway. miRNA-targets can be identified with existing algorithms for prediction of miRNA-targeting [72, 89, 85, 110] or by using databases of validated miRNA-target interactions [65]. Our implementation allows the use of predicted or validated interactions or a mix of both. The user can provide custom interactions our use the default database that we provide with the package which is mirTarBase [65].
2. In the second step, we augment the signaling pathways with the miRNAs identified in the first step. The output is a list of miRNA augmented pathways.

At this point, the data integration problem is mapped to the original pathway analysis problem for which existing methods can be applied. The difference is that here both miRNA and mRNA expression can be taken into consideration.

3. In the third step, we apply any pathway analysis that uses fold change and p-value as input, e.g. Over-representation analysis (ORA) [40] and Impact Analysis [41, 140]. ORA and Impact Analysis are well-known methods developed by our group to identify signaling pathways that are impacted by the effects of diseases. Fig. 3.1 displays the overall pipeline of our approach.

3.2.1 Impact analysis using mRNA and miRNA

Impact Analysis is a topology-aware method that combines two types of evidence: i) the over-representation of DE genes (ORA) [40], and ii) pathway perturbation caused by disease, as measured by propagating expression changes through interactions between the genes. These two types of evidence are captured by two independent p-values: p_{ORA} and p_{PERT} . To calculate p_{ORA} on the miRNA augmented pathways, we use the following information: i) the total number of entities (genes and miRNAs) taken into consideration, ii) the entities belonging to the augmented pathways, iii) the entities that are differentially expressed (DE), and iv) the entities that are differentially expressed in the given pathway. The first input is the total number of genes and miRNAs that were measured. The second input includes the genes of the pathway and the miRNAs that target at least one gene in the pathway. The third and fourth inputs are calculated from the input lists of DE genes and DE miRNAs. We perform a modified t-test [122] on both gene and miRNA expression separately. The significance threshold to determine the genes and miRNAs that are differentially expressed is set to 5%.

To calculate p_{PERT} , we use the following information: i) the entities that are differentially expressed, and ii) a graph that represents the augmented pathway. The

first input can be determined using the modified t-test for each data type [122] while the second input is constructed in the first step of the approach. These two p-values are combined using Fisher’s method [48] to get a single p-value that represents how likely the pathway is impacted by the effect of the disease.

The combination of p-values provides a significant advantage to our approach because we do not require additional cross-platform normalization between mRNA and miRNA data. In other words, our method combines the p-values independently of the technologies or platforms used to measured miRNA and mRNA. In this way, we avoid additional statical error due to cross-platform normalization.

3.3 Proposed algorithm for pathway augmentation

Our method augments the graphical representation of original signaling pathways with interactions between miRNAs and their target genes. The input of this method includes a set of signaling pathways and known miRNA-mRNA interactions (Fig. 3.1a,b). The output is a set of augmented pathways that consists of the original genes, the miRNAs that target those genes and their interactions. Let $P = (V, E)$ denote the graphical representation of the original gene-gene pathway, and $T : M \rightarrow V$ a function that identifies the target genes of miRNAs in M . An edge $e \in E$ can be represented as a 3-tuple $e = (g_1, g_2, interaction)$. We augment the nodes and edges of the original pathway as follows:

$$\begin{aligned}\bar{V} &= V \cup \{m \in M | T(m) \cap V \neq \emptyset\} \\ \bar{E} &= E \cup \{(m, g, inhibition) | m \in V \cap M \wedge g \in T(m)\}\end{aligned}$$

We implemented this algorithm in R and published it as the Bioconductor package named mirIntegrator (<http://bit.ly/mirIntegrator>). mirIntegrator is flexible and

allows users to integrate user-specific pathway databases with user-specific miRNA-mRNA target databases. Additionally, it generates graphical representations of the augmented pathways (see Fig. 6.3). We integrated pathways from Kyoto Encyclopedia of Genes and Genomes (KEGG) [74] (version 73) with miRNA targets from miRTarBase [65] (version 4.5) to generate mirAP, a database of miRNA-augmented pathways (<http://www.cs.wayne.edu/dmd/mirAP>).

Chapter 4: Proposed miRNA augmented pathways database

4.1 Databases needed for integration

The input to generate the augmented pathways includes a database of signaling pathways and a database of miRNA target interactions. Here we describe the databases that we used in our approach.

4.1.1 Database of signaling pathways: KEGG pathways

As described in previously we are using the Kyoto Encyclopedia of Genes and Genomes (KEGG). We downloaded the KEGG pathways release 73.0+/01-03, Jan 2015 (Kanehisa Laboratories). This dataset contains 149 KEGG human signaling pathways. To process the original pathways in R, we parsed them into a list of `graphNEL` objects using the `ROntoTools` package Version 1.2.0. The users can obtain the list of KEGG human signaling pathways with the function `data('kegg_pathways')`. This object contains a list of `graphNEL` objects where each graph represents one KEGG signaling pathway. The name of each pathway is its KEGG pathway identifier. A script that constructs the `kegg_pathways` object may be found in ‘`inst/scripts/get_kegg_pathways.R`’,

The names of the pathways are stored on a different object called `names_pathways`. This object contains a list of KEGG signaling pathways’ names. The names of the pathways in human are obtained with the `ROntoTools` package. This object can be loaded with the instruction `data('names_pathways')`. A script that constructs the `names_pathways` object may be found in ‘`inst/scripts/get_names_pathways.R`’, see the example.

4.1.2 Database of miRNA target interactions: mirTarBase

The second database needed to generate the miRNA augmented pathways is a database of miRNA-target interactions. For this purpose we use mirTarBase [65] which was downloaded from <http://mirtarbase.mbc.nctu.edu.tw/> on 4/1/2015. mirTarBase is a publicly available database of microRNA-target interactions in human. We downloaded mirTarBase release 4.5: Nov. 1, 2013, and made it accessible in our package through the object `mirTarBase`. This object is a data.frame with 39083 interactions and nine variables. This dataset is licensed by its authors (Hsu et al.), see <http://mirtarbase.mbc.nctu.edu.tw/cache/download/LICENSE>.

Even though our package includes mirTarBase, any data.frame with human miRNA-targets interactions can be used to generate the miRNA augmented pathways. For this purpose, the data.frame should contain the following columns:

- `miRNA`: which contains the miRNA ID,
- `Target.ID`: contains the entrez ID of the gene targeted by the miRNA

A script which downloads the file and constructs the mirTarBase object may be found in ‘`inst/scripts/get_mirTarBase.R`’.

4.2 Data representation of the augmented pathways

The primary goal of our mirIntegrator package is to integrate microRNA expression into signaling pathways for pathway analysis. The first step of our pipeline is to augment signaling pathways with miRNA target interactions. Figure 4.1 shows a model of the expected output.

In our package, we include a list of signaling pathways augmented with miRNA in the object `augmented_pathways`. This object is a list of human signaling KEGG pathways augmented with validated miRNA-target interactions from mirTarBase using our algorithm. These interactions represent the biological miRNA repression of

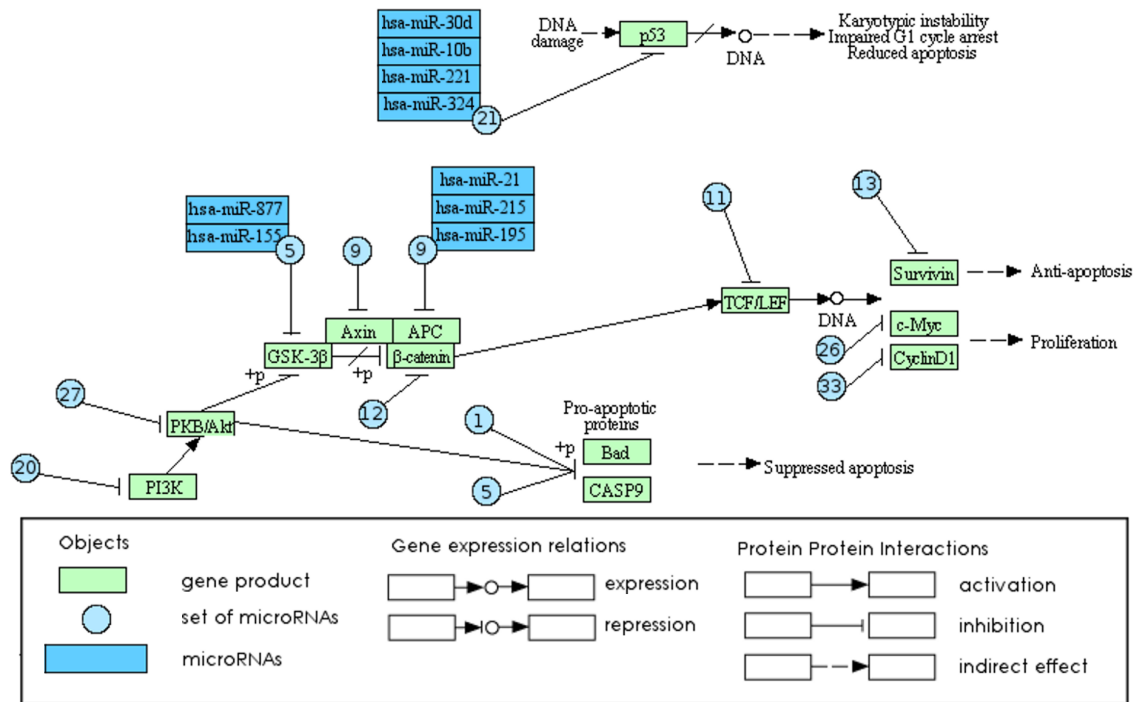


Figure 4.1: Model of a miRNA-augmented pathway. Portion of the *Colorectal Cancer* pathway from KEGG.

its target genes and are included in the model as negative links. The object can be obtained with the instruction `data('augmented_pathways')`. `augmented_pathways` is a list of graphNEL objects where each graph is a pathway that was augmented with miRNA-target interactions. The id of each pathway is its KEGG pathway identifier.

A script that constructs the `augmented_pathways` object may be found in ‘`inst/scripts/get_augmented_pathways.R`’.

4.2.1 Using the augmented pathways: an example

We also include an example that illustrates how to use the augmented pathways. Let us say that we are interested in finding the pathway with a fewer number of nodes among the augmented pathways, i.e. the smallest pathway. This script is included in the function `smallest_pathway`. This simple function is an example of how to navigate the genes on the list of augmented pathways. The parameter is a list of

graph::graphNEL objects, let us call it *pathways*. The output is the index of the pathway with a fewer number of nodes. The instruction to get the number of nodes of the smallest pathway is `length(smallest_pathway(augmented_pathways))`.

4.3 Proposed graphical visualization of the pathways

MIRINTEGRATOR incorporates a functionality to produce a graphical representation of the final pathways. This functionality is useful when researchers need to visualize the nodes that were added to the pathway. For instance, if they need to see how the pathway of “Sulfur relay system” (path:hsa04122) has changed, they can plot the augmented pathway using the function `plot_augmented_pathway`. Here an example, Figure 4.2 is the output of these instructions:

```
data(names_pathways)
plot_augmented_pathway(kpg$"path:hsa04122",
  augmented_pathways$"path:hsa04122",
  names_pathways["path:hsa04122"] )
```

Another useful function is `plot_change` which can be used to see how much the order of the pathways has changed. To demonstrate this functionality, the MIRINTEGRATOR package includes a copy of KEGG human signaling pathways. We obtained these KEGG pathways using the KEGGGGRAPH package. A complete script describing how this dataset was obtained included in this package on ‘/inst/scripts/get_kegg_pathways.R’. An example of the use of the function `plot_change` .

```
data(augmented_pathways)
data(kegg_pathways)
data(names_pathways)
plot_change(kegg_pathways, augmented_pathways, names_pathways)
```

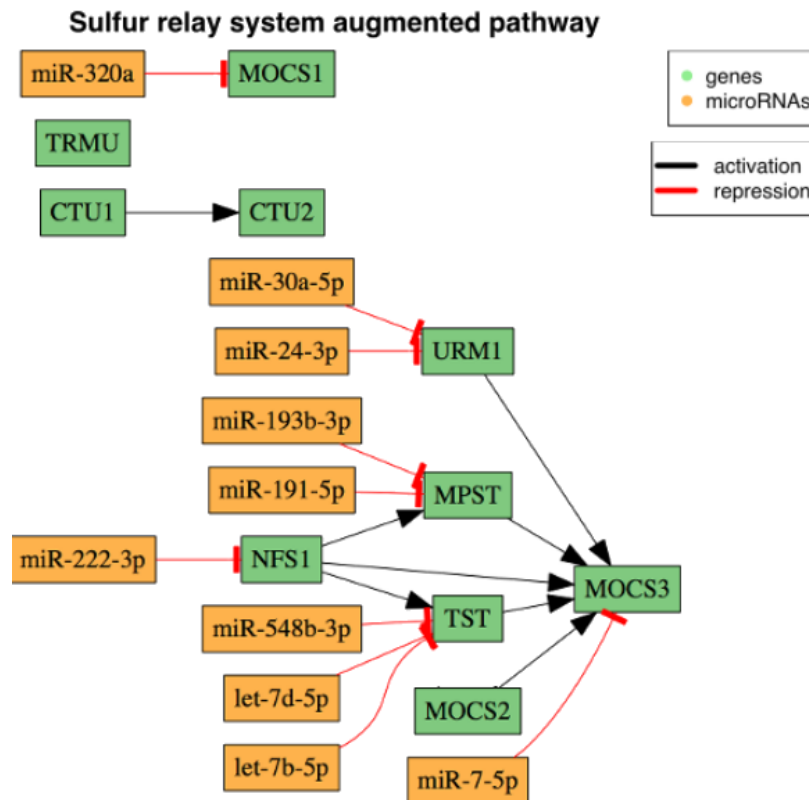


Figure 4.2: Example of augmented pathway. This visualization of the miRNA-augmented *Sulfur Relay System* pathway was generated using the function `plot_augmented_pathway` on our package.

