

**FINDING ASSOCIATIONS BETWEEN
MICRORNAS/PROTEIN DOMAINS AND
DISEASES: A GRAPH-BASED APPROACH**

BY

AHMED MOHAMED MAHMOUD ELMOSELHY

A Thesis Presented to the
DEANSHIP OF GRADUATE STUDIES

KING FAHD UNIVERSITY OF PETROLEUM & MINERALS

DHAHRAN, SAUDI ARABIA

In Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

In

COMPUTER SCIENCE

MAY 2022

KING FAHD UNIVERSITY OF PETROLEUM & MINERALS

DHAHRAN- 31261, SAUDI ARABIA

DEANSHIP OF GRADUATE STUDIES

This thesis, written by **Ahmed Mohamed Mahmoud Elmoselhy** under the direction of his thesis advisor and approved by his thesis committee, has been presented and accepted by the Dean of Graduate Studies, in partial fulfillment of the requirements for the degree of **MASTER OF SCIENCE IN COMPUTER SCIENCE**.



Dr. Hamoud Aljamaan
Department Chairman



Dr. Suliman Saleh Al-Homidan
Dean of Graduate Studies

Date



Dr. Emad Ramadan
(Advisor)



Dr. Moataz Ahmed
(Member)

Imene

Dr. Imane Boudellioua
(Member)



© Ahmed Mohamed Mahmoud Elmoselhy

2022

I dedicate this thesis to my loving family

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my advisor, Dr. Emad Ramadan, for his continuous support, guidance, patience, motivation, and supervision during my research work and completing this thesis. I would also like to thank my committee member Dr. Moataz Ahmed and Dr. Imane Boudelloua for their constructive feedback. Many thanks to my family for their endless support, prayers, and encouragement. Without their continuous support, it was impossible to achieve this goal.

TABLE OF CONTENTS

ACKNOWLEDGMENTS	v
TABLE OF CONTENTS	vii
LIST OF TABLES	xi
LIST OF FIGURES	xiii
LIST OF ABBREVIATIONS	xv
ABSTRACT.....	xvii
ملخص الرسالة.....	xxi
CHAPTER 1 INTRODUCTION	1
1.1 Predicting MicroRNA-Disease Association	1
1.2 Computational and Traditional Techniques in Drug Design/Discovery	3
1.3 Target-based Drug Design.....	6
1.4 Problem Statement.....	7
1.5 Thesis Objectives	7
1.6 List of Contributions.....	7
1.7 The Impact of the Presented Work	8
1.8 Thesis Organization	9
CHAPTER 2 BACKGROUND.....	11
2.1 The Central Dogma of Molecular Biology	11
2.2 Protein Domains	12
2.3 MicroRNAs	12
2.4 SNPs and Diseases	13
2.5 MicroRNA-Disease Associations.....	15
2.6 Multipartite Graph	16
2.7 Graph Databases	17
2.8 Property Graph Model	18
CHAPTER 3 LITERATURE REVIEW	21
3.1 Graph Databases	21
3.2 Graph databases applications in various domains.....	21
3.3 Graph databases applications in biological domains	21
3.4 Finding MicroRNA-Disease Associations	22

CHAPTER 4 PROPOSED APPROACH	27
4.1 Questions Raised by Reviewing the Literature	27
4.2 Proposed Research Approach	28
4.3 Data Sources	29
4.4 Data Collection, Integration, Cleaning, and Preprocessing	31
4.5 Graph Modeling	32
4.6 Graph Analysis	39
4.7 Centrality Measures	39
4.8 Modified Node Degree Centrality	41
CHAPTER 5 EXPERIMENTS, RESULTS & DISCUSSION.....	43
5.1 Protein Domain-Disease Graph Model	43
5.2 Associations using SNPs	48
5.3 Disease-MiRNA Associations	50
5.4 Protein-Domain Data	51
5.5 MicroRNA-Domain Graph Model.....	52
5.6 The Final Graph Model	53
5.7 Analysis of the Final Graph Model.....	54
5.7.1 Node Degrees by Label	55
5.7.2 Protein Hub Nodes	57
5.8 MicroRNA - Disease Node Degree Centrality	59
5.9 MicroRNAs-Domain Associations in Breast Cancer	60
5.10 Calculating the Disease Node Degree Centrality: Breast Cancer as a Case Study.....	61
5.11 Studying the effects of change in the data on the proposed approach	65
5.11.1 Dropping Hub Protein Nodes	65
5.11.2 Dropping Random Edges	70
CHAPTER 6 Validation & Threats to Validity	73
6.1 Validation through <i>InterPro</i> resource	73
6.2 Validation through <i>DAVID</i> data source.....	73
6.3 Validation of Results through Comparison with Experimental Findings in the Biology Literature.....	74
6.4 Threats to Validity	76
CHAPTER 7 CONCLUSION & FUTURE WORK	79

7.1	Conclusion.....	79
7.2	Future Work.....	81
	Appendix.....	83
	References.....	93
	Vitae	99

LIST OF TABLES

Table 1 Main components of property graph model.....	18
Table 2 Data sources for different types of data that were used in the presented work ...	30
Table 3 First graph model statistics	46
Table 4 Sample for SNPs data from Expasy/Uniprot	49
Table 5 Statistics for single amino acid variants	51
Table 6 Data from BioMart and Pfam	51
Table 7 Node statistics for microRNA-domain graph model	52
Table 8 Edge statistics for microRNA-domain graph model.....	52
Table 9 Statistics for the final's graph model relationships.....	54
Table 10 Statistics for the final's graph model nodes.....	54
Table 11 Top five disease, protein domain, protein, genes, miRNA nodes	56
Table 12 A snapshot of gene-miRNA association in breast cancer study	61
Table 13 Top 20 ranked microRNAs with their disease node degree centrality calculated for breast cancer.....	62
Table 14 Disease degree centrality for microRNAs with different percentage of hub protein nodes got removed from the graph	66
Table 15 References from literature for the top 20 MicroRNAs	75
Table 16 Ranked list of microRNAs with their disease node degree centrality calculated for breast cancer.....	83
Table 17 Number of protein nodes with degree above X percentile of the other nodes in the graph	87
Table 18 Change in disease degree centrality for the microRNAs as a result of random edges removal	88

LIST OF FIGURES

Figure 1 High-level flow of drug discovery/design pipeline	4
Figure 2 In silico, in vitro, and in vivo phases in drug discovery pipeline [3]	5
Figure 3 Target-based drug design [5]	6
Figure 4 The central dogma of molecular biology [8]	12
Figure 5 A protein with four domains [10]	12
Figure 6 MicroRNA role in CDMB [12]	13
Figure 7 An example of missense mutation and its effect on the resulting protein	14
Figure 8 MicroRNA role in the central dogma of molecular biology	15
Figure 9 A multipartite graph with P sets of nodes [16]	16
Figure 10 A 4-bipartite graph model [6]	17
Figure 11 Categories of computational microRNA-disease association findings models	25
Figure 12 Workflow of the present work	28
Figure 13 A high-level diagram for the data collection phase	31
Figure 14 Modeling the domain-disease graph, adapted from [6]	33
Figure 15 Modeling the final graph	38
Figure 16 The first graph model.	45
Figure 17 An illustrative example of a domain node	47
Figure 18 The two protein domain nodes with the highest disease node degree centrality score	48
Figure 19 The MicroRNA-Domain graph model	53
Figure 20 The final graph model	53
Figure 21 A pair of microRNA and disease nodes	55
Figure 22 Protein nodes degree distribution	58
Figure 23 An example of microRNA with a score of 16 disease node degree centrality.	59
Figure 24 MicroRNA node with the first largest disease node degree centrality	63
Figure 25 MicroRNA node with the second largest disease node degree centrality	64
Figure 26 Effect of dropping hub protein nodes for different percentile values on the disease degree centrality	69
Figure 27 Effect of dropping [1% - 5%] percentage of the edges on the disease degree centrality	71

LIST OF ABBREVIATIONS

CDMB	:	Central dogma of molecular biology
DNA	:	Deoxyribonucleic acid
RNA	:	Ribonucleic acid
mRNA	:	Messenger RNA
miRNAs	:	Micro ribonucleic acid
SNPs	:	Single-nucleotide polymorphisms
nsSNPs	:	Non-synonymous single nucleotides Polymorphisms
UniProt	:	Universal protein resource
Pfam	:	Protein families
miRTarBase	:	microRNA-target interaction database
miRCancerdb	:	microRNA cancer database
TCGA	:	The cancer genome atlas
DAVID	:	The database for annotation, visualization and integrated discovery

ABSTRACT

Full Name : [Ahmed Mohamed Mahmoud Elmoselhy]
Thesis Title : [Finding associations between microRNAs/protein domains and diseases: A graph-based approach]
Major Field : [Computer Science]
Date of Degree : [May 2022]

MicroRNAs(miRNAs) are small non-coding RNA molecules that participate in gene regulation. They can be part of “gene silencing”, where a gene can be prevented from producing its corresponding product (i.e., protein). Several studies suggested that microRNAs play a crucial role in the development of several diseases, including cancer diseases. Therefore, finding associations between microRNAs and diseases can help in understanding the mechanisms of those diseases, let alone developing treatments for them. As we currently live in the “-omics” era, there are myriad amounts of biological data available in online public databases. One type of biological data among those available ones is data about protein domains. Protein domains are essential biological compounds in the life cycle of almost every cell in living organisms. Relying on that data, several computational approaches have been proposed to find associations between microRNAs and diseases. To our knowledge, graph analysis techniques, coupled with protein domains data, have not been studied in detail in the context of finding associations between diseases and microRNAs. In this thesis, we aimed at providing a novel computational method to solve the problem of finding microRNAs-diseases associations. In doing so, we managed to integrate biological data from multiple resources. We collected different types of data,

including, but not limited to, data about genes, protein domains, and microRNAs. The collected data has gone through a data cleaning, preprocessing, and integration pipeline. Finally, the resulting curated dataset has been encoded and modeled as a multi-bipartite graph. The final graph model is a heterogeneous one, in terms of nodes as well as edges. Having the graph model, we then leveraged the graph representation, which highlights the importance of interconnectivity within the data, to find associations between diseases and microRNAs. To do so, we devised a new node centrality measure, namely the *modified node degree centrality* measure. The *modified node degree centrality* has been used as a similarity measure to find associations between microRNAs and diseases and similarities between protein domains and diseases. Breast cancer was chosen as a case study in this thesis. A ranked list of associated breast cancer-associated microRNAs has been reported, and several resources have been used to validate the results of the proposed work. In order to ensure the reliability and robustness of our approach, we conducted a series of validation experiments. Specifically, we removed a subset of hub protein nodes, which are the nodes with high degree, and a subset of edges from the graph and re-evaluated the modified node degree centrality. By conducting these validation experiments, we were able to demonstrate that our approach is robust to the removal of these key components, indicating that it is likely to produce reliable results even when applied to graphs with different structures or missing information. This further supports the validity of our approach and the confidence we can have in its performance.

Finally, the microRNA-disease graph model is provided, so that it can be utilized to uncover microRNAs associations for other diseases, and it can be further studied to answer more biological questions. Also, the newly devised centrality measure can be used to

calculate the relative importance of entities with different types, given that they can be modeled as a multi-bipartite graph.

|

ملخص الرسالة

الاسم الكامل: احمد محمد محمود المصيلحي

عنوان الرسالة: إيجاد الارتباط بين الميكرو رنا، نطاقات البروتين والأمراض باستخدام المخططات (الشبكات)

التخصص: علوم الحاسوب

تاريخ الدرجة العلمية: مايو ٢٠٢٢ |

الميكرو رنا (miRNA) هي عبارة عن جزيئات احماض نووية ريبوزية صغيرة غير مرمزة للبروتينات وتشارك في عملية التعبير الجيني. من الممكن أن تكون جزءاً من عملية "القمع الجيني" حيث يمكن أن يُمنع جيناً ما من إنتاج البروتين المُعد لإنتاجه. اقترحت العديد من الدراسات البحثية ان الميكرو رنا تلعب دوراً مهماً في تطور العديد من الأمراض، ومنها أمراض السرطان. إن إيجاد علاقات صلة بين بعض من الميكرو رنا وأمراض معينة يمكن ان يساعد في تطوير علاجات لهذه الأمراض.

بما أننا نعيش اليوم في عصر الـ "omics"، توجد هناك كميات هائلة من البيانات البيولوجية في قواعد البيانات العامة، نطاقات البروتين هي مكونات بيولوجية اساسية في دورة حياة أي خلية في الكائنات الحية. بالاعتماد على تلك البيانات، تم تطوير العديد من الطرق الحسابية لإيجاد علاقات صلة بين الميكرو رنا والأمراض. لم يصل الى علمنا أنه تم استخدام تقنيات تحليل الشبكات جنباً الى جنب مع بيانات تخص نطاقات البروتين في سياق إيجاد علاقات صلة بين الأمراض والميكرو رنا. في هذه الأطروحة، قمنا بجمع ودمج العديد من انواع البيانات البيولوجية وكذلك قمنا بنموذج شبكي متعدد الطبقات وغير متجانس والذي يمثل العديد من المركبات البيولوجية. أيضاً، تم تقديم مقياس جديد لقياس اهمية النقاط على الشبكة كونه مقياساً لمعضلة التنبؤ بالرابط لإيجاد علاقات صلة بين الميكرو رنا والأمراض. تم اختيار سرطان الثدي كدراسة حالة لإثبات صلاحية الدراسة. كنتيجة للعمل المقدم، يمكن دراسة الشبكة المقدمة للإجابة على اسئلة اضافية تخص البيانات المتاحة. تم توفير قائمة مرتبة تنازلياً تحتوي على الميكرو رنا المرتبطة بمرض سرطان الثدي. تم التحقق من نتائج هذا العمل بالرجوع الي المراجع الموثوقة ذات الصلة بهذا الموضوع.

للتأكد من ثبات وجودة نهجنا، قمنا بعدة تجارب للتحقق من صحته. على وجه التحديد، قمنا بإزالة مجموعة من العقد الرئيسية التي تعتبر عقد ذات نسبة درجة عالية ومجموعة من العلاقات من الشبكة وإعادة تقييم العقد المعدلة. من خلال هذه التجارب للتحقق من الصحة، كان لدينا القدرة على التثبت من أن نهجنا قابل للتعامل مع إزالة هذه العناصر الرئيسية، مما يدل على احتمالية إنتاج نتائج ثابتة حتى عند التطبيق على الشبكات التي تمتلك بنيات مختلفة أو فقدان من المعلومات. هذا يدعم بعد ذلك صحة نهجنا والثقة التي يمكننا الاعتماد عليها في أدائه. تم كذلك إتاحة النموذج النهائي حتى يتثنى العمل عليه واستخراج روابط لإمراض اخري. من

الممكن أيضاً ان يتم استخدام المقياس العام المقدم لقياس الالهمية النسبية بين نوعين مختلفين من البيانات باعتبار انه يمكن نمذجتهم على شكل شبكة.

CHAPTER 1

INTRODUCTION

1.1 Predicting MicroRNA-Disease Association

One of the most ruthless enemies of humankind is diseases. Diseases can attack humans undetected; they can be lethal, brutal, and without mercy. Cancer stands out as the most ferocious disease of all. It is known to be the second leading reason for death in the world [1]. In nature, cancer is an abnormality in cell growth: when a cell starts to divide uncontrollably and grow maliciously affecting other neighboring cells. The outcomes of such behavior can vary in severity and side effects from a simple mole on the skin to organ failure, and in some cases, death.

In general, diseases are complex, hard to understand, and involve several components. Many diseases have a genetic component where a disfunction on the genetic level may cause severe consequences. Studying diseases on the genetic level involves understanding interactions and associations between the different components of the cellular system. Knowing a new interaction or figuring out a novel association will make us able to understand the mechanism of a disease and hence, allowing for further experiments to leverage this knowledge in developing new therapeutic strategies. In 1993, researchers discovered microRNA, a new biological component in living organisms [2]. From that time miRNA took researchers' interest, and several studies have been published to help in understanding this molecule and its roles and interactions within its surroundings [2][3]. Considerable efforts from researchers in the biology field have shown that

there is an association relationship between diseases and miRNAs. It turns out that discovering these associations will be a valuable finding for some therapeutic applications [4]. However, conducting biological laboratory experiments to achieve that is expensive and time-consuming. Computational techniques offer a suitable way to solve the problem that is less expensive and does not require a long time to conduct.

Throughout the years, many scientists have struggled to understand the unique nature of cancer and tried to find ways to prevent and treat it, especially in recent years, it has shown a steady increase in cancer patients. In 2020, around 19.3 million have been diagnosed with cancer, with female breast cancer as the most common type worldwide. Out of the 19.3 million cancer patients, there was a total of 10 million deaths worldwide [5].

Efforts for cancer treatment have gone through several milestones along with human history. It started from the early ages in the ancient Egyptian era where Imhotep, the Egyptian Pharaoh, used a two-stage process to lead to tumor regression. He used poultice and then he followed it by incision to cause infection in the targeted location. That developed infection would cause regression of the tumor [6]. In the modern era, and as technology advances, cancer is being approached by more and more advanced methods of treatment. In the 19th century, surgeries were used to remove tumors. In the 20th century, more advanced and less invasive methods were used, ranging from endoscopy, freezing the cancer cells by spraying liquid nitrogen (cryosurgery) to radiation, chemo- and immuno-therapy [1].

One of the promising approaches to investigate cancer behavior and genesis is on the genomic level. In such an approach, scientists try to uncover hidden patterns in the relationship between cancer cells and other biological factors that might initiate and nourish cancerous behavior in cells.

Finding hidden associations and relationships that may lead cells to such destructive behavior is important to the process of treatment and prevention.

To study cancer on that micro yet rich level, several types of biological entities, along with their interconnections, need to be taken into account. The problem here is that this data is scattered. With that dispersion, hidden relationships between different entities of this data which may help in further treatment plans, are away of researchers' hands. A core property of those entities is that they are highly interrelated. That interconnectivity needs to be highlighted as this data is being investigated.

Biological data is scattered between a wide spectrum of data sources. As a result, many connections between different biological entities are missing. Revealing hidden relationships/associations could help in discovering new treatment procedures for some of the complex diseases.

1.2 Computational and Traditional Techniques in Drug Design/Discovery

The area of drug design and drug discovery has developed over the years, starting with experiments that are done in wet, experimental labs where the actual experiments are being done either in a living organism (i.e., *in vivo*), outside a living organism (i.e., *in vitro*)[1]. That was a successful way that helped humanity over the years. However, time and cost were the main two obstacles in the development of many drugs, because both *in vitro* and *in vivo* methods are time-consuming and costly. Therefore, and thanks to the advances in the computational sciences, *in silico* techniques became a major player in the pipeline of drug discovery and design in the recent few decades [2]. Figure 1 and Figure 2 show the three main paradigms used in the drug design/discovery field.

