Sudan University of Science and Technology
College of Graduate Studies

A Machine Learning Holistic Strategy for miRNA-mRNA Module Discovery

إستراتيجية شاملة لإكتشاف وحدة الروبيزوم النووي الدقيق وحمض الروبيزوم النووي الناقل بإستخدام تعلم الآلة

Thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy in Computer Science

By
Ghada Ali Mohamed Shommo

Supervised by
Prof. Bruno Apolloni

December 2022

﴿فَسَتَذْكُرُونَ مَآ أَقُولُ لَكُمْۚ وَأُفَوِّضُ أَمْرِيٓ إِلَى ٱللَّهِۚ إِنَّ ٱللَّهَ بَصِيرُۢ بِٱلْعِبَادِ﴾

[غافر :44]

Abstract

A microRNA (abbreviated miRNA) is a small non-coding molecule, that is made up of approximately 22 nucleotides, found in plants, animals, and some viruses. They act as regulators by binding their targets to degrade or suppress the translation of their transcripts. Therefore miRNAs plays an important role in gene regulatory networks, and an improved understanding of miRNAs will widen our knowledge of these networks and their relation with diseases. A single miRNA targets multiple mRNAs, and a single mRNA is targeted by multiple miRNAs to develop a many-to-many relations, known as miRNAmRNA module. Some methods have ignored this relation, by just focusing on miRNA-mRNA pairs. However current methods evolved to consider this relation but unfortunately they focused on some part of the data analyzed, and ignored the other. Unfortunately, they still leave open issues, but the higher benefit is that they provided results, that opened issue of the possibility of widening the scope of module discovery, so as to be extended to a wider disease spectrum, where miRNA-mRNA interactions play a relevant role. The research proposes a holistic procedure for miRNA-mRNA module identification that exploits as much data as possible. It uses machine learning and mathematical approaches to aid in the analysis and implementation. We adopt the strategy of postponing any decision until biological results are exploited. Many statistical tests have been diverted into specially-devised evolving metrics, for sake of possible solutions. Consequently, the Implementation on High Performance Computing (HPC) is crucial, since this strategy is rather expensive in terms of computation. Fortunately, it allows the discovery of modules whose miRNAs and mRNAs are not differentially expressed and the discovery miRNA targets not yet considered, as well. In this research, the procedure, is implemented on a Multiple Myeloma dataset publicly available on Gene Expression Omnibus (GEO) platform, as a case study of diseases, specifically as a cancer instance analysis, and scout some biological issues. The procedure has introduced novel strategies for miRNA-mRNA module discovery.

Main achievements of this thesis work are: 1)we introduce a novel strategy for miRNA-mRNA module discovery; 2) we establish an unprecedented way of jointly using using many metrics to find new links between miRNA and mRNA clusters involving non differentially expressed RNA pairs as well; 3) and finally, we highlight new miRNA-mRNA interactions with a methodology that can be extended to a wide spectrum of diseases.

المستخلص

إن **الروبيزوم النووي الدقيق** microRNA (إختصاره miRNA) هو عبارة عن جزىء صغير غير مرمز مكون من 22 نوكليوتيدات , يوجد في النباتات و الحيوانات و بعض الفيروساتزتعمل هذه الجزيئات كمنظمات لأحماض **الروبيزوم النووي الناقل**) ((**messenger RNA** (**mRNA**)) التي تستهدفها لإفساد أ قمع عملية التحول اي البروتين المعني .لذا فإن **ال** miRNA يلعب دوراكبيرا و هاما جدا في الشبكات التنظيمية للجينات، و لهذا السبب فإن الفهم الجيد و المتطور لعمل ال microRNA سيوسع معرفتنا و إلمامنا بكيفية عمل هذه الشبكات و ماهية علاقتها بالأمراض ومسبباتها.

يستهدف ال (microRNA الواحد عدة (mRNA و في نفس الوقت فإن ال (mRNA يتم استهدافه بواسطة عدة microRNA ليتم خلق علاقة متعددة (many-to-many) تسمى (miRNA -mRNA.

لقد تم تطوير الكثير من الطرق لدراسة هذه العلاقة , و لكنها للأسف تجاهلت التعددية في العلاقة و ركزت على العلاقة الفردية فقط بينهما .بالرغم من ذلك فهنالك طرق حديثة إهتمت بهذه التعددية و لكنها لم تستغل البيانات المتوفرة بشكل كافي,مما أدى الى ترك العديد من الأسئلة المفتوحة في هذا المجال , الا أنها تركت العديد من النتائج التي يمكن استغلالها في توسيع نطاق البحث, و بالتالي تتيح إكتشاف الكثير من الخبايا و الإكتشافات التي توضح أهمية الدور الذي تلعبه تلك العلاقة .

يقترح هذا البحث إستراتيجية شاملة لإكتشاف علاقات جديدة " و ذلك عن طريق استغلال بيانات بقدر الإمكان .هذه الإستراتيجية تمكننا من تأجيل إتخاذ القرار"حتى يتم إستغلال أكبر قدر من البيانات المتاحة .

لقد تم تحويل العديد من الإختبارات الإحصائية الى مقاييس متطورة تساعدنا في البحث عن الحلول, و لذلك , فإن تطبيق مثل هذه الإستراتيجية يحتاج حوسبة ذات أداء عالي .لحسن الحظ أصبح من الممكن إكتشاف علاقات جديدة لم يتم إكتشافها من قبل .

في هذا البحث قمنا بتطبيق هه الإستراتيجية على بيانات تم الحصول عليها من الإنترنت, و هي خاصة ب مرضى المايلوما المتعددة كدراسة حالة لنوع من أمراض السرطان ، واستكشاف بعض المشكلات البيولوجية. قدم الإجراء استراتيجيات جديدة لاكتشاف وحدة miRNA-mRNA. النتائج الرئيسية التي تم الحصول عليها هي: 1) نقدم استراتيجية جديدة لاكتشاف وحدة miRNA-mRNA ؛ 2) أنشأنا طريقة غير مسبوقة للاستخدام المشترك للعديد من المقاييس لإيجاد روابط جديدة بين مجموعات mRNA و mRNA التي تتضمن أزواج miRNA-mRNA غير معبر عنها تفاضليًا أيضًا ؛ 3) وأخيرًا ، نسلط الضوء على تفاعلات miRNA-mRNA الجديدة باستخدام منهجية يمكن توسيعها لتشمل مجموعة واسعة من الأمراض.

Acknowledgement

Firsly, I take this opportunity to pass my great appreciation and gratitude to my supervisor prof Bruno Apolloni, for his prompt response whenever I need help, and continuous advice, support, patience, and high motivation. He has guided me in all the phases of the research, until the production of this thesis. I would also like to thank professor Mohamed Ahmed Salih for helping me during the early stages of my research, and provided me with lots of facilities to succeed in drawing an idea for my research. I would like to sincerely thank Prof Izzeldin Osman who founded this phd program in our country, under the supervision of highly qualified professors. I would like to dedicate this research to my family, my beloved parents; my father who was hardly waiting for this moment; my dead mother's soul, who passed away before attending this moments, God bless her in her grave; My husband who provided all means of support, to finish this work, my kids, my sisters , my colleagues, my friends and everyone who shared me the success of my study. Also special thanks to my university mates at University of Khartoum who provided me with all means and a place where I could finish my study. God bless them all .

Contents

List of Figures

List of Abbreviations

AD	Alzheimer's Disease
BAG-1	BAG family molecular chaperone regulator 1
CART	Classification And Regression Tree
Cdc42	Cell Division Cycle 42
Cdk6	Cyclin-dependent Kinases 6
CML	Chronic Myeloid Leukemia
CRC	Colorectal Cancer
DICORE	DIscovering COllective group RElationships
DNA	Deoxyribonucleic acid
E2F	is a group of genes that encodes a family of transcription
EXP5/XPO5	Exportin 5
FXS	Fragile X-syndrome
G1	First Gap Phase
G2	Second Gap Phase
GC	Gastric Cancer
GC	Guided Clustering
GEO	Gene Expression Omnibus
GEO2R	Interactive tool for GEO Series samples analysis
HOCCLUS2	Hierarchical Overlapping Co-CLUStering2
HPC	High Performance Computing
I	Matrix of ones
KIT	Proto-oncogene, receptor tyrosine kinase
LOOCV	Leave-one-out cross-validation
miRNA	microRNA
MM	Multiple Myeloma

MMRM	miRNA-mRNA Regulatory Module
MRF	Multivariate Random Forest
MRT	Multavariate Random Tree
Myc family	Myelocytomatosis
NCBI	The National Center for Biotechnology Information
NGS	Next Generation Sequencing
PAM.	Partition Around Medoid
PCR	Polymerase Chain Reaction
PFV-1	Primate foamy virus type 1
PM	Proximity Matrix
pre-mi	Pre cursor miRNA
pri-miRNA	Primary miRNA
PTEN	Phosphatase and Tensin
RAS	Resistance to audiogenic seizures
RB	Retinoblastoma
RFCM3	Relevant and Functionally Consistent MiRNA-mRNA module
RH-SAC	Rough Hypercuboid Based Supervised Clustering
RNA	Ribonucleic acid
RT-primer	Reverse Transcript Primer
S	Cell Synthesis Phase
SA	Simulated Annealing
SDM	Significant Dissimilarity Matrix

SigProx	Significant Proximity Matrix
SIX1	Sine oculis homeobox homolog 1
Stag	Small Tumor Antigen
SV40	Simian Virus 40
SV40 LTag	Large T-antigen
TF	factors
TSG	Tumor Suppressor Genes
UTR	Untranslated Region

Chapter 1

Introduction and Background

This chapter is an introduction to the research that has been conducted in discovering micro RNA and RNA modules. Through this chapter, the problem statement, objectives, scope, and how the research has contributed to find a solution to this problem, are all going to be stated.

1.1 Background

A microRNA (abbreviated miRNA) is a small non-coding molecule, that is made up of approximately 22 nucleotides, found in plants, animals, and some viruses. It results in RNA silencing and post-transcriptional regulation of gene expression. They cause transcriptional cleavage or translational repression through binding their target mRNAs. miRNAs expression occurs during a variety of cellular processes, such as development, cell proliferation, apoptosis and stress response, and affects the expression of variety of genes at the post transcriptional level. They act as regulators by binding their targets to degrade or suppress the translation of their transcripts.

1.2 Motivation

Micro RNAs are involved in many molecular interactions, such as defense against viruses and regulation of gene expression during development.

miRNAs interact with 3'UTR regions to repress gene expression. With their action at the post-transcriptional level, they may fine-tune the expression

of as much as 30% of all mammalian protein-encoding genes, they target.

This miRNA-mRNA relationship gives miRNAs the ability to control cellular protein output and function, in more powerful manner. Therefore, methods which are able to discover these miRNA-mRNA regulatory modules (MMRMs), would help to identify the key cellular pathways, contributing to a biological event, such as cancer prognosis [1–3]

1.3 Challenges

Most of miRNA and mRNA researches are done to study disease development and etiologies. Therefore, most of the datasets are related to specific disease conditions, and hence most of the studies concentrates on only differentially expressed miRNAs and mRNAs. Unfortunately, this is not adequate to study miRNA-mRNA regulatory relation in depth, and some information could be neglected. On the other hand, most of the miRNA target prediction databases, produces lots of false positive results. On the same context, many bioinformatics databases have arisen recently to reduce the number of false results, but experimental validation. However, it is unfeasible to carry out this huge number of experiments. What is really challenging, is applying a strong and reliable computation strategy, that needed to exploits as much of the available data as possible, and the choice of suitable miRNA target prediction database. Understanding the basis of these prediction methodologies used in such databases will guide the choice of the most suitable tool as well as output interpretation [4, 5].

1.4 Research direction

The research will not develop a novel target prediction algorithm, but rather exploiting more existing ones. It complies with Jayswal et al framework.

The research focuses on the following:

1. All items available in our dataset are considered as candidates, for module discovery, rather than just focusing on differentially expressed miRNAs and mRNAs We use a measure that fits module discovery, enriched progressively with information.

2. The output is a list of modules sorted according to the level of strength of their discovery.

3. Decision is postponed till the end, when all available information are exploited, rather than taking it based on just partial information.

4. A similar strategy to exploit large amount of data, is made possible, in the presence of large computational resources, that make it possible to manage large amount of data in HPC centers, help us getting the final decisions. In our case, we need to manage 296 x 7325 miRNA-mRNA matrix obtained from Multiple Myeloma dataset available on GEO platform [6].

5. Widening our scope to discover more modules, where unprecedented pairs are not considered. Those pairs are most probably not differentially expressed in the Multiple Myeloma as a disease experiment, and also not targeting according to the available database. However, disease databases may reveal some regulatory relation between each pair.

1.5 Problem statement

There are many factors that affect mRNA regulation such as Transcription factors, proteins, time-course data, and miRNA transfection effects on mRNA and proteins) could the predicted miRNA–mRNA interactions. Therefore in miRNA-mRNA interaction research, more powerful computations are needed since miRNA is not the only factor that interacts with mRNAs. A single miRNA targets multiple mRNAs, and a single mRNA is targeted by multiple miRNAs to develop a many-to-many relations, known as miRNA-mRNA module. Some methods have ignored this relation, by just focusing on miRNA-mRNA pairs. Others do not require the enrichment of miRNA targets and mRNA regulators. Besides, most of the current methods consider down-regulation only, although up-regulation could also be closely related to disease. Consequently, many methods evolved to consider this relation in the form of as miRNA-mRNA module discovery. Unfortunately they focused on some part of the data analyzed, and ignored the other. They still leave open issues, but the higher benefit is that they provided results, that have opened an issue of the possibility of widening the scope of module discovery, so as to be extended to a wider disease spectrum, where miRNA-mRNA interactions play a relevant role.

1.6 Objectives

To investigate the role of non-coding region of the human DNA on the coding region, and hence exploit the Next Generation Sequencing that makes the whole genome sequencing an easy task.

To have a significant contribution in the miRNA-mRNA module discovery area of research by improving the accuracy of finding a strong miRNA-mRNA regulatory relationship that might help scientists to focus on the resulting modules that are involved in many disease etiology. Our research aims to achieve the following:

1. Investigate the many-to-many miRNA-mRNA relation and how methods of identification of miRNA-mRNA modules have resolved this relation complexity.

2. Enhance an existing framework that combines multiple sequence-based miRNA Target Prediction tools with expression data, so as to increase the accuracy of predicting miRNA-mRNA modules.

3. Evaluate and the proposed enhanced framework results, by comparing with disease databases.

4. To have a significant contribution in building ensembles from different datamining algorithms to produce better prediction results.

1.7 Research Methodology

A literature review has been done to introduce different approached that has been used in miRNA-mRNA module discovery. Proposed methods has been introduced and the implementation steps have also been well established. Dataset has been downloaded and preprocessing steps has been done to get data ready for implementing the proposed method. The algorithm has been repeated for getting more accurate results, and needs to be investigated using another algorithms that chooses the most top modules candidates.

miRNA and mRNA clusters have been obtained, then an algorithm is applied to find out which miRNA and mRNA clusters are closely related based on a similarity measure of their expression. At the end those mostly related clusters are the discovered modules.

1.8 Structure of the Thesis

The thesis is made up of five chapters direction. Chapter 1 gives a general overview of the thesis, in terms of research area, and what is expected from its contribution in the this area. It describes the problem statement and the proposed solution. Chapter 2 provides a literature review about what has been done in the area and describes the fundamentals background knowledge, and its state of the art. Chapter 3 describes the details of how the proposed solution has been implemented. Chapter 4 introduces the results obtained from the work done during the methodology. Chapter 5 discusses those results, and provides an open discussion for the scientists in the area of research to give more result interpretation. Chapter 6 is the conclusion and summarizes the contribution of the thesis, and gives some recommendation for future work.

Chapter 2

Literature Review

2.1 Introduction

miRNAs are regulatory endogenous non-coding RNAs, approximately 21-23 nucleotides, found in bacteria, plants and animals. They cause transcriptional cleavage or translational repression through binding their target mRNAs. miRNAs expression occurs during a variety of cellular processes, such as development, cell proliferation, apoptosis and stress response, and affects the expression of a variety of genes at the post transcriptional level. They act as regulators by binding their targets to degrade or suppress the translation of their transcripts.

A single miRNA could simultaneously regulate dozens of mRNAs. On the other hand, any given mRNA sequence could be targeted by multiple different miRNAs [7–15].

Since miRNAs may be involved in many cellular processes, they compose gene regulatory networks Hence, we need to expand our knowledge to have better understanding of these networks.

Because of the above targeting multiplicity, there is a many-to-many relationship, that requires a deep understanding of the biological observations .It gives miRNAs the ability to control cellular protein output and function, in more powerful manner. Therefore, methods which are able to discover these miRNA-mRNA regulatory modules (MMRMs), would help to identify the key cellular pathways, contributing to a biological event, such as cancer prognosis [1–3]Three facts should be considered while discovering MMRMs: First, the expression level of many mRNAs are simultaneously regulated by a single miRNA. Second, mRNA sequence could have many miRNAs binding sites. Third, it is very difficult to conduct experimental studies for finding rela-

tionship between miRNAs and their mRNA targets. Sequence based approach includes seed sequence complementarity, evolutionary conservation, and thermodynamic stability. As a complimentary approach to overcome this complexity, a computationally predictions approach for finding putative miRNA-mRNA targets has been developed to make it easy to characterize relevant miRNAs. Unfortunately this approach produces a high rate of false positive results [16].

To understand the biological function, we need to know that it suffers from instability due to the severely significant changes that occurs in miRNA-mRNA relationships, associated with cancer disease for instance. The large-scale availability of patients matches in miRNA and mRNA expression datasets, made it possible to find miRNA and mRNA biomarkers, that could be helpful to discover disease associated miRNA-mRNA regulatory modules. That means that using sequence and expression data would provide a good opportunity to gain deeper insights into the functions of biological processes, and hence discover disease associated modules [16, 17].

Many methods have been developed in this context, using different algorithms, mostly in different types of cancer studies. Most of them have used mathematical,statistical and datamining approaches. Data used were either sequence data only, expression data only, or combination of both.

Currently, comparing different methods is not an easy task, since different algorithms use different aspects of miRNA-mRNA interactions. An algorithm could be selected according to the what kind of information are available, and what biological question needs to be answered.

Different methods could complement each other, although they are not directly comparable, for the sake of enhancing their functionality due to their integration. However there is a lot of work to do to fine tune the integration and infer as much knowledge as possible from it.

The methodologies discussed in this review may provide better inference to miRNA regulatory mechanism with miRNA-target predictions and expression profiles of miRNA and mRNA. In summary, it may be possible to study functions of miRNAs by integrating genome-wide computational and experimental data and developing sophisticated tools to handle them.

2.2 Role of MicroRNA in human diseases

Studies have proven that miRNA regulatory function play an important role in the pathology and etiology of human diseases. Understanding these functions

will pave the way for the discovery of the crucial role that miRNA could play in disease therapy.

2.2.1 miRNA gene regulatory function

microRNA is a small non-coding RNAs capable of regulating gene expression using a complicated mechanism that result in either mRNA degradation ,inhibiting translation, protein and promoter bindings, or interacting indirectly with another non-coding RNAs.

To resolve the complication of this mechanism, Lytle et al suggested the association of miRNAs with any position of its target mRNA, by binding in 5' UTR rather than just focusing in the 3' UTR Another perspective is that miRNA gene expression regulation occurs at the transcriptional level, by binding to the regulatory elements of the DNA.

2.2.1.1 MicroRNA biogenesis

The small 19-22 nucleotide long miRNAs are single stranded ,non coding and multifunctional RNA responsible for regulating gene and protein functions. They affect translation or post transcriptional regulation of the genes they target, resulting in the regulation of the biological functions of the genes in question. In mammals, these effects are either by mRNA degradation or translational repression. In cancer they could function as oncogenes and tumor suppressor genes [18]. Thus, diverse physiological and pathological processes are functionally affected by the action played by miRNAs regulate gene expression either by translational repression or by mRNA degradation in mammals. There are sequential steps that formulate the canonical miRNA processing pathway see figure 2.1: First, the RNA polymerase II transcribes miRNA into primary miRNA (primiRNA); Second, the pri-miRNA is processed into precursor miRNA (pre-miRNA);Third, Exportin 5 (EXP5/XPO5) protein assists in exporting pre-miRNA from the nucleus to the cytoplasm; Finally, Dicer furtherly processes pre-miRNA into mature miRNA [18, 19].

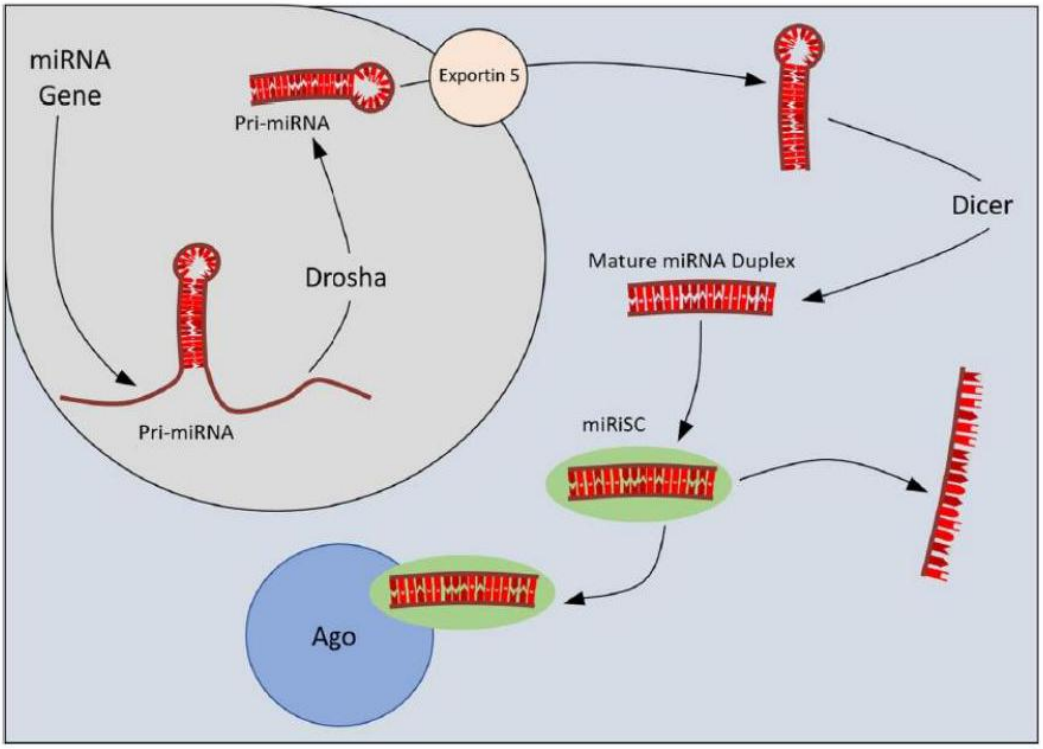


Figure 2.1: Schematic showing the synthesis of miRNAFour canonical miRNA processing pathway :F irst, the RNA polymerase II transcribes miRNA into primary miRNA (pri-miRNA); Second, the pri-miRNA is processed into precursor miRNA (pre-miRNA); Third, Exportin 5 (EXP5/XPO5) protein assists in exporting pre-miRNA from the nucleus to the cytoplasm; Finally, Dicer furtherly processes pre-miRNA into mature miRNA .