See the resulting graph on Figure 6.2.

This package also includes a function to generate a pdf file with the plots of the list of augmented pathways. Here an example of this functionality:

```
data(augmented_pathways)
data(kegg_pathways)
data(names_pathways)
pathways2pdf(kegg_pathways[18:20], augmented_pathways[18:20],
             names_pathways[18:20], "three_pathways.pdf")
```

4.4 Script to augment signaling pathways with miRNA

The main functionality of the `MIRINTEGRATOR` package is the integration of miRNAs into signaling pathways. The input of this functionality are a set of signaling pathways like KEGG pathways [74] or Reactome [26], and a miRNA-target interaction database like mirTarBase [65] or TargetScan [89]. The output is a set of augmented signaling pathways. Each augmented pathway contains the original sets of genes and interactions plus the set of miRNAs involved in the pathways and their miRNA-target interactions. These interactions are the biological miRNA repression to their target genes and are represented in the model as negative interactions.

Here we show an example of the method functionality. Let us say that a researcher needs to integrate the human signaling pathways from KEGG [74] with miRNA interactions from mirTarBase [65]. The researcher must first obtain the list of pathways as a list of `graph::graphNEL` objects. The nodes of each pathway represent the genes involved in the pathway, and the edges represent the biological interactions among those genes (activation or repression). The second step is to obtain a miRNA-target interactions dataset as a `data.frame` with the columns ‘‘miRNA’’ and ‘‘Target.Gene’’. Notice that the symbols used to identify the ‘‘Target.Gene’’ column on the miRNA-target interactions dataset must be the same symbols used on the nodes of the pathways. i.e. If the genes are identified by entrezID on the pathways’ dataset, then the miRNA-targets dataset must identify the genes by entrezID as well. Once the researchers have these two datasets, they can use the function *integrate_mir*.

To demonstrate this functionality, `MIRINTEGRATOR` package includes the object `mirTarBase` which is a copy of the experimentally validated miRNA-target interactions database mirTarBase [65]. We downloaded the mirTarBase database from <http://mirtarbase.mbc.nctu.edu.tw/> on November 11, 2016. A complete script describing how this database was downloaded and formatted is included in this pack-

age on ‘/inst/scripts/get_mirTarBase.R’.

Here is an example of how researchers can generate the list of augmented pathways from five KEGG pathways and mirTarBase interactions using the function *integrate_mir*:

```
require("ROntoTools")
kpg <- keggPathwayGraphs("hsa")
kpg <- kpg[15:20] #delete this line for augmenting all pathways.
require("mirIntegrator")
data(mirTarBase)
augmented_pathways <- integrate_mir(kpg, mirTarBase)
head(augmented_pathways)
```

The result is a list of pathways where each pathway is a `graph::graphNEL` object. When researchers need to see the details of a particular pathway, they can do so by only using the KEGG pathway id of the pathway of interest. For example, the pathway “path:hsa04122” can be reached with the following instruction:

```
augmented_pathways$"path:hsa04122"
```

Chapter 5: Software Implementation

5.1 Software features

The software `mirIntegrator` is implemented using the R programming language. It is publicly available on the Bioconductor website (<http://www.bioconductor.org>). The package has three main functionalities: i) integration of miRNAs into signaling pathways, ii) graphical representations, and iii) pathway analysis using the augmented pathways and both mRNA and miRNA data. These functionalities are documented in the software manual (at Bioconductor.org). We also developed a GUI version for this software using the *shiny* framework [22]. The source code of the graphic user interface is available at <http://datad.github.io/mirIntegrator>.

5.2 Pathway analysis in R and Bioconductor

The primary goal of bioinformatics is to contribute to advances in biology by developing and distributing tools to handle the massive dataset and perform complex analyses. Making these tools available for collaborative improvement is crucial. Projects like Bioconductor, CRAN, Bioperl, and Biopython are repositories of open source bioinformatics and statistical tools to foster collaborative development.

The Bioconductor [54] project aims to reduce the barriers for remote interdisciplinary research and facilitate the reproducibility of research results. Bioconductor packages are primarily written in R although C++ and other programming languages can be incorporated into the packages [39]. There are several well-implemented statistical and visualization tools in R that facilitate and speed the development of new bioinformatic tools. Bioconductor has several advantages over other projects including: supports object-oriented programming for R, promotes modularization at the package level, web connectivity, statistical simulation, modeling, and visualization

support, among others.

In addition to all these advantages, Bioconductor has a peer-review process for publication. This process increases the quality of the packages, reduces duplicity, promotes reusability of software and data structures, and improves the quality of documentations, vignettes, and other manuals that are included in the packages.

We implemented our algorithm in R and published it as the Bioconductor package named `mirIntegrator` (<http://bit.ly/mirIntegrator>). `mirIntegrator` is flexible and allows users to integrate user-specific pathway databases with user-specific miRNA-mRNA target databases. Additionally, it generates graphical representations of the augmented pathways (see Fig. 6.3). We integrated pathways from Kyoto Encyclopedia of Genes and Genomes (KEGG) [74] (version 73) with miRNA targets from miRTarBase [65] (version 4.5) to generate `mirAP`, a database of miRNA-augmented pathways (<http://www.cs.wayne.edu/dmd/mirAP>). Figure 5.1 represents the graphical augmentation of the colorectal cancer pathway in KEGG. Our package has three main functionalities: i) integration of miRNAs into signaling pathways, ii) graphical representations, and iii) pathway analysis using the augmented pathways and both mRNA and miRNA data.

i. Integration of miRNAs into signaling pathways. By default, the package includes KEGG pathways (version 73) [74] and their augmented versions using validated targets from miRTarBase (version 4.5) [65]. However, the software also allows users to download and augment pathways using other databases.

ii. Graphical representations. `mirIntegrator` is suitable to generate a number of graphical representations. For example, Figure 4.2 shows the augmented *Sulfur relay system* pathway. The nodes in green are the genes from the signaling pathways and the nodes in black are the newly added miRNAs. `mirIntegrator` also generates a visualization and statistics of the miRNAs added on each pathway. This chart is particularly useful to analyze the overall impact that miRNA have in a set of

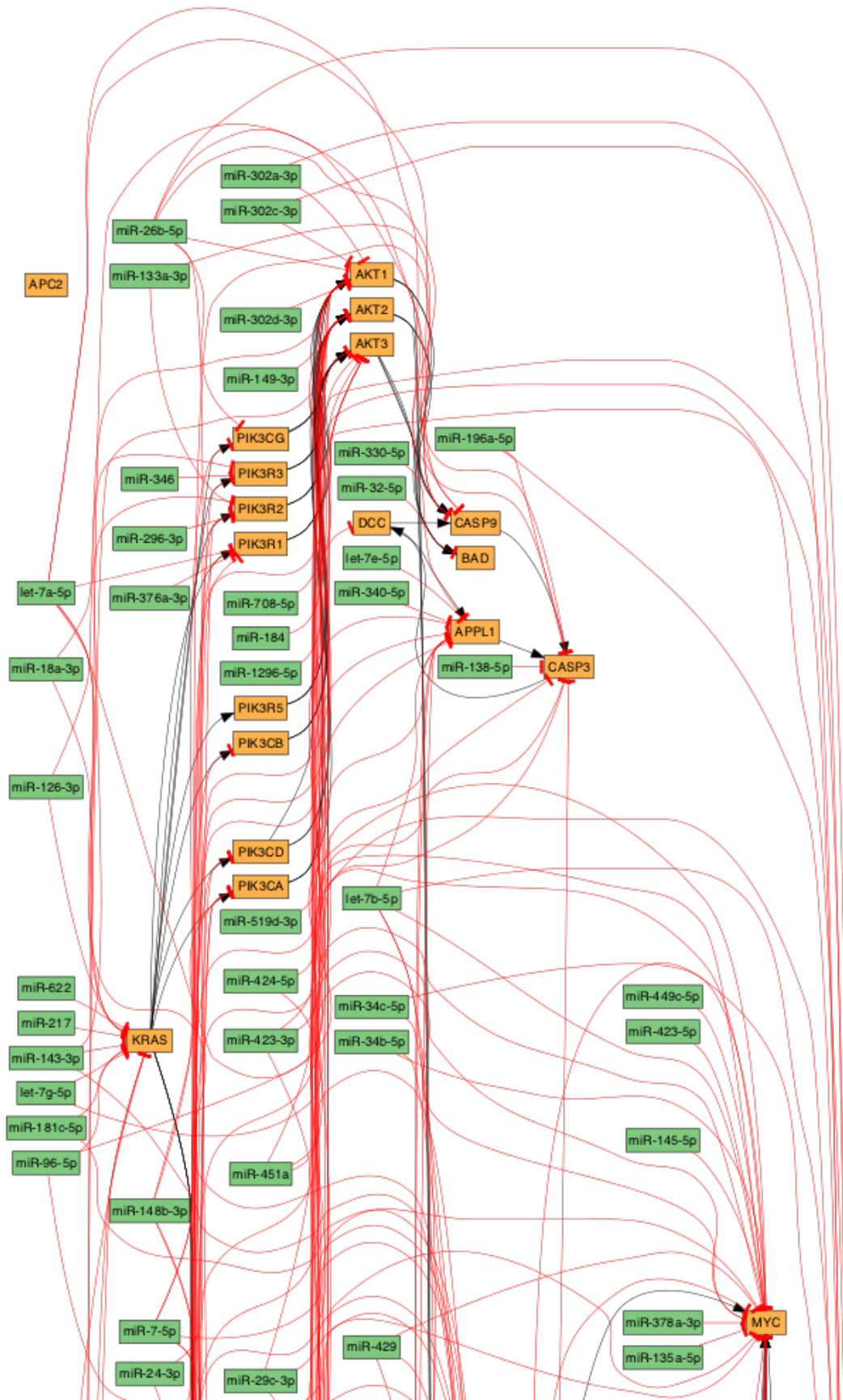


Figure 5.1: Portion of the miRNA-augmented *Colorectal Cancer* pathway.

pathways.

iii. Pathway analysis using augmented pathways. The final goal of the framework is to integrate miRNA and mRNA expression for the purpose of pathway analysis. Figure 5.2 shows a screenshot of the graphical user interface (GUI). First, users are required to upload miRNA-mRNA sample-matched data. The datasets then can be selected using the drop boxes. To perform the analysis, users simply click the button *Run Impact Analysis* to compute the p-value for each pathway. These p-values are then adjusted for multiple comparisons using False Discovery Rate (FDR) [10, 121].