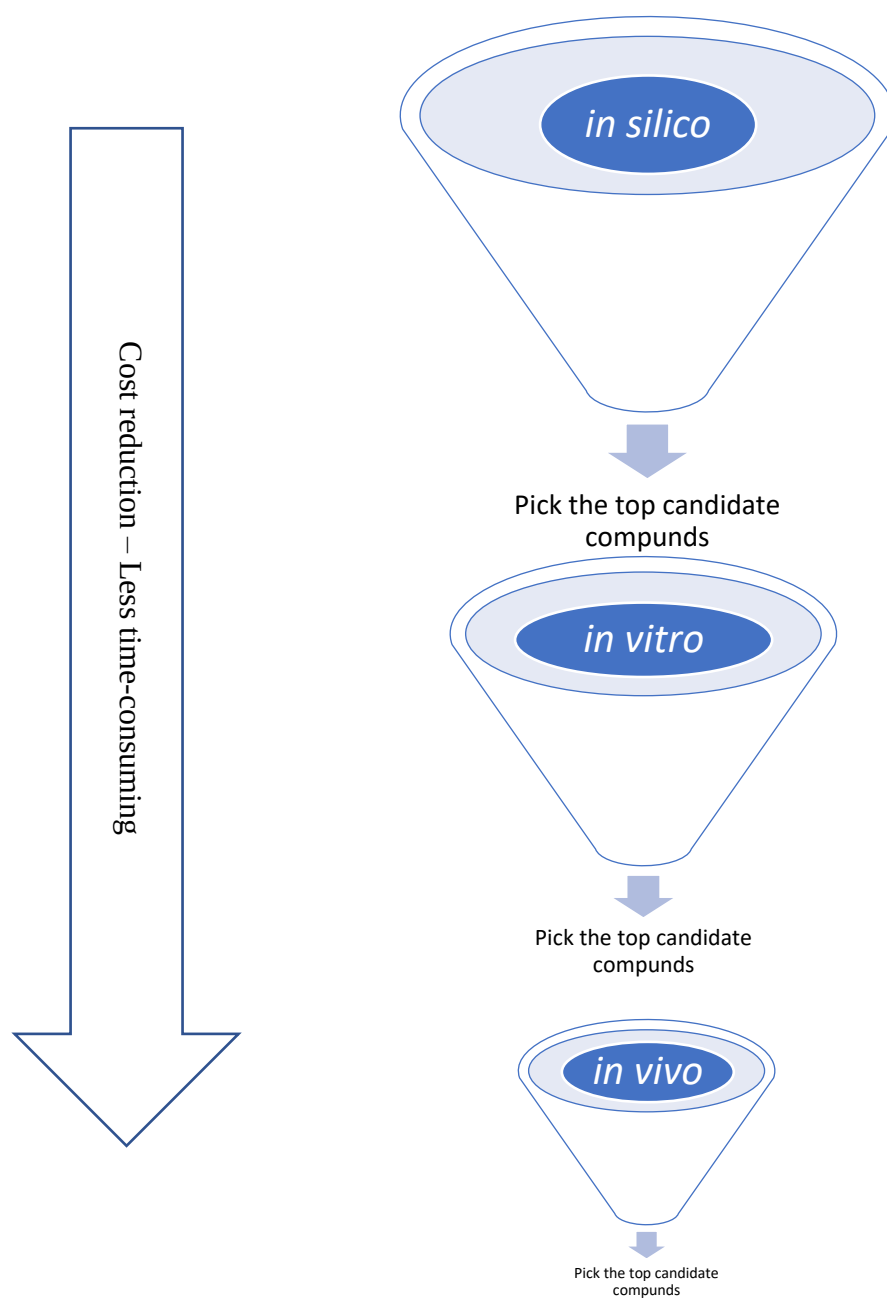


Figure 1 High-level flow of drug discovery/design pipeline

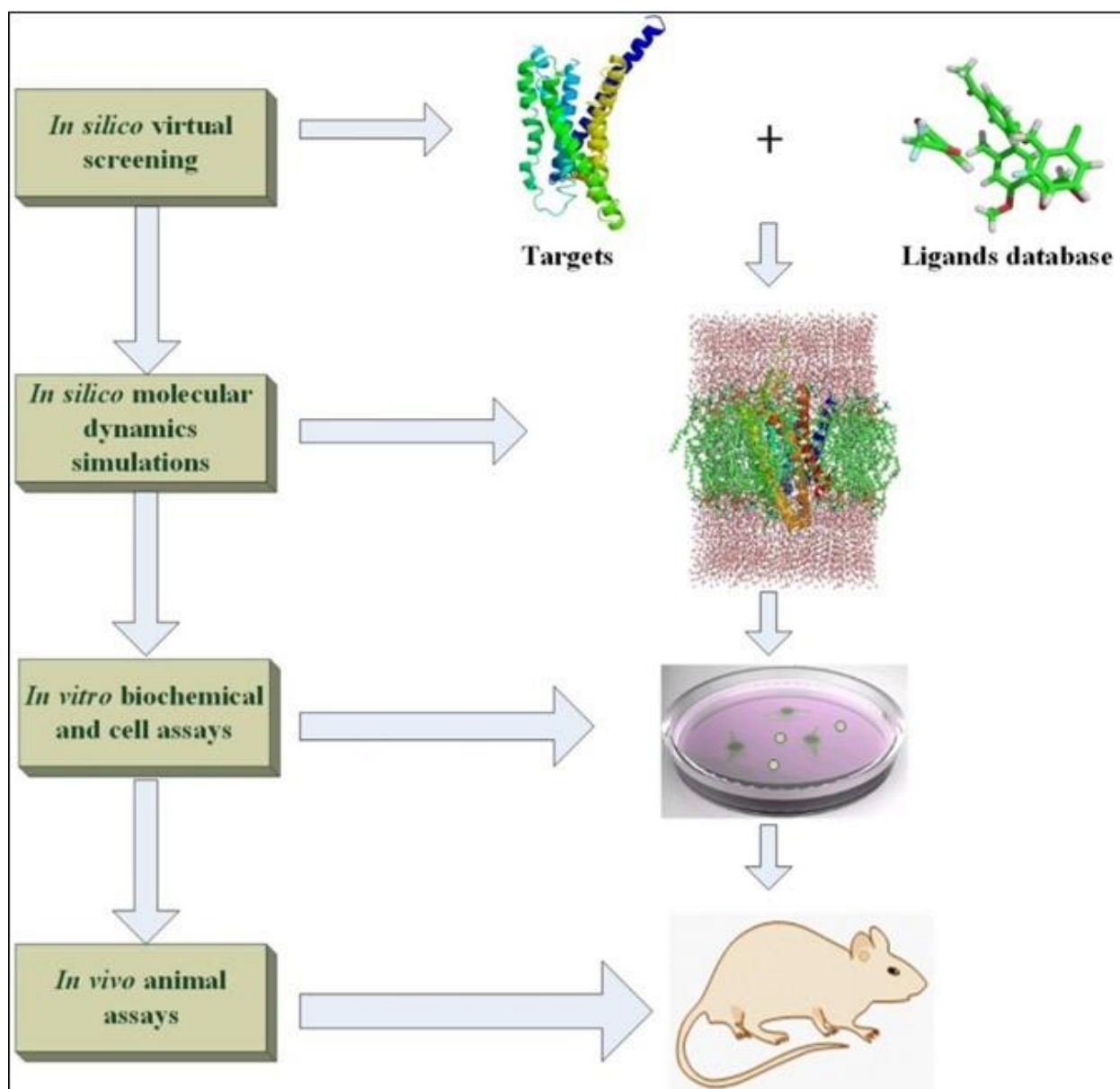


Figure 2 In silico, in vitro, and in vivo phases in drug discovery pipeline [3]

1.3 Target-based Drug Design

Target-based drug design is an example of a field where computational in silico techniques can be highly utilized. Figure 3 shows an example of one of the early processes in the target-based drug design. Where for a drug to be designed to target a certain protein, the building blocks of that protein (protein domains) need to be identified [3]. Knowing that a certain is associated with a certain disease, that information facilitates the job of drug discovery practitioners as now they can consult some of the public data sources (e.g., UniProt) to get the 3D information of that domain, and then, use it while considering the 3D structure of the drug that they will design to bind to that “pocket” in the targeted protein [4].

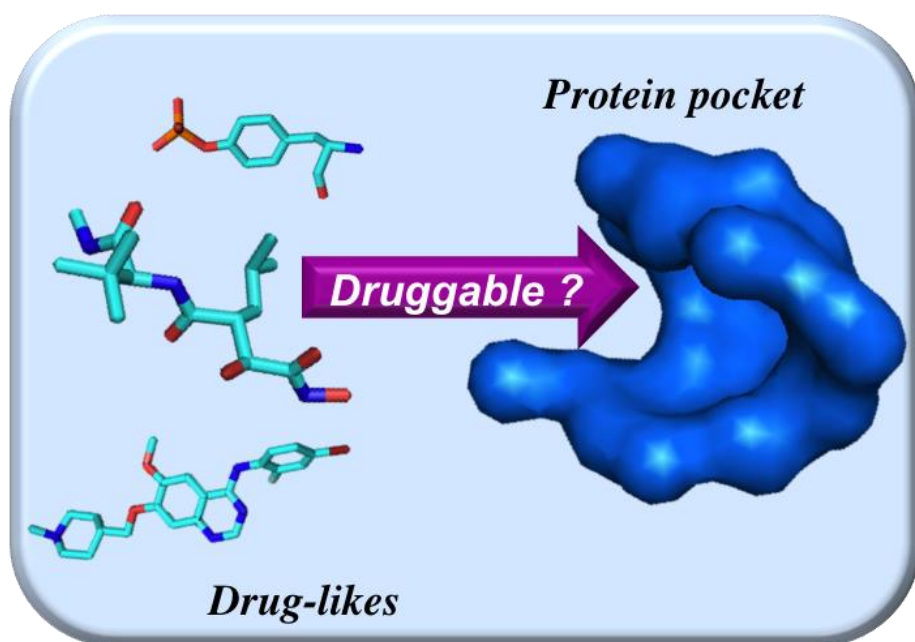


Figure 3 Target-based drug design [5]

1.4 Problem Statement

Utilizing different biological data sources in finding associations between microRNAs and breast cancer disease using graph/network analysis techniques.

1.5 Thesis Objectives

The aim of the presented thesis is enumerated as follows:

1. Integrating biological data from multiple sources.
2. Building a multipartite heterogeneous graph model.
3. Introducing a new node centrality measure.
4. Ranking protein domains based on their association scores with diseases
5. Ranking microRNAs based on their association scores with diseases.

We believe that the process of data collection, cleaning, and preprocessing that led to the building the graph model is a valuable deliverable asset of this work as it is providing to the biological scientists a ready model for further enquiries that help in answering different questions in drug design and drug discovery pipelines.

1.6 List of Contributions

The presented work has resulted in the following set of contributions:

1. A conference article, published at an IEEE conference [6].
2. Integrating biological data from multiple sources into complex multilayer (k-bipartite) heterogeneous graph models.
3. Providing analysis of the designed graph models.
4. Introducing a new generic node centrality measure.
5. Applying the introduced node centrality measure on the modeled graph models.

6. Ranking microRNAs based on their association with diseases.
7. Validating the microRNA ranked list with the literature and suggesting a few novel ones to be further experimentally investigated.

1.7 The Impact of the Presented Work

The impact of the presented work can be shown in the field of graph analysis, and the presented instances of graph models applied on the biological data can contribute to the drug design and drug discovery fields.

Firstly, the provided graph models along with the new generic node centrality measure can be used with different types of data that can be modeled as a k-bipartite (heterogenous or homogenous) graph model. Thus, allowing findings associations between different types of nodes on different layers of the constructed graph. Knowing the associations between two nodes from two different sets of nodes can be considered as an indicator of a missing link between those two nodes. So, the introduced node centrality measure (i.e., *modified node degree centrality*) is a measure of the relative importance between nodes from different independent sets on a multi-bipartite graph. That means it can be used whenever the goal is to relate nodes from different domains to each other relying on the connections between other layers in the network/graph.

Secondly, in the field of drug design, especially in target-based drug design, researchers search for the target protein where the drug is supposed to bind to. Knowing the structure of the target protein, and the target protein domain(s), via consulting public data resources such as UniProt, can greatly facilitate the process of designing a drug. That is because knowing the 3D structure of the target (i.e., pocket) will provide that information to the drug design practitioners to design the candidate drugs that will be able to match (bind to) that 3D structure part of the protein where the drug is supposed to bind.

Additionally, given the evidence from the literature on the important role that microRNAs play in the life cycle of the cell and, therefore, in the development of several diseases, having a list of microRNAs with their associated *disease node degree centrality*, will facilitate the process of choosing the best possible candidates to work on and to proceed with them to the next stage of the drug discovery pipeline.

1.8 Thesis Organization

This thesis is divided into 6 chapters as follows: An introductory chapter, Chapter 1, followed by Chapter 2 that gives an elaborate overview of the fundamentals of main concepts related to the problem from both biology and computer science fields. Chapter 3 reviews the work done by others for solving relevant problems using different computational techniques, followed by Chapter 4 introducing the proposed approach for solving the problem at hand. Chapter 5 shows the experiments, their results, and the discussion of the results. Chapter 6 shows the validation of the results, its limitations, and the threats to validity of the proposed approach. Chapter 7 concludes this thesis by highlighting the conclusions of this work, and future work.

|

CHAPTER 2

BACKGROUND

This chapter will give the necessary background to appreciate the impact of the presented work from the perspective of computer science, biology, and bioinformatics fields. It also helps in clarifying some main concepts from biology. A brief yet informative section is provided for each concept to provide prerequisite knowledge about cells, proteins, protein domains, microRNAs, SNPs and their relationship with diseases, graph databases, property graph models, and multipartite graphs. Firstly, we will start by providing some biology background as it will pave the way for the following discussions. To know what miRNA is, we need to start with a brief discussion about the cell, the basic unit of life.

2.1 The Central Dogma of Molecular Biology

Inside the cell, there is the nucleus which contains deoxyribonucleic acid (DNA). The DNA contains the genetic code of the body. In humans, it is estimated to contain around 3 billion base pairs (nucleotides) which can be considered as the alphabet of DNA. Among these huge numbers of nucleotides, there are some segments known as genes. A gene is a segment of DNA that will code for a protein. A protein is the product of a gene(s) during the process of the central dogma of molecular biology (CDMB). In the CDMB, there are mainly two steps: transcription and translation. In the transcription phase, the gene is converted to another molecule called messenger RNA (mRNA). Then in the translation phase, the mRNA is to be converted to protein. Proteins are functioning molecules in the human body which are responsible for many vital functions and

processes from growing muscles to determining your eye color [7]. Figure 4 shows a simplified illustration of the central dogma of molecular biology.

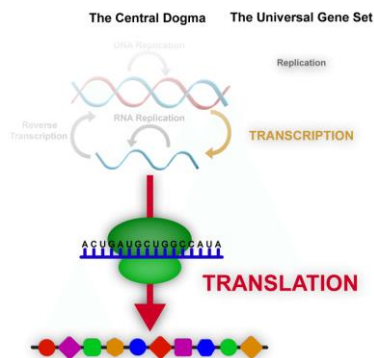


Figure 4 The central dogma of molecular biology [8]

2.2 Protein Domains

Protein domains are a functional part of proteins. A single domain can exist in multiple proteins and can function and exist independently. Figure 5 shows a 3D structure of a protein with four domains [9].

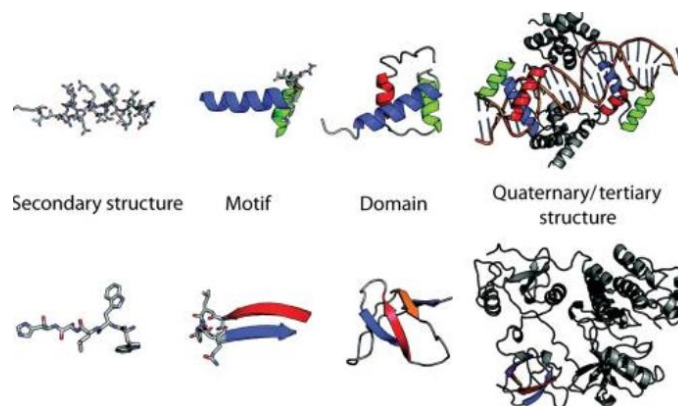


Figure 5 A protein with four domains [10]

2.3 MicroRNAs

MicroRNAs(miRNAs) are small non-coding RNA molecules that participate in the post-transcriptional phase of gene regulation. They can be part of “gene silencing”, where a gene can

be silenced from producing its corresponding product (i.e., protein). Through this process, a miRNA molecule may suppress the intermediate molecule (mRNA) and prevent it from producing a protein. A dysregulation in that process can contribute to causing many diseases including different types of cancer [11].

Figure 6 shows the role of microRNA in the central dogma of molecular biology [7]. The upper part shows the flow of the transcription and the translation processes from the gene to producing the corresponding protein. The lower part shows the scenario when miRNA interferes in the post-transcriptional phase and suppresses the messenger RNA (mRNA) and prevents it from assembling a protein.

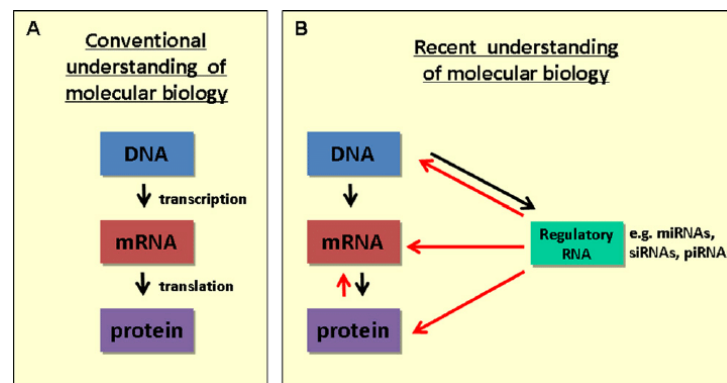


Figure 6 MicroRNA role in CDMB [12]

2.4 SNPs and Diseases

DNA consists of nucleotides (usually abbreviated as A, T, C, and G). A change/substitution of one of these nucleotides is called a single-nucleotide polymorphism.

Single Nucleotides Polymorphisms (SNPs) are a group of variations/mutations that happen in a single base pair in the human genome and some of them are proven to be correlated with certain diseases[13]. A special type of them is Non-synonymous Single Nucleotides Polymorphisms (nsSNPs). nsSNPs are one of the SNPs that occurred in the coding regions of the DNA.