2.2.2 miRNA's involvement in human diseases

miRNAs plays a significant role in disease etiology. Mutations, dysregulation of miRNAs and their targets can lead to the development of many diseases such as cancer. Most of miRNA-disease studies are in cancer, since most of miRNA-targeted human disease genes are oncogenes [20].

In cancer researches miRNAs are divided into two categories: oncogene, and tumor suppressor genes. The most interesting thing is that the same miRNAs can act as oncogene in one cell and TSG in another. For example miR-222 is hyperexpressed in liver cancers targeting PTEN as a suppressor, while is downregulated in erythroblastic leukemias targeting c-KIT as oncogene. Furthermore, miRNAs play a significant role in cancer predisposition and initiation.

miRNAs are very interesting biomarkers for disease diagnosis. This is because it has higher stability than RNA when exposed to severe conditions

such as boiling or very high or low ph level [21].

2.2.3 miRNA functions

miRNA is involved in many essential biological processes such as such as, apoptosis, cell development, proliferation differentiation, signal transduction, nervous system, viral infection, immunity system, angiogenesis, cancer and so on.

2.2.3.1 Function on cell differentiation and development

The overexpression of a miRNA could inhibit a cell proliferation.For instance the overexpression of MiR-128 in glioma cells. Meanwhile, the conserved 3' UTR region of target site of E2F3a region (UTR) of E2F3a, could regulate cell cycle progression. Therefore the protein levels of E2F3 and normal blood cells were negatively correlated to miR-128 expression level.

2.2.3.2 Function on nervous system

miRNA could affect the function of Plasticity which is the ability of the brain to make adaptation to new information. Synaptic pasticity which is the change that occurs in synapses, which are the neurons that allow communication [22].

When miRNAs dysfunction, they are able to maintain the survival of mature neurons and carry on functions on their behalf. This is could form overgrowth, and obstacles to synaptic plasticity, leading to many nervous system diseases such as Alzheimer's disease (AD), fragile X-syndrome (FXS) and autism [23].

2.2.3.3 Function on viral infection

Host cells use miRNAs to defend against RNA and DNA viruses, by targeting the viral function. However viruses could control the host cell , by taking advantage of the miRNAs to control their host cell; while the, host cells reciprocally use miRNAs to target essential viral functions. Experimentally, miRNAs are involved in the innate immunity with which you were born.

The first anti-viral miRNA to be reported was the one, which effectively inhibits the accumulation of the retrovirus primate foamy virus type 1 (PFV-1) found in human cells.

SV40-encoded microRNA (miRNA), miR-S1, impairing LTag and STag functions essential for the viral life cycle, by downregulating them [24]. It limits the exposure of the infected cell to cytotoxic T lymphocytes, which are a type of immune cells capable of killing certain cells, such as cancer cells, and cells infected with a virus [25].

2.2.3.4 Function on immunity

The deregulation of immune related miRNAs could affect immune development and cause immune disorders.

2.2.3.5 Function on cancer

The Abnormal cellular development, in various types of cancer, has also been associated with miRNAs. miRNAs work as regulatory molecules, acting as oncogenes or tumor suppressors [23].

Studies have shown that miRNAs can function as oncogenes or tumor suppressors in the initiation, progression and metastasis of various types of cancers such as lung cancer, breast cancer, ovarian cancer, colon cancer and prostate cancer. The effect of miRNA may differ on the same disease. For example, miR-137 showed both cell multiplication by targeting cell division cycle 42 (Cdc42), cyclin-dependent kinases 6(Cdk6) in lung cancer cells and down-regulation by DNA methylation.

2.2.3.6 Function on Angiogenesis

Angiogenesis is the process of creating new blood vessels, during health and disease It is is an important process that occurs both during health and disease. Blood supply is needed when a new tissue is formed, for its growth and sustenance. When the body loses control of maintaining the balance of angiogenesis regulators, it may result in too much or too little angiogenesis. It would be a serious health issue when it comes to occurs within cancers and tumors. miRNAs such as miR-622 regulates angiogenesis of Corolectar Cancer, and miR-590-5p acts both as an oncogene and tumour suppressor in cervical cancer and renal cancer respectively [26].

2.2.3.7 MicroRNAs and complex diseases

miRNA play a crucial role in complex diseases diagnonsis and treatment, such as in cancer. Their expression profiles represent a useful biomarker in cancer diagnosis , prevention and therapeutics [27]. It could have abnormal activities of proliferation and antiapoptosis that are the main cause of the promotion of oncongenesis . MicroRNAs having these activities could be overexpressed in cancer cells.

MiRNAs act as oncogenes when they have proliferative and antiapoptotic activity ,therefore they may be overexpressed in cancer cells. Examples of such miRNAs is mir-17 cluster that consists of six miRNAs: miRs-17-5p, -18, -19a, $-19b$, -20, and -92. On the other hand microRNAs with antiproliferative and proapoptotic activity, act as tumor suppressor genes and thus leading to underexpression in cancer cells. The family of let-7 miRNAs is a n example for such kind of miRNAs [10] .

2.2.3.7.1 microRNAs as oncogenes

It has been suggested several experiments and clinical analysis, that miRNA may act as oncogene or tumor suppressor.When they are overexpressed in tumors they are considered as oncogenes(oncomirs). They promote the development of tumors by negatively inhibiting tumor suppressors genes, or genes controlling differentiation or apoptosis.mir-17-92 is a good example of oncomirs, that enhances lung cancer cell development as well as accelerating lymphomagenesis [10].

In its relation with E2F-RBpathway, Karina et al have revealed in their study that mir-17-92 functions in collaboration with RB, to contribute in lung cancer etiologies.mir-17-92 regulates the expression of E2F1 repressing its translation. Its overexpression decreases E2F1 protein level. Some findings revealed that E2F1 being modulated by the oncogene c-Myc and mir-17-92, affects the ARE-p53 pathway through which miR-17-92 inhibits Myc induced apoptosis.

One of the mir-17-92 targets is PTEN, which is a tumor suppressor gene in lung cancer promoting apoptosis that prevents upnormal cell growth and division. Mutations of PTEN are the main lung cancer etiologies [28].

2.2.3.7.2 miRNA as a tumor suppressor

In cancer studies, some miRNAs act as tumor suppressors, when their expression is decreased. They prevent tumor development by their negative inhibi-

tion effect on oncogenes and/or genes responsible for cell differentiation or apoptosis.let-7 is an example of such miRNAs. Its upnormal expression may cause oncogenic loss of differentiation. Studies have revealed the negative effect of let-7 on RAS oncogene. In lung tumor tissues for instance, let-7 expression level is reduced, while increased in RAS, when compared to normal lung tissues. This suggests that let-7 acts as tumor suppressor gene.

miR-140 acts a s tumor suppressor in many types of cancers such as hepatocellular cancer, hypopharynx geal and gastric cancer (GC) cancer miR-140-5p induces cell apoptosis and decreases it downregulates SIX1, one of its target, by its in chronic myeloid leukemia(CML) cells [29].

Akiyo Yoshida et al proposed in a study, that IL-6 one of mir-140 direct target; is downregulated by RB through upregulation of mir-140. RB depletion downregulate mir- 140 which in turn, upregulate IL-6 gene. Therefore RB downregulates IL-6 through upregulation of mir-140.

2.2.3.8 The E2F-RB pathway in cancer

E2F is a transcription factors family, which are proteins that bind to DNA to regulate expression either by promoting or suppressing transcription. It is found in eukaryotes [30, 31].

The retinoblastoma protein pRB (gene name is RB or RB1) is a tumor suppressor protein that disfunctions in many types of cancers inhibits cell proliferation ,as a tumor suppressor, it is inactivated in many types of cancers, by inhibiting the TFs of E2F family [32–34].

The E2F-RBpathway regulates apoptosis, and RB inhibition of apoptosis is an important mechanism of tumor suppression whereby cells deficient for RB function can be eliminated by apoptosis. The E2F-RBpathway regulates apoptosis, and RB inhibition of apoptosis is an important mechanism of tumor suppression whereby cells deficient for RB function can be eliminated by apoptosis.

Many studies have revealed the central role played by the pathway controlling RB tumor suppressor protein that regulates E2F transcription factor. This pathway is the main regulator of the initiation of DNA replication, and itit is distrupted in almost all human cancers[13]. Cell proliferation and apoptosis are two main cellular events are required for an organism development.

Many biological processes such as homeostasis and tissue development requires balanced cell proliferation and apoptosis [35]. They are essential for maintaining the 37 million cells that compose human body. When a cell is

Phase One	**G_1 phase**	The cell grows physically larger, copies organelles, and makes the molecular building blocks
	S phase	The cell synthesizes a complete copy of the DNA in its nucleus.
	Ga Phase	More cell growth, proteins and organelles, are made and begins preparation for mitosis
Phase Two	**During the mitotic (M) phase,**	Two new cell are made when the original l cell divides its copied DNA and cytoplasm .

Table 2.1: Cell Division Phases

eliminated by apoptosis, proliferation compensate for cell death by cell proliferation regulated by growth signals, as well as abnormal growth stimulated by overexpression or oncogenes activation that leads to tumor genesis. To overcome tumorgenesis,RB-p53 pathways suppress tumor development by inducing apoptosis. Their inactivation in most of cancers, indicates their role in tumor suppression [20].

2.2.3.8.1 Cell Cycle

In eukaryotic, stages of cell cycle are divided into two main phases: interphase and the mitotic (M) phase. Cell grows and makes its DNA copy During the interphase phase, while During the mitotic (M) phase, the cell forms two new cells, by separating its DNA into two sets as well as dividing its cytoplasm.

The interphase is made of three phases: first gap phase (G1), cell synthesis phase S, and the second gap phase G2. Each phase is monitored by a check point to emphasize that genetic information has been faithfully transmitted to the next generation. The most important checkpoint is G2/M checkpoint to prevent cells with DNA damage from entering the mitosis phase [35]. Table 2.1 below summarizes cell cycle phases [21].

2.2.3.8.2 Role of E2F-RB in proliferation

As a principal target of RB pathway, E2F plays a central role in proliferation. In fact E2F consists of two groups of transcription factors: activators

(E2F1-E2F3a and repressors (E2F3bE2F8). RB family members (pRB and p130) repress their targets by binding to E2F3b-E2F5 on their target promoters (Figure 2). This inhibits E2F's transcriptional activity, then by changing chromatin structure, RB actively represses the expression of E2F target gene. Cyclin D type dependent kinase (CDK4 and 6) are activated,resulting in $\mathrm{p130}$ and pRB inactivation

RB is inactivation occurs during transition from phase G1 to phase S by CDKs, during cell growth. It inhibits binding to E2F repressors, unleashing E2F to promote cell progression. Upon Growth CDK4 and CDK6 are activated to inactivate pRB to inhibit binding to repressors , and hence release E2F from RB suppression inducing its target genes Cyclin E, $\mathrm{E2\ F1-E2\ F3a}$ and Cdc6 which is involved in coordinating the S phase and mitosis. Later, Cyclin E activates CDK2 to inactivate RB, to activate $\mathrm{E2\ F}$ inducing its targets to initiate the S phase.

This functional interaction between E2F and RB elucidates the critical role they play in promoting and retaining cell proliferation. Specifically, defect in RB such as mutation or deletion, could lead to various types of cancer etiologies [20, 36].

2.2.3.8.3 Role of E2F-RB in Apoptosis

E2F promotes apoptosis, during the role it plays in tumor suppression. This has been suggested by E2F1 knockout mice that showed a remarkable increase in tumor formation. When E2F1 is overexpressed, it activates p53 promoting apoptosis. Besides, it activates ARF tumor suppressor gene, which is an upstream p53 activator. Here it is important to discriminate between forced inactivation of pRB as a response to ARF gene induction, and that one induced by cell growth stimulation. The former occurs as a result of oncogenic changes. Stimulation of ARF gene expression by E2F1, serves as a tool to discriminate cancer and normal cells. The role played by RB as proliferation and apoptosis inhibitor of, needs a deep understanding. It is important to know whether RB is inhibited during cell cycle to promote growth by growth stimulation, or RB has lost its function as a result of mutation. Apoptosis is induced by RB loss of function, while cell growth is controlled by RB inhibiters to overcome proliferation resulting in upnormal cell growth, that leads to tumor development. The process of RB inhibition is reversible, and occurs during cell cycle by phosphorylation, while its loss of function such as the one in Rb1-null mice induces apoptosis, hindering potentially malignant cells development. RB could inhibit apoptosis and hence it serves proliferation that

results in tumorgenesis. This RB ability is contradictory to its role as tumor suppressor [37].

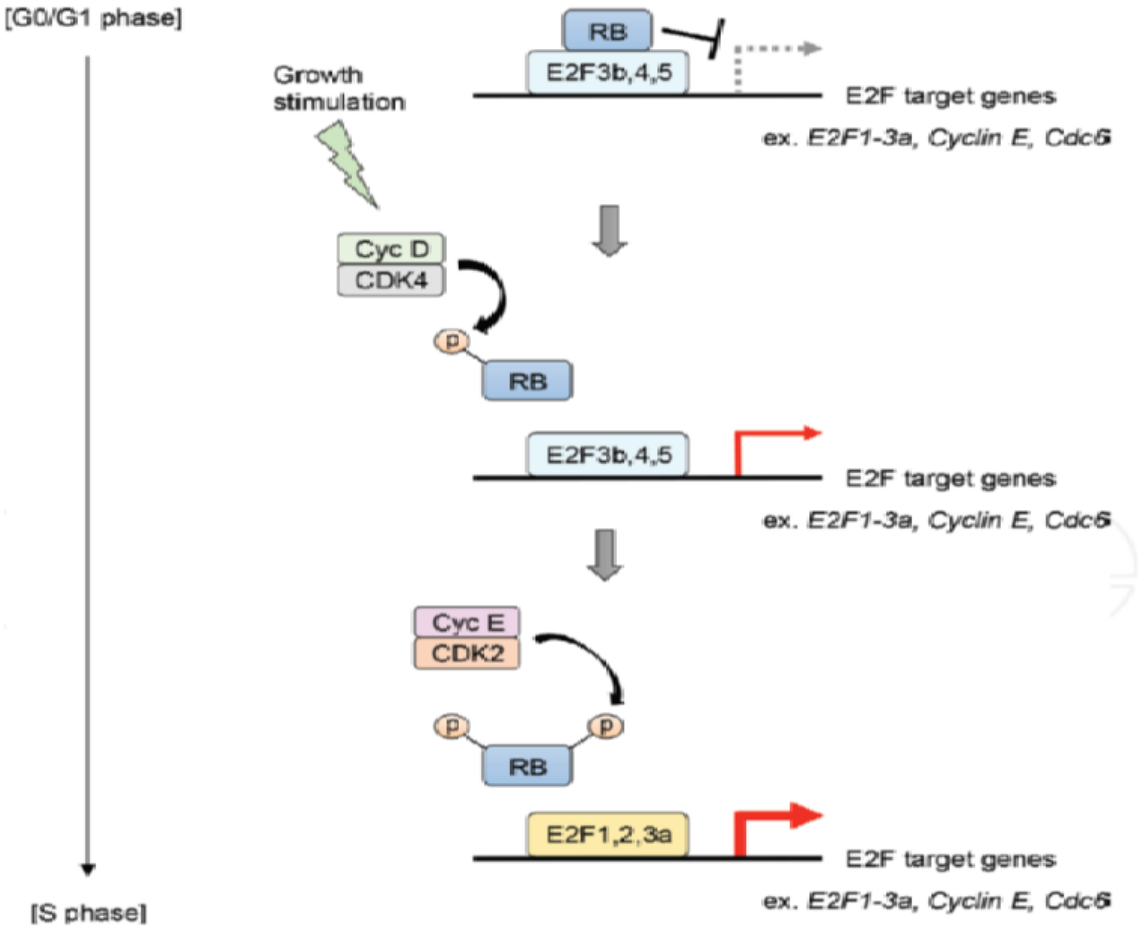


Figure 2.2: Regulatory mechanism of E2F target genes by E2F and RB. RB family members (pRB and p130) repress their targets by binding to E2F3b-E2F5 on their target promoters (Figure 2). This inhibits E2F's transcriptional activity, then by changing chromatin structure, RB actively represses the expression of E2F target.

The RB expression is not lost colorectal cancer (CRC), indicating the role played by its function in colorectal tumor development. As a mechanism in tumor suppressor, RB inhibits apoptosis, to act as an anti-apoptic protein [33].

It then interacts with another anti-apoptic protein BAG-1 that is upregulated in colorectal carcinogenesis. Therefore targeting Rb-BAG-1 complex for sake of promoting apoptosis plays role in therapeutic development [36,38, 39].

2.3 Different approaches in miRNA-mRNA module discovery

The miRNA-mRNA modules are a multidisciplinary topic which has been acquainted by different perspective, depending on the researcher background. The formal statement of the problem is the idenfication of a biclicque in a biparite graph, like the one in Figure 2.3. A bipartite graph is a graph whose vertices can be divided into two disjoint and independent sets U and V such that every edge connects a vertex in U to one in V. In our case, U is ther se of red vertexes associated to miRNA and V is the set of blue ones associated to the mRNA. A biclique, or complete bipartite graph, is a subset of the above bipartite graph, where every vertex of the first set is connected to every vertex of the second set. In our case the subgraphs may be also approximately complete, in the sense that we look for pairs of subsets of U and subsets of V that jointly express a co-regulating mechanism. Hence the problem to solve is the following:

Given a bipartite graph $\mathcal{G}$ with parts U and V find a pair sets $u \subseteq U$ and $v \subseteq V$ such that for each vertex $u_i \in u$ and $v_j \in v$there exist and edge connecting them.

The solution of the formal problem is far from being simple for three reasons:

1. the edges of the graphs are not univocal, as for definition, and generally costly to be identified,

2. the grouping criterion is not univocal as well,

3. once fixed edges and grouping criterion, the analytical solution cannot be expressed in a closed form, while the numerical solution is so costly so as to denote the biclique identification problem to be NP-hard. This requires for approximate solutions of the problem.

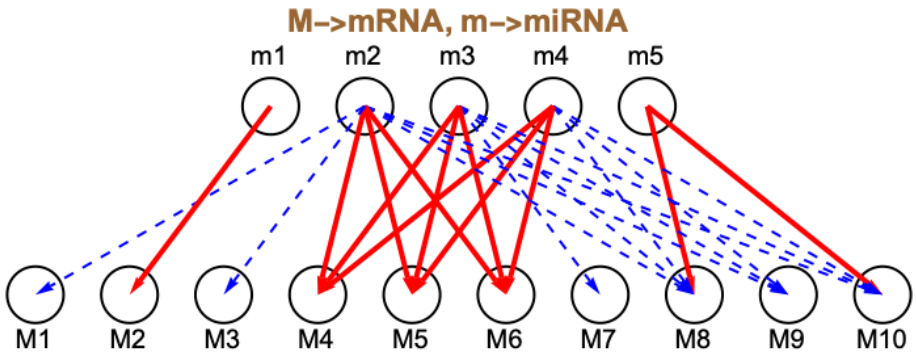


Figure 2.3: Bipartite miRNA-mRNA graphs hosting *modules* aka the biclique emphasized in red.

Hence, various approaches arose in the literature to bypass those complexities by either facing a relaxed formulation of the problem or directly bypass it via alternative formulations. In the following we will surveys the various approaches, with a special focus on the Data Maining Approach, that has been adopted in this thesis work.

2.3.1 Direct Biclustering

- With reference to Figure 2.3 a hinge of the wanted subgraphs may be represented by *seeds* [40]. Namely, for a target mRNA, let's consider the smallest- close Ss subsets of all miRNA that bind it, where smallest refers to the inclusion relation between sets, and closeness to a metric on the elements of S and a threshold over the difference on their values. In particular, the adopted metric is the first principal component of the pairs (local alignment score, duplex free energy) of those miRNAs (see Figure 2.4).

  The key operational tool is a trie, i.e. a uni-directed search tree where branches originated by a same seed overlap and split to recover the above Ss. Since the Ss may derive from different genes, thew overlap crossroads are the glue of the emerging modules (see Figure 2.5).

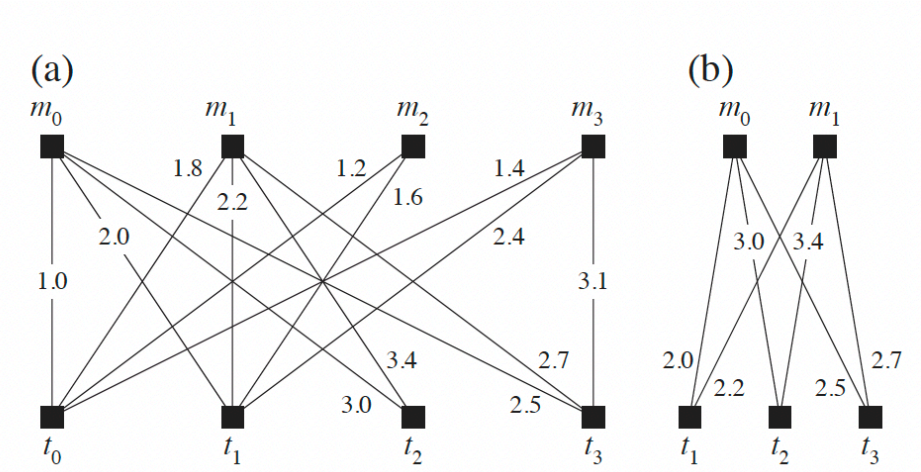


Figure 2.4: (a) Example relation graph G = (M U T, E, w), where M are miRNAs $\{mo, m1, m2, m3\}$, and T are mRNAs $\{to, t1, t2, t3\}$ with some hypothetical weights reporting the above metric. (b) A module found in G.

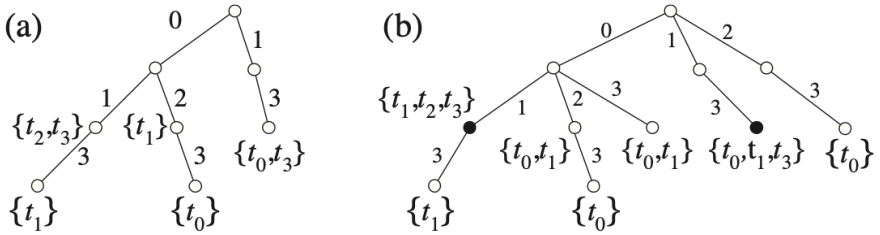


Figure 2.5: (a) The trie representation seeds. The edge labeled i represents miRNA m_i. (b) The merged seeds. The solid-circled vertices represent a candidate for MRMs.

However with current state of the art technologies this method is somewhat obviated.

- A more straightforward method to identify modules is propsed in [41]. Simply, a module is identified by a subset of row and columns of the miRNA$\times$mRNA matrix whose cells globally maximize the sum of cell cross-relation indexes. For instance, the maximization goal g is the function

$$g(I, J) = -\frac{1}{|I||J|} \sum_{i \in I, j \in J} c_{ij} \tag{2.1}$$

where $|I|$ is the size of the subset of rows and $|J|$ is the size of the subset of columns in the sub-matrix and cij is the entry at row i and column j,

given by

$$c_{ij} = min(\rho_i j, 0) \tag{2.2}$$

with $\rho_{ij} =$ the forrelation between row i and colun j when we are interested on the sole inhibitory effects of miRNAs over mRNAs.

According to BUBBLE biclustering method, with proper thresholding $|I|$ and $|J|$ we may identify these clusters in approximate way via Simulated Annealing (SA) algorithm, starting from a seed cluster which is progressively expanded through addition of single rows and column with probabilities ruled by SA. The method deserves further difficulties related ot the visualization of the results (see Figure 2.6).