5.3 Example of pathway analysis of miRNA and mRNA data

The main purpose of the pathways augmentation process is to analyze miRNA and mRNA expressions at the same time. For this reason, we show here how to analyze a multiple sclerosis dataset using the MIRINTEGRATOR package. Jernas et al. [69] published the dataset that we analyzed. They collected heparin-anticoagulated peripheral blood from 21 multiple sclerosis (MS) patients and nine healthy controls. Ten of the 21 samples were used to profile mRNA expression, and the 11 remaining were used to profile miRNA expression. These datasets are accessible at NCBI GEO database [43], with accession GSE43592. We preprocessed the datasets using the LIMMA package [135]. For demonstration purposes, we included the preprocessed datasets on this package.

```
data(GSE43592_miRNA)
data(GSE43592_mRNA)
```

Once researchers have the data and the augmented pathways, they can run the pathway analysis method that they prefer. We suggest using RONTOTOOLS package [150] because it takes into account the topology of the pathways (the method implemented on RONTOTOOLS is explained on [149]). We show here how to perform

mirIntegrator app Pathways Plot Augmented Pathways **Run Pathway Analysis** R package

Choose the mRNA dataset: GSE43592_mRNA Choose the microRNA dataset: GSE43592_miRNA

Run Impact Analysis

Show 10 entries

Pathway_name	pPert	pComb	pPert.fdr	pComb.fdr
path:hsa04062 Chemokine signaling pathway	0.004975124	5.016434e-07	0.01785942	7.023008e-05
path:hsa05134 Legionellosis	0.004975124	4.019915e-06	0.01785942	2.813940e-04
path:hsa04060 Cytokine-cytokine receptor interaction	0.009950249	1.312671e-05	0.02443921	4.896292e-04
path:hsa05202 Transcriptional misregulation in cancer	0.004975124	1.590218e-05	0.01785942	4.896292e-04
path:hsa04064 NF-kappa B signaling pathway	0.019900498	1.748676e-05	0.04286261	4.896292e-04
path:hsa04621 NOD-like receptor signaling pathway	0.009950249	4.505478e-05	0.02443921	1.051278e-03
path:hsa05160 Hepatitis C	0.009950249	7.557872e-05	0.02443921	1.334499e-03
path:hsa05131 Shigellosis	0.004975124	7.625710e-05	0.01785942	1.334499e-03
path:hsa04622 RIG-I-like receptor signaling pathway	0.004975124	2.094673e-04	0.01785942	3.258380e-03
path:hsa05168 Herpes simplex infection	0.004975124	2.728167e-04	0.01785942	3.819434e-03

Showing 1 to 10 of 142 entries Previous 1 2 3 4 5 ... 15 Next

Figure 5.2: A screenshot of the mirIntegrator’s graphical user interface (GUI). Users first upload their sample-matched mRNA and miRNA datasets. They can choose the datasets from the drop boxes. To perform data analysis, users just press the button *Run Impact Analysis*. The output of the software is a list of augmented pathways ranked according to the joint probability of having both the observed level of enrichment as well as the observed level of perturbation just by chance. This probability is corrected for multiple comparisons with the false discovery rate adjustment (pComb.fdr). Other statistics, such as p-values for the observed perturbation (pPert), raw p-value for the combined enrichment and perturbation (pComb), and the FDR-corrected p-value for perturbation alone (pPert.fdr), are also reported.

impact pathway analysis for the augmented pathways:

```

data(GSE43592_mRNA)
data(GSE43592_miRNA)
data(augmented_pathways)
data(names_pathways)
lfcMRNA <- GSE43592_mRNA$logFC
names(lfcMRNA) <- GSE43592_mRNA$entrez
lfcMiRNA <- GSE43592_miRNA$logFC
names(lfcMiRNA) <- GSE43592_miRNA$entrez
keggGenes <- unique(unlist( lapply(augmented_pathways,nodes) ) )
interMi <- intersect(keggGenes, GSE43592_miRNA$entrez)
interM <- intersect(keggGenes, GSE43592_mRNA$entrez)
peRes <- pe(x= c(lfcMRNA, lfcMiRNA ),
           graphs=augmented_pathways, nboot = 200, verbose = FALSE)
message(paste("There are ", length(unique(GSE43592_miRNA$entrez)),
             "miRNAs measured and",length(interMi),
             "of them were included in the analysis."))
message(paste("There are ", length(unique(GSE43592_mRNA$entrez)),
             "mRNAs measured and", length(interM),
             "of them were included in the analysis."))
summ <- Summary(peRes)
rankL <- data.frame(summ,path.id = row.names(summ))
tableKnames <- data.frame(path.id = names(names_pathways),names_pathways)
rankL <- merge(tableKnames, rankL, by.x = "path.id", by.y = "path.id")
head(rankL)

```

Chapter 6: Method Validation

6.1 Validation outline

In this section, we present a detailed quantification of the augmented pathways and the results of our pathway analysis pipeline using the miRNA-augmented pathways (mirAP). For this purpose, we downloaded and augmented signaling pathways with miRNA-gene interaction from KEGG [74] and mirTarBase[65], respectively. To validate our pathway analysis pipeline, we perform pathway analysis of 9 mRNA/miRNA sample-matched datasets using two different methods (Impact Analysis and ORA) and show that mirAP offers a significant improvement over analyzing mRNA data alone. We also compare the obtained results with the state-of-the-art method (microGraphite) [19].

6.2 Descriptive statistics of the augmented pathways

We augmented the KEGG pathways to analyze the change in pathways size after introducing miRNAs. Figure 6.2 shows some statistics about the number of nodes of each pathway before and after augmentation. We sort the pathways in an increasing order of the number of genes in the original pathways for a clear visualization. The red line shows the size (number of genes) of the original pathways. The average number of genes in a KEGG pathway is 102. The blue line indicates the number of miRNAs added to each pathway. On average, each pathway is augmented with 134 miRNAs. The green line shows the total number of both mRNA and miRNA for each pathway after augmentation. On average, an augmented pathway has 234 entities (genes and miRNAs). In essence, the size of each pathway is approximately doubled after augmentation. These results show that pathways are heavily regulated by miRNAs and confirm the crucial role of miRNAs, as well as the importance of

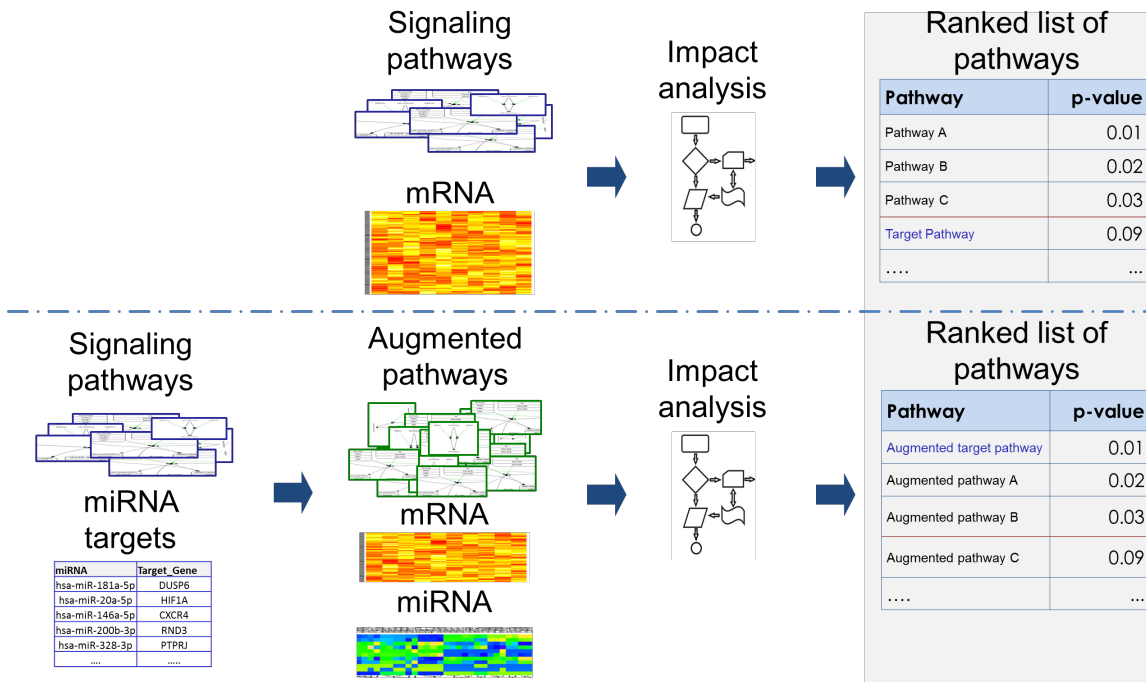


Figure 6.1: Evaluation scheme.

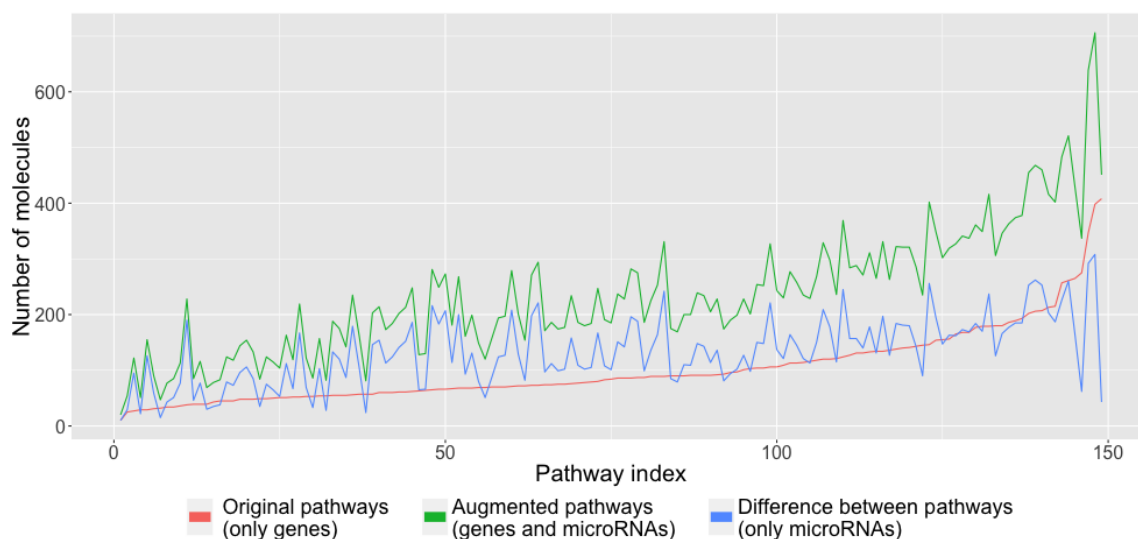


Figure 6.2: Comparison of pathway sizes before and after augmentation. The pathways were sorted by KEGG pathway size. The red line shows the size (number of genes) of the original pathways. The blue line shows the number of miRNAs added each pathway. The green line shows the total number of both mRNA and miRNA for each pathway after augmentation. On average, the size of each pathway is doubled after augmentation. Plot generated using the function `plot_change`.

Table 6.1: Description of the analyzed datasets

Ref.	GEO ID	Pubmed	Disease/Target pathway	KEGGID
[76]	GSE26168	21829658	Type II diabetes	hsa04930
[95]	GSE62699	26381263	Alcoholism	hsa05034
[118]	GSE35834	23987127	Colorectal cancer	hsa05210
[114]	GSE43797	24072181	Pancreatic cancer	hsa05212
[94]	GSE29250	22046296	Non-small cell lung cancer	hsa05223
[36]	GSE32688	22261810	Pancreatic cancer	hsa05212
[69]	GSE43592	23895517	Amyotrophic lateral sclerosis	hsa05014
[155]	GSE35389	23056502	Melanoma	hsa05218
[51]	GSE35982	22703586	Colorectal cancer	hsa05210

being able to analyze miRNA and mRNA data together. Although the framework was demonstrated on KEGG pathways, it can exploit information available in other databases, such as functional modules available in Gene Ontology database [23] or protein-protein interactions available in the STRING database [151].

6.3 Results

We analyzed nine sample-matched datasets from seven different diseases: GSE43592 (multiple sclerosis, 10 controls, 10 cases), GSE35389 (melanoma, 4 controls, 4 cases), GSE35982 (colorectal cancer, 8 controls, 8 cases), GSE26168 (type II diabetes, 8 controls, 9 cases), GSE62699 (alcoholism, 18 controls, 18 cases), GSE35834 (colorectal cancer, 23 controls, 55 cases), GSE43797 (pancreatic cancer, 5 controls, 7 cases), GSE29250 (non-small cell lung cancer, 6 controls, 6 cases), and GSE32688 (pancreatic cancer, 7 controls, 25 cases). For each of these datasets, we used the normalized expression values as found in GEO [7]. The microarray probes were annotated according to their platform’s metadata using GEOquery [29]. Next, we estimated log-fold-change between disease and control groups by fitting to a gene-wise linear model using the `limma` package [122]. We only took into consideration mRNA and miRNA with adjusted p-values lower than 5%. Among these significant mRNAs and miRNAs, we chose the ones that have the highest fold-change as differentially expressed, up to

10% of measured mRNAs and miRNAs.

The nine datasets were selected due to two important reasons. First, these datasets have both mRNA and miRNA measurements for the same set of patients. Second, for each of the underlying diseases, there is a KEGG pathway, henceforth *target pathway*, that was created to describe the underlying mechanisms of the disease. To demonstrate the advantage of the miRNA data integration, we compared the use of the original KEGG pathways with the use of our miRNA augmented pathways (mirAP). We performed two pathway analysis methods that use p-value and fold-change for each set of pathways: impact analysis (IA) [41, 140] and over-representation analysis (ORA) [40]. The input for IA and ORA using KEGG is mRNA expression data. The input for IA and ORA using mirAP includes both mRNA and miRNA expression data. The output of each method is a list of p-values (one per each pathway). These p-values are adjusted for multiple comparisons using the false discovery rate approach (FDR) [10].

We also analyze the nine GEO datasets using microGraphite [19] after quantile normalization [16] to compare with our pipeline. The primary goal of microGraphite is the identification of signal transduction paths correlated with the condition under study. It is implemented in a four-steps recursive procedure as follows: (i) selecting pathways, (ii) best path identification, (iii) meta-pathway construction, and (iv) meta-pathway analysis. Here we only consider the first step of the approach, which is the selection of significant pathways. This selection is based on the significance levels obtained from the test on the mean of the pathways (alpha-mean). The input is the mRNA and miRNA expression data, and it does not take in account fold-changes nor differentially expressed entities.