Figure 7 shows an example of a missense mutation and its effect on the mRNA, and on the produced amino acid, and hence on the resultant protein. Just a single change in the DNA resulted in a completely different protein.

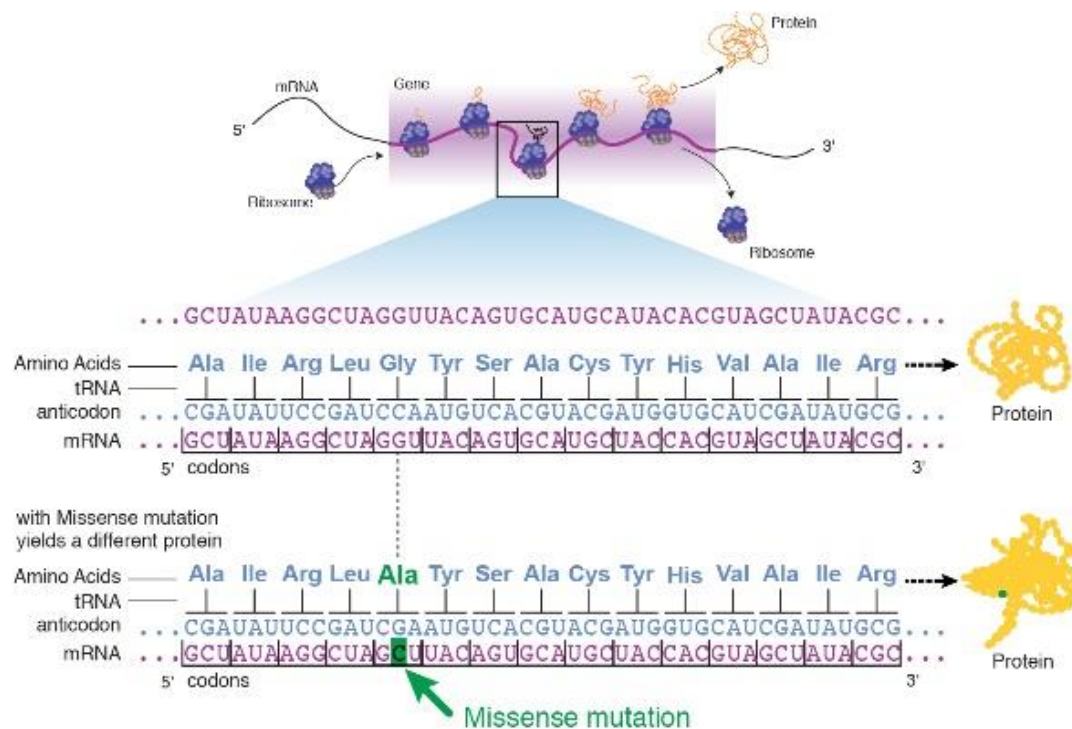


Figure 7 An example of missense mutation and its effect on the resulting protein

2.5 MicroRNA-Disease Associations

MicroRNAs can cause some diseases during CDMB by either binding to the wrong mRNA or by not binding to the mRNA where it is supposed to bind. From this observation, several studies [14], [15] have shown that there is an association relationship between diseases and microRNAs. Figure 8 shows the role that microRNA plays in the central dogma of molecular biology. In that figure, A typical gene is being expressed to produce the messenger RNA through the process of transcription. A normal pattern after that step is to go through the process of translation where the relevant protein will be produced. However, another microRNA gene on the DNA might produce microRNA and that microRNA may bind to the produced messenger RNA in a process known as “gene silencing”. In that process, the microRNA gene is silencing the messenger RNA and hence preventing it from assembling the protein that it should have produced.

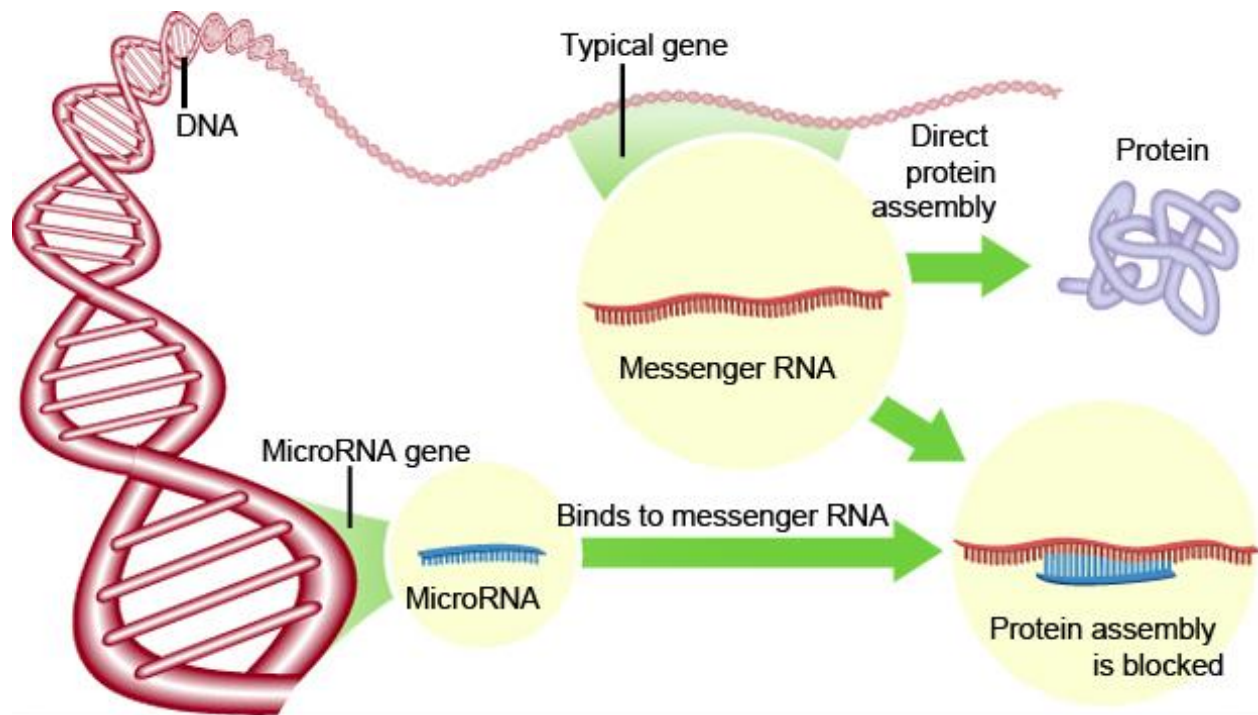


Figure 8 MicroRNA role in the central dogma of molecular biology

2.6 Multipartite Graph

A multipartite graph is a graph of multisets of nodes where the edges exist only between those sets and no edges within the same set of nodes. Figure 9 shows a general structure for a multipartite graph of P sets of nodes.

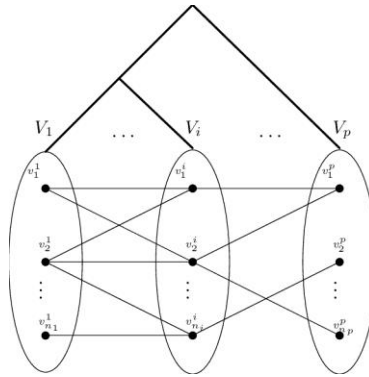


Figure 9 A multipartite graph with P sets of nodes [16]

In Figure 10, a 4-bipartite is shown where the 4 layers are representing 4 independent sets of vertices where no two nodes in one set are connected, but nodes from different sets can be connected together.

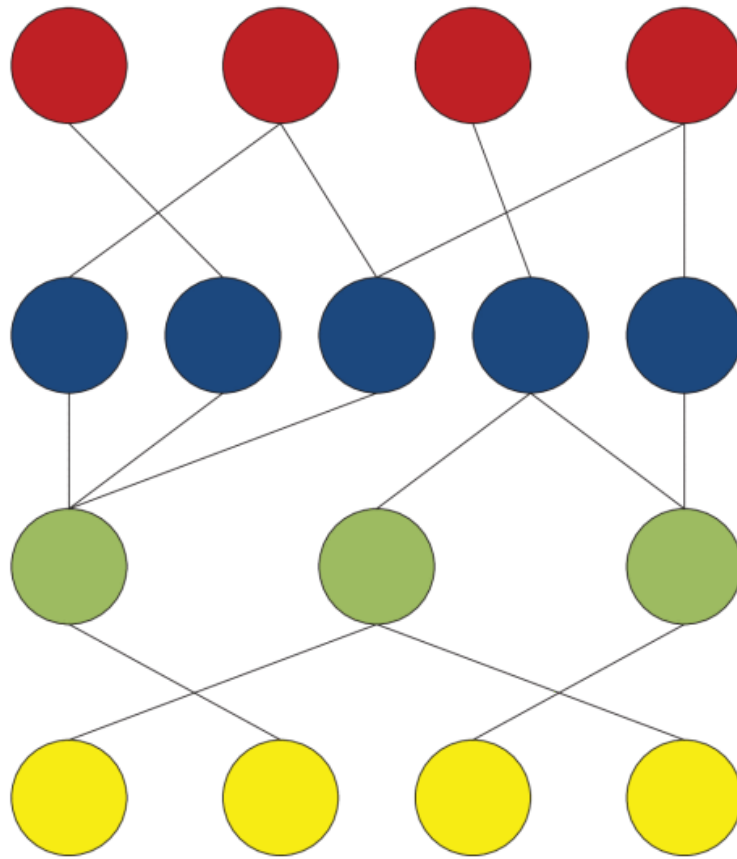


Figure 10 A 4-bipartite graph model [6]

2.7 Graph Databases

A graph database is one of the NoSQL databases models that are designed in a way that promotes the role of relationships in the database, or as stated by Robinson et. al. [17], in a graph database data model, "relationship is a first-class citizen". One of the common graph database models is the property graph model, which contains nodes and relationships (edges). A node represents an entity of data and can have properties (in the form of key-value pairs), and label(s) which, informally, can indicate the type of that node. A relationship (edge) in the graph organizes the nodes and describes the relationship between nodes. It can have a name, properties, and it always has a start and end node [17]. Moreover, graph database management systems store edges as directed pointers which offer a constant time traversal [18].

2.8 Property Graph Model

The property graph model is one of the models that graph databases support. The three main aspects of the property graph model are nodes, relationships, and properties.

Table 1 Main components of property graph model

	Description	Notes	Example
Node	Entity	Can have zero or more labels	Person, drug, gene, disease, etc.
Relationship	Mostly verb that describes the relationship between two entities	Every relationship must have a type	LIKES, INTERACT, PRODUCE, CAUSE, etc.
Properties	In the form of “name-value” pair	Both for nodes and relationships. Both can have as many as the developer wants	Age:20, Ddate:1990, Type: coding, Category: cancer

Many biological components exist in an interconnected form [19], and to study them computationally, a suitable management storage system is vital. Therefore, the nature of representation that a graph database provides is suitable for storing and analyzing biological data.

CHAPTER 3

LITERATURE REVIEW

3.1 Graph Databases

Graph database has been widely used in the last few years in a variety of applications covering a wide spectrum of domains. In this section, we mention some of the efforts that utilized graph databases in different domains.

3.2 Graph databases applications in various domains

Park et al. [20] utilized a graph database within their framework for managing a large-scale healthcare system, and out of eight use cases used in the paper, Neo4j outperformed MySQL in the execution time in five use cases. Neo4j is also used as the graph database to store the USA healthcare data to estimate the fraud risk of the healthcare providers [21].

Another work by Drakopoulos and Kanavos [22] used Neo4j along with Biopython [23] with some other packages to build a tensor-based framework that performs some data mining tasks on the well-known search engine offering access to biomedical literature, PubMed.

3.3 Graph databases applications in biological domains

Lysenko et al. [24] have used Neo4j as the platform to perform graph analysis on multi-omics data. The dataset that was used in that work contains a variety of biological components including data about diseases and protein-protein interactions. The data about diseases was imported from DisGeNET [25], and protein-protein interaction data was imported from EBI IntAct database [26]. Concerning their data and graph modeling, they integrated the data in a heterogeneous graph from

a protein-centric perspective. A protein node is linked to a disease node using a variable-type edge and proteins are connected via a set of connections. Even though the authors integrated multiple types of biological components, they neglected domains that are important biological components and, when investigated, can lead to many discoveries in various fields as mentioned by [27] .

3.4 Finding MicroRNA-Disease Associations

The topic of finding microRNA-disease associations has been studied extensively in the literature from different perspectives. First and foremost, it was studied by biologists where the majority of those types of studies is coming from experimental work in wet labs [28][29].

Next, we highlight a selected set of computational techniques that have been developed to tackle the problem of finding microRNA-disease associations.

Peng et.al. [30] introduced the HNMDA model which starts with applying random walk with restart technique on the disease-to-disease similarity network and on the miRNA-to-miRNA similarity network, and after several iterations of the random walker, it reaches the diffusion state. For a pair of nodes, to if there are similar, they checked if they have a similar diffusion state, which means that their position on the network is close, and they may have similar functions. Then, they performed an optimal projection from miRNA space onto disease space based on the belief that if both vectors (for miRNA and disease) in the same space are close to each other then there is a chance that they have a potential association. They did that by a method that finds the optimal projection from miRNA into disease space. Finally, they transformed the projection problem into a minimization problem. Hepatocellular carcinoma, Lymphoma, Kidney neoplasms, Esophageal neoplasms, and Breast neoplasms were used as case studies.

The authors of [31] approached the problem by providing a network method based on co-neighbors in a bipartite heterogeneous network. They extended the concept of co-neighbor nodes in the general networks, and they defined the concept of neighborhood in the bipartite networks. Simply, consider a set of nodes D that contains diseases nodes and set M that contains miRNA nodes then, and since the network is bipartite, there are no direct co-neighbor nodes in that case between nodes from different categories, thus they presented the concept of neighborhood in the bipartite networks as follows: for a disease node d_i that belongs to set of diseases D , and for a miRNA node m_j that belongs to set of miRNAs nodes M , if they found a path of length 3 that starts from set D to set M , back to set D , and then finally returns to set M with different nodes along the path, then they consider the 2 end nodes (d_i, m_j) as co-neighbor nodes in this bipartite network. Simply put, a path with length 3L: $(d_i \rightarrow m' \rightarrow d' \rightarrow m_j)$. they have used it in calculating an association score between disease nodes d_i and miRNA node m_j . They have utilized some of the well-known indexes for link prediction algorithms for a general network (not K-partite) such as (co-neighbor, Jaccard, PA, Salton, Sorensen, HPI, LHN1, and HDI indices). They integrated each one of those indexes with their definition of co-neighbor in the bipartite network which has been described above. NCN (d_i, m_j) stands for the number of paths with length 3 according to their definition of that path. For the case studies used in that work, they have chosen Breast neoplasms, Colon neoplasms, Breast neoplasms while removing all known breast neoplasms-miRNAs associations, and Colon neoplasms while removing all known colon neoplasms-miRNAs associations.

Chen et.al. [32] provided a network distance analysis approach. The model goes through three steps to get the miRNA-disease predictions: computing the adjusted distance between miRNAs and between diseases. Then, calculating the confidence of diseases and miRNAs. Finally,

converting the confidence score into probabilities. They have used Breast neoplasms, Esophageal Neoplasms, Lymphoma, Prostate neoplasms, and Hepatocellular carcinoma as case studies.

The SACMDA model predicted the miRNA-disease associations with short acyclic connections in a heterogeneous graph [33]. The model has 3 input matrices: disease semantic similarity matrix: contains similarity between diseases, miRNA functional similarity matrix: contains similarity information about all kinds of miRNA, and miRNA-disease association matrix (A): contains the association between diseases and miRNAs, $A(d(i), m(j))$ is equal to 1 when the association is experimentally confirmed. The output of the method is a score matrix F which represents the potential associations between miRNA and disease. Those three matrices, in addition to the transpose of matrix A, have been combined in a huge matrix then they multiplied this huge matrix by its transpose. After that, they got the diagonal elements of this huge matrix and got the inverse of the matrix and combined these values in a new diagonal matrix called WEIGHT. From this matrix, they derived two other matrices which their top-right part is the final output score matrix. Finally, the authors validated their approach on three diseases as case studies: Colon Neoplasms, Carcinoma Hepatocellular, Esophageal Neoplasms.

Figure 11 shows the general categories of computational microRNA-disease association findings models found in the literature. The circle denotes where the proposed model fits at the intersection between complex networks-based models and multiple biological information-based models.

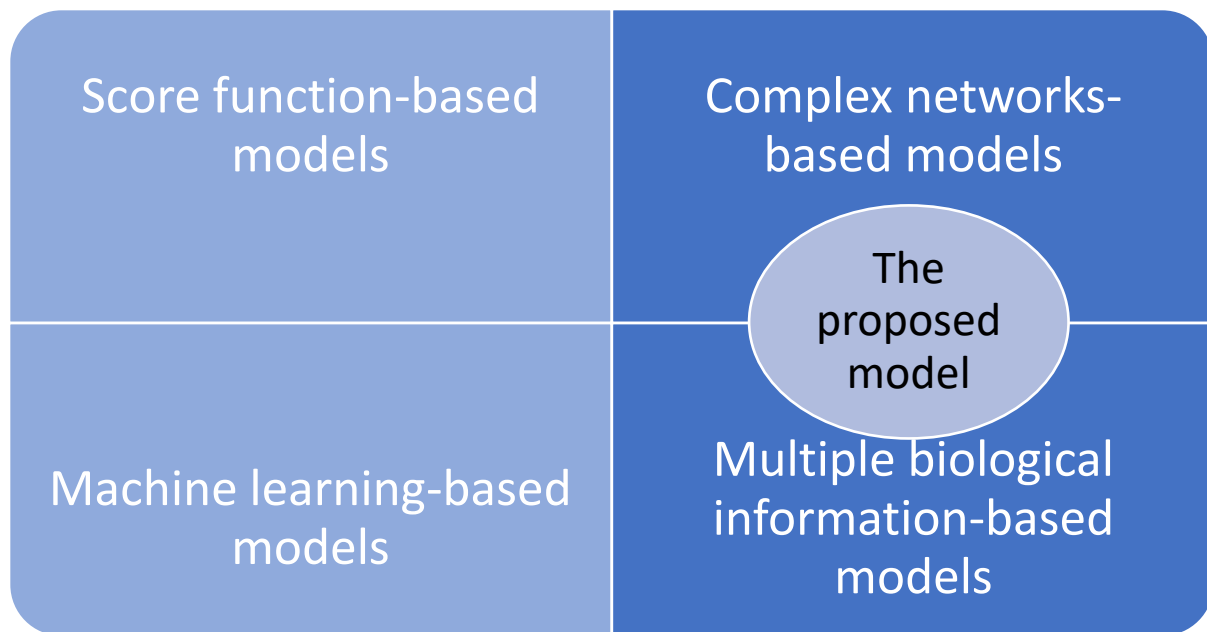


Figure 11 Categories of computational microRNA-disease association findings models

CHAPTER 4

PROPOSED APPROACH

In a nutshell, the proposed approach is based on utilizing graph databases and graph theory techniques in integrating biological data from different sources and discovering relationships between certain biological entities.