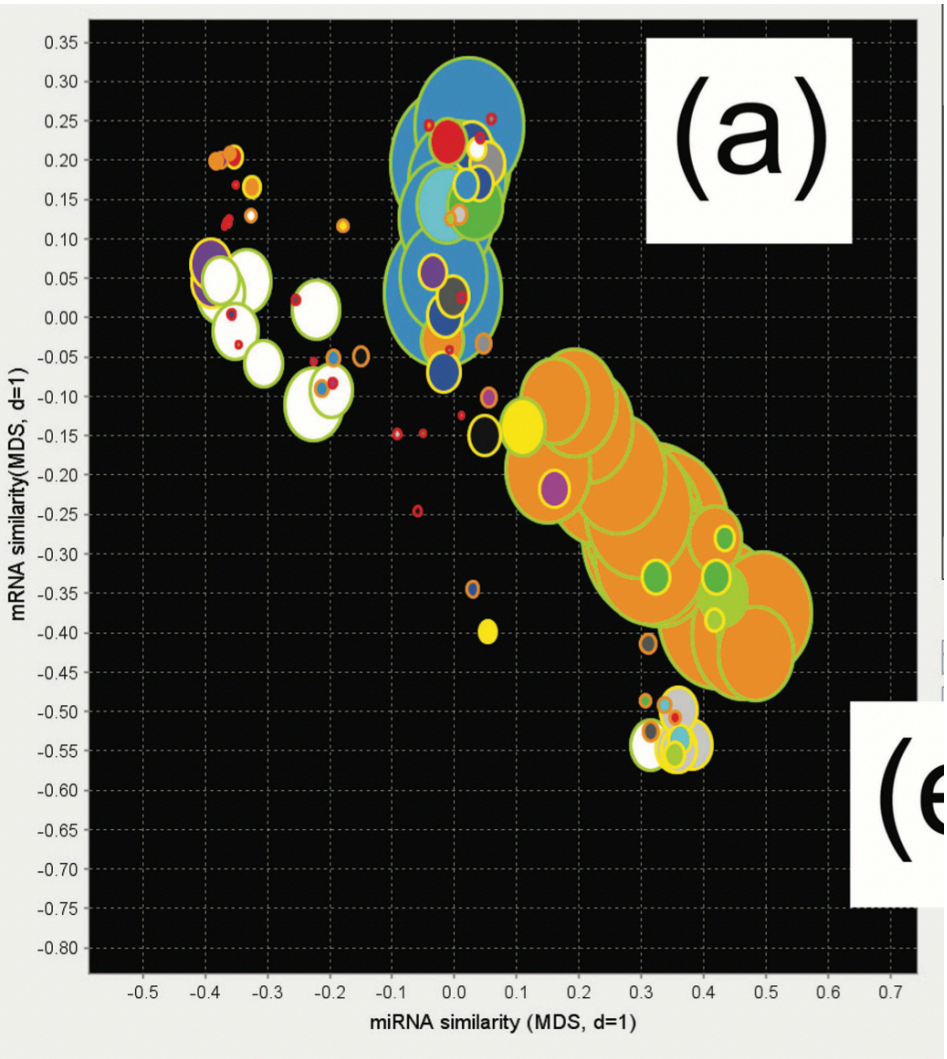


Figure 2.6: bi-cluster visualization in the similarity feture space.

2.3.2 Statistical Approach

- A relaxation of problem 2.3 consists in working with continuous variables in place of discrete ones. This is done in DIscovering COllective group RElationships (DICORE) [42], where edges between $miRNA_i$ and $mRNA_j$ pairs are affected by a weight w_{ij} corresponding to the correlation between the related expressions on a set of patients; edges with weight less that a given thershold are canceled. Then a divide et impera strategy is adopted where these weights are used once to discover homogeneous clusters of miRNAs and of mRNAs, hence (in a special version) to appreciate the functional link between clusters of the two cathecories.

  Namely, a colaboration score between two miRNAs i and j is computed as:

$$v_{ij} = \frac{(\sum_{k=1}^{l} w_{ik} w_{jk})^2}{\sum_{k=1}^{l} w_{ik} \sum_{k=1}^{l} w_{jk}} \tag{2.3}$$

  where l is the number of the possible mRNAs that both $miRNA_i$ and $miRNA_j$ interact with. Analogously for mRNAs. Further, a cohesiveness score $cs(C_i)$of claster C_i is defined as:

$$cs(C_i) = \frac{w_{int}(C_i)}{w_{int}(C_i) + w_{ext}(C_i) + \alpha|C_i|} \tag{2.4}$$

  where $w_{int}(C_i)$ denotes the sum of the collaboration scores of all the internal pairs of variables, i.e. each pair only contains variables within the cluster C_i; $w_{ext}(C_i)$ is the sum of the collaboration scores of all the external pairs, i.e. each pair contains one variable within the cluster C_i and one variable outside the clusterC_i; and $\alpha|C_i|$ is a penalty term.

  A clustering algorithm is aimed to maximize the coesiveness of the computed clusters, so as to accomplish the first phase of the algorithm. The second phase simply sorts the crossed pairs of miRNA and mRNA clusters throug a *canonical correlation* measure *r* beteween a linear combination of the expressions $\overrightarrow{X}$ of the miRNA in a cluster and a linear combination of the expressions $\overrightarrow{Y}$ of the mRNA in a a paired cluster,

computed as

$$\boldsymbol{r} = \max_{a,b} \frac{\overrightarrow{a}' \Sigma_{XY} \overrightarrow{b}}{\sqrt{\overrightarrow{a}' \Sigma_{XX} \overrightarrow{a}} \sqrt{\overrightarrow{b}' \Sigma_{YY} \overrightarrow{b}}} \tag{2.5}$$

where Σ_{XX}, Σ_{YY} and Σ_{XY} are variance of $\overrightarrow{X}$, $\overrightarrow{Y}$, and covariance between $\overrightarrow{X}$ and $\overrightarrow{Y}$. respectively.

The method have been applied to three gene expression disease datasets, and it was found that the miRNAs and mRNAs in highly related miRNA-mRNA groups play important regulation role in the disease in question. That means that the MMRMs identified by DICORE and functionally enriched are the most relevant to the biological conditions of the given datasets. The method obtained good results, although it uses only expression data. Still there are more chances to discover more modules rather than only those associated with specific disease. More data could be exploited for more module discovery.

- Maintaining the biclicque goal, but forgetting the bipatite graph, another kind of approach is based on a sophisticated implementation of a regression model between the marix $X \in \mathbb{R}^{N \times T}$ of the expressions of T mRNAs of N patients and the matrix $Y \in \mathbb{R}^{N \times D}$ of the expressions of D miRNAs of the same patients [43]. From the relation

$$Y = XW + E \tag{2.6}$$

we get the coefficient matrix $W \in \mathbb{R}^{D \times T}$ that we factorize as

$$W = W_x W_y + E \tag{2.7}$$

with $W_x \in \mathbb{R}^{D \times R}$ and $W_y \in \mathbb{R}^{R \times T}$ for a proper R, via a variant the Lasso regularization technique. The authors use these new matrices to identify modules. Namely they:

  1. sort the columns of W_x and the rows of W_y from the largest to the smallest to establish a relevance order of their pairs row-colums,
  2. threshold the coefficients in the above rows and columns to identify the mRNA and miRNA involved in the pairs which so represent modules with a given relevance.

  To check their significance, modules are subjected to functional gene set enrichment and survival analysis. Module passing at least one of the

two tests are considered meaningful. For validation, literature review and miRTarBase computational databases have been used as supporting evidence.

The computational framework has shown its ability to find regulatory modules in different types of cancer, using matched miRNA-mRNA datasets. Although the framework uses only expression data, it did not focus only on differentially expressed (DE) miRNAs and mRNAs, and produced good results that were strongly validated.

- A further step toward a statistical abstraction of the problem is represented by the relevant and functionally consistent miRNA-mRNA module (RFCM3) algorithm [44]. Here the role played by the correlation in [42] is now covered by the Mutual Information. Namely, given two random variable X and Y, the mutual information $I(Y, X)$ is defined as

$$I(Y, X) = H(Y) - H(Y|X) \tag{2.8}$$

where $H(Y)$ denotes the entropy of Y and $H(Y|X)$ the conditional entropy distribution of Y given X. On the basis of mutual information, we identify relations as in Figure 2.7. Namely, for each miRNA X we identify a group of mRNA $Y_i$swhich

  1. have high $I(Y_i|X)$, thus denoting a dependence of the mRNA expression on miRNA expression,
  2. have high $I(Y_i|Y_j)$, thus denoting high relation between each pair of mRNA inside the group.

  To complete the module formation, miRNAs are grouped according to the miRNA similarity matrix (MISIM) [45] which decrees similarity to genes on the basis of association with similar diseases.

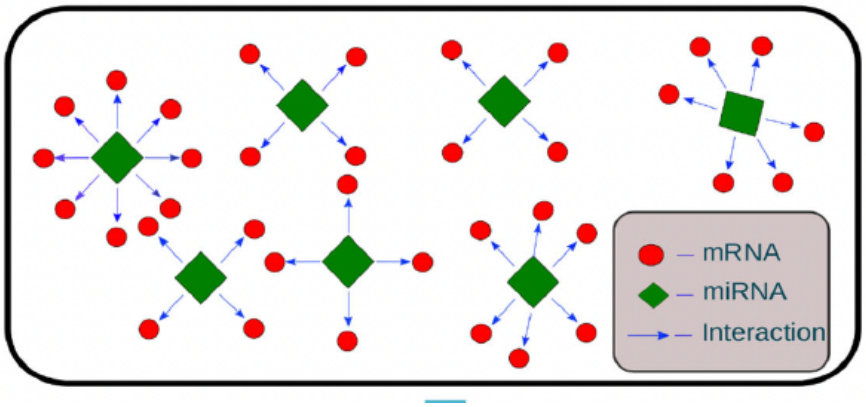


Figure 2.7: Star arrangements of mRNA versus miRNA.

As a result, a module of multiple miRNAs and mRNAs is generated. RFCM3 algorithm outperforms many competitors, whereas the use of heterogeneous data would give a chance for more module discovery.

- Finally, an extreme release of the modules discovering problem from a dry statistical perspective is represented by the Bayesian approach in [46]. The hinge is the notion of conditional probability of a random variable Y given the value x of another random varialeb X (like the probability of a tomorrow jam traffic given that tomorrow is raining). This is used to create a probabilistic chain like in Figure 2.8. Replace $\boldsymbol{a}_d$ with the set of mRNAs, x with one item of the set, z with a module (miRNA regulatory module miRM) and w with the observed gene expression simply normalized in $\{-1, +1\}$, and we obtain the event chain. Its probabilistic features are specified by the distribution θ of $z|x$ (i.e. of miRM given a miRNA) and the distribution ϕ of w given z (i.e of a gene expression given a miRM). Both distributions are symmetric Dirichlet distribution

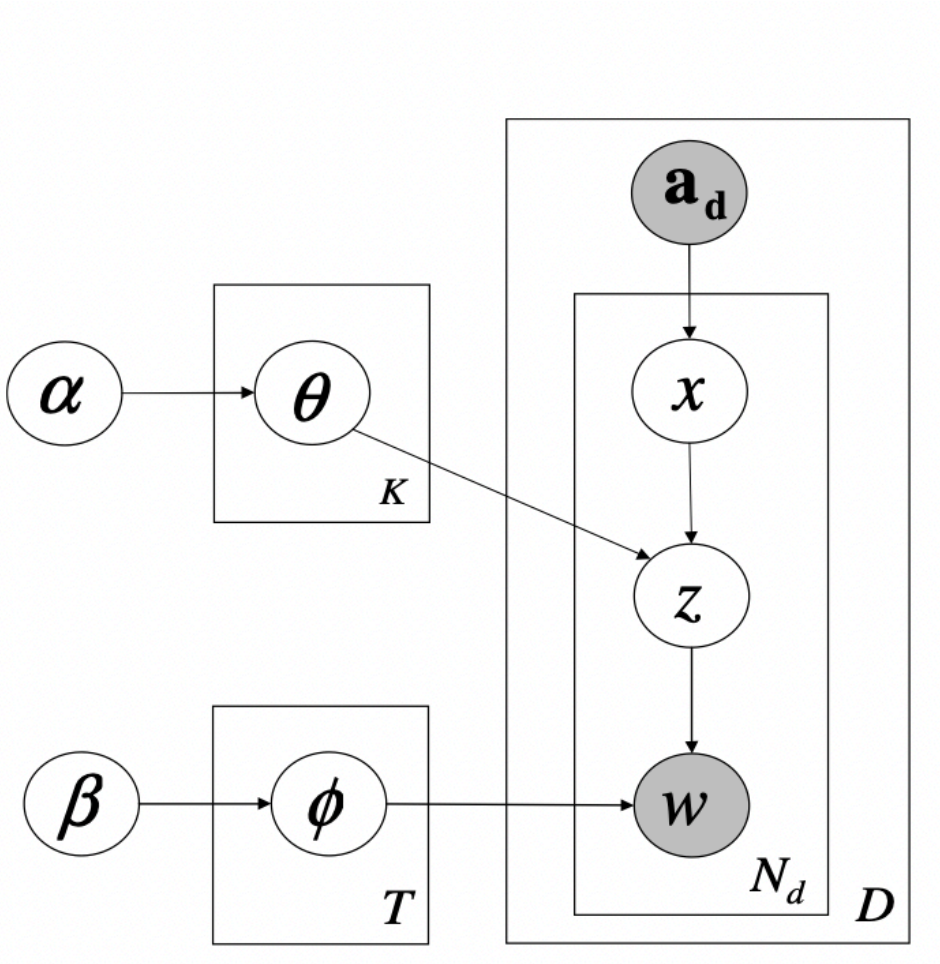


Figure 2.8: The module probabelistic chain.

$$f(x_1, \ldots, x_K; \alpha) = \frac{\Gamma(\alpha K)}{\Gamma(\alpha)^K} \prod_{i=1}^{K} x_i^{\alpha-1}. \tag{2.9}$$

with α replaced by our α and β respectively, so that, once assumed the latter as a-priori hyper-parameters, these distribution are estimated through the statistics:

$$\hat{\theta}_{lk} = \frac{\hat{W}_{lk}^{MT} + \alpha}{\sum_{k'} \hat{W}_{lk'}^{MT} + T\alpha} \tag{2.10}$$

$$\hat{\phi}_{kn} = \frac{\hat{W}_{kn}^{ET} + \beta}{\sum_{n'} \hat{W}_{kn'}^{ET} + Q\beta}, \tag{2.11}$$

where: k, l and n index the miRNA, miRM and the observation, respectively, $\hat{W}_{lk}^{MT}$ is the number of times miRNA l is assigned to the k-th miRM excluding the current instance, $\hat{W}_{kn}^{ET}$ the number of times expression type n is assigned to the k-th miRM excluding the current instance. T is the number of miRMs and Q is the total number of expression types. These values are collected collected via an iterated Gibbs sampling exactly based on a uniform distribution of miRNAs and current estimates of the Dirichlet distributions.

The method succeeded in reducing the number of false positive miRNA targets discovery with respect to conventional bi-clustering in the case studies. It is also open to capture stronger relations between modules and related biological processes.

2.3.3 Evolutionary approach

The joint maximization of within and between groups coherence of the two clusters composing a module may be pursued through a a genetic algorithm as in [47], see Figure 2.9. Namely:

1. the goal function is the fitness F of a module that is expressed as

$$F(M,T) = \alpha BS_{M,T} + \beta EC_M + \gamma EC_T + VOL \tag{2.12}$$

   where $BS_{M,T}$ is the mean binding score of the subset of the target information matrix consisting of (M,T). EC_M and EC_T are the expression coherence (EC) scores of M and T, respectively, computed as the mean of Pearson's correlation coefficient between all miRNAs or mRNA pairs inside M and T. The volume term VOL accounts for the rate of the miRNA-mRNA pairs contained in the module w.r.t. the overall number considered in the module identification task.

2. the initial population of the module is picked randomly from the overall sets of miRNAs and mRNAs.

3. the renewal of the population is based on a probability distribution that is rooted on the above fitness with the following strategy. First we identify a core module of most fitting pairs. The probability of sigle individuals to belong the the new generation is updated with respect to the one of

the previous generation as a function of the connection to the above core via the binding matrix. Namely:

$$p_j^r(q+1) = (1-\delta_r)p_j^r(q) t\delta_r \frac{1}{V}\sum_{k=1}^{V} z_j^k \tag{2.13}$$

where r denotes one of the groups (either od miRNA or of mRNA) and z a corresponding item. $z_j^k \in \{0,1\}$ indicates whether z_j is not binded or it is to any complementary item of the above core module.

4. population renewals proceed until the maximum number of generations is reached.

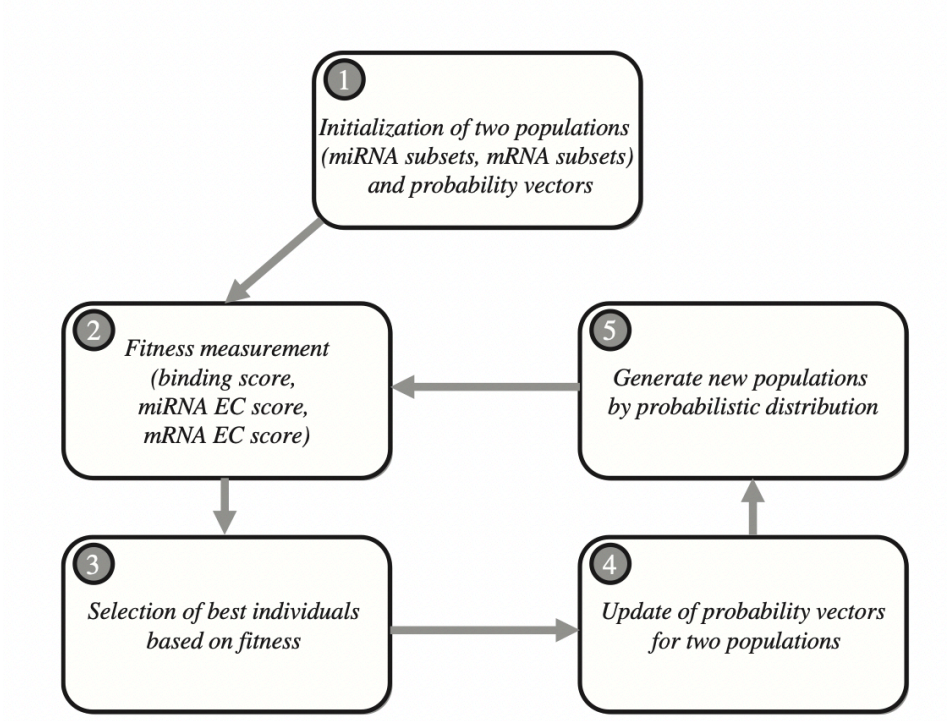


Figure 2.9: The general scheme of the procedure.

Unfortunately, since the algorithm is stochastic in nature and depends mainly on the sensitive parameters in the fit, no guarantee exists for optimal results, although the use of heterogeneous data improves its performance.

2.3.4 Statistical aspects

Data mining approach adopted in this thesis lead us to face the problem of handling correlated samples, which finds a template instance in the problem of computing statistics from truncated samples. Both problems singularly constitute an extensively referenced research track *per se* to which we contribute by stressing some crucial features.

Correlated samples are the typical subject of longitudinal data analysis in medical data, with wide application in epidemiology [48–50]. *Per se*, sample correlation heavily hampers the suitability of the computed statistics. Hence, methods such as mixed effects [51] and generalized estimating equations [52] partition the samples in subsamples that are internally independent and model the dependence among them via regression models. Despite, we introduce a rather empirical method to actually process the sample correlation which we will show in the next sections.

Though in some cases used synonymously, what differentiates censored samples from truncated samples is the knowledge of the sample size [53]. Thus, in a censored sample of size m we may exploit a shorter number of observations because of a threshold on their values (type I censoring) or on their number, for instance we observe only the c smallest values (type II censoring). In a truncated sample we have only r observation available and know the mode (type I or type II, analogously) but we do not know how many data we missed because of truncation, i.e. their count in proportion to the observed samples is also not observed. While methods of inferring from censored data are well developed, especially in survival analysis [54] and insurance [55], truncated samples are dealt with mainly in specific cases [53]. Actually, the problem has been faced since 1760 [56], finding solutions earlier with momentum methods (see for instance [57]) and subsequently with maximum likelihood methods (see for instance [58]). In recent years the problem has taken the features of a machine learning procedure [59, 60], possibly either endowed with approximate *oracles* or affected by some unfeasibility lemmas [61].

2.3.5 Data Mining Approach

Hence, various approaches arose in the literature to bypass those complexities by either facing a relaxed formulation of the problem or directly bypass it via alternative formulations. In the following we will surveys the various approaches, with a special focus on the Data Maining Approach, that has been adopted in this thesis work

2.3.6 Sushmita Paul et al Method

The method proposed by Sushmita Paul et al, to identify miRNA-mRNA modules for Colorectal Cancer is a two-step method, that used a guided clustering datamining approach [62] .Figure 2.10 show the proposed approach for identification of miRNA-mRNA modules. In the first step they applied Rough Hypercuboid Based Supervised Clustering (RH-SAC) algorithm, to generate functionally similar miRNAs/mRNAs biomarkers, having coherent expression for further classification. The Support Vector Machine, has been used later to select the best miRNA/mRNA clusters. In the second step, those best miRNA/mRNA clusters which are highly differentially expressed and have similar functionality, have been integrated to create a miRNA/mRNA regulatory modules. Thereafter, to relate the module to significant pathway associated with CRC disease, functional enrichment analyses, disease association, fisher's test, and correlation analysis have been performed, as well as survival analysis for clinical evaluation. The method was able to identify potential modules, as well as both negative and positive between miRNAs and mRNAs.

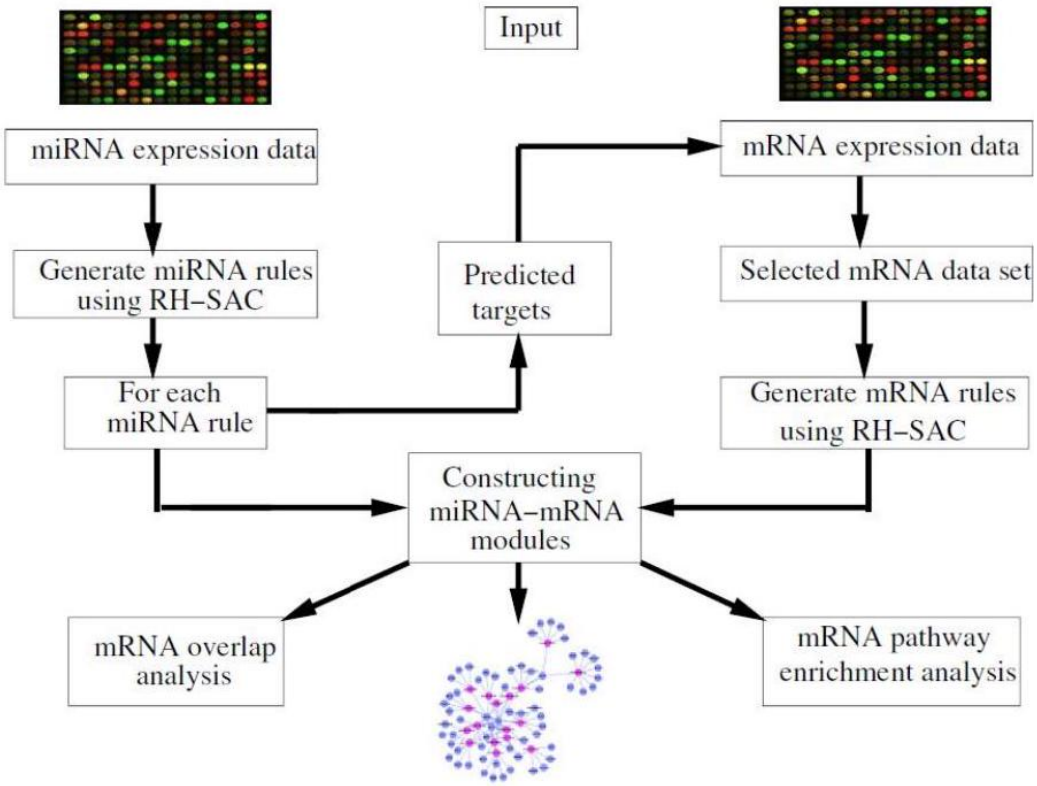


Figure 2.10: Schematic flow diagram of the proposed approach for identification of miRNA-mRNA modules.