For each dataset, we expect a good method to identify the target pathway as significant, as well as to rank it on top. For instance, in the colorectal cancer dataset which compares colorectal cancer tissue vs. normal, the *Colorectal cancer pathway*

Table 6.2: Results of target pathway identification using ORA (column 3), IA (col. 4), ORA on mirAP (col. 5), IA on mirAP (col. 6), microGraphite (col. 7)

GEO ID	Target pathway	ORA	IA	ORAmir	IAmir	microG.
GSE26168	Type II diabetes mellitus	no	no	no	no	yes
GSE29250	Non-small cell lung canc.	no	no	yes	no	no
GSE35982	Colorectal cancer	no	no	no	no	no
GSE32688	Pancreatic cancer	no	no	yes	yes	no
GSE35389	Melanoma	no	no	yes	yes	no
GSE35834	Colorectal cancer	no	no	yes	yes	no
GSE43592	Amyotrophic lateral scl.	no	no	no	yes	no
GSE43797	Pancreatic cancer	no	no	yes	yes	yes
GSE62699	Alcoholism	no	no	no	yes	no

must be shown as significant and should be as close to the top of the ranking as possible since this is the pathway that describes the phenomena involved in colorectal cancer. Based on this evaluation, we compared the rank and p-value of the target pathway in each disease for five methods:

- i) mRNA expression alone using standard KEGG pathways with ORA
- ii) mRNA expression alone using standard KEGG pathways with IA
- iii) mRNA and miRNA expression data using the augmented pathways (mirAP) with ORA
- iv) mRNA and miRNA expression data using mirAP with IA
- v) mRNA and miRNA expression data analyzed with microGraphite.

Table 6.2 shows the target pathways and their significance for the nine datasets. The first and second columns display the datasets and their corresponding target pathways while the other five columns indicate whether the target pathways are identified as significant using the five methods: ORA of mRNA expression on KEGG pathways (ORA+KEGG), IA of mRNA expression on KEGG (IA+KEGG), ORA of miRNA and mRNA expression data on mirRNA-augmented pathways (ORA+mirAP), our

approach IA of miRNA and mRNA expression on mirAP (IA+mirAP), and miRNA and mRNA expression analysis using microGraphite, respectively. The significance threshold is 5% for FDR p-values. IA and ORA fail to identify any target pathway as significant when using just mRNA whereas our approach (IA+mirAP) correctly identify the target pathway in 6 out of 9 datasets (GSE32688, GSE35389, GSE35834, GSE43592, GSE43797, GSE62699). ORA on mirAP correctly identify the target pathway as significant in 5 out of 9 datasets (GSE29250, GSE32688, GSE35389, GSE35834, GSE43797). microGraphite correctly identifies the target pathway as significant in only 2 out of 9 datasets (GSE26168, GSE43797). The results demonstrate that our integration of mRNA and miRNA lifts the statistical power for both pathway analysis techniques (ORA and IA) and outperforms microGraphite in target pathway identification.

Figure 6.4 shows the p-values and rankings of the target pathways using the five methods. The panel (a) shows the FDR corrected p-values of the target pathways. We compare the lists of p-values using t-test and Wilcoxon test. The adjusted p-values produced by IA+mirAP are significantly smaller than those of IA+KEGG ($p=0.002$ using t-test and $p=0.007$ using Wilcoxon test), ORA+KEGG ($p=0.001$ using the t-test, and $p=0.005$ using Wilcoxon test), and microGraphite ($p=0.006$ using t-test and $p=0.009$ using Wilcoxon test).

The panel (b) shows the rankings of the target pathways. Again, the rankings produced by IA+mirAP are significantly smaller than those of IA+KEGG ($p=0.03$ using the t-test, and $p=0.04$ using Wilcoxon test), ORA+KEGG ($p=0.03$ using t-test and $p=0.04$ using Wilcoxon test), and microGraphite ($p=0.0051$ using t-test and $p=0.0058$ using Wilcoxon test). This result confirms that our augmented pathways, mirAP, improve the performance of traditional Impact Analysis and ORA. Also, the results show that the proposed integrative pathway analysis also outperforms microGraphite in both p-values and rankings for target pathway identification. Furthermore, our

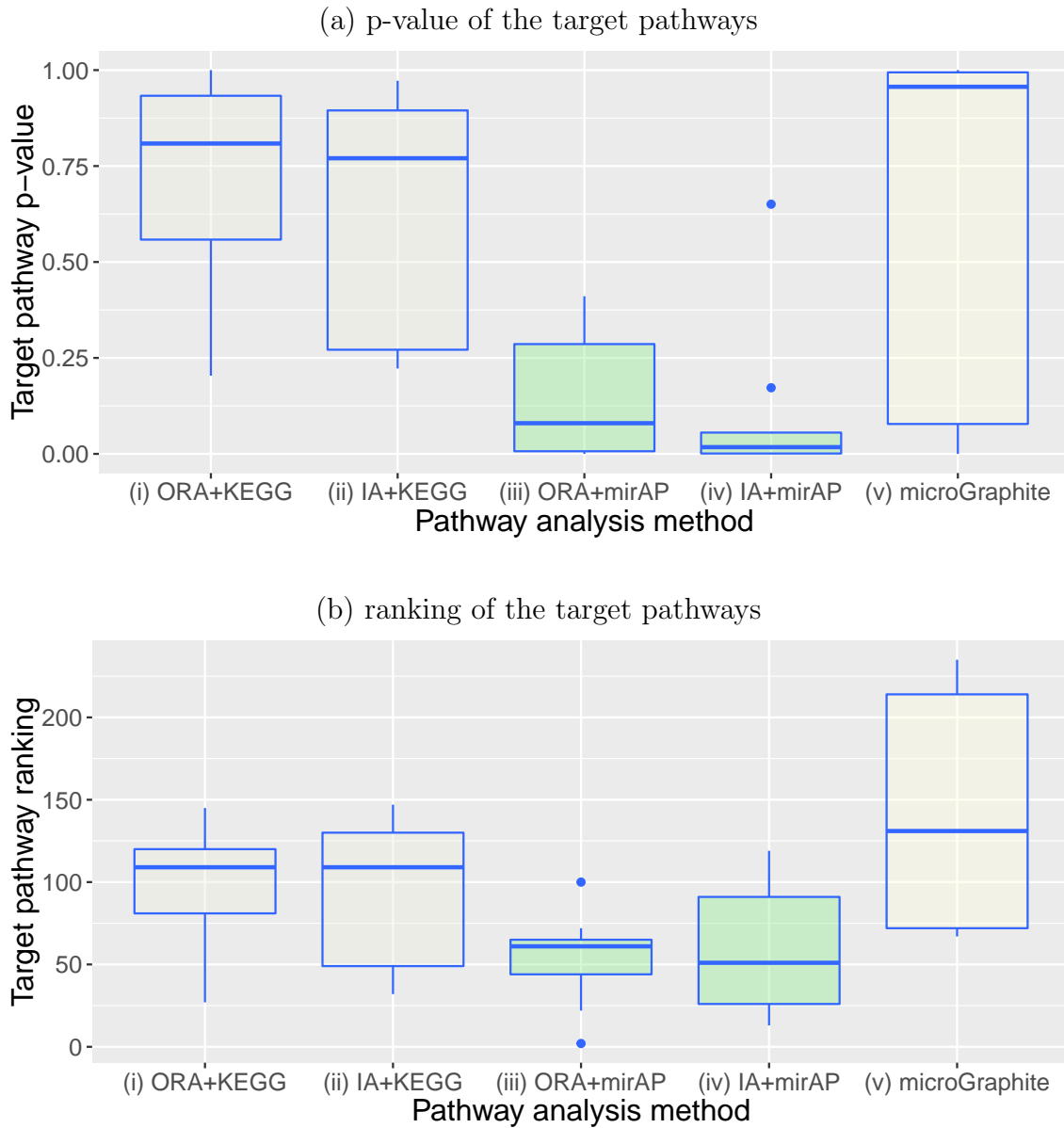


Figure 6.4: Corrected p-values and rankings of the target pathways using different methods.

pathway database (mirAP) is generated with validated miRNA-mRNA interactions, while microGraphite uses predicted interactions which increase the number of false positive miRNA-target interactions. Another drawback of microGraphite is its execution time. A typical analysis with microGraphite takes approximately 22 hours while our approach takes only a few minutes. We ran these experiments on a standard desktop workstation with a 2.6 GHz Intel Core i5, 8GB of RAM, on a single thread, and the OS X 10.11 operative system.

Chapter 7: Reference Manual

The reference manual presented here is also available at Bioconductor (<https://bioconductor.org/packages/release/bioc/html/mirIntegrator.html>).

Package mirIntegrator

Version 1.4.0

Date 2016-07-02

Type Package

Title Integrating microRNA expression into signaling pathways for pathway analysis

Author Diana Diaz (dmd at wayne dot edu)

Maintainer Diana Diaz (dmd at wayne dot edu)

Depends R (≥ 3.3)

Imports graph,ROntoTools, ggplot2, org.Hs.eg.db, AnnotationDbi, Rgraphviz

Suggests RUnit, BiocGenerics

Description Tools for augmenting signaling pathways to perform pathway analysis of microRNA and mRNA expression levels.

License GPL (≥ 3)

URL <http://datad.github.io/mirIntegrator/>

biocViews Network, Microarray, GraphAndNetwork, Pathways, KEGG

NeedsCompilation no

augmented_pathways

Signaling pathways augmented with miRNA.

Description

Human signaling KEGG pathways augmented with validated miRNA-target interactions from mirTarBase using the mirIntegrator package. These interactions represent the biological miRNA repression of its target genes and are included in the model as negative links.

Usage

```
data("augmented_pathways")
```

Value

A list of graphNEL objects where each graph is a pathway that was augmented with miRNA-target interactions. The name of each pathway is its KEGG pathway identifier.

Source

Generated using the mirIntegrator package. A script that constructs the augmented_pathways object may be found in 'inst/scripts/get_augmented_pathways.R', see the example.

References

M. Kanehisa and S. Goto, *KEGG: Kyoto Encyclopedia of Genes and Genomes*, Nucleic Acids Research, vol. 28, pp. 27-30, January 2000.

S.-D. Hsu, Y.-T. Tseng, S. Shrestha, Y.-L. Lin, A. Khaleel, C.-H. Chou, C.-F. Chu, H.-Y. Huang, C.-M. Lin, S.-Y. Ho, T.-Y. Jian, F.-M. Lin, T.-H. Chang, S.-L. Weng, K.-W. Liao, I.-E. Liao, C.-C. Liu, and H.-D. Huang, *miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions*, Nucleic Acids Research, vol. 42, pp. D78 - D85, Jan. 2014.

See Also

mirTarBase and kegg_pathways

Examples

```
data(augmented_pathways)
head(augmented_pathways)

script <- system.file("scripts", "get_augmented_pathways.R",
                      package = "mirIntegrator")

script
readLines(script)
```

GSE43592_miRNA *Top table of preprocessed miRNA of GSE43592 dataset.*

Description

A data.frame with the Log fold change and p-value of preprocessed miRNA expression of GSE43592 dataset.

Usage

```
data(GSE43592_miRNA)
```

Value

A data frame with 881 miRNAs with the following 8 variables: entre, ID, logFC, AveExpr, t, P.Value, adj.P.Val, B.

Source

Raw data obtained from <http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE43592> and preprocessed with the limma package version 3.24.0.

References

M. Jernas, C. Malmstrom, M. Axelsson, I. Nookaew, H. Wadenvik, J. Lycke, and B. Olsson, *MicroRNA regulate immune pathways in t-cells in multiple sclerosis (MS)*, BMC immunology, vol. 14, p. 32, 2013.

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W and Smyth GK (2015). *limma powers differential expression analyses for RNA-sequencing and microarray studies*. Nucleic Acids Research, 43(7), pp. e47.

Examples

```
data(GSE43592_miRNA)
head(GSE43592_miRNA)
```

GSE43592_mRNA	<i>Top table of preprocessed mRNA of GSE43592 dataset.</i>
---------------	--

Description

A data.frame with the Log fold change and p-value of preprocessed mRNA expression of GSE43592 dataset.

Usage

```
data(GSE43592_mRNA)
```

Value

A data frame with 19611 mRNAs with the following 8 variables: entre, ID, logFC, AveExpr, t, P.Value, adj.P.Val, B.

Source

Raw data obtained from <http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE43592> and preprocessed with the limma package version 3.24.0.

References

M. Jernas, C. Malmstrom, M. Axelsson, I. Nookaew, H. Wadenvik, J. Lycke, and B. Olsson, *MicroRNA regulate immune pathways in t-cells in multiple sclerosis (MS)*, BMC immunology, vol. 14, p. 32, 2013.

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W and Smyth GK (2015). *limma powers differential expression analyses for RNA-sequencing and microarray studies*. Nucleic Acids Research, 43(7), pp. e47.

Examples

```
data(GSE43592_mRNA)
head(GSE43592_mRNA)
```

```
integrate_mir      Produce augmented pathways
```

Description

This function takes each pathway of the input list of signaling pathways and adds the miRNAs that are related to it.