4.1 Questions Raised by Reviewing the Literature

The review of the literature on the microRNA-disease association prediction problem revealed a number of gaps in the existing research. In particular, it was found that certain techniques, such as graph modeling and analysis, and certain types of data, such as protein domain data, had not been fully utilized in previous studies. These findings led to the development of the proposed approach, which aims to address these gaps by exploring the potential use of graph modeling and analysis techniques in the context of the microRNA-disease association prediction problem, and by integrating protein domain data with other biological data. To achieve these goals, a number of questions were raised:

- How graph modeling can be used in the context of the microRNA-disease association prediction problem?
- How protein domain data can be integrated with other biological data in the context of the microRNA-disease association prediction problem?

These questions motivated the development of the proposed approach, which is described in the following sub-section. The proposed approach represents a novel approach to the microRNA-

disease association prediction problem that has the potential to significantly improve the accuracy and reliability of prediction by leveraging new techniques and data sources.

4.2 Proposed Research Approach

The core of the proposed research approach is centered around leveraging graph modeling techniques with the different available biological data sources. Having this goal in mind, the approach has been developed as depicted in the workflow shown in Figure 12.

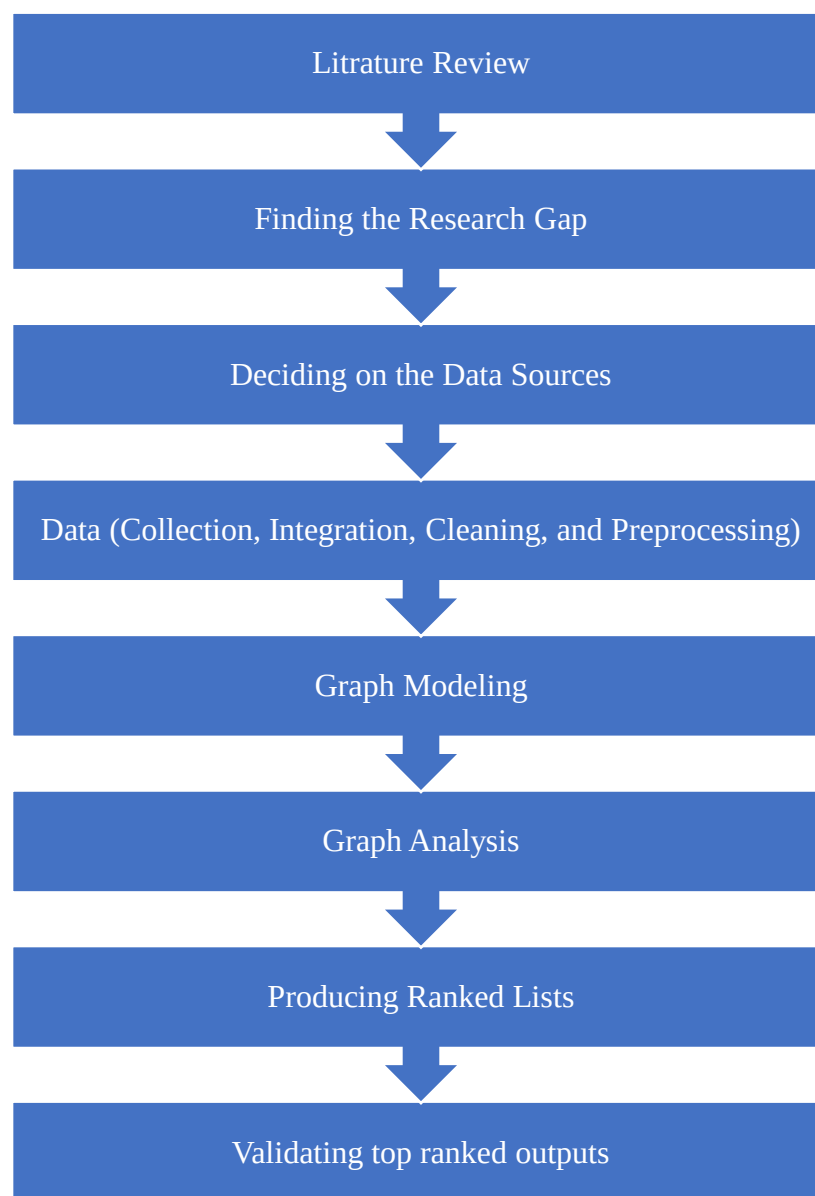


Figure 12 Workflow of the present work

The proposed approach to the microRNA-disease association prediction problem represents an important contribution to the field for a number of reasons. First, the use of graph modeling and analysis techniques allows for a more comprehensive and integrated analysis of the complex interactions between microRNAs and diseases. This can provide a deeper understanding of the mechanisms underlying microRNA-disease associations and may lead to the identification of novel therapeutic targets.

Second, the integration of protein domain data with other biological data allows for a more nuanced and multidimensional approach to prediction. Protein domains are important functional units of proteins that play crucial roles in various cellular processes, and their inclusion in the analysis of microRNA-disease associations can provide additional insights into the mechanisms underlying these associations.

Overall, the proposed approach has the potential to significantly improve the accuracy and reliability of prediction in the microRNA-disease association prediction problem and may inform the development of new therapeutic strategies and the design of more robust biological networks.

4.3 Data Sources

This work tries to explore data sources that have not been fully used to tackle the miRNA-disease association prediction problem. That way, new connections can be found between data entities in those data sources. The data sources utilized in this study, including their relevant details, are listed in Table 2.

Table 2 Data sources for different types of data that were used in the presented work

Data Source	Extracted Data	Brief Description
UniProt	Unique identifiers for proteins	UniProt is one of the main comprehensive biological databases that include a wide spectrum of biological data which provides, among others, annotated sequences of proteins, protein structure, protein function, and more [4].
Pfam	Protein data	Pfam is a database of protein domains. A protein domain is a structural unit of a protein. It is used to retrieve the domain compositions of our intended list of proteins [34].
Ensemble BioMart	Protein domain data	Ensemble BioMart is a web-based tool that allows retrieving biological data from several databases [35].
miRTarBase	miRNA target genes	miRTarBase is a curated database containing thousands of microRNA interactions with their target genes. The homo-sapiens data for microRNA-gene interaction was downloaded from there. It contains 2,600 unique miRNAs and 15,051 unique target genes, and a total of 502,653 interactions [36].

miRCancerdb	Negatively correlated microRNAs with their target genes in the BRCA study	miRCancerdb is a database that contains public data from TCGA [37] and annotations from TargetScan [38].
-------------	---	--

4.4 Data Collection, Integration, Cleaning, and Preprocessing

Figure 13 shows a high-level diagram for the data collection and data processing phases. The specification of each phase is explained in detail in the next chapter.

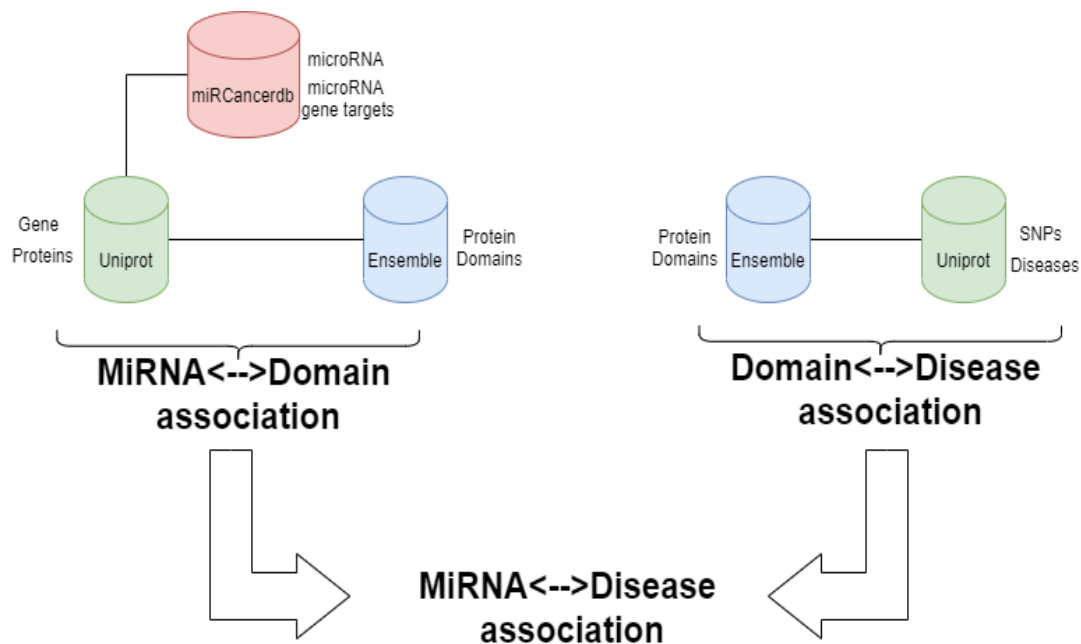


Figure 13 A high-level diagram for the data collection phase

4.5 Graph Modeling

Figure 14 illustrates the modeling of the domain-disease graph following the collection and processing phases described in previous sections. In this model, separate nodes are created for each protein, protein domain, disease, and gene, with each node containing properties such as its name and ID. Relationships between these nodes are then established based on the information collected from various data sources. For example, a gene is connected to a disease if they are found to be associated, and a protein is linked to its constituent genes. Similarly, protein domains are associated with the proteins that contain them.

The resulting graph model, shown in Figure 16, is a multi-bipartite graph with four layers: protein, protein domain, gene, and disease. These layers are connected to each other in the following ways: nodes in the gene layer are connected to nodes in the disease layer, and nodes in the protein layer are connected to nodes in the protein domain layer.

Figure 15 presents the final graph model construction scheme, which includes an additional layer for microRNA data after it has been preprocessed, integrated, and cleansed. This layer is used to perform all analyses related to microRNA-disease associations. Overall, this graph model provides a comprehensive representation of the relationships between proteins, protein domains, genes, diseases, and microRNAs, allowing for a more detailed and nuanced analysis of microRNA-disease associations.

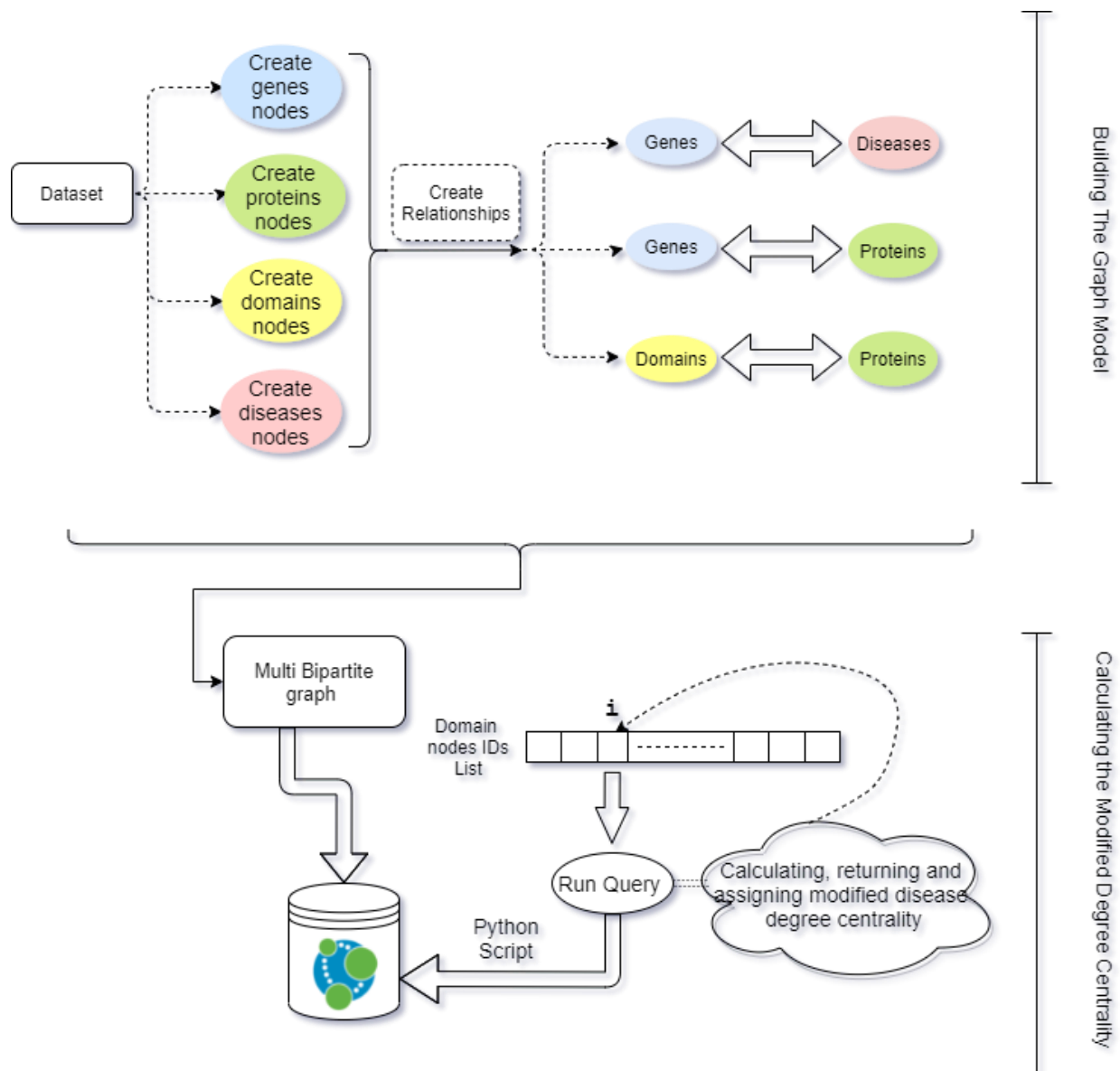


Figure 14 Modeling the domain-disease graph, adapted from [6]

ALGORITHM 1: ALGORITHM TO ENCODE THE DATA AND BUILD A GRAPH MODEL

Input: Preprocessed dataset

Output: Multi-bipartite heterogenous graph

```
1  While ! EOF
2      //Nodes
3      Create a unique node for every occurrence of gene identifier
4      Create a unique node for every occurrence of protein identifier
5      Create a unique node for every occurrence of protein domain identifier
6      Create a unique node for every occurrence of disease identifier
7      Create a unique node for every occurrence of microRNA identifier
8      // Relationships
9      Create relationship (protein domain  $\Leftrightarrow$  protein)
10         Encode the protein domain position(s) on the relationship
11      Create relationship (gene  $\Leftrightarrow$  protein)
12      Create relationship (microRNA  $\Leftrightarrow$  gene)
13         Encode the microRNA-gene correlation value on the relationship
14      Create relationship (disease  $\Leftrightarrow$  protein domain)
End
```

Algorithm 1 takes as input a preprocessed dataset and generates a multi-bipartite heterogenous graph as output. Inside the loop, the algorithm creates a unique node for each occurrence of a gene identifier, protein identifier, protein domain identifier, disease identifier, and microRNA identifier. These nodes represent the different types of entities present in the dataset. Next, the algorithm creates relationships between these nodes to represent the relationships and interactions between the entities. Specifically, it creates a relationship between protein domains and proteins, encoding the protein domain position(s) on the relationship. It also creates relationships between genes and proteins, microRNAs and genes (encoding the microRNA-gene correlation value on the relationship), and diseases and protein domains.

ALGORITHM 2: ALGORITHM TO CALCULATE THE DISEASE NODE DEGREE CENTRALITY

Input: Multi-bipartite graph (heterogenous or homogenous)
Output: Disease node degree centrality measure for protein domain/ microRNA of interest

- 1 *List* $L \leftarrow []$
- 2 *List* $X \leftarrow$ Protein domain or microRNA nodes
- 3 $D \leftarrow$ Disease of interest
- 4 **For** $i \in X$ **do**
- 5 $L.append(ID(i))$
- 6 **End**
- 7 $Path_length \leftarrow 3$ (protein domain case) or 5 (microRNA case)
- 8 **For** $i \in L$ **do**
- 9 $i.disease_degree_centrality \leftarrow$ number of paths $[i -(*0...Path_length)- D]$
- 10 **End**
- 11 Rank L
- 12 Report the top K nodes

Algorithm 2 calculates the disease node degree centrality measure for protein domains or microRNAs of interest in a multi-bipartite graph, which can be either heterogenous or homogenous. The input for the algorithm is the multi-bipartite graph, and the output is the disease node degree centrality measure for the protein domain or microRNA of interest. The algorithm first initializes two lists: L , which will be used to store the IDs of the protein domains or microRNAs, and X , which contains all of the protein domain or microRNA nodes. It then selects the disease of interest, represented by D .

Next, the algorithm iterates through each protein domain or microRNA in X and appends its ID to the list L . It then sets the path length to 3 for protein domains or 5 for microRNAs. The algorithm then iterates through each element in L and calculates the disease node degree centrality measure for each protein domain or microRNA. This is done by counting the number of paths that exist

between the protein domain or microRNA and the disease of interest, with a length of either 3 or 5, depending on the type of node. Finally, the algorithm ranks the elements in L and reports the top K nodes, which are the protein domains or microRNAs with the highest disease node degree centrality measure.

The workflow started with the curated dataset, and then we used it for building the graph model by creating five different types of graph nodes, namely (gene, protein, protein-domain, microRNA, and disease). Then, we created relationships between them as follows: relationships between genes nodes and diseases nodes, relationships between genes nodes and proteins nodes, relationships between proteins nodes and domains nodes, and relationships between microRNAs and genes. After that, this model is implemented in the Neo4j graph database. Then, the python script is used to calculate and assign the modified disease node degree centrality for each protein domain/microRNA node with respect to the disease of interest. To elaborate more on the two algorithms, we list below the key steps:

1. For each gene in the dataset, create a unique node that represents that gene and contains information about it.
2. Similarly, for each protein, protein domain, disease, and microRNA in the dataset, create a unique node for each one.
3. Create a relationship between each gene node and the corresponding protein node.
4. Similarly, for each corresponding (gene, disease), for each corresponding (protein, domain), and for each (microRNA-gene).