This is the first step for generating miRNA-mRNA modules, which is selecting miRNA clusters/rules whose average expression has 100% accuracy of classifying the samples using SVM classifier. The rough hypercuboid based supervised clustering (RH-SAC) algorithm has been used to select potential miRNA/mRNA rules/clusters. It discovers groups of miRNA/mRNAs, which are functionally similar, and their average expression values are capable of discriminating samples. The similarity measure generated using this algorithm helps grouping miRNA/mRNA.

Next is the description of $\mathrm{RH}-\mathrm{SAC}$ algorithm:

Let $\mathbb{C} = \{\mathcal{M}_1, \ldots, \mathcal{M}_i, \ldots, \mathcal{M}_j, \ldots, \mathcal{M}_m\}$ set of features and $\mathbb{D}$ is the class label.

Let $R_{\mathcal{M}_i}(\mathbb{D})$ be the relevance of feature $\mathcal{M}_i \in \mathbb{C}$ with respect to class label $\mathbb{D}$. It is a metric, whose values fall in the range 0-1, that shows to what extent the feature is relevant to class label.

The supervised clustering algorithm starts with feature $\mathcal{M}_i$ that has the highest relevance value. An initial cluster of featrures $\mathbb{V}_i$ is created, selecting set of features $\{\mathcal{M}_j\}$ that have similarity value $\psi(\mathcal{M}_i, \mathcal{M}_j)$ with $\mathcal{M}_i$ which is greater than a predefined threshold value δ.$\mathbb{V}_i$ is defined as

$$\mathbb{V}_i = \{\mathcal{M}_j \mid \psi(\mathcal{M}_i, \mathcal{M}_j) \geqslant \delta; \mathcal{M}_j \neq \mathcal{M}_i \in \mathbb{C}\}$$

two augmented cluster representatives could be generated,by averaging $\mathcal{M}_j$ with features, or $\mathcal{M}_j$ compliment with $\mathbb{V}_i$ features, using the two equations below respectively

$$\overline{\mathcal{M}}^{+}_{i+j} = \frac{1}{|\overline{\mathbb{V}}_i| + 1}\left\{\sum_{\mathcal{M}_k \in \mathbb{V}_i} \mathcal{M}_k + \mathcal{M}_j\right\}$$

$$\overline{\mathcal{M}}^{-}_{i+j} = \frac{1}{|\nabla_i| + 1}\left\{\sum_{\mathcal{M}_k \in \mathbb{V}_i} \mathcal{M}_k - \mathcal{M}_j\right\}$$

Another way of computing the augmented feature $\overline{\mathcal{M}}_i$ is by considering a subset of features $\nabla_i \subset \mathbb{V}_i$ that increases the relevance value rather than considring the whole set $\mathbb{V}_i$. To find more clusters of features, the same procedure is repeated after discarding the ∇_i set of features from $\mathbb{V}$.

After generating miRNA and mRNA clusters using RH-SAC algorithm, further investigation is needed to identify miRNA-mRNA modules. Let us consider only miRNA/mRNA clusters that achieved 100% leave-one-out cross-validation (LOOCV), although only mRNAs that are targets for those miR-

NAs are considered. Consequently, experimentally validated miRNA target database, miRTarBase37 was further used to select targets for each miRNA.

Only mRNAs that are targets to the selected miRNAs were considered together with their expression. The same procedure was applied using the RH-SAC algorithm to find mRNA clusters that are functionally similar and differentially expressed. Finally the miRNA and their target mRNA rules were combined to form a miRNA-mRNA regulatory network.

2.3.7 Jayaswal et al Method

Jayaswal et al have proposed a two step miRmR module identification method. The first step is to create mRNA and miRNA clusters. The second step is to find association between those miRNA and mRNA clusters, so that miRNA-mRNA clusters having significant association are the potential miRmR module [11]. The author has introduced the guided and unguided clustering approach. The unguided clustering uses only information from target prediction databases, to create the dissimilarity matrix as an input to the clustering algorithm. On the other hand, guided clustering combines information from miRNA target prediction database that uses sequence data, with miRNA and mRNA expression profile, to create the dissimilarity matrix. In the first step, the author has used guided clustering. He created a dissimilarity matrix using Multivariate Random Forest classification algorithm, and Partition around medoid as a clustering algorithm (PAM). That means a clustering algorithm has been guided by a classification algorithm that created the input.

The author was comparing clustering results using guided and unguided clustering, using bootstrap test to find which miRNA and mRNA clusters are enriched, and found that guided clustering produced higher number of enriched clusters.

2.3.7.1 The MRF algorithm

The guided clustering method Figure 2.11 is based on multivariate random forest (MRF) [63]to cluster mRNAs or miRNAs. In case of clustering mRNAs, these two matrices are:Y $\times X$ miRmR, where Y is represents the mRNAs and X represents the miRNAs. Each [i,j) entry of $Y \times X$ takes the value 1 if miRNA j targets mRNA i, and 0 otherwise. Second, the $Y \times T$ matrix which is the mRNA expression and T represents the number of conditions require four parameters:

a)Node size: which is the least number of element in a leaf node. This is a user defined value. b)numcov:to build a regression tree

[64–66] either we use one variable to split the node, or more than one variable, where the best is chosen to split the node.The term univariate is used when we have a single variable, and multivariate otherwise in the $Y \times T$ matrix T represents the no of conditions in the expression matrix. That is why our tree is a multivariate regression tree. We refer to the number of variables to be taken each time the node is split as numcnd. The value

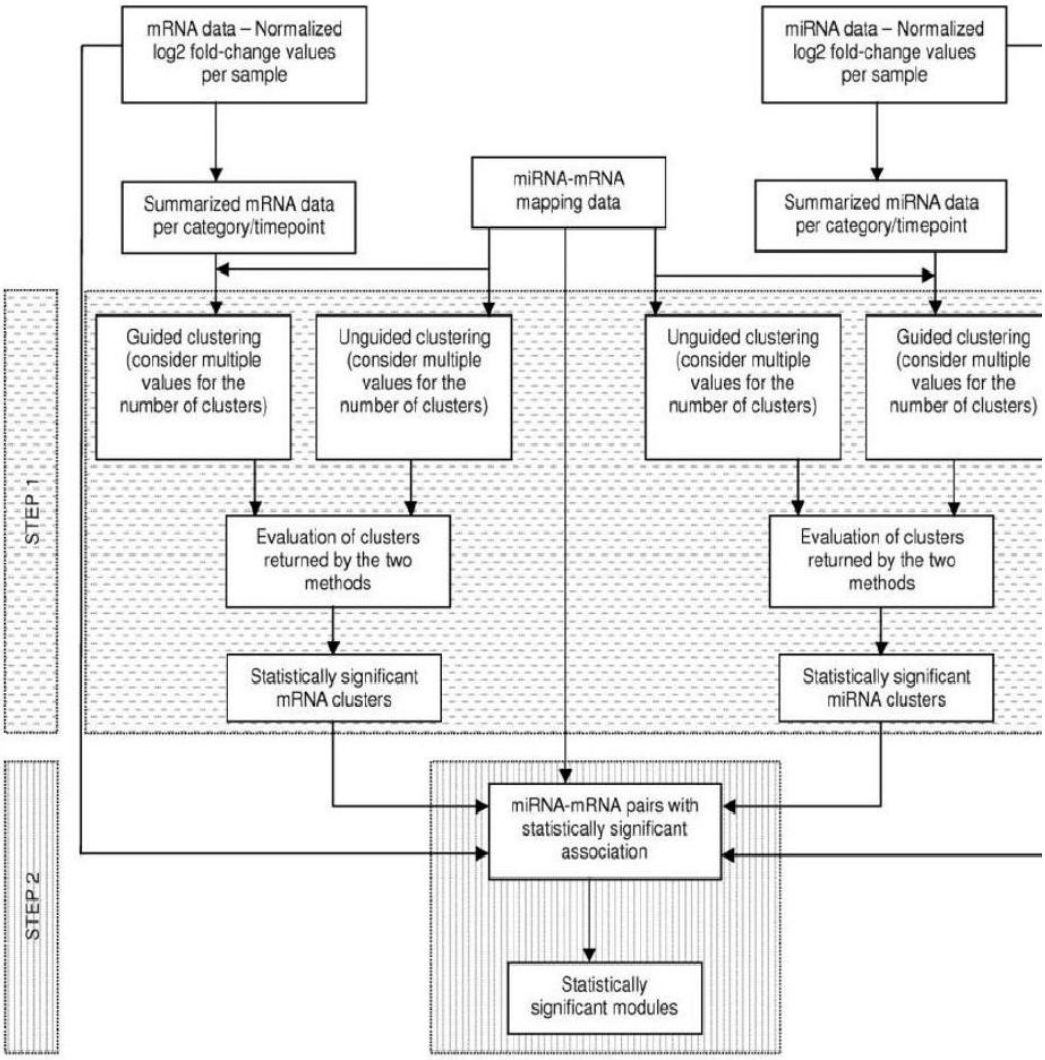


Figure 2.11: Two steps method for the identification of miRmR modules. Step1: The clustering of miRNAs or mRNAs requires two input parameters - miRNAs or mRNAs dissimilarity between and the number of clusters, K. Step 2: Identification of statistically significant modules. A miRmR pair is considered to have an association if the pair is computationally predicted, and have associated change expression.

of numcnd was suggested by Segal to be the square root of the whole number of miRNs/mRNA in the dataset. By defining $d\left(\mathbf{x}_1, \mathbf{x}_{\text{avg}}\right)$ as the Euclidean distance between the vectors $\mathbf{x}_1$ and $\mathbf{x}_{\text{avg}}$ and $\mathrm{S}(\theta) = \sum_{\mathrm{l}\in \mathrm{Q_m}} \left\{\mathrm{d}\left(\mathbf{x}_1, \mathbf{x}_{\text{avg}}\right)\right\}^2$ to be a measure of node homogeneity, i.e. how miRNA expression are close in their values. Change in node homogeneity is defined as: $f\left(\mathrm{BN}, DN_1, DN_2\right) = S(BN) - S\left(DN_1\right) - S\left(DN_2\right)$.the function returns different values for each miRNA, therefore we refer to h as the miRNA with highest value.

c) No of trees: which represents how many trees that build our MRF.

The output of the MRF algorithm is YxY matrix (XxX in case of miRNAs) proximity matrix. Each entry takes the value 1 if the two mRNAs are targeted by the same miRNA, and 0 otherwise.

The second step is the identification of statistically significant modules using miRNA and mRNA clusters. Statistical significance is denoted, if for each miRNA-mRNA pair is computationally predicted, and a change in each miRNA expression is associated with a change in mRNA expression. The latter conditioned is tested using a linear model.

The method has the ability to identify negative miRNA-mRNA pair associated having miRNA targeting mRNA and positive association in which a miRNA indirectly target mRNA in the miRNA-mRNA pair. It also succeeded in decreasing the false positive miRNA-mRNA association into a smaller one, enabling biologists to proceed with further analysis.

2.3.8 Malik Yousif et al Method

Malik Yousif et all have developed a new tool called miRcorrNet for discovering miRNA-mRNA module by analyzing their expression data by performing machine-learning based integration [67]. miRcorrNet generates groups of a single miRNA with mRNAs whose expression level are correlated to it. The development of this tool has been inspired from previously developed tools such as Recursive Cluster Elimination (RCE) for classification and feature selection SVM-RCE [68] [69], [68, 69].

The general approach used in such tools is using two components: a grouping function $G()$ and ranking function $R()$. In the The grouping function $G()$ component mRNA and miRNAs are grouped using either computational algorithm such as k-means clustering algorithm; grouping based on biological information; or both computational and biological. The tool is strong in the sense that identified set of genes are capable of distinguishing cases versus control classes. Those genes together with the associated miRNAs, serve as biomarkers for some diseases (see Fig. 2.12).

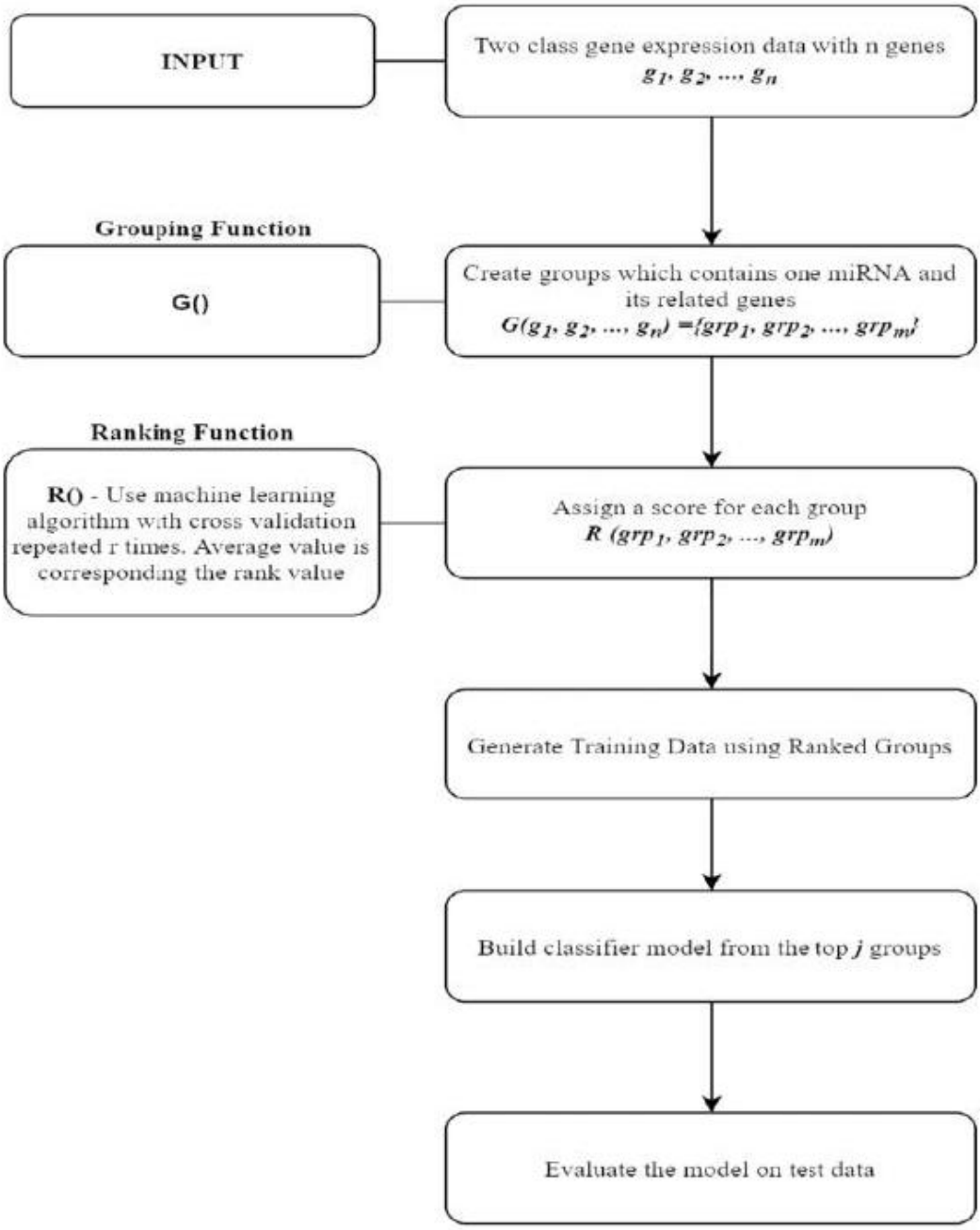


Figure 2.12: The flow chart of the proposed procedure.

After generating groups using G(), the ranking function $R()$ is used to assign scores from each group, to express the ability of each relevant group to distinguish between case and control cases. The $R()$ function uses cross validation together with a classification algorithm to compute ranking scores. The scores are then ranked from largest to smallest, and then the tool is tested on the top y ranks, and therefore using genes associated with those top y groups,

sub data. The performance evaluation was also conducted using biological information from the literature, and using independent datasets.

2.3.9 Gianvito et al Method

Gianvito et al has proposed (Hierarchical Overlapping Co-CLUStering2),(HOCCLUS2) data mining algorithm that biclusters miRNAs and target mRNAs, using information from experimentally-verified and/or predicted interactions to extract significant biclusters, rather than using gene expression [70].

4. A set of initial non hierarchical clusters are extracted.

5. Creating an iterative algorithm, where in each iteration, overlap and merging are identified. Merging is performed when some heuristic criteria are satisfied. Several biclusters can be merged at the merging phase. This merging may lead to adding or removing additional level of hierarchy. The process continues until no more overlaps and merging are identified (see the algorithm described in figure 2.13 lines 9-23, and Figure 2.14).

6. Finally, to identify which biclusters are the most significant, ranking of extracted biclusters is done, based on statistical tests that compares intra-and inter-functional similarity of each bicluster, according to Gene Ontology classification.

Unfortunately, although HOCCLUS2, represents a good potential for miRNA-mRNA module identification, it has a poor performance on very large datasets.

Input: V_r (mRNA), V_c (miRNA); $A^{n\times m}$; function $q(\cdot,\cdot)$; parameters β; values avg_mirna, abs_min_mrna, min_mrna;

$L \leftarrow get\_set\_of\_one\_miRNA\_bicliques(V_r, V_c, A, \beta)$;
$aggregationCandidates \leftarrow \emptyset$;
for all pairs of biclusters $C', C'' \in L$ **do**
  if $C'_r \cup C''_r \geq min\_mrna$ **and** $C'_c \cap C''_c \leq avg\_mirna$ **then**
    $agg\_qual \leftarrow jaccard(C'_r, C''_r) * q(aggregate(C', C''), A)$;
    $aggregationCandidates \leftarrow aggregationCandidates \cup \{\langle C', C'', agg\_qual\rangle\}$;
  end if
end for
while $(aggregationCandidates \neq \emptyset)$ **do**
  $cand = \langle C', C'', agg\_qual\rangle \leftarrow getBest(aggregationCandidates)$; {according to agg_qual}
  if $(C' \in L)$ **and** $(C'' \in L)$ **then**
    $L \leftarrow L \setminus \{C', C''\}$;
    $C''' \leftarrow aggregate(C', C'')$;
    for all biclusters $C \in L$ **do**
      if $C_r \cup C'''_r \geq min\_mrna$ **and** $C_c \cap C'''_c \leq avg\_mirna$ **then**
        $agg\_qual \leftarrow jaccard(C_r, C'''_r) * q(aggregate(C, C'''), A)$;
        $aggregationCandidates \leftarrow aggregationCandidates \cup \{\langle C, C''', agg\_qual\rangle\}$;
      end if
    end for
    $L \leftarrow L \cup \{C'''\}$;
  end if
  $aggregationCandidates \leftarrow aggregationCandidates \setminus \{cand\}$;
end while
return L;

Figure 2.13: Initial biclustering algorithm for miRNA-mRNA module detection.

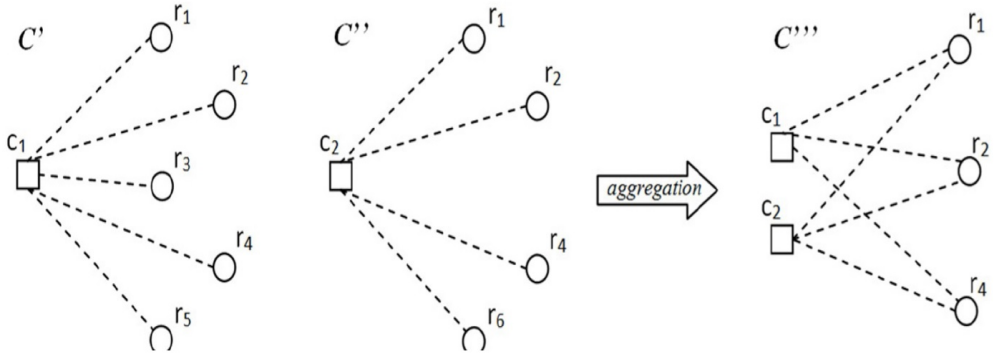


Figure 2.14: An example illustrating how two bicliques are aggregated (C' and C'') into a new biclique (C'''').

Chapter 3

Methodology

This chapter focuses on the approaches followed and the pipelines implemented while the research is conducted. It gives a clear description for these pipelines, and how they have been implemented during the study. It is worth mentioning how clusters have been derived and hence modules have been sorted in terms of strngth of association, leaving a wide research area for further investigation by the researchers of this domain in the future.

3.1 Holistic procedure to identify miRNA-mRNA modules

The aim of this research is to implement a framework that comprises a holistic procedure to identify miRNA-mRNA modules within a population of candidate pairs. It is obvious that current methods has proved to produce better results than those in the past, however they still leave open issues. In the previous studies, the focus was on analyzing differentially expressed miRNAs and mRNA only. Even the choice of DE mRNA and mRNA candidates in some studies was on the basis focusing on miRNA targets that were downregulated [71], and on those upregulated in others [72] . It is very important to understand that miRNA-mRNA relation is many-to-many relation. That means that miRNA could target a non DE miRNA and vis versa, and that is why when trying to discover more modules, this relation should be considered and there is a need to widen the scope of module discovery by exploiting the whole datasets and the biological results obtained from analyzing these data. Therefore, we adopt the strategy of postponing taking any decision until all the biological results are exploited. Instead of just implementing a pipeline of statistical tests,

we adopted a sequences of specially-devised evolving metrics to find possible solutions. In terms of computation, this strategy is rather expensive since there is a need for using High Performance Computing (HPC) for implementations. However, it allows the discovery of new modules, possibly hosting non differentially expressed miRNAs and mRNAs and pairs containing genes that currently are not discovered to be miRNA targets. In this research we implement our strategy on cancer instance analysis, using a Multiple Myeloma dataset publicly available on the Gene Expression Omnibus (GEO) platform, as a template to hazard some biological issues.

3.2 The strategy

Recall the approaches used in miRNA-mRNA module discovery, some authers adopted a divide et impera strategy that divide the problem into subtasks. Jayaswal for instance started with clustering miRNAs and mRNAs individually, then looked for more related pairs [11]. Complying with Jayaswal et al. pipeline, as an operational aspect, our strategy adopts a holistic approach, that postpone last decision to be as late as possible. Namely:

1. Unlike the common practice of focusing on only differentially expressed miRNAs and mRNAs, all available data is exploited to be as initial candidates for modules to be discovered.