Usage

```
integrate_mir(original_pathways, targets_db)
```

Arguments

original_pathways

A list of `graph::graphNEL` objects where each of the nodes is named with '`<gene.ID>`'. Gene IDs used to identify the nodes must be the same gene IDs used to identify the genes on the miRNA-target interactions data.frame, `targets_db`. i.e. If the genes are identified by Entrez ID on the `original_pathways` `graph::graphNEL` list, then the `targets_db` data.frame must identify the genes by Entrez ID as well. Nodes of each `graph::graphNEL` represent the genes involved in the pathway and edges represent the biological interactions (activation or repression) among those genes (activation or repression).

targets_db A data.frame with columns: 'miRNA' which names the miRNAs and 'Target.ID' which gives the gene ID of the target gene. The Gene IDs used to identify the "Target.ID" column must be the same gene IDs used on the nodes of the `original_pathways`. i.e. If the genes are identified by Entrez ID on the `original_pathways` `graph::graphNEL` list, then the `targets_db` data.frame must identify the genes by Entrez ID as well.

Value

Gene signaling pathways augmented with miRNA interactions. The augmented pathways are contained in a list of `graph::graphNEL` objects where each of the nodes is named with '`<gene.ID>`'. Nodes of each `graph::graphNEL` represent

genes and miRNAs involved in the pathway and edges represent the biological interactions (activation or repression) among them.

Author(s)

Diana Diaz

Examples

```
data(kegg_pathways)
data(mirTarBase)
kegg_pathways <- kegg_pathways[1:5] #delete this for augmenting all pathways.
augmented_pathways <- integrate_mir(kegg_pathways, mirTarBase)
```

<code>kegg_pathways</code>	<i>List of KEGG signaling pathways of human.</i>
----------------------------	--

Description

This dataset contains 149 KEGG human signaling pathways. The original pathways were parsed to a list of graphNEL objects using the ROntoTools package. The original KEGG pathways were published by Kanehisa Laboratories, release 73.0+/01-03, Jan 2015.

Usage

```
data("kegg_pathways")
```

Value

A list of graphNEL objects where each graph represents one KEGG signaling pathway. The name of each pathway is its KEGG pathway identifier.

Source

Obtained using the ROntoTools package Version 1.2.0 with KEGG database release 73.0+/01-03, Jan 2015. A script that constructs the kegg_pathways object may be found in 'inst/scripts/get_kegg_pathways.R', see the example.

References

M. Kanehisa and S. Goto, *KEGG: Kyoto Encyclopedia of Genes and Genomes*, Nucleic Acids Research, vol. 28, pp. 27-30, January 2000.

C. Voichita, M. Donato, and S. Draghici, *Incorporating gene significance in the impact analysis of signaling pathways*, in 2012 11th International Conference on Machine Learning and Applications (ICMLA), vol. 1, pp. 126-131, Dec. 2012.

Examples

```
data(kegg_pathways)
head(kegg_pathways)

script <- system.file("scripts", "get_kegg_pathways.R",
                      package = "mirIntegrator")
script
readLines(script)
```

mirTarBase	<i>MicroRNA-target interactions in human.</i>
------------	---

Description

Dataset of miRNA-target interactions in human obtained from mirTarBase release 4.5: Nov. 1, 2013.

Usage

```
data(mirTarBase)
```

Format

A data.frame with 39083 interactions and 9 variables. The columns needed for this package are:

- `miRNA`: which contains the miRNA ID,
- `Target.ID`: contains the entrez ID of the gene targeted by the miRNA

Details

This dataset is licensed by its authors (Hsu et al.), see <http://mirtarbase.mbc.nctu.edu.tw/cache/download/LICENSE>.

Value

A data.frame with human miRNA-targets interactions

Source

Downloaded from <http://mirtarbase.mbc.nctu.edu.tw/> on 4/1/2015. A script which downloads the file and constructs the `mirTarBase` object may be found in `'inst/scripts/get_mirTarBase.R'`, see the example.

References

S.-D. Hsu, Y.-T. Tseng, S. Shrestha, Y.-L. Lin, A. Khaleel, C.-H. Chou, C.-F. Chu, H.-Y. Huang, C.-M. Lin, S.-Y. Ho, T.-Y. Jian, F.-M. Lin, T.-H. Chang, S.-L. Weng, K.-W. Liao, I.-E. Liao, C.-C. Liu, and H.-D. Huang, *miRTarBase*

update 2014: an information resource for experimentally validated miRNA-target interactions, Nucleic Acids Research, vol. 42, pp. D78 - D85, Jan. 2014.

Examples

```
data(mirTarBase)
head(mirTarBase)

script <- system.file("scripts", "get_mirTarBase.R",
                      package = "mirIntegrator")
script
readLines(script)
```

`names_pathways` *List of KEGG signaling pathways' names.*

Description

Names of the KEGG signaling pathways in human obtained with the ROntoTools package. The original KEGG pathways were published by Kanehisa Laboratories, release 73.0+/01-03, Jan 2015.

Usage

```
data("names_pathways")
```

Value

A list of KEGG signaling pathways' names.

Source

Obtained using the ROntoTools package Version 1.2.0 with KEGG database release 73.0+/01-03, Jan 2015. A script that constructs the `names_pathways` object may be found in ‘`inst/scripts/get_names_pathways.R`’, see the example.

References

M. Kanehisa and S. Goto, *KEGG: Kyoto Encyclopedia of Genes and Genomes*, Nucleic Acids Research, vol. 28, pp. 27-30, January 2000.

C. Voichita, M. Donato, and S. Draghici, *Incorporating gene significance in the impact analysis of signaling pathways*, in 2012 11th International Conference on Machine Learning and Applications (ICMLA), vol. 1, pp. 126-131, Dec. 2012.

Examples

```
data(names_pathways)
head(names_pathways)

script <- system.file("scripts", "get_names_pathways.R",
                      package = "mirIntegrator")
script
readLines(script)
```

`pathways2pdf` *Export augmented pathways to pdf*

Description

This function creates a pdf file with plottings of a list of augmented pathways.

Usage

```
pathways2pdf(original_pathways, augmented_pathways, pathway_names, file)
```

Arguments

original_pathways

A list of `graph::graphNEL` objects where each of the nodes is named with '`<gene_ID>`'. Nodes of each `graph::graphNEL` represent the genes involved in the pathway and edges represent the biological interactions (activation or repression) among those genes (activation or repression).

augmented_pathways

A list of `graph::graphNEL` objects where each of the nodes is named with '`<gene_ID>`'. Nodes of each `graph::graphNEL` represent genes and miRNAs involved in the pathway and edges represent the biological interactions (activation or repression) among them.

pathway_names

A list of names of the pathways identified by '`<pathway_ID>`'.

file

The name of the file where the plots will be saved.

Value

A pdf file with the plottings of the augmented pathways.

Author(s)

Diana Diaz

Examples

```
data(augmented_pathways)
```

```

data(kegg_pathways)
data(names_pathways)
#The following instruction writes a pfd with three pathways
pathways2pdf(kegg_pathways[18:20],augmented_pathways[18:20],
             names_pathways[18:20], "three_pathways.pdf")
#The following instruction writes a pfd with all the pathways:
#NOTE: It may take time.
# pathways2pdf(kegg_pathways,augmented_pathways,
#             names_pathways, "all_pathways.pdf")

```

```
plot_augmented_pathway
```

Plotting of augmented pathway

Description

Functions for plotting a particular augmented pathway. In the plot, miRNAs that were added to the original pathway are differentiated from proteins that were originally in the pathway. Blue boxes represent the proteins that were part of the original pathway, and black boxes represent the miRNAs that were added during augmentation.

Usage

```

plot_augmented_pathway(original_pathway, augmented_pathway,
                       pathway_name = " ", ...)

```

`plot_change`*Plotting the change in pathways order*

Description

Function for plotting a lines plot of the difference in pathways' order. The resultant plot shows the comparison between the order of the original pathways and the order of the augmented pathways. It also contains a line with the order difference (order of the augmented pathways minus order of the original pathways). The order of a biological pathway is the number of genes that are involved in it.

Usage

```
plot_change(original_pathways, augmented_pathways, pathway_names, ...)
```

Arguments

`original_pathways`

A list of `graph::graphNEL` objects where each of the nodes is named with '<gene_ID>'. Nodes of each `graph::graphNEL` represent the genes involved in the pathway and edges represent the biological interactions (activation or repression) among those genes (activation or repression).

`augmented_pathways`

A list of `graph::graphNEL` objects where each of the nodes is named with '<gene_ID>'. Nodes of each `graph::graphNEL` represent genes and miRNAs involved in the pathway and edges represent the biological interactions (activation or repression) among them.

`pathway_names`

A list of names of the pathways identified by '`<pathway_ID>`'.

... Other arguments for the '`<plotLines>`' function.

Value

A lines plot of the comparison of pathways order.

Author(s)

Diana Diaz

Examples

```
data(augmented_pathways)
data(kegg_pathways)
data(names_pathways)
plot_change(kegg_pathways, augmented_pathways, names_pathways)
```

`smallest_pathway` *Get the smallest pathway*

Description

Find the pathway with the fewer number of nodes among a list of pathways. This simple function is an example of how to navigate the genes on a list of pathways.

Usage

```
smallest_pathway(pathways)
```

Arguments

`pathways` A list of `graph::graphNEL` objects.

Value

The index of the pathway with fewer number of nodes.

Author(s)

Diana Diaz

Examples

```
data(augmented_pathways)
smallest_pathway(augmented_pathways)
smallest_pathway
```

Chapter 8: Discussion and Conclusion

In this thesis, we present the background for integrating multiple types of data for the purpose of pathway analysis and propose a method to augment signaling pathways with miRNA-target interactions. We also show that miRNA-augmented pathways (mirAP) offer a more comprehensive view and a deeper understanding of complex diseases. Our contributions include a pipeline that use mirAP to integrate miRNA and mRNA expression data for the purpose of pathway analysis, a publicly available database of miRNA-augmented pathways, and an open source Bioconductor package (mirIntegrator).

We describe methods for integrating multiple types of data for the purpose of pathway analysis. We classified these methods into two broad categories: topology based and non-topology based approaches. Topology-based methods take into consideration the topology and interactions between genes while non-topology based approaches treat a pathway as a set of genes or entities. When integrating multi-omics data, the existing pathways (designed for gene expression) are expanded to include other data types. Integrative pathway analysis methods extend the current graphs by adding new nodes, relations, and interactions. The added nodes and links represent the new data types. Then, the integrated model can be analyzed using the classical pathway analysis methods. The input for integrative pathway analysis includes the expanded pathways and multi-omics data. However, one major drawback of the existing integrative pathway analysis methods is that they are slow due to the intensive computation required by the graphical models. Graphical models were initially designed for graphs with less than a hundred nodes. However, the number of entities in biological data can include thousands of genes or molecules. Additionally, existing approaches combine expression values that are measured on different technologies or platforms. Combining expression values requires cross-platform normalization which

increases the statistical error. To fill this gap, we integrate the data using p-values of differentially expressed entities instead of expression values. The combination of p-values provides a significant advantage as p-values and do not require cross-platform normalization.

As miRNA expression data are becoming freely accessible, miRNA-mRNA integrative analyses are likely to become a routine. Our pathway analysis pipeline augments gene-gene signaling pathways with miRNA-target interactions. Then we perform a topology-based pathway analysis taking into consideration both types of molecular data. To demonstrate the power of the integrative analysis, we compared our approach (mirIntegrator) with the state of the art method, microGraphite, using nine sample-matched datasets that were assayed in independent labs. While microGraphite failed to identify target pathways as significant for 7 out of 9, our approach correctly identified the target pathway as significant in 6 out of 9 datasets. Also, mirIntegrator produced significantly smaller p-values and rankings of the target pathways. In summary, our pipeline outperforms the state of the art method for identifying target pathways (smaller p-values and rankings of the target pathways).