5. Initialize an empty list to store nodes' IDs in the next step.
6. Collect the IDs of protein domain/disease nodes in list L
7. Set the path length to three (in case of protein domain nodes), and to five (in case of microRNA nodes)
8. For each protein domain/microRNA node, calculate the number of paths with length three or five from it to the disease node of interest (D).
9. Assign the disease node degree centrality score that was calculated at step 8 to the corresponding protein domain/microRNA node.
10. Rank list L
11. Report the K top nodes (protein domains or microRNAs)

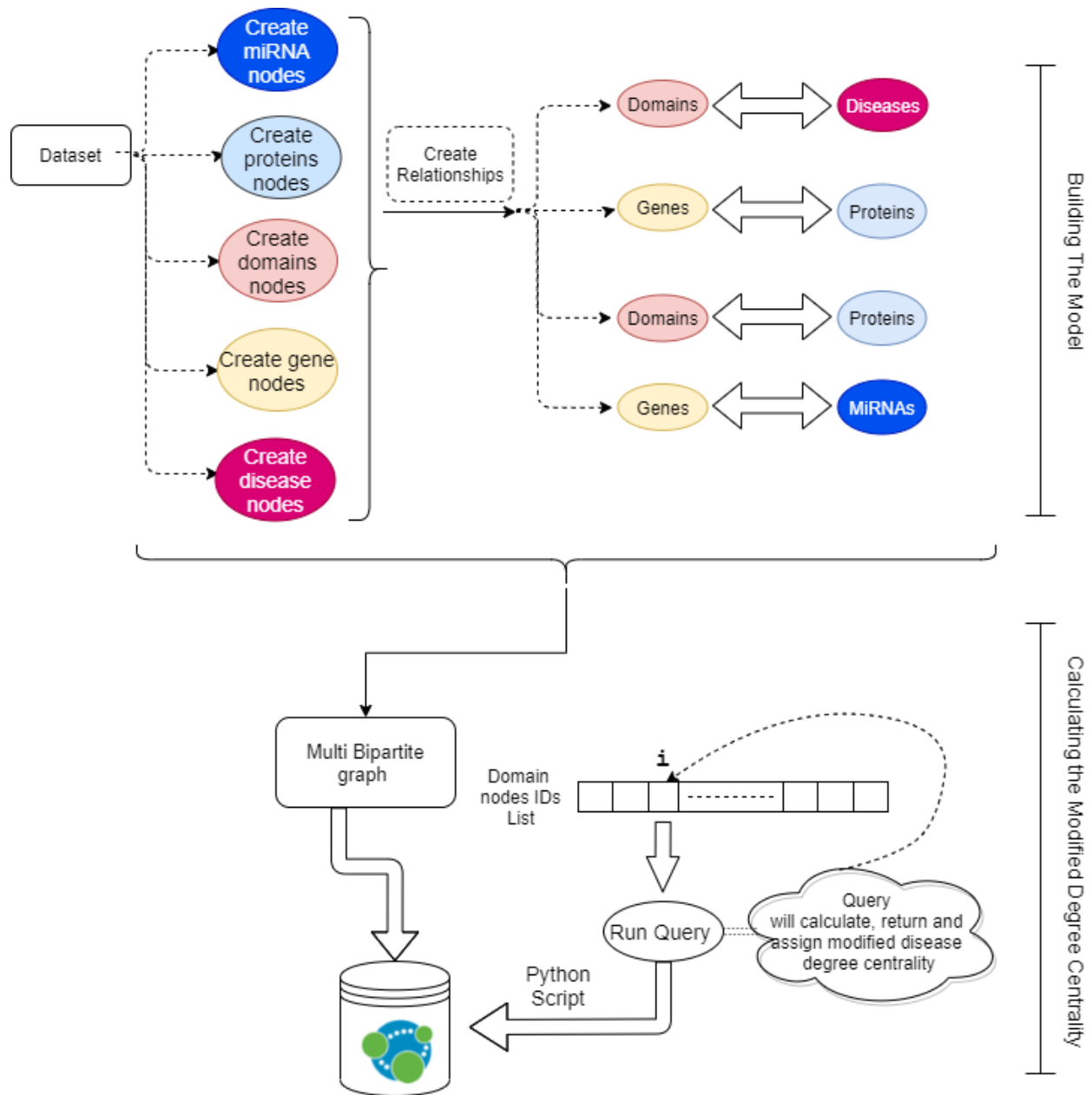


Figure 15 Modeling the final graph

Figure 15 shows the modeling of the microRNA-disease graph after the collection, integration, and processing phases as mentioned in the previous sections. That graph model is considered as an extension of the previous one. The addition of microRNA data enriched the graph model as it has widened the scope of the current work to include the connections between microRNAs and their target genes. It is worth mentioning here that we focused the work on the set of microRNAs

related to breast cancer disease. As an extension of the previous algorithm mentioned above, we included the microRNA layer in the graph and the set of connections between microRNA nodes and their target genes.

4.6 Graph Analysis

In this study, network analysis techniques and node centrality measures were employed in order to analyze the relationships between proteins, protein domains, genes, diseases, and microRNAs. Network analysis involves the study of complex systems composed of interconnected elements, and is commonly used in fields such as biology, sociology, and computer science. Network analysis techniques can provide insights into the structure and function of these systems by examining the patterns of connections between the elements.

Node centrality measures, on the other hand, are used to assess the importance or influence of individual nodes within a network. There are several different types of node centrality measures, including degree centrality, betweenness centrality, and closeness centrality, each of which captures different aspects of node importance. By applying network analysis techniques and node centrality measures to the proposed graph models, it is possible to identify key nodes and understand the underlying structure of the system. These techniques and measures are described in more detail in the following sections.

4.7 Centrality Measures

In this study, a new centrality measure was developed in order to better capture the interrelated nature of the different biological components in the proposed graph models. Centrality measures are used to assess the importance or influence of individual nodes within a network, and there are several different types of centrality measures available. However, the existing measures may not

adequately capture the complex relationships between the various elements in the domain-disease graph. To address this issue, a new centrality measure was developed based on the intuition that the different biological components are interrelated and that their relationships should be taken into account when assessing their importance within the system. This new centrality measure was designed to consider not only the direct connections between nodes, but also the indirect relationships that may exist through other nodes in the network. By using this new centrality measure, it was possible to gain a more comprehensive understanding of the role played by each node in the proposed graph models. This new measure was then used in the next chapter to analyze the microRNA-disease associations.

The *modified node degree centrality* measure is a centrality measure developed specifically for this study to assess the relative importance of protein domain nodes and microRNA nodes in relation to disease nodes in the domain-disease graph. It is based on the concept of node degree centrality, which is a measure of the number of edges connected to a node in a graph. In this case, the modified node degree centrality measure is used to identify the associations between protein domains (or microRNAs) and diseases by traversing the different layers of the graph model and counting the number of paths between the protein domain (or microRNA) nodes and the disease nodes they are linked to.

This measure is not specific to the particular graph model used in this study but can be applied and generalized to any other multi-bipartite graph model. It allows for the identification of the relative importance of protein domain nodes and microRNA nodes in relation to disease nodes and can be used to inform the development of therapeutic strategies targeting specific microRNAs or protein domains. Overall, the modified node degree centrality measure provides a useful tool for

understanding the complex relationships between different biological components in the domain-disease graph.

The *modified node degree centrality* measure is used to show the relative importance between two layers in the graph model. In the case of a protein domain node, for each protein domain node in the protein domain layer in the graph, we traverse the other layers of the graph to find all the disease nodes that are linked to that protein domain node, then we return the total number of paths between that protein domain node and all the disease that it is linked to. In general, that kind of traversing is not fixed with a certain length as the number of layers in the graph can be variable. One of the features of that measure is that it is not specific to the graph model that we are building here, however, it can be used and generalized to any other multi-bipartite graph model.

4.8 Modified Node Degree Centrality

The modified node degree centrality measure is a centrality measure that can be generalized for use in any heterogeneous network or graph. We implemented it using the Neo4j graph database and its graph query language, Cypher, to calculate the relative importance of a node in one set of nodes (S_i) with respect to another set of nodes (S_j). This measure is particularly useful in situations where it is necessary to understand the relationships between nodes in different sets and can be applied to a wide range of scenarios where the concept of modified node degree centrality is relevant.

To calculate the modified node degree centrality measure for a node in S_i , the variable-length path support provided by the graph structure in the database is used to traverse the graph and count the number of paths between the selected node and all the nodes in S_j that it is linked to. This process can be repeated for all nodes in S_i , resulting in a modified node degree centrality measure for each node that can be compared to assess the relative importance of different nodes within the network.

By using the modified node degree centrality measure, it is possible to gain a more comprehensive understanding of the relationships between nodes in different sets within a heterogeneous network or graph.

CHAPTER 5

EXPERIMENTS, RESULTS & DISCUSSION

Two main experiments were conducted as part of this study. The first experiment focused on the relationships between protein domains and diseases, while the second experiment built upon the findings of the first by extending the graph model to include additional data and considering the interconnected nature of biological components. In the second experiment, we also explored different variations in the data by manipulating subsets of the data and analyzing the results based on the changes in the graph model. Both experiments utilized a graph database to store and manage the data, as it is well-suited for representing and analyzing complex, interconnected systems.

The details and results of each experiment are described below. The first experiment examined the relationships between protein domains and diseases, and the second experiment extended the analysis to include additional data about microRNAs and consider the interconnected nature of these biological components. Overall, these experiments provide valuable insights into the complex relationships between different biological components and inform the development of therapeutic strategies targeting specific microRNAs or protein domains. [39], [40].

5.1 Protein Domain-Disease Graph Model

A significant area for research in the biomedical field shows that domain-disease associations are important in many different ways [41], [42]. A dysfunction in domains may affect the function of proteins in some diseases. Moreover, drug-domain interactions play a big role in determining protein-drug interactions [42]. Based on the foregoing, we found that identifying associations

between diseases and domains may contribute to certain areas including, drug design and discovery [43].

Data on diseases and their related genes were collected from DisGeNet [25], for this study. A total of 9,413 genes corresponding to 10371 diseases were collected from DisGeNet. Data on the relationships between genes and proteins, and proteins and domains, were also obtained from the UniProt database[44] this resulted in a total of 20,405 proteins and 33,741 domains. These data were used to build the domain-disease graph model and analyze the relationships between different biological components.

The process of collecting the data goes as follows. Firstly, the curated gene-disease associations have been downloaded from DisGeNet, then the related proteins were retrieved from the UniProt database using the genes from DisGeNet. Finally, the domains corresponding to the resultant proteins have also been retrieved from the UniProt database. The data has been cleaned and imported to a graph database management system as mentioned below.

We used Neo4j as the graph database management system. Neo4j allows the implementation of the property graph model. Our graph can be viewed as a 4-partite heterogeneous graph. The graph contains four groups of nodes (genes, proteins, domains, and diseases). The edges of the graph are of three different types. An edge (relationship) between gene node and protein node (modeled as GPR, gene-to-protein relation) exists if this gene contributes to the production of that protein. An edge between gene and disease GDR, (gene-to-disease relation) exists if this gene causes that disease, particularly, if a mutation in that gene causes that disease. An edge between protein and domain PDR, (protein-to-domain relation) exists if this protein contains that domain.

Figure 16 shows the graph model. In that model, we have four types of nodes; each one represents a set of nodes of its category. In another word, the red node represents a whole set of 7,446 red nodes, each one representing a disease from the DisGeNet database. The blue node is representative of a whole set of 3,910 blue nodes, each one represents a gene from the DisGeNet database.

The green node is a representative of a whole set of 3,907 nodes, each one represents a protein from the UniProt database. The orange node is a representative of a whole set of 3,895 nodes each one represents a domain from the UniProt database.

Eventually, we have a huge heterogeneous graph of four different types of nodes represented as shown in Figure 16. In that figure, red nodes indicate diseases; blue nodes indicate genes; green nodes for proteins and orange ones indicate protein domains.

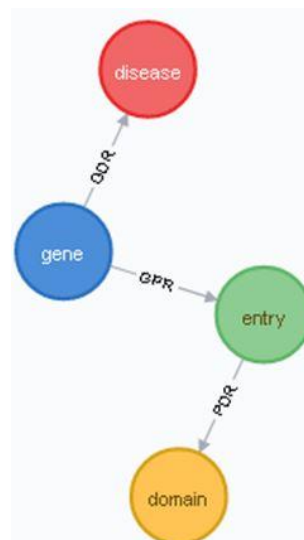


Figure 16 The first graph model.

Neo4j supports the concept of “Variable Length Path” which allows retrieving patterns between any two nodes in the graph without the need to specify the length of that path, i.e., no need to mention if those two nodes are one, two, or even X edges apart. However, one can limit the length of the retrieved path once there was a need to. We are showing an example of how we leveraged this feature later in this thesis.

After importing and integrating the collected data, the domain-disease graph model was constructed with 3,907 protein nodes, 7,446 disease nodes, 3,910 gene nodes, and 3,985 domain nodes. The graph also contained 101,034 edges connecting gene nodes to disease nodes, 7,820 edges connecting gene nodes to protein nodes, and 8,242 edges connecting protein nodes to domain nodes. The statistics of the graph, as retrieved from Neo4j, are shown in Table 3.

Table 3 First graph model statistics

	Type	Count
Nodes	Protein	3,907
	Disease	7,446
	Gene	3,910
	Domain	3,985
Edges	Gene-disease (GDR)	101,034
	Gene-protein (GPR)	7,820
	Protein-domain (PDR)	8,242

One of the objectives of this study was to identify associations between diseases and domains. To achieve this, we focused on domains and developed a measure of their importance relative to diseases. Specifically, we sought to understand how important a domain node is in relation to all disease nodes in the graph. To this end, we introduced a modified version of the well-known concept of node degree centrality. This modified node degree centrality measure was used to assess the relative importance of domain nodes in relation to disease nodes in the domain-disease graph. For each domain node, we calculate what we called, “*disease node degree centrality*” which shows how many diseases are linked to a certain domain.

Figure 17 shows an illustrative example for a domain node with a disease node degree centrality score equal to eleven. In the figure, we have the domain 32 181 *IIRG-type G domain* that is part of the *A1A4Y4* protein which is a product of the *IRGM* gene. Edges between the *IRGM* gene and diseases, which are from DisGeNet, indicate that some mutation(s) in the *IRGM* gene can cause any of the eleven connected diseases.

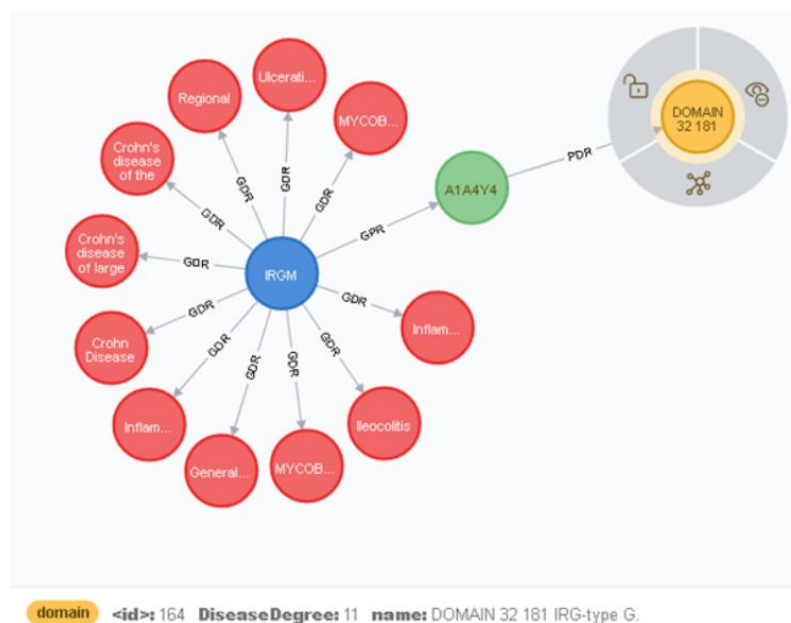


Figure 17 An illustrative example of a domain node.

Further investigation of the domain identified in this study could involve the following approaches:

- Given that the eleven diseases in question are all related to this domain, it is possible that they share some similarities. This could potentially make a known drug for one of the diseases a good candidate for treating another disease, possibly with some modification.
- It is also possible that this domain has not previously been considered as a drug target for any of the eleven diseases. The results of this study highlight the potential value of exploring this domain as a candidate drug target for one or more of these diseases.

Figure 18 shows a snapshot from the graph showing the top two protein domain nodes with a score of 144 disease node degree centrality for both of them.

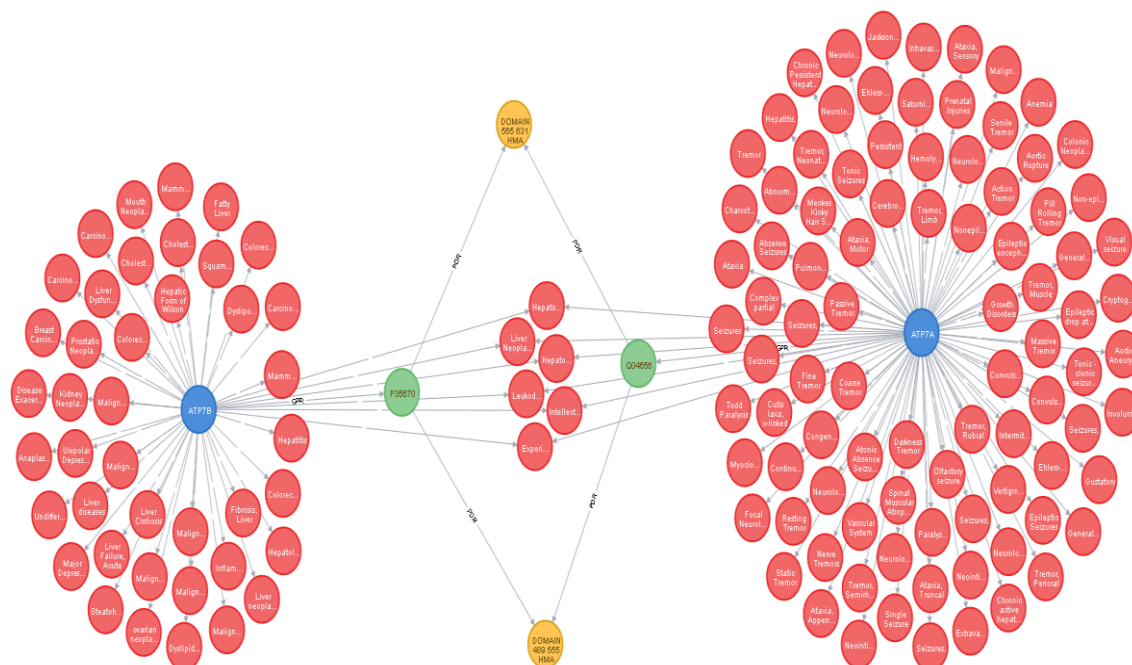


Figure 18 The two protein domain nodes with the highest disease node degree centrality score

5.2 Associations using SNPs

To further strengthen the associations identified between protein domains, it is possible to incorporate additional data, such as Single Nucleotide Polymorphisms (SNPs) and non-synonymous SNPs (nsSNPs). This type of data can be obtained from sources such as UniProt

(version: October 2022) [4]. and imported into the analysis. Table 4 provides an example of the data that can be obtained from UniProt. By incorporating this type of data, it may be possible to, more confidently, identify associations between protein domains and diseases.