2. Beside maintaining all of candidates, we consider information that enrich our discovery and fits with the module discovery goal.

3. The procedure produces a list of modules which is sorted according to their the strength of their relation based on an optimality criteria to their discovery. Our strategy was made feasible by managing a large amount of data using HPC centers. To reach our goal. Working with Multiple Myeloma dataset available on GEO platform, we produced 296×7325 miRNA-mRNA pairs. Therefore it was possible to widen the scope looking for new modules that considered unprecedented mRNAmRNA pairs that were non-differentially expressed. Nevertheless, many disease databases has supported the candidate relevance of being having regulatorily related.

3.3 The pipeline

Several miRNAs co-regulate a single mRNA, and multiple mRNAs are regulated by a single miRNA. Therefore to identify regulatory modules based on this fact is a major challenge As mentioned in the Introduction, to resolve complications that evolved from this many-to-many relation, we used some statistics to order the candidate modules according to some statistics. Based on this, we: 1) looked for computationally feasible solutions, and 2) enable the researcher to exploit the result for further analysis for more module discovery.

3.3.1 Metrics

Based on the statistics mentioned above, we used as metrics which is a combination of two datatypes: expression data and binding motifs from databases that uses sequence data.

We rely on $Y \times X$ and its transpose XxY binary matrices (map matrices) derived from databases and tools that use binding information based on sequence data, found on the GEO website.Y denotes mRNAs, X denote miRNAs and each matrix cell has value of 1 if crossing row and column bind, 0 otherwise. From these matrices a metric for partitioning tree is derived .Each node (column) k out of a predefined number of nodes h, is considered as a binary vector, where rows fall in the same partition, if the k' s bits coincide in this way rows are partitioned into two. Another metric which we relied on is the $Y \times T$ and $X \times T$mRNA and miRNA expression matrices respectively, which were also downloaded from the GEO website. These values of these matrices are floating values that have been normalized using standard steps. These matrices are used individually to derive an introductory similarity measure sm1 between the rows of node k mentioned above, considering possibly either norm L_2 Euclidean distance, or norm L_1 Manhattan distance. They are also used to derive sm2 a similarity significance measure between every two rows i and j, as 1 - pvalue of the linear regression of row i on row j.this results in a Y x X or X x X significance matrix.

3.3.2 Elbow Method

Knowing the best number of clusters before running a clustering algorithm, is very essential to give better data modelling. The Elbow Method is based upon the idea of choosing a number of clusters such that if one more cluster is added,

will not give a better data model [73,74].Thus it calculates the highest number of clusters suitable to model our data. We used the significant dissimilarity matrix as an input to the Elbow method.

3.3.3 Hausdroff Distance

It is really worth including the implementation of Hausdroff Distance in our pipeline [75–77]. The Hausdorff distance $d_H(A, B)$ is computed between two subsets A to represent mRNAs and B to represent miRNAs. These subsets are mRNAs and miRNAs clusters, respectively, so that the distance is defined as

$$d_{\mathrm{H}}(A, B) = \max \left\{ \max_{y \in A} d(y, B),\ \max_{x \in B} d(A, x) \right\} \tag{3.1}$$

where $d(y, B) = \min_{x \in B} d(y, x)$, $d(A, x) = \min_{y \in A} d(y, x)$, min and max are the usual minimum and maximum operators over sets, respectively, and the external maximum in (3.1) is the maximum of $\max_{y \in A} \{\min_{x \in B} d(y, x)\}$ and $\max_{x \in B} \{\min_{y \in A} d(y, x)\}$. Referring to figure 4.1-left, $d_H(A, B)$ represents the length of the longest double arrow (the blue one). Besides, differently from what has been represented in the above figure, we have implemented $d(y, x)$ as the Euclidean distance between y and the linear regression value of y over x. Therefore the linear regression function ℓ of y over x, is then computed, as well as regressed value $\nu_i = l(x_i)$, where i ranges over the number of components of y and x. We denoted distance d_l, such that $d_l(y, x) = \|y - y^{\sim}\|_2$. In figure 3.1-right $y_i - y_i$ is represented by the double arrow in orange [78].

3.4 Algorithms

We implemented three clustering procedures: Hierarchical divisive clustering, Agglomerative clustering [79], and Hausdorff linkage [76], which we are going to describe below.

3.4.1 Hierarchical divisive clustering

Hierarchical divisive clustering yields dendrograms that results in a binary tree. At the beginning we have a set $q \leqslant 2^m$ binary strings, where m is the number of partitions (nodes) in the tree, and the strings have fixed length n.

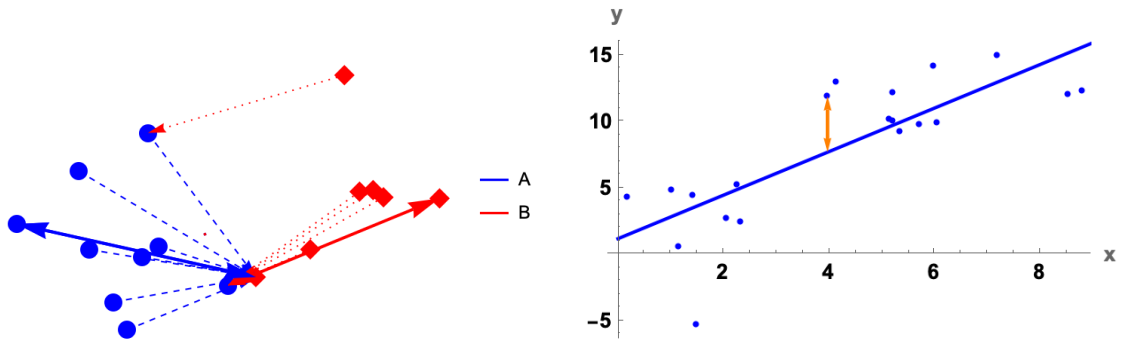


Figure 3.1: A sketch of the Hausdorff distance computation in our implementation. Left picture: bullets → elements of set A, rhombuses → elements of set B; blue dashed lines → minimal distances of bullets from the set B, red dotted lines → minimal distances of rhombuses from the set A; thick double arrows Maximal distances of A from B (in blue) and of B from A (in red). Right picture: points → experimental (x,y) pair; line → their regression line; orange double arrow → the difference $\tilde{y}_i - y_i$ relative to the pair of components (x_i, y_i).

We have q different ways to split these binary strings based on the value of the ith bit, for $i \in \{1, \ldots, n\}$. In principle, we take only a subset of those q strings to proceed with splitting. Thereafter, the decision of which string in the subset represents the best tree to be considered for further splitting, is based on some threshold that evaluates the former split to create a dendrogram. The process of splitting continues until the partition meets some condition that decides the continuity of the split. This condition could be testing whether the splitting could derive a node whose size is less than a predefined user number, so that, when this is met, the partition will not be split, and is considered as a cluster (leaf node) .This splitting procedure is performed for every partition in the tree, so that this binary tree is actually a dendrogram .The overall procedure is repeated many times to produce many trees, which represent a random forest which [80] is an ensemble of these dendrograms. Clustering results produced by those dendrograms, are then merged properly for further investigations.

3.4.2 Agglomerative clustering

Agglomerative clustering is a clustering family, within which the most famous one is the k-means algorithm. Starting with k as the number of clusters, k random values are used to act as centroids. for each point in the set, its distance is calculated from every centroid .Clusters grows as each centroids attracts the nearest points to it. New centroids are updated by calculating the mean value of the points attracted and, therefore new points are attracted to the nearest

centroid. In our clustering we used kmedoid clustering algorithm, where the centroids are chosen from the set points rather than randomly in the k-means algorithm. A centroid value x is chosen such that it minimizes the sum of its distances from the other points in the set In formulas: x medoid $=$ arg min $y \in \chi X|\chi|i = 1d(y, xi)$

3.4.3 Hausdorff linkage

Finding an association between miRNA and mRNA clusters is essential to find miRNAmRNA modules. Hence we found a hausdorff linkage [76], using the Hausdorff distance to rank the links of miRNA and mRNA clusters in order of their significance.

3.5 The holistic procedure

Our holistic procedure follows Guided Clustering (GC) strategy.GC is s a hybrid classification algorithm. which combines elements of both supervised" and "unsupervised" algorithms [81].As a data integration strategy, it is able to combine both experimental and clinical high-throughput data (i.e sequence and expression data). It could identify an outstanding sets of genes in experimental data, which are enriched with coherent expression in clinical data .It performs a single joint analysis for both sequence and expression datasets.

Our procedure exploits three input matrices Y $\times$ X, Y $\times$ T and X $\times$ T, which have been downloaded from NCBI, to cluster both miRNA and mRNA .The procedure consists of three phases: phases, pre-processing, individual clustering and module detection, which have been illustrated in figure 3.2. 3.

3.5.1 Data Preprocessing

During this phase we have downloaded a disease conditions data from NCBI database [6], that consists of miRNA and mRNA profiling. We used mirWalk online database to create a YxX matrix by finding miRNA targets intersecting with those available with the disease data. Y x T and X x T expression data has been normalized. They belong to a set of patients with different disease conditions.

Unlike the convention of miRNA and mRNA studies, all miRNAs and mRNAs have been considered in our study, but to avoid computation com-

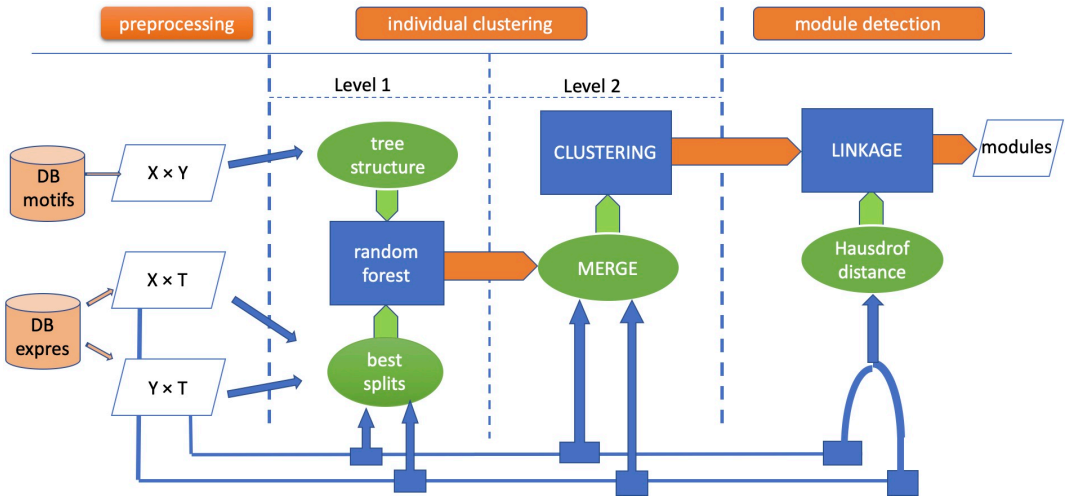


Figure 3.2: The flow chart of the proposed procedure

plexity, we used p value greater than 0.05 to release a larger number of differentially expressed items.

3.5.2 Individual Clustering

This is the first phase, in our Guided Clustering based procedure. This is divided into two tasks: i) creating a good metric for clustering, ii) exploiting the metric results to obtain final clusters.

Our procedure method is based on multivariate random forest (MRF) [82] to cluster mRNAs or miRNAs. It is based on two input matrices; YxX and YxT in case of clustering mRNA, and XxY and XxT in case of clustering miRNAs. We are focusing on implementing the guided clustering method on mRNAs, bearing in mind that it would be implemented on miRNAs clustering as well. Therefore, in case of clustering mRNAs, these two matrices are: $\mathrm{Y} \times \mathrm{X}$ miRmR and $\mathrm{Y} \times \mathrm{T}$, where Y is represents the mRNAs and X represents the miRNAs and T represents the number of conditions.

3.5.2.1 Creating a good metric

We built an MRF which results from supervised clustering using multivariate regression trees, which is a supervised machine learning algorithms.

3.5.2.2 Preparing ingredients

Our goal is to translate the two ways edges sketched in the bipartite graph in figure 3.3 into dendrograms that expresses miRNA-mRNA cotargeting relation .We are looking for Y clusters of mRNAs having common miRNAs co-targeting them with high frequency, as well as having similar expression values. We called these dendrograms trees as mentioned before. X clusters are collected in the same way.

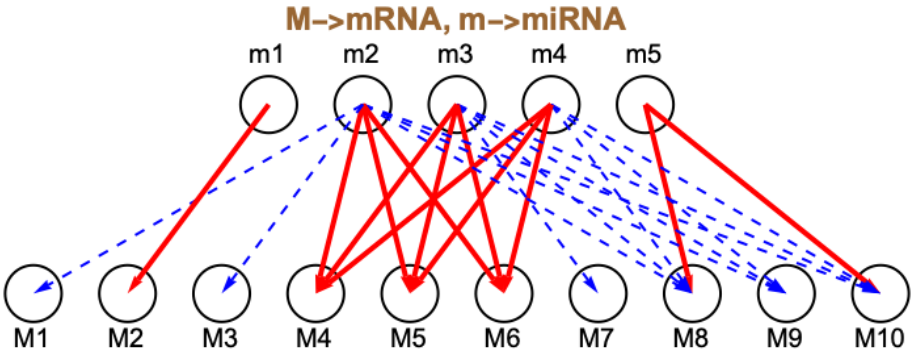


Figure 3.3: Bipartite miRNA-mRNA graphs hosting *modules* aka the biclicque emphasized in red.

Every terminal node in the final tree, represents the mRNAs that have been targeted by the same miRNAs. The miRNA column in the Y $\times$ X values are labeled 0 or 1, and they represent the splitting bit that separates mRNA rows. As a starting point, to choose the best splitting column, another homogeneity criteria is considered, and finally only one column is chosen to be the tree, that continues splitting.

3.5.2.3 Measuring homogeneity

Suppose we have a son Q in the tree contains $\{y1, \ldots, ym\}$, that is splitted into two prongs P1 and P2 containing mRNAs $\{y1, \ldots, yk\}$ and $\{yk+1, \ldots, ym\}$, respectively. Each row i in the YxT matrix represent the expression values of mRNA i. To calculate the homogeneity measure among mRNAs falling in the same prong, we firstly calculate $\omega(A)$ for set of rows A as follows:

$$\omega(A) = \sum_{i \in |A|} S_l, avg \tag{3.2}$$

where S_l,avg $= d\,(X_l, X_{\text{avg}}\,)$ such that Xl is the expression vector of mRNA I in the YxT expression matrix.X_{avg} is a row vector whose elements represent

the average of expression values of $mRNAs$ in set A, for each $t \in T$ in the $Y \times T$ matrix

The homogeniuty measure is a metric used to find the best tree, performing the best split. Suppose that b_k is the bit pattern of column k in the YxX matrix .In fact column k is the parent node which is going to be split based on bk. Change of homogeneity, that resulted from splitting the nodes based on its pattern is calculated, for $k_1, k_2, \ldots, k_h$, where h is chosed to be the cardinality order of the square root of the total number of X [82].

The change in homogeneity measure of parent node Q being split into P_{k1} and P_{k2} child nodes, based on b_k is given by :

$$\Delta\omega(Q, b_k) = \omega(Q) - \omega(P_{k_1}) - \omega(P_{k_2}) \tag{3.3}$$

These value are calculated for all b_kS, and are then sorted to choose the highest value. Therefore the split of the parent node Q, whose b_k returned the highest value is the best split. The same process of splitting is used, every time a candidate node is split, choosing h columns using sampling without replacement. Therefore column k(miRNA) appears once in the tree. This ends up with having sets of leaves, as a result of unsupervised clustering.

This procedure have been followed by Xiao and Segal [83],but as a multivariate regression tree [84], a member of the CART family [85], Y × X rows or columns are used as independent variables, and the X × T and Y × T expression columns t_i 's as dependent variables, that aids in unsupervised clustering, by providing an input for a clustering algorithm. Questier et al., exploited the t_i ' s in MRT together with CART to obtain a homogeneity measure allow both supervised and unsupervised feature selection [86].

Of course in our case, as splitting is based on bits, there is a drawback although a homogeneity measure is used. Thus, this has been treated by creating a forest, rather than just a single tree, by repeating the procedure many times based on a predefined parameters, such as the size of the which is depending on the size of candidate-bits sample. The parameter of this forest is:

1. the threshold τ to the size of the prongs
2. the number of coverage $\boldsymbol{\mu}$ of the candidate subsets
3. the number v of the trees.

3.5.2.4 Decision of Splitting a node

A bootstrap sampling without replacement is done, using τ as a sample size. With "withuot replacement" here we mean without replacing the b_k that retuned the highest change of homogeneity value. Therefore the other b_k 's are returned to the population.Algorithm 1 explains how an MRT is built.

A node is a candidate for splitting, if two conditions are satisfied:1)The size of the prongs is greater than or equal to 2τ. 2))None of the prongs have a size less than τ.

3.5.2.5 Exploiting the metric for final Clustering

As mentioned before our procedure ouputs leaves as an initial result of unsupervised clustering.These MRT leaves are merged and translated to a YxY matrix using the proximity function [87] $\delta_h(x_1, x_2)$ as follows:

$$\delta_h(x_1, x_2) = \begin{cases} 1 & \text{if } (x_1, x_2) \text{ belong to a same leaf of the } h- \text{ th tree} \\ 0 & \text{otherwise.} \end{cases} \tag{3.4}$$

Algorithm 2 describes how a Y x Y proximity matrix of an MRT is created.

This proximity matrix yielded by the function above, has a low prediction accuracy. Therefore it should be increased, and therefore a forest is used instead of a single tree, then the proximity matrix is calculated for the MRF. Another factor that improves the accuracy is to enrich each $(\mathrm{x}_1, \mathrm{x}_2)$ pair with their significant values using regression.

Algorithm 3 describes how to create Y x Y significance matrix to combine it with MRT proximity matrices.Each $(\mathrm{x}_1, \mathrm{x}_2)$ pair in the significance matrix, is combined with the actual $(\mathrm{x}_1, \mathrm{x}_2)$ in the YxY proximity matrix to be enriched before proceeding the next step of our clustering procedure. Therefore we end up with a significant proximity matrix for each MRT.

3.5.2.6 MRF Dissimilarity Proximity Matrix (PM)

As mentioned before in our procedure, the output of the multivariate regression tree is a significant $Y \times Y$ proximity matrix. MRF is an ensemble of N trees.

Algorithm 1 Generating Multivariate Random Tree

Input : Y $\times$ X$-$ $>$ map matrix, Y $\times$ T$-$ $>$ expression matrix, $\tau-$ $>$ node size, and $\mu-$ $>$ number of coverage.

1. Let initial candidates to be the mIRNAs of the map matrix if we look for trees which split mRNAs and vice versa.

2. Take a sample of size $\boldsymbol{\mu}$, without replacement from the current total number of candidates available in the dataset.

3. For each element in $\boldsymbol{\mu}$, select the corresponding bit patterns (0's and 1 's) b_kS, in the $Y \times X$.

4. Define node Q as the bit pattern b_k. Q is considered as the root node when the whole dataset are available, otherwise it is a leaf node.

5. Before starting the split check whether the size of Q is greater than or equal 2τ.

6. Split each node Q such that:

   a. Left child node (left prong) P_{k1} represents "0" elements b. Right child node P_{k2} represents the "1" elements.

7. To test the success of a split check the condition that, the size of each child node is greater than or equal to τ.

8. Add every successfully splitting candidate to the set of successful candidates.

9. For each successful candidate Q :

   a. Select the corresponding rows in the $Y \times T$ expression matrix to have an expression matrix $Q \times T$. b. Split the $Q \times T$ matrix into two matrices: $P_{k1} \times T$ and $P_{k2} \times T$. c. Calculate the homogeneity measure $\omega(Q), \omega\left(P_{k1}\right)$ and $\omega\left(P_{k2}\right)$. d. Calculate the change of homogeneity $\Delta\omega\left(Q, b_k\right)$.

10. Choose the candidate that returned the highest $\Delta\omega\left(Q, b_k\right)$ value, and assign it as the best tree.

11. Delete the candidate from the whole list of candidates.

Algorithm 2 Creating the Proximity Matrix of an MRT

Input : MRT leaves, and $Y->$ set of all elements in the leaves.
For each MRT :

1. Given a blank $Y \times Y$ matrix that has a dimension $m \times m$, where m is the total number of elements of the root node Q.
2. Insert entry $\delta_h\,(y_1, y_2)$ as an entry assigned the value 1, if y_1 and y_2 fall in the same leaf and 0 otherwise.

Algorithm 3 Create Y $\times$ Y significance matrix

Input : $Y \times T->$ expression matrix

1. Given a YxT expression matrix
2. Create a $Y \times Y$ regression matrix such that :
   (a) Each entry $\beta\,(y_1, y_2)$ is the regression coefficient of y_1 on y_2.
   (b) If $y_1 = y_2, \beta\,(y_1, y_1) = 0$, by definition.
3. Subtract each entry from 1 to get an entry $\propto (y_1, y_2)$ of $Y \times Y$ significance matrix.

Therefore, if the MRF has N trees, then the element $\psi(x_1, x_2)$ of the MRF significant proximity matrix is obtained by this formula:

$$\psi(\mathrm{x}_1, \mathrm{x}_2) = \frac{1}{v} \sum_{h=1}^{v} (\delta h(\mathrm{x}_1, \mathrm{x}_2) * \propto(\mathrm{x}_1, \mathrm{x}_2)) \tag{3.5}$$

Now, the last step to prepare the final input of a clustering algorithm such as PAM. The MRF significant proximity matrix SigProx accounts for measuring the similarity between mRNAs. On the other hand, $1-$ SigProx matrix, where 1 denotes a Y $\times$ Y matrix having all its elements equal to 1, is the matrix that represent the dissimilarity between them. This significant dissimilarity matrix (SDM) is provided as an input for k$-$ medoid clustering, that needs a predefined number of clusters which is decided based on the elbow method which will be discussed later. Algorithm 4 shows how an SDM is

built. Figure 3.4 illustrates how the SDM is created as an input to the clustering algorithm.