REFERENCES

- [1] AKULA, S. P., MIRIYALA, R. N., THOTA, H., RAO, A. A., AND GEDELA, S. Techniques for integrating omics data. *Bioinformatics* 3, 6 (Jan. 2009), 284–286.
- [2] AL-SHAHROUR, F., DÍAZ-URIARTE, R., AND DOPAZO, J. Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information. *Bioinformatics* 21, 13 (2005), 2988–2993.
- [3] ASHBURNER, M., BALL, C. A., BLAKE, J. A., BOTSTEIN, D., BUTLER, H., CHERRY, J. M., DAVIS, A. P., DOLINSKI, K., DWIGHT, S. S., EPPIG, J. T., HARRIS, M. A., HILL, D. P., ISSEL-TARVER, L., KASARSKIS, A., LEWIS, S., MATESE, J. C., RICHARDSON, J. E., RINGWALD, M., RUBIN, G. M., AND SHERLOCK, G. Gene Ontology: tool for the unification of biology. *Nature Genetics* 25 (2000), 25–29.
- [4] BACKES, C., MEESE, E., LENHOF, H.-P., AND KELLER, A. A dictionary on microRNAs and their putative target pathways. *Nucleic Acids Research* 38, 13 (2010), 4476–4486.
- [5] BALDI, P., AND LONG, A. D. A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics* 17, 6 (2001), 509–519.
- [6] BARRELL, D., DIMMER, E., HUNTLEY, R. P., BINNS, D., O'DONOVAN, C., AND APWEILER, R. The goa database in 2009—an integrated gene ontology annotation resource. *Nucleic Acids Research* 37, suppl 1 (2009), D396–D403.
- [7] BARRETT, T., SUZEK, T. O., TROUP, D. B., WILHITE, S. E., NGAU, W. C., LEDOUX, P., RUDNEV, D., LASH, A. E., FUJIBUCHI, W., AND

- EDGAR, R. NCBI GEO: mining millions of expression profiles—database and tools. *Nucleic Acids Research* 33, Database Issue (2005), D562–6.
- [8] BARRY, W. T., NOBEL, A. B., AND WRIGHT, F. A. Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics* 21, 9 (May 2005), 1943–1949.
- [9] BEISSBARTH, T., AND SPEED, T. P. GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics* 20 (June 2004), 1464–1465.
- [10] BENJAMINI, Y., AND YEKUTIELI, D. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics* 29, 4 (August 2001), 1165–1188.
- [11] BERGER, B., PENG, J., AND SINGH, M. Computational solutions for omics data. *Nature Reviews Genetics* 14, 5 (May 2013), 333+. 333.
- [12] BERRIZ, G. F., KING, O. D., BRYANT, B., SANDER, C., AND ROTH, F. P. Characterizing gene sets with FuncAssociate. *Bioinformatics* 19, 18 (2003), 2502–2504.
- [13] BEWICK, V., CHEEK, L., AND BALL, J. Statistics review 12: Survival analysis. *Critical Care* 8, 5 (2004), 389–394.
- [14] BIO CARTA. BioCarta - Charting Pathways of Life. <http://www.biocarta.com>.
- [15] BOJA, E. S., KINSINGER, C. R., RODRIGUEZ, H., SRINIVAS, P., AND \$AUTHOR.LASTNAME, A. F. Integration of omics sciences to advance biology and medicine. *Clinical Proteomics* 11, 1 (Dec. 2014), 45.

- [16] BOLSTAD, B. M., IRIZARRY, R. A., ÅSTRAND, M., AND SPEED, T. P. A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on variance and bias. *Bioinformatics* 19, 2 (January 2003), 185–193.
- [17] BRAZMA, A., PARKINSON, H., SARKANS, U., SHOJATALAB, M., VILO, J., ABEYGUNAWARDENA, N., HOLLOWAY, E., KAPUSHESKY, M., KEMMEREN, P., LARA, G. G., OEZCIMEN, A., ROCCA-SERRA, P., AND SANSONE, S.-A. ArrayExpress—a public repository for microarray gene expression data at the EBI. *Nucleic Acids Research* 31, 1 (2003), 68–71.
- [18] BRESLIN, T., EDEN, P., AND KROGH, M. Comparing functional annotation analyses with Catmap. *BMC Bioinformatics* 5, 1 (2004), 193.
- [19] CALURA, E., MARTINI, P., SALES, G., BELTRAME, L., CHIORINO, G., D’INCALCI, M., MARCHINI, S., AND ROMUALDI, C. Wiring miRNAs to pathways: a topological approach to integrate miRNA and mRNA expression profiles. *Nucleic Acids Research* 42, 11 (2014), e96.
- [20] CAMON, E., MAGRANE, M., BARRELL, D., LEE, V., DIMMER, E., MASLEN, J., BINNS, D., HARTE, N., LOPEZ, R., AND APWEILER, R. The gene ontology annotation (goa) database: sharing knowledge in uniprot with gene ontology. *Nucleic Acids Research* 32, suppl 1 (2004), D262–D266.
- [21] CHALISE, P., KOESTLER, D. C., BIMALI, M., YU, Q., AND FRIDLEY, B. L. Integrative clustering methods for high-dimensional molecular data. *Translational cancer research* 3, 3 (June 2014), 202–216.
- [22] CHANG, W., CHENG, J., ALLAIRE, J., XIE, Y., AND MCPHERSON, J. *shiny: Web Application Framework for R*, 2016. R package version 0.13.2.
- [23] CONSORTIUM, G. O., ET AL. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research* 32, suppl 1 (2004), D258–D261.

- [24] COX, D. R. Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)* 34, 2 (1972), 187–220.
- [25] CREIGHTON, C. J., NAGARAJA, A. K., HANASH, S. M., MATZUK, M. M., AND GUNARATNE, P. H. A bioinformatics tool for linking gene expression profiling results with public databases of microRNA target predictions. *RNA* 14, 11 (Nov. 2008), 2290–2296.
- [26] CROFT, D., MUNDO, A. F., HAW, R., MILACIC, M., WEISER, J., WU, G., CAUDY, M., GARAPATI, P., GILLESPIE, M., KAMDAR, M. R., JASSAL, B., JUPE, S., MATTHEWS, L., MAY, B., PALATNIK, S., ROTHFELS, K., SHAMOVSKY, V., SONG, H., WILLIAMS, M., BIRNEY, E., HERMJAKOB, H., STEIN, L., AND D’EUSTACHIO, P. The Reactome pathway knowledgebase. *Nucleic Acids Research* 42, D1 (2014), D472–D477.
- [27] CURTIS, C., SHAH, S. P., CHIN, S.-F., TURASHVILI, G., RUEDA, O. M., DUNNING, M. J., SPEED, D., LYNCH, A. G., SAMARAJIWA, S., YUAN, Y., ET AL. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* 486, 7403 (2012), 346–352.
- [28] DAS, J., PODDER, S., AND GHOSH, T. C. Insights into the miRNA regulations in human disease genes. *BMC Genomics* 15 (2014), 1010.
- [29] DAVIS, S., AND MELTZER, P. S. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics* 23, 14 (2007), 1846–1847.
- [30] DE KEERSMAECKER, S. C. J., THIJS, I. M. V., VANDERLEYDEN, J., AND MARCHAL, K. Integration of omics data: how well does it work for bacteria? *Molecular Microbiology* 62, 5 (Dec. 2006), 1239–1250.

- [31] DEMPSTER, A. P., LAIRD, N. M., AND RUBIN, D. B. Maximum-likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B.* 39 (1977), 1–39.
- [32] DIAZ, D., DONATO, M., NGUYEN, T., AND DRAGHICI, S. MicroRNA-augmented pathways (mirAP) and their applications to pathway analysis and disease subtyping. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* (2016), vol. 22, p. 390.
- [33] DIAZ, D., AND DRAGHICI, S. *mirIntegrator: Integrating miRNAs into signaling pathways*, 2015.
- [34] DIAZ, D., NGUYEN, T., AND DRAGHICI, S. A systems biology approach for unsupervised clustering of high-dimensional data. In *The Second International Workshop on Machine Learning, Optimization and Big Data* (2016).
- [35] DINU, I., POTTER, J. D., MUELLER, T., LIU, Q., ADEWALE, A. J., JHANGRI, G. S., EINECKE, G., FAMULSKI, K. S., HALLORAN, P., AND YASUI, Y. Improving gene set analysis of microarray data by SAM-GS. *BMC Bioinformatics* 8, 1 (2007), 242.
- [36] DONAHUE, T. R., TRAN, L. M., HILL, R., LI, Y., KOVOCHICH, A., CALVOPINA, J. H., PATEL, S. G., WU, N., HINDOYAN, A., FARRELL, J. J., ET AL. Integrative survival-based molecular profiling of human pancreatic cancer. *Clinical Cancer Research* 18, 5 (2012), 1352–1363.
- [37] DONIGER, S. W., SALOMONIS, N., DAHLQUIST, K. D., VRANIZAN, K., LAWLOR, S. C., AND CONKLIN, B. R. MAPPFinder: using Gene Ontology and GenMAPP to create a global gene expression profile from microarray data. *Genome biology* 4, 1 (2003), R7.

- [38] DRĂGHICI, S., CHEN, D., AND REIFMAN, J. Applications and challenges of DNA microarray technology in military medical research. *Military Medicine* 169, 8 (2004), 654–659.
- [39] DRĂGHICI, S. *Statistics and Data Analysis for Microarrays using R and Bioconductor*. Chapman and Hall/CRC Press, 2011.
- [40] DRĂGHICI, S., KHATRI, P., MARTINS, R. P., OSTERMEIER, G. C., AND KRAWETZ, S. A. Global functional profiling of gene expression. *Genomics* 81, 2 (2003), 98–104.
- [41] DRĂGHICI, S., KHATRI, P., TARCA, A. L., AMIN, K., DONE, A., VOICHIȚA, C., GEORGESCU, C., AND ROMERO, R. A systems biology approach for pathway level analysis. *Genome Research* 17, 10 (2007), 1537–1545.
- [42] DWEEP, H., AND GRETZ, N. miRWalk2. 0: a comprehensive atlas of microRNA-target interactions. *Nature Methods* 12, 8 (2015), 697–697.
- [43] EDGAR, R., DOMRACHEV, M., AND LASH, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research* 30, 1 (2002), 207–210.
- [44] EFRON, B., AND TIBSHIRANI, R. On testing the significance of sets of genes. *The Annals of Applied Statistics* 1, 1 (2007), 107–129.
- [45] EIN-DOR, L., KELA, I., GETZ, G., GIVOL, D., AND DOMANY, E. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics* 21, 2 (2005), 171–178.
- [46] EIN-DOR, L., ZUK, O., AND DOMANY, E. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *In Proceedings of the National Academy of Sciences* 103, 15 (2006), 5923–5928.

- [47] EMMERT-STREIB, F., AND V. GLAZKO, G. Pathway Analysis of Expression Data: Deciphering Functional Building Blocks of Complex Diseases. *PLoS Computational Biology* 7, 5 (2011), e1002053.
- [48] FISHER, R. A. *Statistical methods for research workers*. Oliver & Boyd, Edinburgh, 1925.
- [49] FREY, B. J., AND MACKAY, D. J. A revolution: Belief propagation in graphs with cycles. *Advances in Neural Information Processing Systems* (1998), 479–485.
- [50] FRIEDMAN, N. Inferring Cellular Networks Using Probabilistic Graphical Models. *Science* 303, 5659 (Feb. 2004), 799–805.
- [51] FU, J., TANG, W., DU, P., WANG, G., CHEN, W., LI, J., ZHU, Y., GAO, J., AND CUI, L. Identifying microRNA-mRNA regulatory network in colorectal cancer by a combination of expression profile and bioinformatics analysis. *BMC Systems Biology* 6 (2012), 68.
- [52] G. S. FIRESTEIN, D. S. P. DNA microarrays: boundless technology or bound by technology? Guidelines for studies using microarray technology. *Arthritis and Rheumatism* 46, 4 (2002), 859–861.
- [53] GE, H., WALHOUT, A. J. M., AND VIDAL, M. Integrating ‘omic’ information: a bridge between genomics and systems biology. *Trends in Genetics* 19, 10 (Oct. 2003), 551–560.
- [54] GENTLEMAN, R. C., CAREY, V. J., BATES, D. M., BOLSTAD, B., DETTLING, M., DUDOIT, S., ELLIS, B., GAUTIER, L., GE, Y., GENTRY, J., HORNIK, K., HOTHORN, T., HUBER, W., IACUS, S., IRIZARRY, R., LEISCH, F., LI, C., MAECHLER, M., ROSSINI, A. J., SAWITZKI, G., SMITH, C.,

- SMYTH, G., TIERNEY, L., YANG, J. Y., AND ZHANG, J. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* 5, 10 (2004), R80.
- [55] GLAAB, E., BAUDOT, A., KRASNOGOR, N., SCHNEIDER, R., AND VALENCIA, A. EnrichNet: network-based gene set enrichment analysis. *Bioinformatics* 28, 18 (2012), i451–i457.
- [56] GLAAB, E., BAUDOT, A., KRASNOGOR, N., AND VALENCIA, A. TopoGSA: network topological gene set analysis. *Bioinformatics* 26, 9 (2010), 1271–1272.
- [57] GOEMAN, J. J., VAN DE GEER, S. A., DE KORT, F., AND VAN HOUWELINGEN, H. C. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics* 20, 1 (2004), 93–99.
- [58] GOSSET, W. S. The Probable Error of a Mean. *Biometrika* 6 (1908), 1–25.
- [59] GUTIERREZ-ARCELUS, M., LAPPALAINEN, T., MONTGOMERY, S. B., BUIL, A., ONGEN, H., YUROVSKY, A., BRYOIS, J., GIGER, T., ROMANO, L., PLANCHON, A., ET AL. Passive and active dna methylation and the interplay with genetic variation in gene regulation. *Elife* 2 (2013), e00523.
- [60] HARRELL, F. E., CALIFF, R. M., PRYOR, D. B., LEE, K. L., AND ROSATI, R. A. Evaluating the yield of medical tests. *Jama* 247, 18 (1982), 2543–2546.
- [61] HARTLEY, H. O. Maximum Likelihood Estimation from Incomplete Data. *Biometrics* 14, 2 (June 1958), 174–194.
- [62] HATFIELD, G., HUNG, S.-P., AND BALDI, P. Differential analysis of dna microarray gene expression data. *Molecular microbiology* 47, 4 (2003), 871–877.
- [63] HENEGAR, C., CANCELLO, R., ROME, S., VIDAL, H., CLÉMENT, K., AND ZUCKER, J.-D. Clustering biological annotations and gene expression data to