Table 4 Sample for SNPs data from Expasy/Uniprot

Main Gene Name	Swiss-Prot AC	FTID	AA Change	Variant Category	dbSNP	Disease Name
A1BG	P04217	VAR_018369	p.His52Arg	LB/B	rs893184	-
AAAS	Q9NRG9	VAR_012804	p.Gln15Lys	LP/P	rs121918549	Achalasia-addisonianism-alacrima syndrome (AAAS) [MIM:231550]
ABCA3	Q99758	VAR_084242	p.Arg280Cys	US	-	Pulmonary surfactant metabolism dysfunction 3 (SMDP3) [MIM:610921]

The data used in this section is summarized in Table 4. Before analysis, the data was cleaned to remove any invalid or incomplete records. Only records that were classified as "likely pathogenic or pathogenic" were retained, resulting in a total of 23,352 records. This cleaned and processed data was then used to identify associations between protein domains and diseases.

As an example, from the data presented in the table above, the nsSNP rs121918549 causes a change from the amino acid His to Arg at position 52 of the protein with entity ID "P04217". To integrate nsSNP data with domain data, we sought to determine which protein domains are located on the protein "P04217" and, specifically, which domain contains location 52. This information was obtained using BioMart [45], a database that provides data on proteins and their associated protein domains, including their locations on the proteins. By linking nsSNPs to the protein domains in which they occur, we can further investigate the associations between protein domains and diseases.

5.3 Disease-MiRNA Associations

To identify associations between microRNAs (MiRNAs) and diseases, particularly cancer diseases, we are utilizing several types of intermediate biological data, including protein domains, non-synonymous Single Nucleotide Polymorphisms (nsSNPs), genes, and proteins. By analyzing these intermediate data points, we can better understand the relationships between MiRNAs and diseases.

To begin this analysis, data on nsSNPs and the diseases they may cause is collected and cleaned. Next, the impact of these mutations on the protein chain is traced based on the domain coordinates within that protein chain. If the effect of an nsSNP on a disease can be identified within the coordinates of a particular domain in a protein, that domain is linked to the disease. In this way, new associations between protein domains and diseases are identified.

MicroRNA data is more complex and requires multiple attempts to obtain reliable, experimentally validated information. MiRNA is a gene suppressor, meaning it inhibits the expression of a gene and prevents it from participating in the central dogma of molecular biology by blocking the production of proteins. As a result, there is a strong connection between miRNA and proteins. To

further our analysis, a set of miRNAs with negative correlations with their target genes was identified. In addition, data on human genes and the proteins they produce, including their building blocks (i.e., domains), was included in the analysis. This allowed us to establish associations between domains and miRNA. By combining this information with the previous analysis on protein domains and diseases, we were able to identify associations between miRNA and diseases. As an example, a sample of the data collected on diseases and nsSNPs is provided below.

Table 5 shows the statistics for single amino acid variants.

Table 5 Statistics for single amino acid variants

LP/P	LB/B	US	Total
32081	39607	9684	81372

5.4 Protein-Domain Data

As we have mentioned above, the domain is the structural building block for proteins. A change in one of the protein's domains can lead to a change in the final structure of the protein. So, it was necessary to consider the relationship between domains and proteins. For that, data about domains and proteins have been collected from BioMart [45] and Pfam [34].

In Table 6, data about protein domains are represented. Each protein domain is being presented as a Pfam ID with its coordinates (Pfam start, and Pfam end).

Table 6 Data from BioMart and Pfam

Gene stable ID	Gene name	UniProtKB/Swiss-Prot ID	Pfam ID	Pfam start	Pfam end
ENSG00000198888	MT-ND1	P03886	PF00146	6	304
ENSG00000198763	MT-ND2	P03891	PF00361	23	284
ENSG00000198763	MT-ND2	P03891	PF06444	290	343
ENSG00000198804	MT-CO1	P00395	PF00115	14	460

5.5 MicroRNA-Domain Graph Model

Figure 19 shows the microRNA-domain multipartite graph model. In that graph model, the red node represents a layer of protein domain nodes (2,556 distinct nodes), the cyan node represents a layer of protein nodes (3,651 distinct nodes), the orange node represents a layer of gene nodes (3,646 distinct nodes), and the brown node represents microRNA nodes (148 distinct nodes).

The microRNA-domain graph model is characterized by the statistical data presented in tables 7 and 8.

Table 7 Node statistics for microRNA-domain graph model

Node Type	Count
Protein	3,651
Protein domain	2,556
Gene	3,646
MicroRNA	148

Table 8 Edge statistics for microRNA-domain graph model

Edge Type	Count
PART_OF	15,517
BRCA_VE_CORR	14,487
PRODUCE	3,651

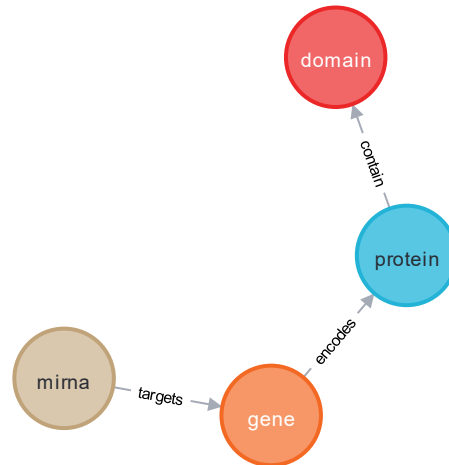


Figure 19 The MicroRNA-Domain graph model

5.6 The Final Graph Model

Figure 20 shows the schema for the final graph model. A model of a 5-layer (multi-bipartite) graph. The blue nodes/layer represents microRNAs; the orange nodes represent genes; the cyan nodes represent proteins; the red nodes represent protein domain; the pink nodes represent diseases. Below are some statistics for the graph.

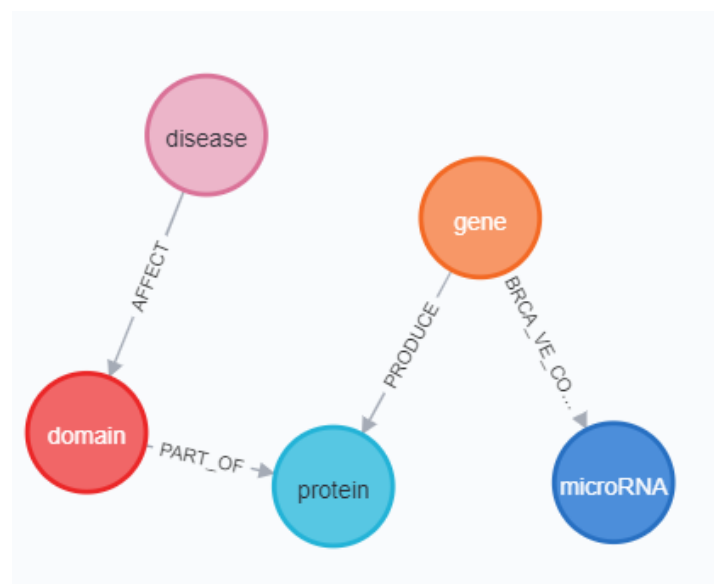


Figure 20 The final graph model

5.7 Analysis of the Final Graph Model

After constructing the graph model, several analyses were performed, among them is applying the modified node degree centrality on the graph to find the relevant importance of specific nodes, which can be considered as an indicator to predicting a link between pairs of nodes in the graph.

Table 9 and Table 10 show some statistics about the final graph model and Figure 21 shows an example of applying the modified node degree centrality on a pair of miRNA and disease nodes. In the figure, the number of nodes has been limited to 20 for the sake of clear visualization.

Table 9 Statistics for the final's graph model relationships

Relationship Type	Count	From	To
PART_OF	10,495	Domain	Protein
PRODUCE	2,617	Gene	Protein
BRCA_VE_CORR	10,441	Gene	microRNA
AFFECT	3,288	Disease	Domain
Total	26841		

Table 10 Statistics for the final's graph model nodes

Node Type	Count
Protein	2,617
microRNA	148
Domain	932
Gene	2,616
Disease	2,522
Total	8,835

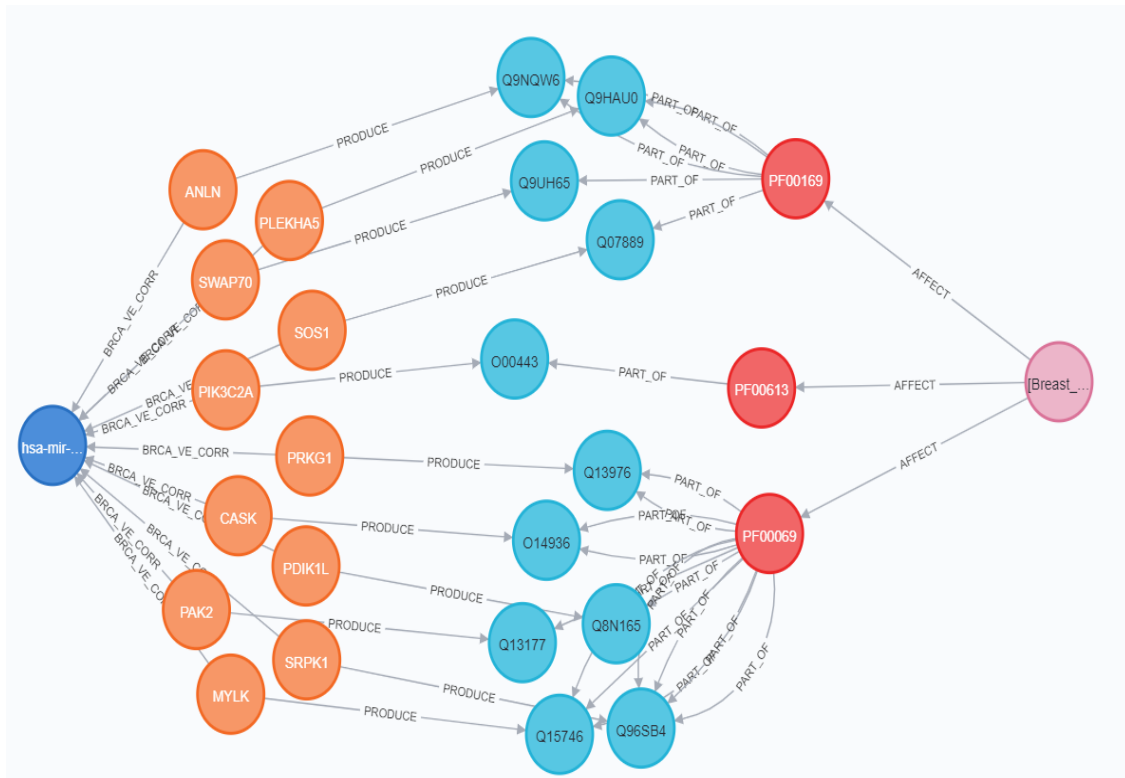


Figure 21 A pair of microRNA and disease nodes

5.7.1 Node Degrees by Label

One of the measures that determine the importance of a node in the graph is its degree. Node degrees can be divided into two categories: in-degree and out-degree.

To find the importance of nodes in the graph, taking into consideration the type of the node, the degree for all the nodes has been calculated. Table 11 shows the top five nodes with the highest degrees are listed for the five types of nodes in the graph.

Table 11 Top five disease, protein domain, protein, genes, miRNA nodes

Label	Node Name	Degree
Disease	[Breast_cancer_(BC)_[MIM:114480]]	8
	[Glanzmann_thrombasthenia_(GT)_[MIM:273800]]	7
	[Colorectal_cancer_(CRC)_[MIM:114500]]	7
	[Malignant_hyperthermia_1_(MHS1)[MIM:145600]]	7
	[[Central_core_disease_of_muscle_(CCD)_[MIM:117000]]	6
Label	Node Name	Degree
Domain	PF00096	924
	PF00400	431
	PF00041	311
	PF12796	270
	PF00069	256
Label	Node Name	Degree
Protein	P02751	82
	Q96RW7	57
	Q14766	53
	P35555	53
	P98164	53
Label	Node Name	Degree
Gene	CBX5	30
	ANKRD52	30
	STRN	29
	STYX	28
	UBE2W	28

Label	Node Name	Degree
MiRNA	hsa-mir-15a	313
	hsa-mir-200c	305
	hsa-mir-429	299
	hsa-mir-200b	298
	hsa-mir-19a	291

5.7.2 Protein Hub Nodes

Imbalanced datasets are a common challenge in the field of biology, particularly when dealing with protein data. This is often due to the fact that some proteins have received more attention from researchers, leading to a disproportionate amount of data being collected about them. As a result, biological networks often exhibit a skewed distribution of protein node degrees, with a few nodes having significantly higher degrees compared to the majority of the nodes. These nodes, known as "hubs," play a crucial role in the structure of the overall network due to their high number of connections.

There is no widely accepted definition or threshold for determining whether a protein node should be considered a hub in a biological network. Different studies have used various methods and criteria to identify hub proteins, and there is no one-size-fits-all approach. In fact, many researchers have relied on network or graph analysis to determine a suitable threshold value for identifying hub proteins in a specific network. This approach involves analyzing the characteristics of the network and determining a threshold based on the specific features of that particular network.[46]. In this context, it is important to note that the definition of a hub protein may vary depending on the specific context and the goals of the analysis. For example, a protein node with a high degree may be considered a hub in one study, while in another study it may not be considered as a hub.

Figure 22 show the degree distribution of protein nodes. It is apparent that there are few hub protein nodes in the graph.

From analyzing the nodes in the top percentiles of node degrees, we found that there are 14 protein nodes that have connections more than 99.5% of the other protein nodes in the graph. Thus, those nodes have been considered as hubs in the first experiment setup. The same process was repeated for different percentiles, as shown in section 5.11.

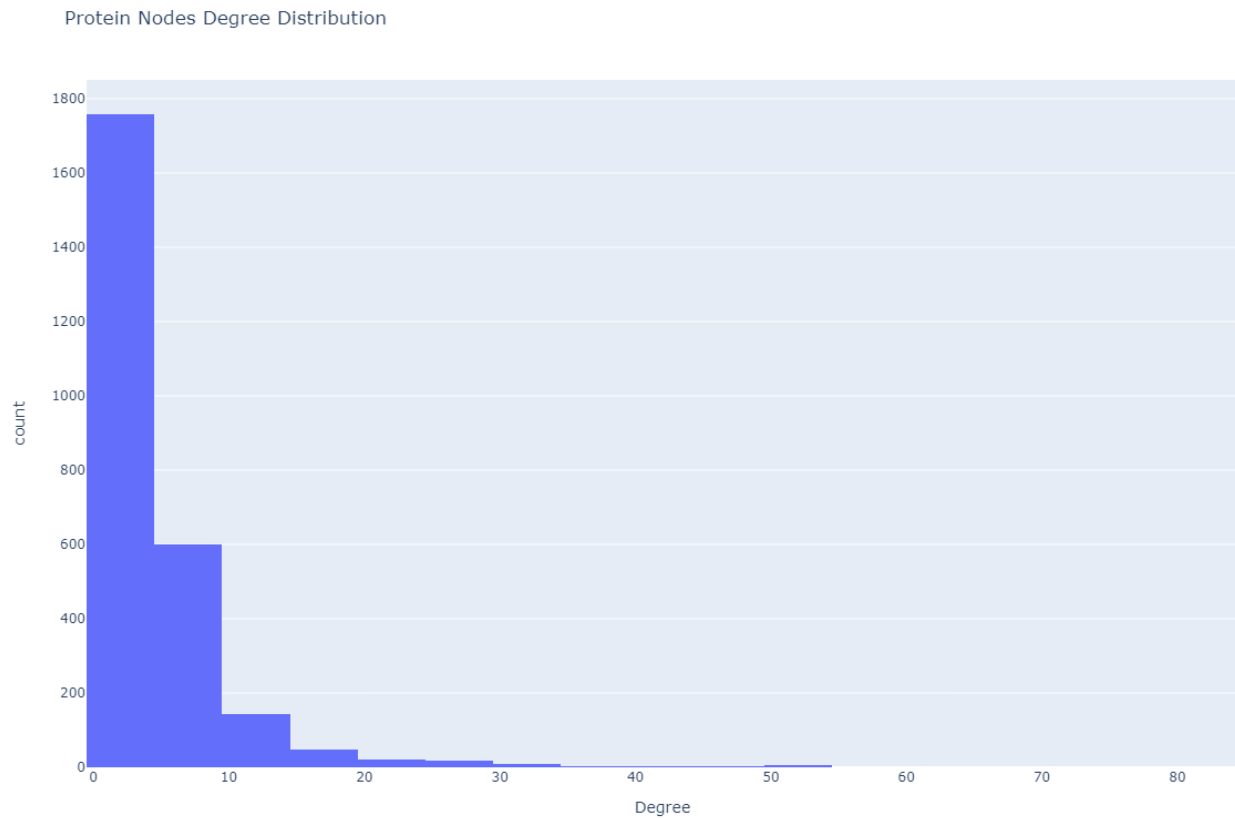


Figure 22 Protein nodes degree distribution

5.8 MicroRNA - Disease Node Degree Centrality

The devised measure has been used to get the relationships between microRNAs and diseases.

Figure 23 shows an example from the graph where the presented microRNA (hsa-mir-148a) has a disease node degree centrality of sixteen, and it is connected to the *Conotruncal_heart_malformations* disease.