Algorithm 4 Calculating the MRF SDM

Input : MRT PM, (SigProx) matrix, I matrix and $v- >$ number of trees.
For each entry $\psi\left(\mathrm{y}_1, \mathrm{y}_2\right)$ in the SigProx matrix:

1. Let $\delta h(y1, \mathrm{y2})$ be an MRT PM entry.
2. let $\propto(y1, y2)$ be the significance matrix entry.
3. Compute $\psi\left(\mathrm{y}_1, \mathrm{y}_2\right) = 1 - \left(\left(\frac{1}{v}\sum_{h=1}^{v} \quad \delta h(\mathrm{y1}, \mathrm{y2})\right) * \propto(\mathrm{y1}, \mathrm{y2})\right).$

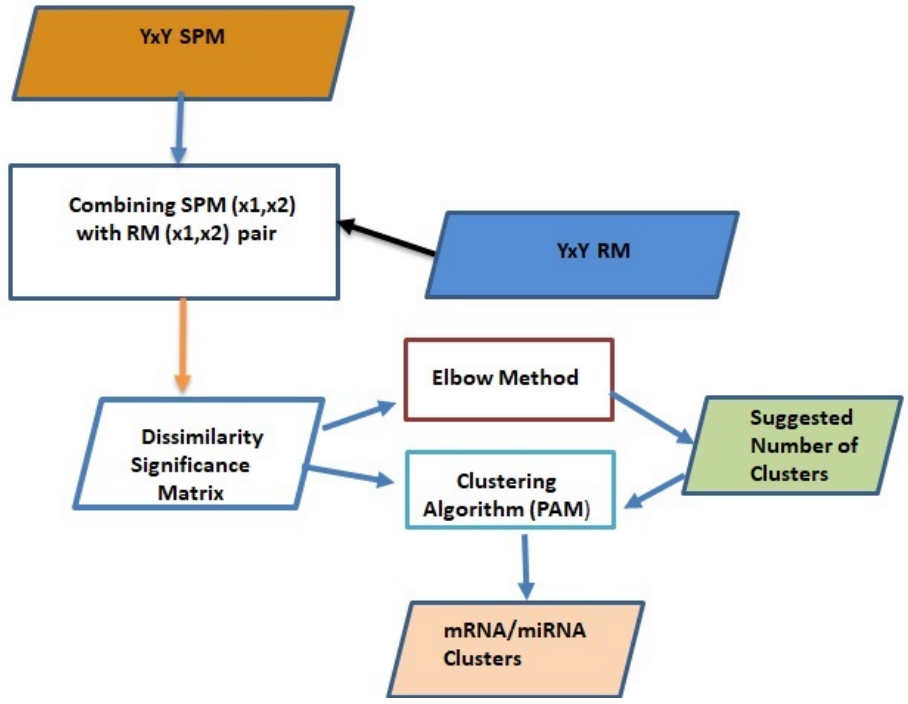


Figure 3.4: Creating a dissimilarity matrix, from the product of pair (x1,x2) from SPM and (x1,x2) from RM.

3.5.2.7 Example to illustrate the MRF method

Let us have this miRmR map matrix 3.5: Using sampling without replacement, from the X dataset (the miRNAs), a sample of size square root of the total

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}
X_1	1	1	0	1	0	1	1	0	0	1	1	1	0	1	0	1
X_2	0	0	0	1	1	0	1	0	0	0	1	1	0	1	1	1
X_3	1	1	0	0	1	0	1	1	0	0	0	1	1	0	1	0
X_4	0	1	0	1	0	0	1	1	1	1	0	0	1	0	0	1
X_5	1	0	0	0	1	0	0	1	1	1	1	1	1	0	1	1
X_6	0	1	0	0	1	0	1	0	0	1	0	0	0	0	0	1
X_7	1	1	0	1	0	1	0	1	0	0	1	0	1	1	0	1

Figure 3.5: an example of a Y x X map matrix generated from miRNA target prediction database

number in the dataset is taken. The sample generated is $[X_2, X_5, X_8, X_{12}]$. A matrix is generated, by selecting the corresponding columns, in the miRmR YxX matrix, (see figure 3.6 (a). Figure 3.6 (b) is the Y x T expression matrix. Figure 3.7 shows all the trees generated from the sample. Hence the best tree is chosen based on the highest homogenuity distance.

3.5.2.7.1 Calculating the best homogeneity distances:

$$\begin{aligned}
\mathrm{S}\,(Y_1 - Y_7) =& \mathrm{d}\,(X_{\mathrm{y1}}, X_{\mathrm{avg}})^2 + \mathrm{d}\,(X_{\mathrm{y1}}, X_{\mathrm{avg}})^2 + \mathrm{d}\,(X_{\mathrm{y1}}, X_{\mathrm{avg}})^2 + \mathrm{d}\,(X_{\mathrm{y1}}, X_{\mathrm{avg}})^2 + \mathrm{d}\,(X_{\mathrm{y1}}, X \\
& + \mathrm{d}\,(X_{\mathrm{y1}}, X_{\mathrm{avg}})^2 + \mathrm{d}\,(X_{\mathrm{y1}}, X_{\mathrm{avg}})^2 \\
\mathrm{d}\left(X_{\mathrm{y1},X_{\mathrm{avg}}}\right)^2 =& (-0.016169273 - (-0.333161531))^2 + (-0.121961247 - 0.015610902)^2 + \\
& (-0.121801284 - (-0.199665763))^2 + (-0.072215758 - (-0.331770051)) \\
=& 0.192841496 \\
\mathrm{d}\,(X_{\mathrm{y2}}, X_{\mathrm{avg}})^2 =& 0.037154382 \\
\mathrm{d}\,(X_{\mathrm{y3}}, X_{\mathrm{avg}})^2 =& 0.063581213 \\
\mathrm{d}\,(X_{\mathrm{y4}}, X_{\mathrm{avg}})^2 =& 0.040491495 \\
\mathrm{d}\,(X_{\mathrm{y5}}, X_{\mathrm{avg}})^2 =& 0.064981797
\end{aligned}$$

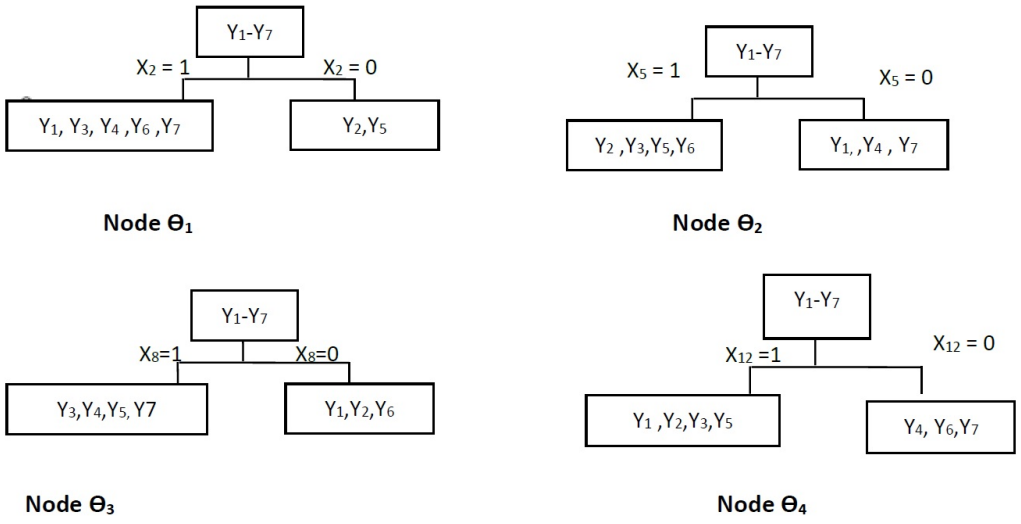


Figure 3.7: Performing the first split

Node θ_1 split $_1 = S(y1{-}y7){-}S(y1$ y3 y4 y6 y7)-S(y2 y5) $= 0.016510396$
Node θ_2 split $_2 = S(y1{-}y7){-}S(y2, y3, y5$ y6) $-S(y1$ y4 $y7) = 0.048113193$
Node θ_3 split $_3 = S(y1{-}y7){-}S(y3$ y4 y5 y7)- S(y1 y2 y6) $= \underline{0.484875685}$
Node θ_4 split $_4 = S(y1{-}y7){-}S(y1$ y2 y3 y5)-S(y4 y6 y7) $= 0.368752878$

From these calculations, it is obvious that Node Θ_3 has the highest homogeneity distance, therefore, it would be a tree in our random forest.

3.5.2.7.2 Decision of performing a further split

From the calculations above, the maximum increase in expression homogeneity is **0.484875685**, and miRNA W_8 is used for the actual split and the child nodes returned by the MRF method are (Y_1, Y_2, Y_6) and (Y_3, Y_4, Y_5, Y_7).

Once the child nodes have been obtained, the validity of performing a further split should be checked, for each child node.

Recall using sampling without replacement, using the same num-cov, another sample is taken. Given the value of the node-size, the validity of performing a further split should be checked, for each child node.

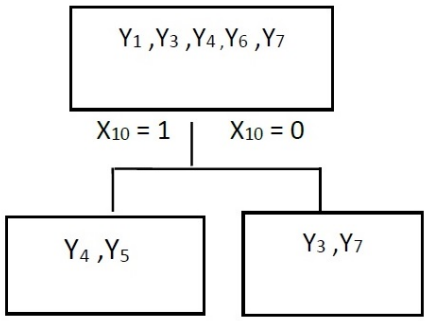
In this example, the node size is 2, therefore node (Y_3, Y_4, Y_5, Y_7), is checked for validity since the split could result in child nodes with 2 mRNAs, equal to the node size.

While checking the validity of a node for further split, a miRNA sample, is taken without replacement. Note that after performing the first split, we obtain a new map matrix with the column $\mathbf{X_8}$ omitted since it was chosen as a best tree. If the sample taken could not return a valid split, try another sample

until the split is achieved, or return no split if all the samples have been tested .Each time a sample is chosen, a new miRmR map matrix would be composed of the new chosen miRNAs, and the mRNAs at the node to be split. See figure 3.8(a).

	X_6	X_{10}	X_{13}	X_{15}
Y_3	0	0	1	1
Y_4	0	1	1	0
Y_5	0	1	1	1
Y_7	1	0	1	0

(a)



(b)

Figure 3.8: The resulting further split

The stopping criteria of the algorithm, is when no more child is valid for split, and hence the final tree is obtained 3.9.

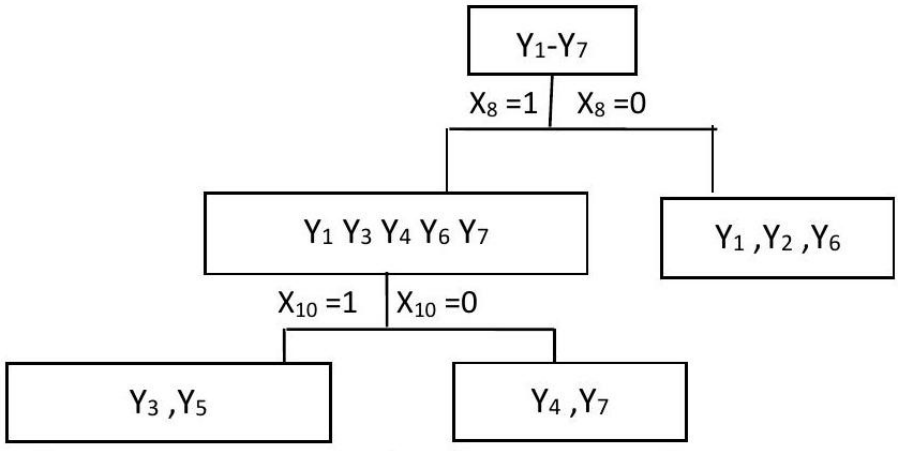
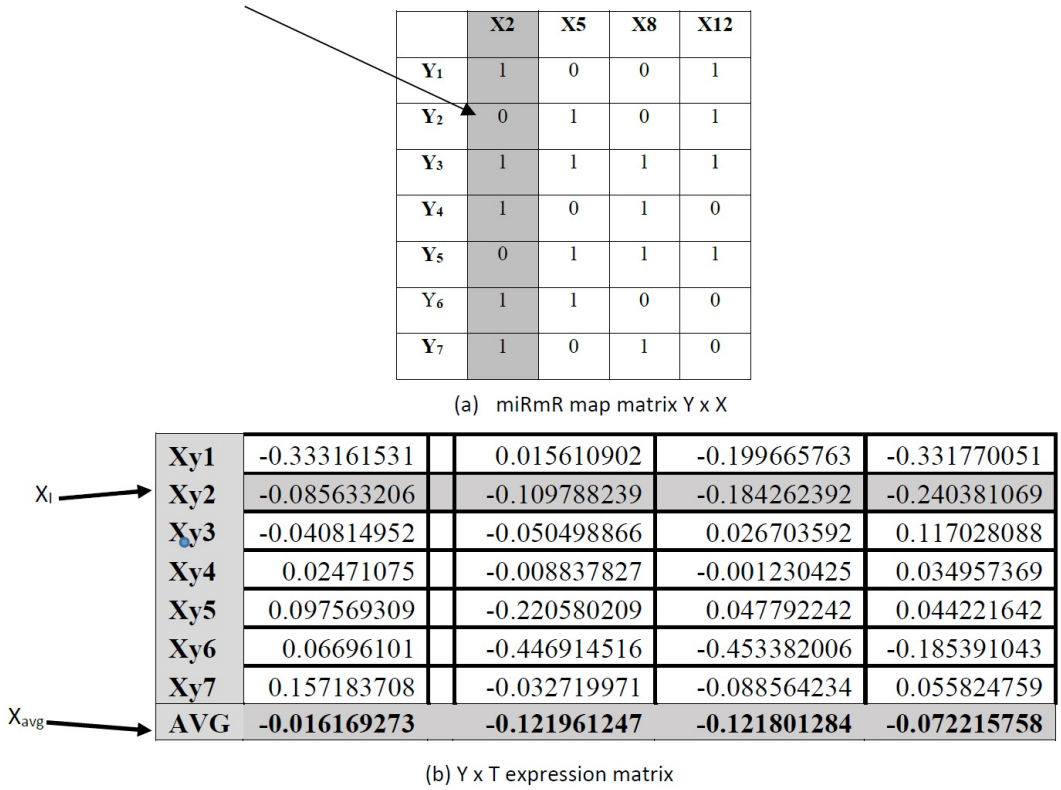


Figure 3.9: The resulting further split

3.5.2.7.3 Obtaining YxY Proximity Matrix

After obtaining the final tree, the YxY proximity matrix of the tree, is created such that if two mRNAs i and j are on the same node the entry Y_{ij} is 1 and 0 otherwise. The final matrix would look like **??**:



	X2	X5	X8	X12
Y1	1	0	0	1
Y2	0	1	0	1
Y3	1	1	1	1
Y4	1	0	1	0
Y5	0	1	1	1
Y6	1	1	0	0
Y7	1	0	1	0

(a) miRmR map matrix Y x X

Xl →	Xy1	-0.333161531	0.015610902	-0.199665763	-0.331770051
	Xy2	-0.085633206	-0.109788239	-0.184262392	-0.240381069
	Xy3	-0.040814952	-0.050498866	0.026703592	0.117028088
	Xy4	0.02471075	-0.008837827	-0.001230425	0.034957369
	Xy5	0.097569309	-0.220580209	0.047792242	0.044221642
	Xy6	0.06696101	-0.446914516	-0.453382006	-0.185391043
	Xy7	0.157183708	-0.032719971	-0.088564234	0.055824759
Xavg →	**AVG**	**-0.016169273**	**-0.121961247**	**-0.121801284**	**-0.072215758**

(b) Y x T expression matrix

Figure 3.6: The Y x X map matrix sample and Y x T expression matrix. The shaded vectors are Xl, is row l in YxT. Xavg is the vector of the average expression of column in the YxT matrix.

This YxY matrix is created for each tree, according to how many they are in the forest. To create the proximity of the MRF, each entry (i, j) is the average of all corresponding entries of the YxY proximity matrix of the trees inside the forest.

3.5.2.7.4 The Significance Matrix

Using the Yx T expression matrix, performing regression analysis using Microsoft Excel, the YxY Significance Matrix. The Regression matrix for the above example is 3.11:

	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7
Y_1	0.235262	0.000618	0.532068	0.456858	0.478794	0.212838	0.307097
Y_2	0.000618	1.86E-05	0.875682	0.572823	0.286565	0.721605	0.015608
Y_3	0.532068	0.875682	0.281373	0.068094	0.013218	0.25009	0.87987
Y_4	0.456858	0.572823	0.068094	0.380857	0.52988	0.510997	0.756358
Y_5	0.478794	0.286565	0.013218	0.52988	0.023284	0.431394	0.622282
Y_6	0.212838	0.721605	0.25009	0.510997	0.431394	5.02E-06	0.063492
Y_7	0.307097	0.015608	0.87987	0.756358	0.622282	0.063492	5.58E-06

Figure 3.11: The regression matrix

Each entry is a p-value which expresses how each two mRNAs are related to each other. To show how significant is the difference between each two mRNAs, the significance matrix is created such that each entry is equal to 1-p-value. Therefore the significance matrix is obtained as shown below 3.12:

	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7
Y_1	0.764738	0.999382	0.467932	0.543142	0.521206	0.787162	0.692903
Y_2	0.999382	0.999981	0.124318	0.427177	0.713435	0.278395	0.984392
Y_3	0.467932	0.124318	0.718627	0.931906	0.986782	0.74991	0.12013
Y_4	0.543142	0.427177	0.931906	0.619143	0.47012	0.489003	0.243642
Y_5	0.521206	0.713435	0.986782	0.47012	0.976716	0.568606	0.377718
Y_6	0.787162	0.278395	0.74991	0.489003	0.568606	0.999995	0.936508
Y_7	0.692903	0.984392	0.12013	0.243642	0.377718	0.936508	0.999994

Figure 3.12: The resulting significance matrix

3.5.2.7.5 Creating MRF Significant Proximity Matrix

The MRF significance proximity matrix would be the product of the MRF proximity matrix, and the YxY significance matrices. Therefore to create MRF significant dissimilarity matrix, each entry is subtracted from 1. This MRF SDM is an entry to the clustering algorithm.

3.5.3 Module Detection

From the second phase of the procedure, we came out with miRNA individual clusters, whose elements co-express and are convoluted by elements in mRNA clusters(see figure 3.2) miRNA and mRNA sides.)

To find a link between clusters on the two different sides, there is a need for a similarity or distance measure to find how clusters from both sides could be related. We found that the hausdroff distance we mentioned earlier is a good option, to be used as a linkage between clusters to group according to their closeness. Hence the clusters could be sorted according to their closeness, to provide a direction for candidate module analysis. Analysis starts from the closest clusters pairs, until optimum results are obtained.

Since our modules analysis include all miRNAs and mRNAs, they could be very large. The user then decides, which modules are interesting.

3.5.3.1 Some background about the DataSets

Before describing how miRNA and mRNA datasets have been preprocessed, let us have some background about them. Data has been downloaded from NCBI GSE16558 from GEO website [6]. It corresponds to a Multiple Myeloma (MM) study, where five healthy donors and sixty patients categorized as : no cytogenetic abnormality, cytogenetic abnormality t$(4; 14)$ (with or without RB deletion), cytogenetic abnormality t$(11; 14)$ (with or without RB deletion), and RB deletion as a unique cytogenetic abnormality [88]. miRNA profiling

According to the TaqMan MicroRNA Reverse Transcription Kit (PE Applied Biosystems, Foster City, CA). compliment (cDNA) was synthesized from total RNA, using hair-pin RT-primer DNA. Reverse transcriptase reactions were incubated in an Applied Biosystems GeneAmp PCR System 9700. Accurate quantification of 365 human miRNAs and three endogenous controls (RNU48, RNU48 and RNU6B) to aid in data normalization, had been enabled. Real-time PCR was performed using an Applied Biosystems 7900 HT Fast Real Time PCR Sequence Detection system. Note that each miRNA corresponded one or more probests. Therefore expression values were provided as probe expression data.

3.5.3.2 mRNA gene expression profiling

RNA labeling and microarray hybridization have been previously reported. RNA were amplified and labeled using the WT Sense Target labelling and control reagents kit (Affymetrix, Santa Clara, CA, USA), and then hybridized to Human Gene 1.0 ST Array (Affymetrix). It worths to mention that each mRNA corresponded to one or more probesets on the Affymetrix array. Washing and scanning were carried out using GeneChip System of Affymetrix. Expression value for each probe set was calculated using RMAExpress program

that uses RMA (Robust Multi-Array Average) algorithm. The total number of probes was 33297

3.5.3.3 Dataset Preprocessing

In this phase we preprocessed mRNA data, miRNA data and finally we prepared the Y x X, Y x T, X x T as input matrices. The table figure 3.1 below summarizes the preprocessing results for both mRNA and miRNA datasets.

mRNA preprocessing		miRNA preprocessing	
Name	**Total Number**	**Name**	**Total Number**
All probes	33252	miRNAs probes	365
Probes for which genes were found	27427	miRNAs	330
Genes found	25252	miRNA with found Targets	296
Targeted Genes	19137		
DE probes	9788		
DE probes for which genes found	9553		
Targeted DE genes found	7325		

Table 3.1: Multiple Myeloma miRNA and mRNA preprocesssing Result

3.5.3.4 Preprocessing the mRNA expression dataset

As mentioned before, mRNA datasets comes in probesets, therefore we actually preprocessed probeset expression dataset, where finally we obtained expression for every single mRNA. Preprocessing has been carried for two things: first, to prepair the whole mRNA expression matrix for our experiment; second, to determine the differentially expressed mRNA.

3.5.3.5 Prepairing the mRNA expression for the whole dataset

This is done in two phases. The first one included the application of RMA background correction algorithm [89], quantile normalization [90] and applying median polish algorithm to summarize the mRNA expression values. In the second one we obtained the log 2 fold-change values for MM patients with respect to healthy donors. The p-values were calculated and adjusted for multiple comparisons using the Benjamini and Hochberg (BH) [91] method

Our research considers all mRNA and miRNAs in the analysis. Here, loosely DE (LDE) probesets, were detected using adjusted p-values less than 0.1 .LDE allowed us to exploit as much data as possible, and at the same time avoiding computation complexity that could arise when the whole dataset is used.

We obtained the log2 fold-change values for 33252 probes expression values, for MM patients with respect to healthy donors. We extracted the expression values for the 9788LDE probesets. To calculate the expression values for the genes (mRNAs), the median value of the of the equivalent LDE probeset log2 fold-change values, were considered to be the log2fold-change value of mRNAs of interest.

3.5.3.6 Detecting Differentially Expressed Probeset

We used NCBI GEO2R [92] tool from the Gene Expression Omnibus to obtain the DE probes and taken p-value < 0.1.The GEO2R uses RMAEXpress tool that implement both phase one and two preprocessing steps. We found gene names for 27427 out of 33252 in the probe dataset. 9788 DE probes were detected. Furtherly, using R bioMart [93] we found 9553 gene names out of 9788 probes (98%). Each mRNA corresponds to one or more probeset, that is why the number of mRNAs is not equal to the number of probeset. Table 3.2 below shows how probes are mapped to gene names.The light shaded part shows that there could be more than one gene in a probeset, while the darker shaded part shows that a single gene could correspond to more than a probeset. Therefore to calculate the expression value for a single gene, the median expression of the probes are calculated. Finally the total number of DE genes was 7325.

3.5.3.7 Preprocessing the miRNA expression dataset

Since one miRNA corresponds to one or more probesets, we defined 330 miRNAs out of 365 miRNA probesets. Expression values have been normalized using Log2 fold change values.

Probe	Gene
7893919	RPS27A
7893919	RPS27AP1
7893919	RPS27AP16
7893919	RPS27AP3
7893973	EIF3D
7894019	EIF3D
7894125	RPL21
7894125	RPL21P11
7894125	RPL21P119
7894125	RPL21P120
7894125	RPL21P16
7894125	RPL21P28
7894125	RPL21P75
7894125	RPL21P93
7894125	SNORA27
7899134	CEP85
7899153	CEP85
7910030	DNAH14
7910047	DNAH14
7910054	DNAH14

Table 3.2: Multiple Myeloma miRNA and mRNA preprocesssing Result

Next the expression values for miRNAs having more than one probeset were calculated using the median of their expression value. All miRNAs and their expression values have been included in the analysis. We could find targets for 296 miRNAs using miRWalk database (98%).