- identify putatively co-regulated biological processes. *Journal of bioinformatics and computational biology* 4, 04 (2006), 833–852.
- [64] HSU, S.-D., LIN, F.-M., WU, W.-Y., LIANG, C., HUANG, W.-C., CHAN, W.-L., TSAI, W.-T., CHEN, G.-Z., LEE, C.-J., AND CHIU, C.-M. miRTarBase: a database curates experimentally validated microRNA–target interactions. *Nucleic Acids Research* (2010), D163–D169.
- [65] HSU, S.-D., TSENG, Y.-T., SHRESTHA, S., LIN, Y.-L., KHALEEL, A., CHOU, C.-H., CHU, C.-F., HUANG, H.-Y., LIN, C.-M., HO, S.-Y., JIAN, T.-Y., LIN, F.-M., CHANG, T.-H., WENG, S.-L., LIAO, K.-W., LIAO, I.-E., LIU, C.-C., AND HUANG, H.-D. miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions. *Nucleic Acids Research* 42, D1 (Jan. 2014), D78–D85.
- [66] HUANG, D. W., SHERMAN, B. T., AND LEMPICKI, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research* 37, 1 (2009), 1–13.
- [67] ISCI, S., OZTURK, C., JONES, J., AND OTU, H. H. Pathway analysis of high-throughput biological data within a Bayesian network framework. *Bioinformatics* 27, 12 (2011), 1667–1674.
- [68] JACOB, L., NEUVIAL, P., AND DUDOIT, S. Gains in power from structured two-sample tests of means on graphs. *Arxiv preprint arXiv:1009.5173* (2010).
- [69] JERNÅS, M., MALMESTRÖM, C., AXELSSON, M., NOOKAEW, I., WADENVIK, H., LYCKE, J., AND OLSSON, B. MicroRNA regulate immune pathways in T-cells in multiple sclerosis (MS). *BMC Immunology* 14, 1 (2013), 32.
- [70] JIANG, Q., WANG, Y., HAO, Y., JUAN, L., TENG, M., ZHANG, X., LI, M., WANG, G., AND LIU, Y. miR2Disease: a manually curated database for

- microRNA deregulation in human disease. *Nucleic Acids Research* 37, suppl 1 (2009), D98–D104.
- [71] JIANG, Z., AND GENTLEMAN, R. Extensions to gene set enrichment. *Bioinformatics* 23, 3 (2007), 306–313.
- [72] JOHN, B., ENRIGHT, A. J., ARAVIN, A., TUSCHL, T., SANDER, C., AND MARKS, D. S. Human MicroRNA Targets. *PLOS Biology* 2, 11 (Oct. 2004), e363.
- [73] JOSHI-TOPE, G., GILLESPIE, M., VASTRIK, I., D’EUSTACHIO, P., SCHMIDT, E., DE BONO, B., JASSAL, B., GOPINATH, G., WU, G., MATTHEWS, L., LEWIS, S., BIRNEY, E., AND STEIN, L. REACTOME: a knowledgebase of biological pathways. *Nucleic Acids Research* 33, Database issue (2005), D428–432.
- [74] KANEHISA, M., AND GOTO, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* 28, 1 (2000), 27–30.
- [75] KANEHISA, M., GOTO, S., HATTORI, M., AOKI-KINOSHITA, K. F., ITOH, M., KAWASHIMA, S., KATAYAMA, T., ARAKI, M., AND HIRAKAWA, M. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Research* 34, suppl 1 (2006), D354–D357.
- [76] KAROLINA, D. S., ARMUGAM, A., TAVINTHARAN, S., WONG, M. T. K., LIM, S. C., SUM, C. F., AND JEYASEELAN, K. MicroRNA 144 impairs insulin signaling by inhibiting the expression of insulin receptor substrate 1 in type 2 diabetes mellitus. *PloS One* 6, 8 (2011), e22839.
- [77] KHATRI, P., AND DRĂGHICI, S. A comparison of existing tools for ontological analysis of gene expression data. In *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics*. Wiley Online Library, 2005, ch. 4. 4.5:54.

- [78] KHATRI, P., DRĂGHICI, S., OSTERMEIER, G. C., AND KRAWETZ, S. A. Profiling gene expression using Onto-Express. *Genomics* 79, 2 (2002), 266–270.
- [79] KHATRI, P., SIROTA, M., AND BUTTE, A. J. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Computational Biology* 8, 2 (2012), e1002375.
- [80] KIM, S.-Y., AND VOLSKY, D. J. Page: parametric analysis of gene set enrichment. *BMC bioinformatics* 6, 1 (2005), 144.
- [81] KIM, T. Y., KIM, H. U., AND LEE, S. Y. Data integration and analysis of biological networks. *Current Opinion in Biotechnology* 21, 1 (Feb. 2010), 78–84.
- [82] KONG, S. W., PU, W. T., AND PARK, P. J. A multivariate approach for integrating genome-wide expression data and biological knowledge. *Bioinformatics* 22, 19 (2006), 2373–2380.
- [83] KOSMADAKI, M., NAIF, A., AND HEE-YOUNG, P. Recent progresses in understanding pigmentation. *Giornale italiano di dermatologia e venereologia: organo ufficiale, Società italiana di dermatologia e sifilografia* 145, 1 (2010), 47–55.
- [84] KOTELNIKOVA, E., SHKROB, M. A., PYATNITSKIY, M. A., FERLINI, A., AND DARASELIA, N. Novel approach to meta-analysis of microarray datasets reveals muscle remodeling-related drug targets and biomarkers in Duchenne muscular dystrophy. *PLoS Computational Biology* 8, 2 (2012), e1002365.
- [85] KREK, A., GRÜN, D., POY, M. N., WOLF, R., ROSENBERG, L., EPSTEIN, E. J., MACMENAMIN, P., DA PIEDADE, I., GUNSALUS, K. C., STOFFEL, M., ET AL. Combinatorial microRNA target predictions. *Nature genetics* 37, 5 (2005), 495–500.

- [86] KRISTENSEN, V. N., LINGJÆRDE, O. C., RUSSNES, H. G., VOLLAN, H. K. M., FRIGESSI, A., AND BØRRESEN-DALE, A.-L. Principles and methods of integrative genomic analyses in cancer. *Nature Reviews Cancer* 14, 5 (May 2014), 299–313.
- [87] LAY JR., J. O., LIYANAGE, R., BORGMANN, S., AND WILKINS, C. L. Problems with the “omics”. *TrAC Trends in Analytical Chemistry* 25, 11 (Dec. 2006), 1046–1056.
- [88] LEE, Y. S., AND DUTTA, A. MicroRNAs in cancer. *Annual Review of Pathology* 4 (2009).
- [89] LEWIS, B. P., BURGE, C. B., AND BARTEL, D. P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* 120, 1 (2005), 15–20.
- [90] LEWIS, B. P., BURGE, C. B., AND BARTEL, D. P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* 120, 1 (2005), 15–20.
- [91] LI, Y., QIU, C., TU, J., GENG, B., YANG, J., JIANG, T., AND CUI, Q. HMDD v2.0: a database for experimentally supported human microRNA and disease associations. *Nucleic Acids Research* 42, Database issue (Jan. 2014), D1070–D1074.
- [92] LINN, R. J., HALL, D. L., AND LLINAS, J. Survey of multisensor data fusion systems. In *Data Structures and Target Classification* (1991), vol. 1470, pp. 13–29.
- [93] LU, M., ZHANG, Q., DENG, M., MIAO, J., GUO, Y., GAO, W., AND CUI, Q. An analysis of human microRNA and disease associations. *PloS One* 3, 10 (2008), e3420.

- [94] MA, L., HUANG, Y., ZHU, W., ZHOU, S., ZHOU, J., ZENG, F., LIU, X., ZHANG, Y., AND YU, J. An integrated analysis of miRNA and mRNA expressions in non-small cell lung cancers. *PloS One* 6, 10 (2011), e26502.
- [95] MAMDANI, M., WILLIAMSON, V., MCMICHAEL, G. O., BLEVINS, T., ALIEV, F., ADKINS, A., HACK, L., BIGDELI, T., VAN DER VAART, A. D., WEB, B. T., BACANU, S.-A., KALSI, G., COGA CONSORTIUM, KENDLER, K. S., MILES, M. F., DICK, D., RILEY, B. P., DUMUR, C., AND VLADIMIROV, V. I. Integrating mRNA and miRNA Weighted Gene Co-Expression Networks with eQTLs in the Nucleus Accumbens of Subjects with Alcohol Dependence. *PloS One* 10, 9 (2015), e0137671.
- [96] MARTIN, D., BRUN, C., REMY, E., MOUREN, P., THIEFFRY, D., AND JACQ, B. GOToolBox: functional analysis of gene datasets based on Gene Ontology. *Genome Biology* 5 (2004), R101.
- [97] MARTINI, P., SALES, G., MASSA, M. S., CHIOGNA, M., AND ROMUALDI, C. Along signal paths: an empirical gene set approach exploiting pathway topology. *Nucleic Acids Research* 41, 1 (2013), e19–e19.
- [98] MATTHEWS, L., GOPINATH, G., GILLESPIE, M., CAUDY, M., CROFT, D., DE BONO, B., GARAPATI, P., HEMISH, J., HERMJAKOB, H., JASSAL, B., ET AL. Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Research* 37, suppl 1 (2009), D619–D622.
- [99] MCLACHLAN, G., AND KRISHNAN, T. *The EM Algorithm and Extensions*. John Wiley & Sons, Nov. 2007.
- [100] MI, H., LAZAREVA-ULITSKY, B., LOO, R., KEJARIWAL, A., VANDERGRUFF, J., RABKIN, S., GUO, N., MURUGANUJAN, A., DOREMIEUX, O., CAMP-

- BELL, M. J., ET AL. The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Research* 33, suppl 1 (2005), D284–D288.
- [101] MI, H., MURUGANUJAN, A., AND THOMAS, P. D. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Research* 41, D1 (2013), D377–D386.
- [102] MITREA, C., TAGHAVI, Z., BOKANIZAD, B., HANOUDI, S., TAGETT, R., DONATO, M., VOICHIȚA, C., AND DRĂGHICI, S. Methods and approaches in the topology-based analysis of biological pathways. *Frontiers in Physiology* 4 (2013), 278.
- [103] MONTI, S., TAMAYO, P., MESIROV, J., AND GOLUB, T. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning* 52, 1-2 (2003), 91–118.
- [104] MOOTHA, V. K., LINDGREN, C. M., ERIKSSON, K.-F., SUBRAMANIAN, A., SIHAG, S., LEHAR, J., PUIGSERVER, P., CARLSSON, E., RIDDERSTRÅLE, M., LAURILA, E., HOUSTIS, N., DALY, M. J., PATTERSON, N., MESIROV, J. P., GOLUB, T. R., TAMAYO, P., SPIEGELMAN, B., LANDER, E. S., HIRSCHHORN, J. N., ALTSHULER, D., AND GROOP, L. C. PGC-11 α -responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics* 34, 3 (2003), 267–273.
- [105] MORENO-RISUENO, M. A., BUSCH, W., AND BENFEY, P. N. Omics meet networks — using systems approaches to infer regulatory networks in plants. *Current Opinion in Plant Biology* 13, 2 (Apr. 2010), 126–131.
- [106] NAGARAJ, N., WISNIEWSKI, J. R., GEIGER, T., COX, J., KIRCHER, M., KELSO, J., PÄÄBO, S., AND MANN, M. Deep proteome and transcriptome

- mapping of a human cancer cell line. *Molecular Systems Biology* 7, 1 (2011), 548.
- [107] NAM, S., LI, M., CHOI, K., BALCH, C., KIM, S., AND NEPHEW, K. P. MicroRNA and mRNA integrated analysis (MMIA): a web tool for examining biological functions of microRNA expression. *Nucleic Acids Research* 37, suppl 2 (May 2009), W356–W362.
- [108] NG, A. Y., JORDAN, M. I., WEISS, Y., ET AL. On spectral clustering: Analysis and an algorithm. In *NIPS* (2001), vol. 14, pp. 849–856.
- [109] NGUYEN, T., DIAZ, D., TAGETT, R., AND DRAGHICI, S. Overcoming the matched-sample bottleneck: an orthogonal approach to integrate omic data. *Nature Scientific Reports* 6 (2016), 29251.
- [110] NIELSEN, C. B., SHOMRON, N., SANDBERG, R., HORNSTEIN, E., KITZMAN, J., AND BURGE, C. B. Determinants of targeting by endogenous and exogenous microRNAs and siRNAs. *RNA* 13, 11 (Nov. 2007), 1894–1910.
- [111] OGATA, H., GOTO, S., SATO, K., FUJIBUCHI, W., BONO, H., AND KANEHISA, M. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* 27, 1 (1999), 29–34.
- [112] PALSSON, B., AND ZENGLER, K. The challenges of integrating multi-omic data sets. *Nature Chemical Biology* 6, 11 (Nov. 2010), 787–789.
- [113] PARASKEVOPOULOU, M. D., GEORGAKILAS, G., KOSTOULAS, N., VLACHOS, I. S., VERGOULIS, T., RECZKO, M., FILIPPIDIS, C., DALAMAGAS, T., AND HATZIGEORGIU, A. DIANA-microT web server v5.0: service integration into miRNA functional analysis workflows. *Nucleic Acids Research* 41, Web Server issue (July 2013), W169–W173.