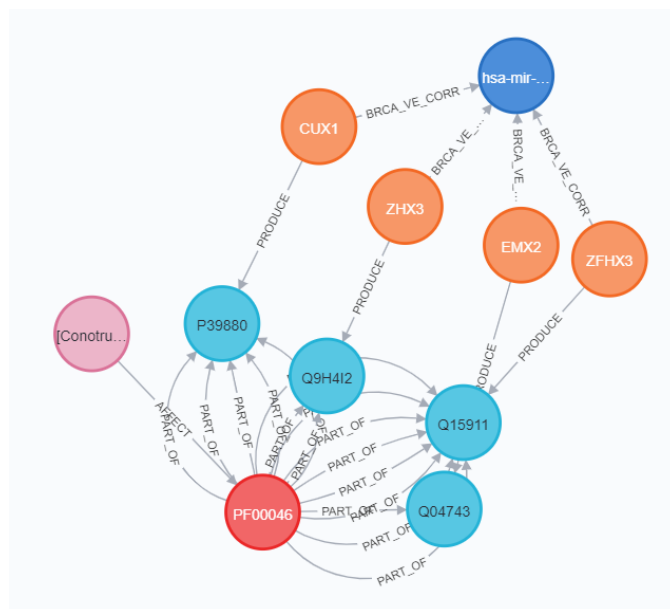


Figure 23 An example of microRNA with a score of 16 disease node degree centrality

The intuition behind the use of the disease node degree centrality measure here is that having a high disease node degree centrality for certain microRNAs can be an indicator of the strength of the relationship between this microRNA and that disease. Thus, that specific microRNA should be proposed to be on the top of the priority list of microRNAs to be further investigated by the drug design practitioners in the drug discovery field who are trying to understand the mechanism of that disease.

5.9 MicroRNAs-Domain Associations in Breast Cancer

Our aim from this step is to reach a list of pairs of genes and miRNAs wherein one pair of the microRNA is targeting the corresponding gene, and both are found to be negatively correlated in the breast cancer study. We achieved that, and we reached 15,276 pairs.

Following, we are describing the steps that we followed to achieve that. Firstly, we headed towards National Cancer Institute Genomic Data Commons (NCI GDC) [47] to collect and analyze the data, then we found that paper [48] that performed an elegant correlation analysis to the data there. So, we relied on it for that step.

MiRCancerDB [48] was used to retrieve the microRNAs and their target genes in the breast cancer study. Among different tables, there is the ‘cor_mir’ table where the columns are as follows: feature stands for the gene/protein, mirna_base stands for the microRNA ID and the rest of the columns are for different types of cancer (e.g., breast cancer). Each row represents the correlation between that pair of (gene and miRNA) with the correlation value that corresponds to the related cancer type. Of course, the correlation value can be positive or negative).

A statical description will be given later to analyze the data set. For now, we are giving the result that we reached after preprocessing and analyzing the data. Below is a snapshot from the final list of 15,276 pairs.

In Table 12 the *hsa-mir-98* microRNA is targeting the *AAK1* gene, and they have been found to be negatively correlated in the context of breast cancer study with a correlation value of -13.

Table 12 A snapshot of gene-miRNA association in breast cancer study

Feature	Mirna_base	BRCA
AAK1	hsa-mir-98	-13

Next, we show the statistics for microRNA data. The dataset contains a total of 677,459 associations between 15,277 distinct genes and 723 distinct microRNAs. Also, there are a total of 234,701 target relationships between 10,476 distinct genes and 186 distinct microRNAs. As an output of the integration phase, we had 15,276 associations between 3,870 distinct genes and 148 distinct microRNAs.

Next, we show the statistics for the protein data. The integration phase here started with 3,782 distinct genes and 3,789 proteins distinct, and we had a total of 3,789 gene-to-protein relationships. Next, we show the statistics for protein domains. We started with 6,499 unique protein domains and 19,300 unique proteins. Then we have 113,702 relationships between the proteins and domains. Then we integrated genes from the previous integration step with the protein and protein domains, we got a total of 15,517 relationships between 3,646 unique genes, 3,651 unique proteins, and 2,556 unique protein domains.

5.10 Calculating the Disease Node Degree Centrality: Breast Cancer as a Case Study

As a demonstration of the proposed work, breast cancer was used as a case study. The disease node degree centrality was calculated for its node in the graph against the 148 microRNAs.

Table 13 shows the top 20 microRNAs with the highest disease node degree centrality. The complete list of disease node degree centrality scores for each microRNA node with respect to breast cancer is provided in Table 16 under the appendix section.

Table 13 Top 20 ranked microRNAs with their disease node degree centrality calculated for breast cancer

Index	microRNA	Disease node degree centrality
1	hsa-mir-93	59
2	hsa-mir-15a	58
3	hsa-mir-429	50
4	hsa-mir-20a	50
5	hsa-mir-19a	48
6	hsa-mir-200c	46
7	hsa-mir-17	43
8	hsa-mir-340	42
9	hsa-mir-106b	39
10	hsa-mir-200b	36
11	hsa-mir-182	34
12	hsa-mir-29c	31
13	hsa-mir-32	31
14	hsa-mir-96	30
15	hsa-mir-27a	27
16	hsa-mir-130b	26
17	hsa-mir-30a	25
18	hsa-mir-301b	25
19	hsa-mir-454	24
20	hsa-mir-181d	24

5.11 Studying the effects of change in the data on the proposed approach

While the biological interactions within a cell set limits on our graph modeling, it is important to follow these rules while still adjusting certain parts of the data. This ensures that the modeling adheres to the principles of biology. Several studies in literature demonstrate the impact of altering certain aspects of a biological network's structure on certain properties of the network. This highlights the importance of considering the effects of partial changes in biological network structure.

An earlier assumption was that the measure in question would be effective for both balanced and imbalanced datasets. To test this assumption, we employed a common technique used in similar computational studies by identifying a threshold for node degrees and considering nodes with degrees above this threshold as hub nodes. This allowed us to filter out these hub nodes and examine the effectiveness of the measure in both balanced and imbalanced datasets.

5.11.1 Dropping Hub Protein Nodes

Hub nodes and edges play important roles in the structure and function of a graph, and their removal can potentially impact the results obtained using our approach. The protein nodes with degree above a certain threshold represent the protein nodes that have a high number of connections to other nodes in the graph. These nodes are often referred to as "hub" nodes. Table 17 in the appendix section shows the number of (hub) protein nodes with degree greater than different nodes percentile values.

Therefore, to explore the influence of hub protein nodes on the disease degree centrality of microRNAs, we conducted several variations of the experiment by dropping different percentages of hub protein nodes and analyzed the resulting changes. The summary of results of these

experiments are presented in Table 14 which shows the disease degree centrality of various microRNAs (rows) at different percentile thresholds of protein nodes with degree above that threshold (column headers).

The percentile thresholds in the columns of the table represent the degree threshold above which a certain percentage of protein nodes are removed from the graph. For example, the 99.50% column represents the disease degree centrality of the microRNAs when the protein nodes with degree above the 99.50th percentile were removed from the graph. This means that only the protein nodes that have a higher degree than 99.50% of the nodes in the graph are removed.

The values in the table represent the disease degree centrality of the microRNAs after the protein hub nodes with degree above the specified percentage of the rest of graph nodes were removed from the graph. For example, the value of 59 in the hsa-mir-93 row and the 99.50% column indicates that the disease degree centrality of hsa-mir-93 was 59 after the protein nodes with degree above the 99.50th percentile were removed from the graph. This means that only the protein nodes that have a higher degree than 99.50% of the nodes in the graph were removed.

Table 14 Disease degree centrality for microRNAs with different percentage of hub protein nodes got removed from the graph

	Protein Nodes with Degree Above X Percentile Got Removed									
X Percentile MicroRNA	99.50%	99.00%	98.50%	98.00%	97.50%	97.00%	96.50%	96.00%	95.50%	95.00%
hsa-mir-93	59	59	59	59	59	59	59	59	59	59
hsa-mir-15a	58	58	58	55	55	55	55	55	55	55
hsa-mir-429	50	50	50	47	47	47	47	47	47	47
hsa-mir-20a	50	50	50	50	50	50	50	50	50	50
hsa-mir-19a	48	48	48	48	48	48	48	48	48	48
hsa-mir-200c	46	46	46	43	43	43	43	43	43	43
hsa-mir-17	43	43	43	43	43	43	43	43	43	43
hsa-mir-340	42	42	42	42	42	42	42	42	42	42
hsa-mir-106b	39	39	39	39	39	39	39	39	39	39
hsa-mir-200b	36	36	36	33	33	33	33	33	33	33
hsa-mir-182	34	34	34	34	34	34	34	30	30	30

hsa-mir-29c	31	31	31	31	31	31	31	31	31	31
hsa-mir-32	31	31	31	31	31	31	31	31	31	31
hsa-mir-96	30	30	30	30	30	30	30	26	26	26
hsa-mir-27a	27	27	27	27	27	27	27	27	27	27
hsa-mir-130b	26	26	26	26	26	23	23	23	23	23
hsa-mir-30a	25	25	25	25	25	25	22	22	22	22
hsa-mir-301b	25	25	25	25	25	22	22	22	22	22
hsa-mir-454	24	24	24	24	24	24	24	24	24	24
hsa-mir-181d	24	24	24	24	24	24	24	24	24	24
Disease Degree Centrality										

Upon examining the table, it can be observed that some microRNAs have more consistent disease degree centrality values across different percentile thresholds, while others show more variability. For example, the disease degree centrality of hsa-mir-93 remains consistently at 59 across all percentile thresholds, indicating that the association of hsa-mir-93 with disease is relatively stable regardless of the percentage of hub protein nodes removed from the graph. On the other hand, the disease degree centrality of hsa-mir-182 exhibits more variability, with a value of 34 at most percentile thresholds but a lower value of 30 at the 95.00% - 96.00% thresholds. This suggests that the association of hsa-mir-182 with disease is more sensitive to the removal of hub protein nodes from the graph.

Figure 26 also shows the results of protein hub nodes removal with a wider range of thresholds. This line plot visualizes how the disease degree centrality of each microRNA changes as the percentage of hub protein nodes removed from the graph increases. The x-axis represents different percentiles of dropped protein nodes in the graph, while the y-axis represents different values of the disease degree centrality measure. Each horizontal-like line reflects the change of the disease degree centrality value after dropping the hub protein nodes with degree higher than the corresponding x percentile value of the rest of the protein nodes in the graph. The legend of the figure is color-coding different microRNAs with different colors to ease its interpretation.

We can see how the disease degree centrality values vary across the different percentile thresholds for each microRNA and how the lines for different microRNAs compare to each other. For example, if one microRNA's line is consistently higher than another microRNA's line, it may indicate that the first microRNA is more strongly associated with disease. On the other hand, if one microRNA's line is relatively flat while another microRNA's line has more steep slopes, it may suggest that the first microRNA's association with disease is more consistent across different protein node removal scenarios, while the second microRNA's association is more sensitive to changes in the graph structure.

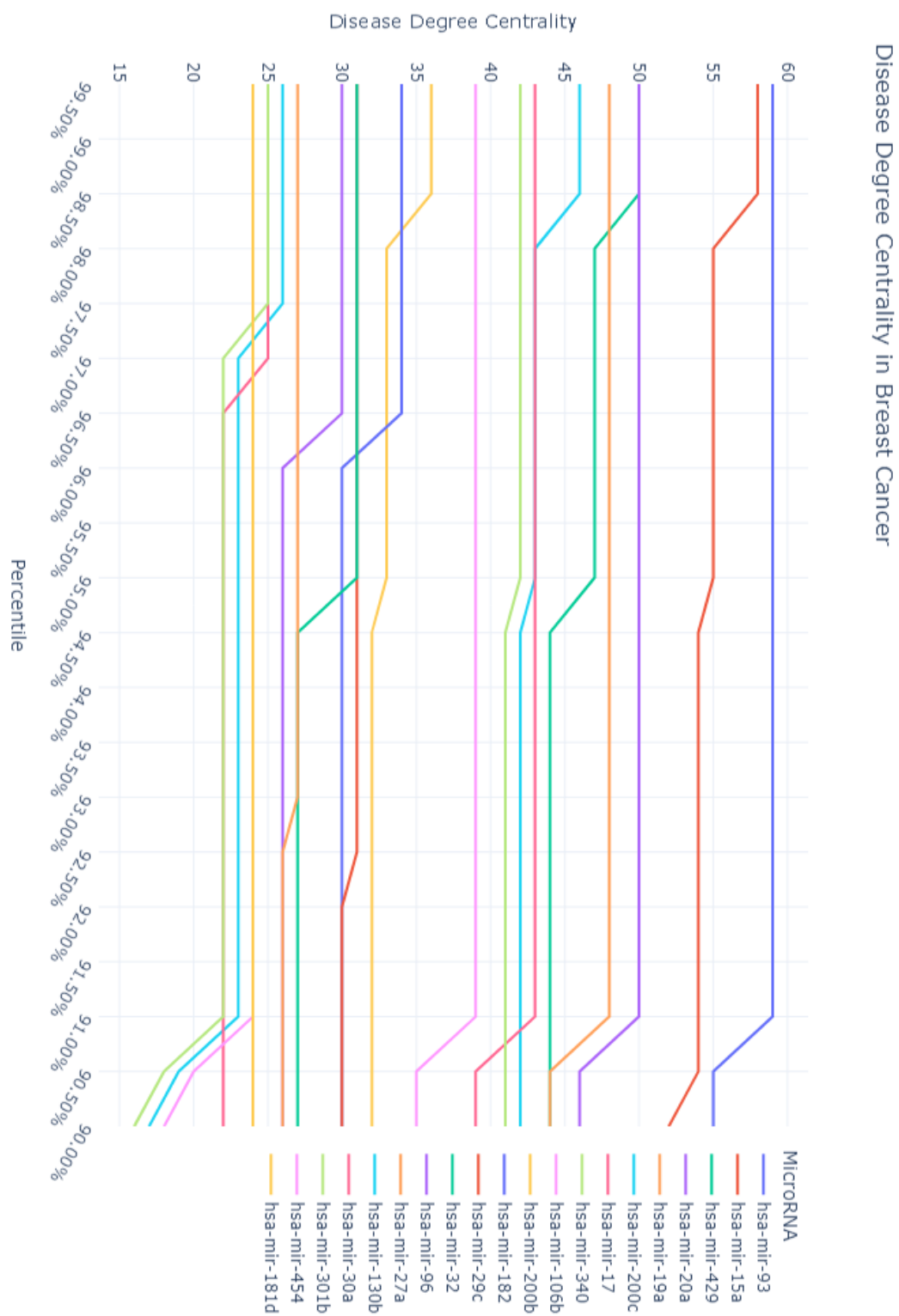


Figure 26 Effect of dropping hub protein nodes for different percentile values on the disease degree centrality

5.11.2 Dropping Random Edges

In computational biology and bioinformatics, it is common to investigate the effects of removing certain edges from a graph on various properties of the graph. This is often done by randomly selecting a percentage or number of edges to remove [49]. In our research, we applied edge pruning with different percentages and examined the impact on the disease degree centrality measure values for various microRNAs. This allowed us to see how the removal of different percentages of edges affects these measure values.

Table 18 in the appendix section presents the change in disease degree centrality for microRNAs as a result of the removal of random edges from the graph. Specifically, nodes with indices: 83, 90, 93-95, 98, 100, 101, 103-105, 107, and 109 are the most affected nodes as with high percentage of edges (starting from 3%) are removed, as they became disconnected from the network.

Figure 27 shows the effect of the edge pruning process on the disease degree centrality. In the figure, it is evident that the measure is affected by edge removal. The rationale behind that is that the disease degree centrality is highly dependent on nodes connections within the graph. So, it was expected that the disease degree centrality values of most of the microRNAs to change as more edges are removed from the graph. The figure shows that the disease degree centrality values of most microRNAs change as more edges are removed from the graph. This suggests that the association of most microRNAs with disease is sensitive to changes in the graph structure.

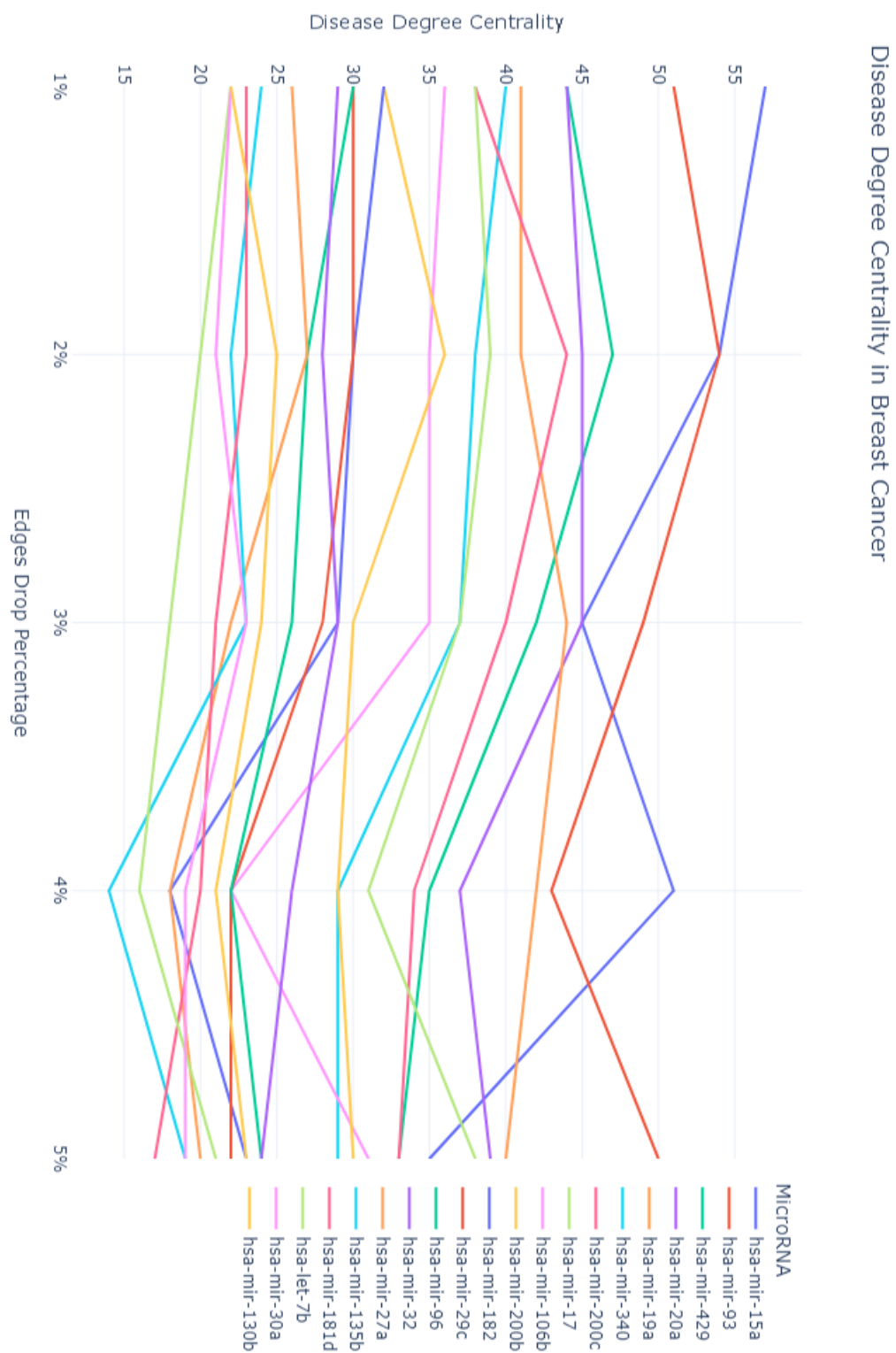


Figure 27 Effect of dropping [1% - 5%] percentage of the edges on the disease degree centrality

It can be observed from the two line plots that these two types of graph modifications have different impacts on the overall structure and connectivity of the graph, which could lead to different effects on the disease degree centrality measure. The effect of dropping edges from the graph on the disease degree centrality measure is more significant than the effect of dropping hub protein nodes. This means that the disease degree centrality values of the microRNAs are more sensitive to changes in the graph structure when edges are removed compared to when hub protein nodes are dropped. One possible reason is that the connections between nodes in the graph are more important for the disease degree centrality measure than the specific nodes themselves. Dropping edges would disrupt these connections more than dropping individual nodes, which could lead to a greater impact on the disease degree centrality measure.