3.5.3.8 Creating input matrices

In this phase we use a target prediction database to prepair our Y $\times$ X and Y $\times$ T matrices as an input to the supervised learning algorithm. We chose miRWalk database since it provides both predicted and validated information about miRNA-target interaction. It also enables researchers to validate new targets on the other regions of all known genes, rather than just on 30 -UTR [94]. miRNA names were converted from version 17 to version 20, to be identified by miRWalk .We were able to find 19137 targets for only 296 miRNAs, out of which 7325 were differentially expressed. Therefore, we could

obtain the plausible miRmR pairs for the 7325 LDE mRNA, and 296 miRNAs. We converted this miRmR relation into 7325 x 296 map matrix such that the rows and columns correspond to mRNAs and miRNAs, respectively. An elememnt $[i, j]$ of the map matrix takes the value of 1 if j targets i, and 0 otherwise. As mentioned before, in the preprocessing step we calculated the log2 fold change value for both mRNA and miRNA expression values. Next we filtered the expression matrix for the corresponding 7325 LDE mRNAs.

3.6 Parameters

Recall that our procedure tries to avoid computational complexities, we adopted tuning optimum values for the thresholds, upon which the successful solution depends.

For Hierarchical divisive clustering, the first level of clustering, the following parameters were tuned as follows:

1. Threshold τ. We established $\tau = 10$ to be the least node size, such that at the end each leaf size is not less than 10. This vale is always tested on the resulting nodes. That means we have al least 10 miRNAs or mRNA in a cluster

2. numcov μ. This value is calculated to be equal to the square root of the number of the binary strings bks to be split, in the YXX or X $\times$ Y matrices. In our case μ, is 85 for mRNAs and 17 for miRNAs.

3. numtrees v. Here $v = 100$ is established, to compensate for any obstruction resulting from an inadequate value of μ.

In agglomerative clustering, getting to know the value of k, the number of cluster, is essential. In fact this issue is the cause of one of the complexity problems in the overall process of clustering. Adopting the elbow method could aid overriding this problem [73].

Clustering here is based on the measure of distortion among the elements to be clustered. Distortion is measured as the sum of the square of the distances of each element from the closest centroid. Based on the elbow of the graph of figure 3.13, we identified a value for k miRNA cluster is chosen to be 8. For mRNA clustering, an equivalent elbow of a similar graph identifies k to be 20 for mRNA clustering, to suit the split of the huge mRNA dataset.

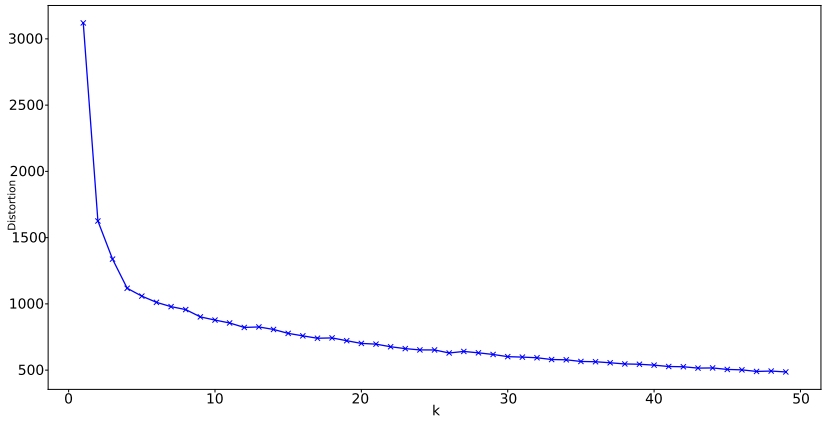


Figure 3.13: The elbow graph to identify a suitable number k for clustering miRNAs

3.7 Implementation tools

3.7.1 Programming Languages

The following programming languages have been used:

1. R language have been to write a code that uses R biomart package, find the gene names matching the probes
2. Matlab has been used during the first preprocessing steps
3. Python Programming Language has been used to generate codes for :

- downloading miRNA targets for all miRNAs from miRWALK database.
- writing parrallel computing suitable for algorithm implementation.

All codes are found in Github repository https://github.com/ghadashummo/Multivariate-Random-Forest.

3.7.2 High performance Computing

RECAS High performance Computer resources have been used to generate the parallel codes.

Chapter 4

Results

The results consists of a set of mRNA clusters and miRNA clusters, a Hausdorff distance matrix between the clusters of the two (mRNA and miRNA), a sorted list of modules and miRNA-mRNA module analysis.

4.1 mRNA clusters and miRNA clusters

Using the Elbow method, we detected the number of miRNA and mRNA clusters.Next the PAM clustering algorithm have been implemented to obtain both clusters, in such a way that the items in each cluster are close to each other on the basis of a similarity measure that produces an enhanced input to the clustering algorithm using random forest. Table 4.1 shows those miRNA and mRNA clsutering result. In fact mRNA clusters are 20 according to the Elbow method, but three clsuters were omitted since they contain just 1 element ot of 7325 mRNA.

cls_name	mi0	mi1	mi2	mi3	mi4	mi5	mi6	mi7	
cls_size	4	32	23	5	58	20	23	130	
cls_name	mr0	mr1	mr2	mr3	mr6	mr7	mr9	mr10	mr11
cls_size	7	3925	1244	8	60	326	52	34	38
cls_name	mr12	mr13	mr14	mr15	mr16	mr17	mr18	mr19	
cls_size	31	1397	14	64	28	81	5	8	

Table 4.1: miRNA clusters (*miXX*) and mRNA clusters (*mrYY*) generated by our procedure and their sizes.

It is remarkable that the number of mRNA clusters is different from that of the miRNAs, according the data provided.

4.2 Hausdorff distance matrix between the clusters of the two (mRNA and miRNA)

Table 4.2 shows the Hausdorff distance matrix between mRNA and miRNA clusters. By looking at the histograms of figure 4.1 left, it is clear that these distances have been categorized, using the value 0.030 as a threshold, into two groups, such that the distances values greater than this threshold are large distances, and the ones below are small distances.

Figure 4.1 right on the other hand shows samples miRNA-mRNA pairs, having d_l distances we defined before based on their expression values.

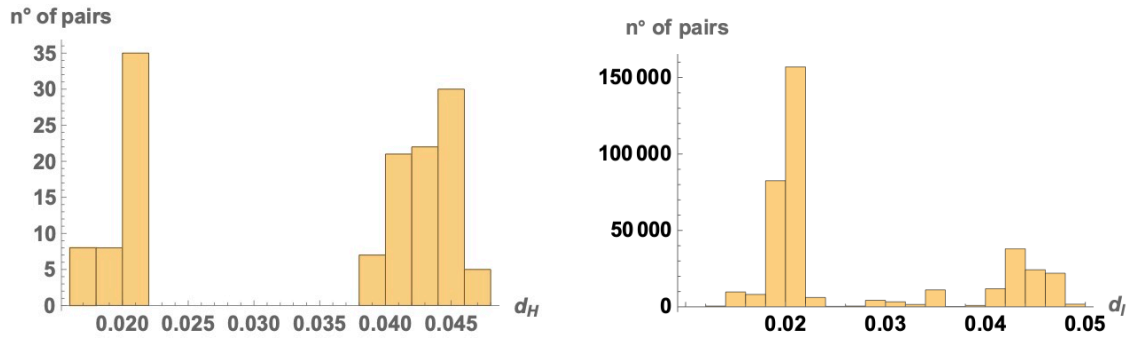


Figure 4.1: Histograms of Hausdorff distances d_H and distances d_l in our case study. The former refers to all miRNA-mRNA clusters, the latter to a downsample of the miRNA-mRNA pairs. The distances have been divided by $\sqrt{n}$, where $n = 60$ is the number of the *case* patients.

	mr0	mr1	mr2	mr3	mr6	mr7	mr9	mr10	mr11	mr12	mr13	mr14	mr15	mr16	mr17	mr18	mr19
mi0	0.3346	0.3003	0.3003	0.322	0.3216	0.3093	0.322	0.3229	0.3165	0.3274	0.3052	0.3284	0.3122	0.316	0.3178	0.3373	0.3277
mi1	0.3597	0.3356	0.3346	0.3562	0.345	0.3318	0.3438	0.3562	0.3552	0.3508	0.3344	0.3427	0.3501	0.3541	0.3438	0.356	0.3541
mi2	0.1545	0.1302	0.1329	0.1527	0.1414	0.1302	0.1465	0.1371	0.1427	0.1374	0.1302	0.1532	0.1369	0.1474	0.1388	0.1572	0.1463
mi3	0.1605	0.1605	0.1605	0.1603	0.1605	0.1605	0.1605	0.1605	0.1605	0.1605	0.1605	0.1605	0.1605	0.1604	0.1605	0.1601	0.1603
mi4	0.3605	0.3237	0.3378	0.3637	0.346	0.3407	0.3457	0.3533	0.3434	0.3465	0.3252	0.3643	0.3444	0.3519	0.3509	0.361	0.3538
mi5	0.1605	0.1605	0.1605	0.1603	0.1605	0.1605	0.1605	0.1605	0.1605	0.1605	0.1605	0.1605	0.1605	0.1604	0.1605	0.1601	0.1603
mi6	0.3277	0.3007	0.2959	0.3267	0.3125	0.3141	0.3277	0.3153	0.3143	0.3159	0.3073	0.3219	0.3221	0.3207	0.3174	0.3348	0.3193
mi7	0.3532	0.3267	0.3156	0.352	0.3401	0.3339	0.3466	0.3435	0.3499	0.3389	0.3335	0.3512	0.346	0.3404	0.3355	0.352	0.3503

Table 4.2: Hausdorff distances of the pairs miRNA-mRNA clusters

4.3 Sorting Modules

miRNA and mRNA modules hausdrof distances in table 4.2 are sorted using the value 0.25 as a threshold to define small distances marked with gray circles shown in figure 4.2 center. The choice of the threshold value depends upon how are the values could be grouped into small and large values. Using this threshold value, those distances are found to refer to miRNA 2, 3 and 5 modules. For each of the three mentioned above miRNA modules the three shortest distances are marked with red to obtain 9 closest miRNA-mRNA modules. On the other hand, miRNA 1 and 4 modules have the furthest distances, therefore nine furthest modules are chosen from their modules. Let us describe the three parts of figure 4.2 .The upper part shows the 3D shape of table 4.2.The center part is a representation of table 4.2, and the different colours are locations inside the table. Red circles are the location of the 9 closest modules out of the and gray circles, that show the location of the small distance determined by the threshold 0.25.

The hausdroff distances shown in table 4.3 could then be normalized , such that each distance value between , a miRNA cluster and an mRNA cluster in the table, is divided by the inverse of the square root of mi x mr such that mi and mr are the number of elements inside miRNA and mRNA clusters respectively . Finally the 9 blue boxes are the location of the new 9 closest distances obtained after normalization. Finally, the lower shape of figure 4.2 is for the new values generated by normalizing the values in table 4.2.

4.4 Far and Close Modules Analysis

Let us illustrates the analysis results of far and close modules in table 4.3. Each column in the table represents a module analysis result. Inside each module, some statistical features have been considered, for both miRNAs and mRNAs. These are mainly, the size of the module in terms of the number of its pairs, and how many DE and non-DE miRNAs and mRNAs in both target and non-target pairs. That is according to the Multiple Myeloma dataset , how many miRNAs/mRNAs were differentially expressed. Let us define a pair of miRNA and its mRNA target as target pair “t pair”, and a pair of a miRNA and mRNA that is not targetd by this miRNA, a non-target pair “nt pair”, according to miRNA targeting information, obtained from a target prediction database, which is miRWalk in our case. Therefore the analysis also includes the number

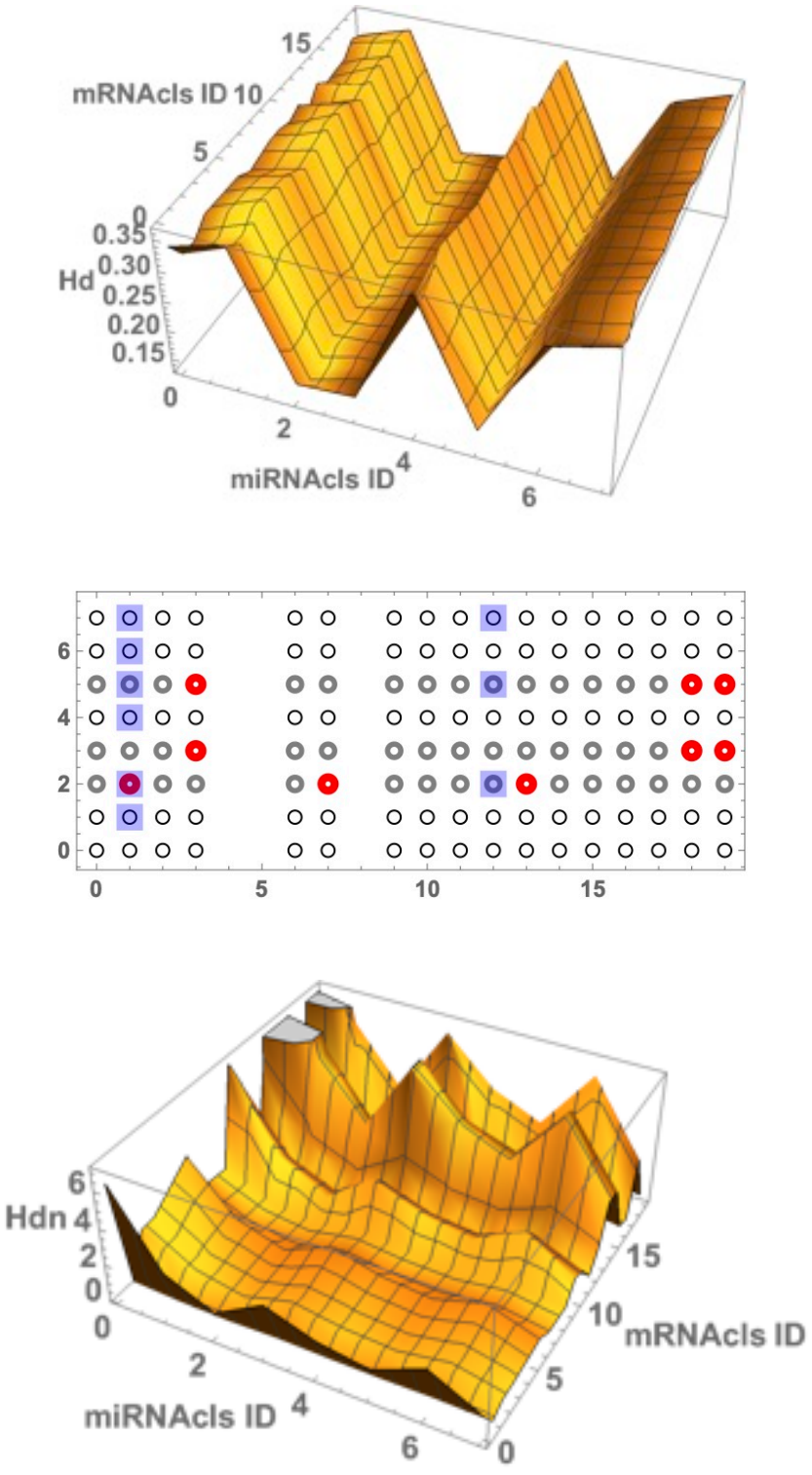


Figure 4.2: Synopses of the H_d distances. Upper: the H_d landscape over the cluster pairs. Center: Marking the closest cells of table 4.2: gray circles→ small distance cells; red circles → smallest distance within the gray cell rows; blue squares→ the 9 smallest distances according to the landscape on the bottom. Lower: same as on the upper but with reference to a *normalized* H_d.

of "t " and "nt" pairs inside the module. The first nine columns correspond to the analysis of the most nine close modules, while the other nine ones belong to the nine furthest modules. Let us furtherly analyze the close modules, since they are more interesting than the far modules for miRNA discovery. They have been classified into low, intermediate and highly populated modules in term of the number of miRNA-mRNA pairs they contain. Another interesting statistical feature is the rate of t pairs ,that tell us how many other targets have been discovered inside the module. For instance in the close module in 2,7,and the far module 4,14 module analysis, this value is 0.875,and 0.824 respectively which is a high value. On the other hand it is very low in low populated modules. These statistical features, could help biologists to find answers to many questions, and act as opening keys for future researches. For instance the diagnosing an "nt" pair could lead to a new indirect miRNA target, that may lead to a module discovery.

It is really worth to mention that new far and close modules, obtained after normalization,can be analyzed in the same manner, using the same statistical features.

4.5 Compliance with other Studies

Among the 9 closest modules we obtained after normalization, are modules $\{1,1\},\{4,1\},\{6,1\}$ and$\{7,1\}$, it is quite obvious that the most mRNA cluster involved is cluster 1. Let us find a linkage between these modules and previous disease studies, for instance COVID-19. Many studies has been conducted in relation with this disease. A study has been carried out ,in September 2020 by Erola Pairo-Castineira, et al on intensive care patients COVID-19 patients, found that low expression of IFNAR2, or high expression of TYK2, are associated with life-threatening. Another study has been conducted by Xin Gao et al. They constructed a six-gene model (comprising IFIT3, OASL, USP18, XAF1, IFI27, and EPSTI1and found that they were highly expressed in patients with COVID-19 and positively correlated with the expression of SARS-CoV-2. IFNAR2 and OASL fall on the same mRNA cluster 1. Also another study revealed that patient who did not respond to tocilizumab COVID treatment, have low level of miR-146a-5p (cluster7). Another study conducted in August 2020, on human patients with COVID-19,by Caixia Li ,et al has elucidated a total of 73/390 were differentially expressed . 35 miRNAs were upregulated and 38 miRNAs were downregulated. Out of the upregulated miRNAs we have: hsa-miR-18a-5p,hsa-miR-17-5p,hsa-miR-618, hsa-miR-16-2-

Module	{2,7}	{2,1}	{2,13}	{3,3}	{3,18}	{3,19}	{5,3}	{5,18}	{5,19}	{4,14}	{4,3}	{4,18}	{4,0}	{1,0}	{1,10}	{1,3}	{1,18}	{1,11}
No of C miRNAs	23	23	23	5	5	5	20	20	20	58	58	58	58	32	32	32	32	32
No of C mRNAs	326	3925	1397	8	5	8	8	5	8	38	8	5	7	7	34	8	5	38
No of M pairs	7498	90275	32131	40	25	40	160	100	160	2204	464	290	406	224	224	256	160	1216
Rate of t pairs	0.875	0.608	0.846	0.125	0.040	0.025	0.287	0.34	0.500	0.824	0.838	0.751	0.830	0.727	0.727	0.769	0.750	0.754
No miRNAs in t pairs	23	23	23	2	1	1	17	16	19	58	58	57	57	32	32	31	32	31
No DE miRNAs in t pairs	1	1	1	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
No DE miRNAs in nt pairs	1	1	1	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
No mRNAs in t pairs	326	3925	1397	4	1	1	8	5	8	38	8	5	7	7	7	8	5	38
No mRNAs in nt pairs	315	3925	1356	8	5	8	8	5	8	38	8	5	7	7	7	8	5	38
No DE mRNAs in t pairs	178	1948	737	3	1	0	4	5	5	19	4	5	3	3	3	4	5	19
No DE mRNAs in nt pairs	171	1948	717	4	5	5	4	5	5	19	4	5	3	3	3	4	5	19

Table 4.3: Main statistical features of extremal modules. C miRNAs → miRNAs inside the cluster, C mRNAs → mRNAs inside the cluster; M pairs → pairs inside the module, t pairs → targeted pairs inside the module; DE miRNA/mRNA → differentially expressed miRNA/mRNA.

Module {7,1}					
	hsa-miR-146b-5p	hsa-miR-20a-5p	hsa-miR-21-5p	hsa-miR-30c-5p	hsa-miR-627-5p
IFNAR2	nt	nt	t	t	nt
OASL	nt	t	nt	nt	nt

Module {6,1}	
	hsa-miR-618
IFNAR2	t
OASL	t

Module {4,1}		
	hsa-miR-17-5p	hsa-miR-18a-5p
IFNAR2	nt	t
OASL	t	t

Table 4.4: COVID-19 Modules in compliance with the closest modules.Here the valuew of IFNAR2 and OASL inside each module indicates either direct targeting ”t” or indirect targeting ”nt”.

5p, hsa-miR-30c-5p, hsa-miR-21-5p, hsa-miR-20a-5p, hsa-miR-146b-5p and out of the downregulated were hsa-miR-627-5p,hsa-miR-183-5p . Table 4.4 above shows the modules that contains these miRNAs and mRNAs related to COVID.

Cancer disease is a very complicated disease that needs continueous and deep invesitigation, either to prevent it, or to treat it with drugs with least side effects.E2F-RB pathway as mentioned before plays a basic role in cancer development. That is why miRNA related to this pathway are the focus in cancer researches. A study conducted by a Conkrite et al.shows its relation with miR-17 92 cluster [95]. Table 4.5 shows some modules that encapsulate E2f-RB pathway and miRNAs.

Module{7,1}			
mRNA	hsa-miR-19a-3p	hsa-miR-19b-3p	hsa-miR-20a-5p
E2F1	nt	nt	nt
RB1	nt	nt	nt

Module {1,1}	
mRNA	hsa-miR-92a-3p
E2F1	t
RB1	t

Module {4,1}			
mRNA	hsa-miR-17-3p	hsa-miR-17-5p	hsa-miR-18a-5p
E2F1	t	t	t
RB1	t	t	t

Table 4.5: E2F-RB pathway modules

Chapter 5

Discussion

Recall the clique of figure 2.3, finding modules means finding a method of discovering regulatory relations between group of miRNAs and groups of mRNAs. To evaluate and interpret the results using some statistics, we obtained, starting from table 4.3, the hausdroff distance matrix. We have then concluded nine red circled modules (see figure 4.2 center) to represent the closest modules, together with 9 modules with the highest hausdroff distance(furthest modules): the closer the modules the stronger regulatory relation they have.

In earlier stages, we exploited both expression and binding motif information to explore these modules. We used an associations of both Y $\times$ X map matrix reflecting binding information, together with expression matrix, to separately group miRNAs and mRNAs into clusters.

Now by finding both miRNA and mRNA clusters, we associated their clusters based on their hausdroff distances, that have been computed based on miRNA-mRNA pairs inside each miRNA and mRNA cluster.

To indicate how our procedure is fruitful, let us look for representative for more crowded close and far modules. Starting with close modules in figure 4.2 center, let us $(\{2,7\},\{2,1\},\{2,13\})$ among the closest modules and $(\{4,14\},\{4,3\},\{4,0\},\{1,11\})$ among the furthest ones. Two features are considered to characterize our modules:highly populated module and low populated module. The first feature permits the discovery of new direct and indirect targets according to Plotnikova et Al [96], or the discovery of binding mechanism such the combinatorial binding [97], that have not yet come out to light. Some mRNA inside such modules, are not considered as targets to the miRNAs according to available miRNA target prediction databases, especially those that are experimentally validated. On the other hand, modules that have very low population, such as $(\{3,3\},\{3,18\},\{3,19\})$, discovers miRNA-

mRNA regulatory relations that have been neglected. Intermediate modules such as $(\{5, 3\}, \{5, 18\}, \{5,\ 19\})$ are analyzed using the same features as that of the high and low populated modules. In compliance with these findings, let us consider the relation of mRNA GEMIN5 and EXOC8, with the miRNAs hsa-miR-584d in the closest modules $\{3, 19\}$, hsa-miR-99a-5p in the closest module $\{5, 19\}$ and hsa-miR-145-5p in the twelfth furthest modules $\{4, 19\}$, respectively.