- [114] PARK, M., KIM, M., HWANG, D., PARK, M., KIM, W. K., KIM, S. K., SHIN, J., PARK, E. S., KANG, C. M., PAIK, Y.-K., AND KIM, H. Characterization of gene expression and activated signaling pathways in solid-pseudopapillary neoplasm of pancreas. *Modern Pathology* 27, 4 (2014), 580–593.
- [115] PAVLIDIS, P., QIN, J., ARANGO, V., MANN, J. J., AND SIBILLE, E. Using the Gene Ontology for Microarray Data Mining: A Comparison of Methods and Application to Age Effects in Human Prefrontal Cortex. *Neurochemical Research* 29, 6 (June 2004), 1213–1222.
- [116] PEARSON, E. S., AND HARTLEY, H. O. Biometrika tables for statisticians. *Biometrika Trust* (1976).
- [117] PEĆINA-ŠLAUS, N., AND PEĆINA, M. Only one health, and so many omics. *Cancer Cell International* 15, 1 (June 2015), 64.
- [118] PIZZINI, S., BISOGNIN, A., MANDRUZZATO, S., BIASIOLO, M., FACCIOLLI, A., PERILLI, L., ROSSI, E., ESPOSITO, G., RUGGE, M., PILATI, P., ET AL. Impact of microRNAs on regulatory networks and pathways in human colorectal carcinogenesis and development of metastasis. *BMC Genomics* 14, 1 (2013), 589.
- [119] QUITADAMO, A., TIAN, L., HALL, B., AND SHI, X. An integrated network of microRNA and gene expression in ovarian cancer. *BMC Bioinformatics* 16, Suppl 5 (Mar. 2015), S5.
- [120] RAHNENFÜHRER, J., DOMINGUES, F. S., MAYDT, J., AND LENGAUER, T. Calculating the Statistical Significance of Changes in Pathway Activity From Gene Expression Data. *Statistical Applications in Genetics and Molecular Biology* 3, 1 (2004).

- [121] REINER, A., YEKUTIELI, D., AND BENJAMINI, Y. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics* 19, 3 (2003), 368–375.
- [122] RITCHIE, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., AND SMYTH, G. K. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* 43, 7 (2015), e47.
- [123] ROBINSON, S. W., FERNANDES, M., AND HUSI, H. Current advances in systems and integrative biology. *Computational and Structural Biotechnology Journal* 11, 18 (Aug. 2014), 35–46.
- [124] ROUSSEEUW, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics* 20 (1987), 53–65.
- [125] RUSTICI, G., KOLESNIKOV, N., BRANDIZI, M., BURDETT, T., DYLAG, M., EMAM, I., FARNE, A., HASTINGS, E., ISON, J., KEAYS, M., KURBATOVA, N., MALONE, J., MANI, R., MUPO, A., PEREIRA, R. P., PILICHEVA, E., RUNG, J., SHARMA, A., TANG, Y. A., TERNENT, T., TIKHONOV, A., WELTER, D., WILLIAMS, E., BRAZMA, A., PARKINSON, H., AND SARKANS, U. ArrayExpress update—trends in database growth and links to data analysis tools. *Nucleic Acids Research* 41, D1 (2013), D987–D990.
- [126] SARANGARAJAN, R., AND BOISSY, R. E. TYRP1 and oculocutaneous albinism type 3. *Pigment cell research* 14, 6 (2001), 437–444.
- [127] SCHAEFER, C., ANTHONY, K., KRUPA, S., BUCHOFF, J., DAY, M., HANNAY, T., AND BUETOW, K. PID: the Pathway Interaction Database. *Nucleic Acids Research* 37, Database issue (2009), D674–D679.

- [128] SERPEDIN, E., CHEN, T., AND RAJAN, D. *Mathematical Foundations for Signal Processing, Communications, and Networking*. CRC Press, Dec. 2011.
- [129] SETHUPATHY, P., CORDA, B., AND HATZIGEORGIOU, A. G. TarBase: A comprehensive database of experimentally supported animal microRNA targets. *RNA* 12, 2 (Feb. 2006), 192–197.
- [130] SHAI, R., SHI, T., KREMEN, T. J., HORVATH, S., LIAU, L. M., CLOUGHESY, T. F., MISCHEL, P. S., AND NELSON, S. F. Gene expression profiling identifies molecular subtypes of gliomas. *Oncogene* 22, 31 (2003), 4918–4923.
- [131] SHARMA, S., WAGH, S., AND GOVINDARAJAN, R. Melanosomal proteins—role in melanin polymerization. *Pigment cell research* 15, 2 (2002), 127–133.
- [132] SHEN, R., OLSHEN, A. B., AND LADANYI, M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* 25, 22 (2009), 2906–2912.
- [133] SHI, Z., WANG, J., AND ZHANG, B. NetGestalt: integrating multidimensional omics data over biological networks. *Nature Methods* 10, 7 (2013), 597–598.
- [134] SHOJAIE, A., AND MICHAILIDIS, G. Analysis of Gene Sets Based on the Underlying Regulatory Network. *Journal of Computational Biology* 16, 3 (2009), 407–426.
- [135] SMYTH, G. K. Limma: linear models for microarray data. In *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, and W. Huber, Eds. Springer, New York, 2005, pp. 397–420.

- [136] STESSMAN, H. A., BERNIER, R., AND EICHLER, E. E. A Genotype-First Approach to Defining the Subtypes of a Complex Disease. *Cell* 156, 5 (Feb. 2014), 872–877.
- [137] SUBRAMANIAN, A., TAMAYO, P., MOOTHA, V. K., MUKHERJEE, S., EBERT, B. L., GILLETTE, M. A., PAULOVICH, A., POMEROY, S. L., GOLUB, T. R., LANDER, E. S., AND MESIROV, J. P. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceeding of The National Academy of Sciences of the United States of America* 102, 43 (2005), 15545–15550.
- [138] TAN, P. K., DOWNEY, T. J., SPITZNAGEL JR, E. L., XU, P., FU, D., DIMITROV, D. S., LEMPICKI, R. A., RAAKA, B. M., AND CAM, M. C. Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Research* 31, 19 (2003), 5676–5684.
- [139] TARCA, A. L., DRĂGHICI, S., BHATTI, G., AND ROMERO, R. Down-weighting overlapping genes improves gene set analysis. *BMC Bioinformatics* 13, 1 (2012), 136.
- [140] TARCA, A. L., DRĂGHICI, S., KHATRI, P., HASSAN, S. S., MITTAL, P., KIM, J.-S., KIM, C. J., KUSANOVIC, J. P., AND ROMERO, R. A novel signaling pathway impact analysis. *Bioinformatics* 25, 1 (2009), 75–82.
- [141] TAVAZOIE, S., HUGHES, J. D., CAMPBELL, M. J., CHO, R. J., AND CHURCH, G. M. Systematic determination of genetic network architecture. *Nature Genetics* 22 (1999), 281–285.
- [142] TCGA RESEARCH NETWORK. The Cancer Genome Atlas. <http://cancergenome.nih.gov/>.

- [143] THERNEAU, T. M., AND GRAMBSCH, P. M. *Modeling Survival Data: Extending the Cox Model*. Springer, 2000.
- [144] TIAN, L., GREENBERG, S. A., KONG, S. W., ALTSCHULER, J., KOHANE, I. S., AND PARK, P. J. Discovering statistically significant pathways in expression profiling studies. *Proceeding of The National Academy of Sciences of the USA* 102, 38 (2005), 13544–13549.
- [145] TSENG, G. C., GHOSH, D., AND FEINGOLD, E. Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic Acids Research* 40, 9 (2012), 3785–3799.
- [146] TSENG, G. C., GHOSH, D., AND ZHOU, X. J. *Integrating Omics Data*. Cambridge University Press, Aug. 2015.
- [147] VASKE, C. J., BENZ, S. C., SANBORN, J. Z., EARL, D., SZETO, C., ZHU, J., HAUSSLER, D., AND STUART, J. M. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics* 26, 12 (2010), i237–i245.
- [148] VLACHOS, I. S., ZAGGANAS, K., PARASKEVOPOULOU, M. D., GEORGAKILAS, G., KARAGKOUNI, D., VERGOULIS, T., DALAMAGAS, T., AND HATZIGEORGIOU, A. G. DIANA-miRPath v3. 0: deciphering microRNA function with experimental support. *Nucleic Acids Research* 43, W1 (2015), W460–W466.
- [149] VOICHIȚA, C., DONATO, M., AND DRĂGHICI, S. Incorporating gene significance in the impact analysis of signaling pathways. *Proceedings of the International Conference on Machine Learning Applications (ICMLA)* (Dec. 2012).

- [150] VOICHIȚA, C., AND DRĂGHICI, S. *ROntoTools: R Onto-Tools suite*. R package, <http://bioconductor.org/packages/release/bioc/html/ROntoTools.html>.
- [151] VON MERING, C., HUYNEN, M., JAEGGI, D., SCHMIDT, S., BORK, P., AND SNEL, B. STRING: a database of predicted functional associations between proteins. *Nucleic Acids Research* 31, 1 (2003), 258–261.
- [152] WANG, B., MEZLINI, A. M., DEMIR, F., FIUME, M., TU, Z., BRUDNO, M., HAIBE-KAINS, B., AND GOLDENBERG, A. Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods* 11, 3 (2014), 333–337.
- [153] WATSON, J. D., AND CRICK, F. H. The structure of dna. In *Cold Spring Harbor symposia on quantitative biology* (1953), vol. 18, Cold Spring Harbor Laboratory Press, pp. 123–131.
- [154] XIA, J., AND WISHART, D. S. MetPA: a web-based metabolomics tool for pathway analysis and visualization. *Bioinformatics* 26, 18 (2010), 2342–2344.
- [155] XIAO, D., OHLENDORF, J., CHEN, Y., TAYLOR, D. D., RAI, S. N., WAIGEL, S., ZACHARIAS, W., HAO, H., AND MCMASTERS, K. M. Identifying mRNA, microRNA and protein profiles of melanoma exosomes. *PloS One* 7, 10 (2012), e46874.
- [156] XIAO, F., ZUO, Z., CAI, G., KANG, S., GAO, X., AND LI, T. miRecords: an integrated resource for microRNA-target interactions. *Nucleic Acids Research* 37, Database issue (Jan. 2009), D105–110.
- [157] XIE, Y., AND AHN, C. Statistical Methods for Integrating Multiple Types of High-Throughput Data. *Methods in molecular biology (Clifton, N.J.)* 620 (2010), 511–529.

ABSTRACT

INTEGRATIVE PATHWAY ANALYSIS PIPELINE FOR miRNA AND mRNA DATA

by

DIANA DIAZ

Advisor: Dr. Sorin Draghici

Major: Computer Science

Degree: Master of Science

The identification of pathways that are involved in a particular phenotype helps us understand the underlying biological processes. Traditional pathway analysis techniques aim to infer the impact on individual pathways using only mRNA levels. However, recent studies showed that gene expression alone is unable to capture the whole picture of biological phenomena. At the same time, MicroRNAs (miRNAs) are newly discovered gene regulators that have shown to play an important role in diagnosis, and prognosis for different types of diseases. Current pathway analysis techniques do not take miRNAs into consideration. In this project, we investigate the effect of integrating miRNA and mRNA expression in pathway analysis. In order to analyze biological pathways using miRNA expression data, we developed a novel method that augments KEGG pathways with microRNAs targeting genes. To validate our method, we analyzed nine GEO datasets. We also performed the analyses using just mRNA as well as using the integrative state-of-the-art method (microGraphite) to compare the results. In each case, we monitored the position of the pathway describing the given condition. We observed that our method outperforms the state-of-the-art approach.

AUTOBIOGRAPHICAL STATEMENT

DIANA DIAZ

EDUCATION

- Master of Science (Computer Science), 2017
Wayne State University, Detroit, MI, USA
- Master of Science (Computer Systems Engineering), 2009
The Andes University, Colombia
- Bachelor of Engineering (Computer Systems Engineering), 2006
The Piloto University of Colombia, Colombia

PUBLICATIONS

1. **Diaz D**, DONATO M, NGUYEN T, DRAGHICI S. MicroRNA-Augmented Pathways (mirAP) and their applications to pathway analysis and disease subtyping. In proceedings of *Pacific Symposium on Biocomputing (PSB) 2017*. (2017).
2. **Diaz D**, NGUYEN T, DRAGHICI S. A Systems Biology Approach for Unsupervised Clustering of High-Dimensional Data. In proceedings of *International Workshop on Machine Learning, Optimization and Big Data*. (2016).
3. NGUYEN T, **Diaz D**, TAGETT R, DRAGHICI S. Overcoming the matched-sample bottleneck: an orthogonal approach to integrate omic data. *Scientific Reports*, 6, 29251 (2016).
4. **Diaz D**, DRAGHICI S. mirIntegrator: Integrating miRNAs into signaling pathways. *Bioconductor* (2015).