Although those genes were targeted by miRNA hsa-miR-145-5p in the furthest module, which is not differentially expressed, a biological relation has been found for them with miRNA hsa-miR-584d and miR-99a-5p in the closest modules.

During the study conducted about CD2+ T lymphocytes GEMEN5 was significantly differentially overexpressed. Meanwhile hsa-miR-584d was differentially upregulated, unlike hsa-miR-99a-5p that was downregulated [98].

A similar study in breast cancer, miR-99a was significantly downregulated, but EXOC8 was significantly upregulated [99]. mRNA disease database such as Malacards [100], and HMDD miRNA Disease associated databases [101]have been used to find disease associated with both GEMEN5 and EXOC8 genes, and the three miRNAs. Breast cancer was found to be a common disease among all of them. Table 4.4 shows how miRNAs and mRNAs related to COVID-19,belong to the closest modules created after normalization.The choice of the closest modules gives a chance to focus on specific modules that are closely related to a disease, and paves the way for more investigation of other miRNAs and mRNA and might also be related to that disease ,or may be other diseases such as cancer. CYP46A1 is a pharmacogene that falls in mRNA cluster 1, so this might also shed some light on drug discovery to treat that diseases whose related genes are found in the same cluster.

It worths to mention the E2F-RB pathway in cancer, and the role they play in cell developement. E2F and RB1 were found also in cluster 1. It is also very important to know the miRNAs that target them during their function and cancer development. Table 4.5 shows the modules that contains E2F-RB and some miRNAs targeting them. They are among the closest modules. This also open some biological research questions to explore more miRNA-mRNA direct and indirect relation.

Our research is quantitative, since it is based on a tight processing of certified data through advanced statistical algorithms. In fact, to evaluate the meaningfulness of a module, we provide a quantitative tool, rather than proposing a statistical test to decide whether a module is meaningful or not.

Still better, this research proposes a quantitive tool to prioritize the researches of biologists, so that they can discover meaningful modules.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

Adopting a holistic strategy, in this research, paves the way to introduce a procedure to discover new miRNA-mRNA interactions, that would be omitted or neglected in the common literature. It starts with following the conventional procedure followed in most of miRNA-mRNA researches, bearing in mind the fact that miRNA-mRNA relation is not restricted for analysis from just one experiment. That is why this procedure uses tools to derive new metrics so as to be implemented using HPC utilities.

The procedure derives from a general strategy and uses standard tools that are properly devised in order to exploit new metrics to be implemented on HPC utilities. Basically those utilities are applied for standard biological databases.

To give more strength to the procedure, biological intervention is needed for further elaborations. The discovered interactions that refer to diseases, has opened a new biological issue to study such diseases in a more wide spectrum of pathologies, in the sense that miRNAmRNA interactions play a relevant role in disease etiology.

6.2 Future work

The procedure being followed could be improved. Firstly, the procedure being applied on greater datasets from different experiments, would enrich the discovered modules, and allow more discovery.

Secondly, would be very beneficial to follow a standard and more reliable strategy to maximize the number of enriched modules is needed, to avoid experimental validation, which is very expensive . The standards could also include the size and the method the sample is taken, to avoid any undesired biasedness in the analysis of the results.

The strategy could be applied, using other datamining classification, and clustering algorithms to be included in the ensemble, as well as other statistical, and mathematical metrics used in the miRNA-mRNA module detection.

Bibliography

[1] Rishav Ray and Priyanka Pandey. Surveying computational algorithms for identification of miRNA–mRNA regulatory modules. *Nucleus (India)*, 60(2):165–174, 2017.

[2] Harpreet K. Saini, Anton J. Enright, and Sam Griffiths-Jones. Annotation of mammalian primary microRNAs. *BMC Genomics*, 9:1–19, 2008.

[3] X. Pan, A. Wenzel, .LJ. Jensen, and J. Gorodkin. Genome-wide identification of clusters of predicted microrna binding sites as microrna sponge candidates. *PLoS One.*, 2018.

[4] Sarah M Peterson, Jeffrey A Thompson, Melanie L Ufkin, Pradeep Sathyanarayana, Lucy Liaw, and Clare Bates Congdon. Common features of microrna target prediction tools. *Frontiers in genetics*, 5:23, 2014.

[5] Most Mauluda Akhtar, Luigina Micolucci, Md Soriful Islam, Fabiola Olivieri, and Antonio Domenico Procopio. Bioinformatic tools for microrna dissection. *Nucleic acids research*, 44(1):24–44, 2016.

[6] GEO Accession viewer.

[7] Victor Ambros. The functions of animal microRNAs. *Nature*, 431(7006):350–355, 2004.

[8] David P. Bartel. MicroRNAs: Target Recognition and Regulatory Functions. *Cell*, 136(2):215–233, 2009.

[9] Alexander Hüttenhofer and Jörg Vogel. Experimental approaches to identify non-coding RNAs. *Nucleic Acids Research*, 34(2):635–646, 2006.

[10] H-W Hwang and J T Mendell. MicroRNAs in cell proliferation, cell death, and tumorigenesis. *British Journal of Cancer*, 94:776–780, 2006.

[11] Vivek Jayaswal, Mark Lutherborrow, David D.F. Ma, and Yee H. Yang. Identification of microRNA-mRNA modules using microarray data. *BMC Genomics*, 12, 2011.

[12] M. Jovanovic and M. O. Hengartner. miRNAs and apoptosis: RNAs to die for. *Oncogene*, 25(46):6176–6187, 2006.

[13] Wigard P. Kloosterman and Ronald H.A. Plasterk. The Diverse Functions of MicroRNAs in Animal Development and Disease. *Developmental Cell*, 11(4):441–450, 2006.

[14] S. M. Masud Karim, Lin Liu, Thuc Duy Le, and Jiuyong Li. Identification of miRNA-mRNA regulatory modules by exploring collective group relationships. *BMC Genomics*, 17(1), 2016.

[15] Sushmita Paul, Petra Lakatos, Arndt Hartmann, Regine Schneider-Stock, and Julio Vera. Identification of miRNA-mRNA Modules in Colorectal Cancer Using Rough Hypercuboid Based Supervised Clustering. *Scientific Reports*, 7, 2017.

[16] Ramanjulu Sunkar and Jian Kang Zhu. Novel and stress regulated microRNAs and other small RNAs from Arabidopsis w inside box sign. *Plant Cell*, 16(8):2001–2019, 2004.

[17] Hui Wang, Honghong Wang, Xinrui Duan, Chenghui Liu, and Zhengping Li. Digital quantitative analysis of microRNA in single cell based on ligation-depended polymerase colony (Polony). *Biosensors and Bioelectronics*, 95:146–151, 2017.

[18] Anjan K Pradhan, Luni Emdad, Swadesh K Das, Devanand Sarkar, and Paul B Fisher. Chapter Two - The Enigma of miRNA Regulation in Cancer. In Carlo M Croce and Paul B Fisher, editors, *miRNA and Cancer*, volume 135 of *Advances in Cancer Research*, pages 25–52. Academic Press, 2017.

[19] S. Swarbrick, N. Wragg, S. Ghosh, and Alexandra Stolzing. Systematic Review of miRNA as Biomarkers in Alzheimer's Disease. *Molecular Neurobiology*, 56(9):6156–6167, 2019.

[20] Bassam Abdul Rasool Hassan, Zuraidah Binti Mohd Yusoff, Mohamed Azmi Hassali Othman, Saad Bin, Additional information is available at the end of the Chapter, and Http://dx.doi.org/10.5772/55358. We are IntechOpen , the world ' s leading publisher of Open Access books Built by scientists , for scientists TOP 1 *Intech*, page 13, 2012.

[21] Phases of the cell cycle (article) | Khan Academy.

[22] What is synaptic plasticity? - Queensland Brain Institute - University of Queensland.

[23] Yong Huang, Xing Jia Shen, Quan Zou, Sheng Peng Wang, Shun Ming Tang, and Guo Zheng Zhang. Biological functions of microRNAs: A review. *Journal of Physiology and Biochemistry*, 67(1):129–139, 2011.

[24] Misa Tokorodani, Hirona Ichikawa, Katsutoshi Yuasa, Tetsuyuki Takahashi, and Takao Hijikata. SV40 microRNA miR-S1-3p downregulates the expression of t antigens to control viral DNA replication, and tnfα and IL-17F expression. *Biological and Pharmaceutical Bulletin*, 43(11):1715–1728, 2020.

[25] Definition of cytotoxic T cell - NCI Dictionary of Cancer Terms - National Cancer Institute.

[26] What is Angiogenesis?

[27] Baohong Zhang, Xiaoping Pan, George P. Cobb, and Todd A. Anderson. microRNAs as oncogenes and tumor suppressors. *Developmental Biology*, 302(1):1–12, 2007.

[28] PTEN (gene) - Wikipedia.

[29] Akiyo Yoshida, Shunsuke Kitajima, Fengkai Li, Chaoyang Cheng, Yujiro Takegami, Susumu Kohno, Yuan Song Wan, Naoyuki Hayashi, Hayato Muranaka, Yuuki Nishimoto, Naoko Nagatani, Takumi Nishiuchi, Tran C. Thai, Sawako Suzuki, Shinji Nakao, Tomoaki Tanaka, Osamu Hirose, David A. Barbie, and Chiaki Takahashi. MicroRNA-140 mediates RB tumor suppressor function to control stem cell-like activity through interleukin-6. *Oncotarget*, 8(8):13872–13885, 2017.

[30] Transcription factor - Wikipedia.

[31] E2F - Wikipedia.

[32] Retinoblastoma protein - Wikipedia.

[33] M.M. Wilson, M.S. , Metink-Kane. □□□□□NIH Public Access. *Bone*, 23(1):1–7, 2012.

[34] The retinoblastoma gene: role in cell cycle control and cell differentiation; The retinoblastoma gene: role in cell cycle control and cell differentiation. Technical report.

[35] Penghui Li, Hongxin Fei, Lihong Wang, Huiyu Xu, Haiyan Zhang, and Lihong Zheng. Pdcd5 regulates cell proliferation, cell cycle progression and apoptosis. *Oncology Letters*, 15(1):1177–1183, 2018.

[36] Patrick Viatour and Julien Sage. Newly identified aspects of tumor suppression by RB. *DMM Disease Models and Mechanisms*, 4(5):581–585, 2011.

[37] Paola Indovina, Francesca Pentimalli, Nadia Casini, Immacolata Vocca, and Antonio Giordano. RB1 dual role in proliferation and apoptosis: Cell fate control and implications for cancer therapy. *Oncotarget*, 6(20):17873–17890, 2015.

[38] Erick J Morris, Jun-Yuan Ji, Fajun Yang, Luisa Di Stefano, Anabel Herr, Nam-Sung Moon, Eun-Jeong Kwon, Kevin M Haigis, Anders M Näär, and Nicholas J Dyson. E2F1 represses β-catenin transcription and is antagonized by both pRB and CDK8.

[39] T. J. Collard, B. C. Urban, H. A. Patsos, A. Hague, P. A. Townsend, C. Paraskeva, and A. C. Williams. The retinoblastoma protein (Rb) as an anti-apoptotic factor: Expression of Rb is required for the antiapoptotic function of BAG-1 protein in colorectal tumour cells. *Cell Death and Disease*, 3(10):e408–9, 2012.

[40] Sungroh Yoon and Giovanni De Micheli. Prediction of regulatory modules comprising microRNAs and target genes. *Bioinformatics*, 21(suppl$_2$) : $ii93 - -ii100, 092005$.

[41] Kenneth Bryan, Marta Terrile, Isabella M. Bray, Raquel Domingo-Fernandéz, Karen M. Watters, Jan Koster, Rogier Versteeg, and Raymond L. Stallings. Discovery and visualization of miRNA–mRNA

functional modules within integrated data using bicluster analysis. *Nucleic Acids Research*, 42(3):e17–e17, 12 2013.

[42] L Masud, Kand Liu, TD Le, and J Li. Identification of mirna-mrna regulatory modules by exploring collective group relationships. *BMC Genomics.*, 11(17), Jan 2016.

[43] Milad Mokhtaridoost and Mehmet Gönen. An efficient framework to identify key miRNA–mRNA regulatory modules in cancer. *Bioinformatics*, 36(Supplement$_2$) : $i592 - -i600, 122020$.

[44] Sushmita Paul and . Madhumita. Rfcm3: Computational method for identification of mirna-mrna regulatory modules in cervical cancer. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 17:1729–1740, 2020.

[45] Dong Wang, Juan Wang, Ming Lu, Fei Song, and Qinghua Cui. Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. *Bioinformatics*, 26(13):1644–1650, 05 2010.

[46] Je-Gun Joung and Zhangjun Fei. Identification of microRNA regulatory modules in Arabidopsis via a probabilistic graphical model. *Bioinformatics*, 25(3):387–393, 12 2008.

[47] Je-gun Joung, Kyu-baek Hwang, Jin-wu Nam, Soo-jin Kim, and Byoung-tak Zhang. Discovery of microRNA – mRNA modules via population-based probabilistic learning. 23(9):1141–1147, 2007.

[48] R A Kaslow. The multicenter aids cohort study: rationale, organization, and selected characteristics of the participants. *American journal of epidemiology*, 2(162):161–169, 1987.

[49] Hanley. J.A., A. Negassa, M.D. Edwardes, and .J.E. Forrester. Statistical analysis of correlated data using generalized estimating equations: an orientation. *Am J Epidemiol.*, 4(157):364–375, 2003.

[50] A.S. Coletti, P.J. Heagerty, A.R. Sheon, M. Gross, B.A. Koblin, D.S. Metzger, and G.R. Seage. Randomized, controlled evaluation of a proto- type informed consent process for hiv vaccine efficacy trials. *Journal of Acquired Immune Deficiency Syndrome*, (32):161–169, 2003.

[51] Robert A. McLean, William L. Sanders, and Walter W. Stroup. A unified approach to mixed linear models. *The American Statistician*, 45(1):54–64, 1991.

[52] KUNG-YEE LIANG and SCOTT L. ZEGER. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 04 1986.

[53] A. Clifford Cohen. *Truncated and Censored Samples*. CRC Press, Boca Raton, 1991.

[54] David G. Kleinbaum and Mitchel Klein. *Introduction to Survival Analysis*, pages 1–54. Springer New York, New York, NY, 2012.

[55] M. Denui, O. Purcaru, and I. V. Keilegom. Patterns of neuronal migration in the embryonic cortex. *Journal of Actuarial Practice*, 13:5–329, 2006.

[56] D. Bernoulli. Essai dÕune nouvelle analyse de la mortalit caus e par la petite v role, et des avantages de lÕinoculation pour la pr venir. *Histoire de lÕAcad., Roy. Sci.(Paris)*, pages 1–45, 1760.

[57] A. Lee. Table of the gaussian ÒtailÓ functions; when the ÒtailÓ is larger than the body. *Biometrika*, 23(10):208–214, 1914.

[58] A. Fisher. Estimation of the parameters in a truncated normal distribution. *Mathematical tables*, (1):815– 852, 1931.

[59] M. H. Laxman and C. D. Ram. Estimation of the parameters in a truncated normal distribution. *Published online*, 2(162):4177–4195, 2007.

[60] Constantinos Daskalakis, Themis Gouleakis, Christos Tzamos, and Manolis Zampetakis. Efficient statistics, in high dimensions, from truncated samples. In Mikkel Thorup, editor, *59th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2018, Paris, France, October 7-9, 2018*, pages 639–649. IEEE Computer Society, 2018.

[61] Vasilis Kontonis, Christos Tzamos, and Manolis Zampetakis. Efficient truncated statistics with unknown truncation. In David Zuckerman, editor, *60th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2019, Baltimore, Maryland, USA, November 9-12, 2019*, pages 1578–1595. IEEE Computer Society, 2019.

[62] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, volume 86, pages 2278–2324, 1998.

[63] Mark Segal and Yuanyuan Xiao. Multivariate random forests. *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, 1(1):80–87, 2011.

[64] A Liaw and M Wiener. Classification and Regression by randomForest. *R News*, 2(3):18–22, 2002.

[65] LEO BREIMAN. Random Forest. 2001.

[66] Tu Minh Phuong, Doheon Lee, and Kwang Hyung Lee. Regression trees for regulatory element identification. *Bioinformatics (Oxford, England)*, 20(5):750–757, mar 2004.

[67] Malik Yousef, Gokhan Goy, Ramkrishna Mitra, Christine M Eischen, Amhar Jabeer, and Burcu Bakir-Gungor. mircorrnet: machine learning-based integration of mirna and mrna expression profiles, combined with feature grouping and ranking. *PeerJ*, 9:e11458, 2021.

[68] Malik Yousef, Segun Jung, Louise C. Showe, and Michael K. Showe. Recursive Cluster Elimination (RCE) for classification and feature selection from gene expression data. *BMC Bioinformatics*, 8:1–12, 2007.

[69] Malik Yousef, Loai Abdallah, and Jens Allmer. maTE: discovering expressed interactions between microRNAs and their targets. *Bioinformatics (Oxford, England)*, 35:4020–4028, 2019.

[70] Gianvito Pio, Michelangelo Ceci, Domenica D'Elia, Corrado Loglisci, and Donato Malerba. A novel biclustering algorithm for the discovery of meaningful biological correlations between microRNAs and their target genes. *BMC Bioinformatics*, 14(SUPPL7), 2013.

[71] Paulína Pidíkova, Richard Reis, and Iveta Herichova. mirna clusters with down-regulated expression in human colorectal cancer and their regulation. *International journal of molecular sciences*, 21(13):4633, 2020.

[72] Jacob O'Brien, Heyam Hayder, Yara Zayed, and Chun Peng. Overview of microrna biogenesis, mechanisms of actions, and circulation. *Frontiers in endocrinology*, 9:402, 2018.

[73] Purnima Bholowalia and Arvind Kumar. EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN. *International Journal of Computer Applications*, 105(9):975–8887, 2014.

[74] Trupti M Kodinariya and Prashant R Makwana. Review on determining of cluster in K-means. *International Journal of Advance Research in Computer Science and Management Studies*, 1(6):90–95, 2013.

[75] D. P. Huttenlocher, W. J. Rucklidge, and G. A. Klanderman. Comparing images using the Hausdorff distance under translation, 1992.

[76] Nicolas Basalto, Roberto Bellotti, Francesco De Carlo, Paolo Facchi, Ester Pantaleo, and Saverio Pascazio. Hausdorff clustering, 2008.

[77] Hausdorff distance - Wikipedia, 2021.

[78] Ghada Shommo and Bruno Apolloni. A holistic mirna-mrna module discovery. *Non-coding RNA Research*, 6(4):159–166, 2021.

[79] Xiao-Tong Yuan, Bao-Gang Hu, and Ran He. Agglomerative mean-shift clustering. *IEEE Transactions on Knowledge and Data Engineering*, 24(2):209–219, 2010.

[80] Yanjun Qi. Random forest for bioinformatics. In *Ensemble machine learning*, pages 307–323. Springer, 2012.

[81] Matthias Maneck, Alexandra Schrader, Dieter Kube, and Rainer Spang. Genomic data integration using guided clustering. *Bioinformatics (Oxford, England)*, 27(16):2231–2238, aug 2011.

[82] Feifei Xiao, Zhixiang Zuo, Guoshuai Cai, Shuli Kang, Xiaolian Gao, and Tongbin Li. miRecords: An integrated resource for microRNA-target interactions. *Nucleic Acids Research*, 37(SUPPL. 1):105–110, 2009.

[83] Yuanyuan Xiao and Mark R. Segal. Identification of yeast transcriptional regulation networks using multivariate random forests. *PLoS Comput. Biol.*, 5(6), 2009.

[84] Glenn De'ath. Multivariate regression trees: a new technique for modeling species-environment relationships. *Ecology*, 83(4):1105—1117, 2002.

[85] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth and Brooks, 1984.

[86] Frederik Questier, Raf Put, Danny Coomans, Beata Walczak, and Yvan Vander Heyden. The use of cart and multivariate regression trees for supervised and unsupervised feature selection. *Chemometrics and Intelligent Laboratory Systems*, 76:45–54, 2005.

[87] LEO BREIMAN. Random Forest. 2001.

[88] Norma Carmen Gutiérrez, María Eugenia Sarasquete, I Misiewicz-Krzeminska, M Delgado, J De Las Rivas, FV Ticona, E Ferminan, P Martin-Jimenez, C Chillon, A Risueno, et al. Deregulation of microrna expression in the different genetic subtypes of multiple myeloma and correlation with gene expression profiling. *Leukemia*, 24(3):629–637, 2010.

[89] Rafael A. Irizarry, Bridget Hobbs, Francois Collin, Yasmin D. Beazer-Barclay, Kristen J. Antonellis, Uwe Scherf, and Terence P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Selected Works of Terry Speed*, pages 601–616, 2003.

[90] B. M. Bolstad, R. A. Irizarry, M. Åstrand, and T. P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, 2003.

[91] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[92] GEO2R - GEO - NCBI.

[93] Steffen Durinck, Paul T Spellman, Ewan Birney, and Wolfgang Huber. Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomart. *Nature protocols*, 4(8):1184–1191, 2009.

[94] Harsh Dweep, Carsten Sticht, Priyanka Pandey, and Norbert Gretz. MiRWalk - Database: Prediction of possible miRNA binding sites by ” walking” the genes of three genomes. *Journal of Biomedical Informatics*, 44(5):839–847, 2011.

[95] Karina Conkrite, Maggie Sundby, Shizuo Mukai, J. Michael Thomson, David Mu, Scott M. Hammond, and David MacPherson. Mir-17∼92 cooperates with

RB pathway mutations to promote retinoblastoma. *Genes and Development*, 25(16):1734–1745, 2011.

[96] Olga Plotnikova, Ancha Baranova, and Mikhail Skoblov. Comprehensive analysis of human microrna-mrna interactome. *Frontiers in Genetics*, 10:933, 2019.

[97] R. Murugan. Theory on the mechanisms of combinatorial binding of transcription factors with dna. *arXiv: Subcellular Processes*, 2016.

[98] Yevgeniy A. Grigoryev, Sunil M. Kurian, Traver Hart, Aleksey A. Nakorchevsky, Caifu Chen, Daniel Campbell, Steven R. Head, John R. Yates, and Daniel R. Salomon. Microrna regulation of molecular networks mapped by global microrna, mrna, and protein expression in activated t lymphocytes. *The Journal of Immunology*, 187(5):2233–2243, 2011.

[99] Xinghua Long, Yu Shi, Peng Ye, Juan Guo, Qian Zhou, and Yueting Tang. Microrna-99a suppresses breast cancer progression by targeting fgfr3. *Frontiers in Oncology*, 9:1473, 2020.

[100] Noa Rappaport, Michal Twik, Inbar Plaschkes, Ron Nudel, Tsippi □Stein, Jacob Levitt, Moran Gershoni, C. Paul Morrey, Marilyn Safran, and Doron Lancet. MalaCards: an amalgamated human disease compendium with diverse clinical and genetic annotation and structured search. *Nucleic Acids Research*, 45(D1):D877–D887, 11 2016.

[101] Zhou Huang, Jiangcheng Shi, Yuanxu Gao, Chunmei Cui, Shan Zhang, Jianwei Li, Yuan Zhou, and Qinghua Cui. HMDD v3.0: a database for experimentally supported human microRNA□disease associations. *Nucleic Acids Research*, 47(D1):D1013–D1017, 10 2018.