CHAPTER 6

Validation & Threats to Validity

In this chapter, we present the experiments we conducted to evaluate the robustness and scope of application of the work presented in this thesis. Additionally, the second section of this chapter discusses potential threats to the validity of our work.

In this chapter, we present the analyses we conducted to assess the robustness and scope of application of the work presented in this thesis. These analyses were designed to provide evidence for the validity and applicability of our results, as well as to highlight any limitations of the proposed approach. By doing so, we hope to provide a clearer understanding of where the proposed approach is appropriate to use.

6.1 Validation through *InterPro* resource

We consulted *InterPro* [72] to investigate on the two protein domains with the highest disease node degree centrality. The results of *InterPro* analysis on those two protein domains showed that they are involved in two of the inherited human syndromes (Menkes and Wilson diseases). We matched those results with our results where we found that our measure has reported those two diseases as associated with those top two domains. Hence, the results of *InterPro* analysis matched our findings.

6.2 Validation through *DAVID* data source

DAVID [73] is an online bioinformatics database that provides analysis and functional interpretation of genes. We used DAVID to investigate the two genes that are associated with our

top two protein domains. The result of DAVID analysis showed that there is evidence in the literature denoting the association between those two genes and the two human syndromes (Menkes and Wilson diseases).

6.3 Validation of Results through Comparison with Experimental Findings in the Biology Literature

As a way to validate the results of the current work, an example of an association found from the graph analysis has been checked with the literature in that area. The *miR-15a* microRNA has been chosen as the microRNA example and breast cancer disease as the disease example. Our analysis shows that they are associated. We checked on the literature and we found the results of the work by Luo et.al [50] support our findings as they show that under-expression of *miR-15a* microRNA and its inhibition to *CCNE1* gene contributes to breast cancer disease.

Further, following the work of *Yu Li* and *Kris V. Kowdley* [51], we compared the results of this work with the literature.

Table 15 presents a comparison of the top 20 microRNAs ranked using our technique with the results from academic reference papers that have experimentally verified the association of these microRNAs with breast cancer disease. This cross-checking allows for the validation of our method for ranking microRNAs based on their potential involvement in breast cancer. By comparing our results with those from experimental studies, we can gain confidence in the accuracy and reliability of our technique for identifying microRNAs that may be relevant to the development or progression of breast cancer. This information can be used to inform the design of future experiments and the development of therapeutic strategies targeting specific microRNAs.

Table 15 References from literature for the top 20 MicroRNAs

MicroRNA	Reference
hsa-mir-93	[52], [53]
hsa-mir-15a	[54]
hsa-mir-429	[55], [56]
hsa-mir-20a	[57]
hsa-mir-19a	Unconfirmed
hsa-mir-200c	[58]
hsa-mir-17	Unconfirmed
hsa-mir-340	[59][60][61]
hsa-mir-106b	[62]
hsa-mir-200b	[63]
hsa-mir-182	[64]
hsa-mir-29c	[65]
hsa-mir-32	[66]
hsa-mir-96	[64]
hsa-mir-27a	Unconfirmed
hsa-mir-130b	[67]
hsa-mir-30a	[68]
hsa-mir-301b	[69]
hsa-mir-454	[70]
hsa-mir-181d	[71]

The proposed approach has produced results that are consistent with findings from other biology researchers who have studied the association between microRNAs and breast cancer experimentally. This can provide additional support for the validity and reliability of your results.

It is important to consider the experimental validation status of these microRNAs, as those that have not been experimentally confirmed may not yet be suitable for use on an industrial scale. While these microRNAs may be thought to be associated with breast cancer based on our ranking technique, their association has not yet been definitively established through biological experiments. Therefore, it is necessary to further investigate these microRNAs in order to determine their potential utility in the diagnosis, treatment, or prevention of breast cancer.

6.4 Threats to Validity

As is the case with most -if not all- computational biology techniques, the results of that work should not be taken as is and directly implemented on an industrial scale. However, the results need to be handled with caution if they are not validated with biological experiments (e.g., in wet labs) to give stronger evidence or, possibly, to nullify the claim that the results have come up with. We may refer the reader to Figure 1 and Figure 2 that shows the conventional pipeline that is utilized to further validate the results of an *in-silico* technique by backing it with *in-vitro* followed by *in-vivo* techniques.

Analysis of the results shows that one potential threat to the validity of this work is the sensitivity of the disease degree centrality measure to changes in the graph structure. The structure of the graph can be modified in various ways, such as by dropping random edges or hub protein nodes. These modifications can change the overall connectivity and organization of the graph in ways that may not be immediately apparent.

Although the changes in the top ranked list of microRNAs are not severe for slight modifications in the graph structure, especially with protein hub nodes removal, we provide a suggested technique that is expected to mitigate that effect.

One potential way to mitigate the threat of change in the measure with the change in graph structure is to use techniques such as network aggregation or resampling, and to use multiple measures to verify the results. By aggregating the data from multiple related graphs into a single graph, and by repeatedly sampling from the original graph, it may be possible to reduce the impact of noise or random variations on the measure that are results of the graph modifications and provide a more stable estimate of the association between microRNAs and disease.

CHAPTER 7

CONCLUSION & FUTURE WORK

In this chapter, we conclude the presented work while highlighting its contributions and their impact. Then we mention the future work that will handle its limitations.

7.1 Conclusion

Given the wide spectrum of available biological data from different public resources, in this thesis we managed to collect, clean, and integrate different types of biological data representing more than five different types of biological molecules.

From the data collection and integration phase, we found that a suitable way to model this data, in order to gain hidden insights from it, is to model it as a graph. The rationale behind that is that the data was found to be highly interconnected, thus the modeling phase is instantiated by considering a multi-bipartite heterogeneous graph model of 4 layers (protein-domain association graph model) and another multi-bipartite heterogeneous graph model of 5 layers (microRNA-disease association graph model). Those two proposed heterogeneous multipartite graph models were leveraged to infer associations between protein domains and diseases as well as associations between microRNAs and diseases. Breast cancer disease was used as a case study.

The proposed work can be leveraged as one of the *in-silico* techniques at the early stages of drug design and drug discovery pipeline. Given a certain disease, the drug design practitioners can pick the top k protein domains with the highest disease node degree centrality to proceed with them to the next stage in the pipeline. Similarly, drug discovery practitioners can consider the top k

microRNAs with the highest disease node degree centrality to further study them to understand the mechanism of the chosen disease.

A new node centrality measure has been introduced to calculate the relevant importance of protein domain nodes to disease nodes and microRNA nodes to disease nodes. That measure can be considered as an indicator of missing link(s) on the graph. Thus, the higher the *disease node degree centrality*, the more likely a link to be formed between those two entities (whether they are a protein domain node and a disease node, or a microRNA node and a disease node). Given the flexibility of the proposed node centrality measure, it is not restricted to the current set of data or the current problem, however, it can be generalized and used in similar scenarios where the relative importance of two far away entities on a multi-bipartite (heterogenous or homogenous) graph model is concerned. Thus, the proposed node centrality measure can be used in other contexts where the data can be modeled with a similar, or slightly different, graph structure.

To check the graph model robustness against nodes or edges removal, several experiments were conducted and the effect of that was observed in the produced top-k ranked list of the microRNAs. The microRNAs with high consistency across those different experimental setups seem to be the ones that are not much affected, and therefore their rank may be more reliable than the other inconsistent ones.

To verify the results of the proposed work, the top 20 ranked microRNAs were compared with relevant work by biology research. A subset of the results was claimed by the proposed work to be associated with breast cancer even though they have not yet been confirmed (mentioned) in the literature to have associativity with breast cancer. Those unconfirmed ones need further investigation from biology researchers in order to confirm or reject our proposed claim about them being associated with breast cancer or otherwise. It is also worth noting that the fact that our

results are consistent with other findings does not necessarily mean that they are the only possible explanation for the association between microRNAs and breast cancer. There may be other factors or mechanisms at play that have not yet been fully understood or identified. It is important to continue to conduct research and gather evidence to better understand the complex relationships between microRNAs and diseases in general.

7.2 Future Work

Future work can be done to handle the limitations of this work. That can be defined in two main areas, namely biological data, and graph analysis techniques plus one interdisciplinary area, namely, bioinformatics. Additionally, the validation of the results can be further solidified.

Therefore, the following list of future work can be made to improve those areas.

- From the biological data side, new types of data should be incorporated into the analysis. Given we will have a new type of data in the new graph model, the holistic view that we aim for will be further extended and therefore, more complex patterns can be inferred within that new graph model that will model that data.
- From the graph analysis side, more robust techniques can be combined with the current technique, and extra analysis on the graph model can be done.
- From the bioinformatics side, further analysis and additional experiments can be done. One of them is concerned with the protein nodes that tend to have a high degree, usually referred to as protein hubs. Those proteins are known to have an impact on the structure of the graph and hence an effect on the connections between different entities within the graph. So, a question that would be answered as an outcome of running an experiment is what the effect of would be removing some of those protein hubs, from the graph, on the results.

- To further solidify the results, different combinations of edge pruning and hub protein nodes dropping percentages can be explored to reach a good balance between both of them. That balance point is the one that may result in the highest number of confirmed microRNAs in the top-k list produced by that experimental setup.
- Another interesting extension of this work is the study of the measure for different versions of the biological graph over time. That can have applications in many fields such as genetic-based personalized medicine.

Appendix

Table 16 Ranked list of microRNAs with their disease node degree centrality calculated for breast cancer

Index	microRNA	Disease node degree centrality
1	hsa-mir-93	59
2	hsa-mir-15a	58
3	hsa-mir-429	50
4	hsa-mir-20a	50
5	hsa-mir-19a	48
6	hsa-mir-200c	46
7	hsa-mir-17	43
8	hsa-mir-340	42
9	hsa-mir-106b	39
10	hsa-mir-200b	36
11	hsa-mir-182	34
12	hsa-mir-29c	31
13	hsa-mir-32	31
14	hsa-mir-96	30
15	hsa-mir-27a	27
16	hsa-mir-130b	26
17	hsa-mir-30a	25
18	hsa-mir-301b	25
19	hsa-mir-454	24
20	hsa-mir-181d	24
21	hsa-mir-141	24
22	hsa-mir-135b	24
23	hsa-mir-30d	23

24	hsa-let-7b	22
25	hsa-mir-98	22
26	hsa-mir-223	22
27	hsa-mir-374a	22
28	hsa-let-7i	21
29	hsa-mir-320a	21
30	hsa-mir-200a	20
31	hsa-mir-195	20
32	hsa-mir-155	19
33	hsa-mir-374b	19
34	hsa-mir-29a	18
35	hsa-mir-30b	17
36	hsa-mir-497	17
37	hsa-let-7c	16
38	hsa-mir-181c	16
39	hsa-mir-149	16
40	hsa-mir-27b	15
41	hsa-mir-21	15
42	hsa-mir-107	15
43	hsa-mir-148b	14
44	hsa-mir-145	13
45	hsa-mir-22	13
46	hsa-mir-301a	12
47	hsa-mir-92b	12
48	hsa-mir-505	11
49	hsa-mir-221	11
50	hsa-mir-203	10

51	hsa-mir-183	10
52	hsa-mir-33a	10
53	hsa-mir-26b	10
54	hsa-mir-33b	10
55	hsa-mir-150	10
56	hsa-mir-152	9
57	hsa-mir-30e	9
58	hsa-mir-1271	9
59	hsa-mir-190b	9
60	hsa-mir-15b	9
61	hsa-mir-381	8
62	hsa-let-7d	8
63	hsa-mir-25	8
64	hsa-mir-18a	8
65	hsa-mir-134	7
66	hsa-mir-214	7
67	hsa-mir-185	7
68	hsa-mir-130a	6
69	hsa-mir-224	6
70	hsa-mir-421	6
71	hsa-mir-34a	6
72	hsa-mir-143	6
73	hsa-let-7e	5
74	hsa-mir-377	5
75	hsa-mir-202	5
76	hsa-mir-106a	5
77	hsa-let-7g	4

78	hsa-mir-31	4
79	hsa-mir-146a	4
80	hsa-mir-758	4
81	hsa-mir-212	4
82	hsa-mir-410	4
83	hsa-mir-132	4
84	hsa-mir-10a	4
85	hsa-mir-503	4
86	hsa-mir-192	3
87	hsa-mir-197	3
88	hsa-mir-222	3
89	hsa-mir-23b	3
90	hsa-mir-494	2
91	hsa-mir-335	2
92	hsa-mir-148a	2
93	hsa-mir-23a	2
94	hsa-mir-363	2
95	hsa-mir-193b	2
96	hsa-mir-376c	2
97	hsa-mir-186	2
98	hsa-mir-18b	2
99	hsa-mir-378	2
100	hsa-mir-378c	2
101	hsa-mir-210	1
102	hsa-mir-495	1
103	hsa-mir-137	1
104	hsa-mir-20b	1

105	hsa-mir-10b	1
106	hsa-mir-191	1
107	hsa-mir-190	1
108	hsa-mir-424	1
109	hsa-mir-326	1

Table 17 Number of protein nodes with degree above X percentile of the other nodes in the graph

Percentile	Number of Protein Nodes
99.50%	14
99.00%	27
98.50%	40
98.00%	53
97.50%	66
97.00%	79
96.50%	92
96.00%	105
95.50%	118
95.00%	131
94.50%	144
94.00%	158
93.50%	171
93.00%	184
92.50%	197
92.00%	210

91.50%	223
91.00%	236
90.50%	249
90.00%	262

Table 18 Change in disease degree centrality for the microRNAs as a result of random edges removal

Index	Percentage of edges dropped	1%	2%	3%	4%	5%
	microRNA	Disease Degree Centrality				
1	hsa-mir-15a	57	54	45	51	35
2	hsa-mir-93	51	54	49	43	50
3	hsa-mir-429	44	47	42	35	33
4	hsa-mir-20a	44	45	45	37	39
5	hsa-mir-19a	41	41	44	42	40
6	hsa-mir-340	40	38	37	29	29
7	hsa-mir-200c	38	44	40	34	33
8	hsa-mir-17	38	39	37	31	38
9	hsa-mir-106b	36	35	35	22	31
10	hsa-mir-200b	32	36	30	29	30
11	hsa-mir-182	32	30	29	18	23
12	hsa-mir-29c	30	30	28	22	22
13	hsa-mir-96	30	27	26	22	24
14	hsa-mir-32	29	28	29	26	24
15	hsa-mir-27a	26	27	22	18	20
16	hsa-mir-135b	24	22	23	14	19
17	hsa-mir-181d	23	23	21	20	17
18	hsa-let-7b	22	20	18	16	21
19	hsa-mir-30a	22	21	23	19	19

20	hsa-mir-130b	22	25	24	21	23
21	hsa-mir-223	22	19	19	14	13
22	hsa-mir-98	21	19	18	17	13
23	hsa-mir-141	21	20	22	18	20
24	hsa-mir-30d	21	21	19	23	16
25	hsa-mir-374a	21	20	19	19	17
26	hsa-mir-301b	20	24	22	21	19
27	hsa-mir-195	20	19	18	18	13
28	hsa-mir-454	19	21	18	14	17
29	hsa-mir-374b	19	17	19	14	13
30	hsa-mir-200a	17	17	16	17	17
31	hsa-mir-29a	17	16	16	16	15
32	hsa-mir-497	17	14	15	15	10
33	hsa-mir-320a	17	20	20	18	14
34	hsa-let-7c	16	14	16	9	13
35	hsa-let-7i	16	18	21	13	14
36	hsa-mir-30b	16	15	17	15	11
37	hsa-mir-149	15	12	13	14	10
38	hsa-mir-155	15	19	12	16	9
39	hsa-mir-107	15	13	14	11	13
40	hsa-mir-27b	14	15	11	13	13
41	hsa-mir-148b	14	10	10	11	10
42	hsa-mir-21	14	15	15	12	8
43	hsa-mir-181c	13	15	13	12	14
44	hsa-mir-145	13	12	10	10	10
45	hsa-mir-22	13	8	12	12	7
46	hsa-mir-505	11	10	10	6	10
47	hsa-mir-221	11	9	10	11	6

48	hsa-mir-33a	10	9	9	8	8
49	hsa-mir-33b	10	9	8	9	10
50	hsa-mir-152	9	8	8	8	8
51	hsa-mir-203	9	10	9	9	5
52	hsa-mir-30e	9	8	9	8	9
53	hsa-mir-1271	9	8	9	8	8
54	hsa-mir-15b	9	7	8	8	6
55	hsa-mir-150	9	9	10	7	8
56	hsa-let-7d	8	7	8	5	6
57	hsa-mir-301a	8	12	8	10	9
58	hsa-mir-92b	8	10	11	10	8
59	hsa-mir-26b	8	8	8	9	8
60	hsa-mir-381	7	7	7	5	6
61	hsa-mir-25	7	7	7	4	6
62	hsa-mir-183	7	10	10	8	6
63	hsa-mir-190b	7	5	9	6	5
64	hsa-mir-18a	7	7	6	6	6
65	hsa-mir-130a	6	3	6	4	5
66	hsa-mir-224	6	6	6	4	6
67	hsa-mir-421	6	5	6	4	3
68	hsa-mir-134	6	7	7	6	6
69	hsa-mir-34a	6	6	5	4	5
70	hsa-mir-214	6	6	7	6	7
71	hsa-mir-106a	5	5	4	3	4
72	hsa-let-7g	4	3	4	3	2
73	hsa-let-7e	4	4	5	4	4
74	hsa-mir-377	4	4	4	5	3
75	hsa-mir-202	4	5	5	3	4

76	hsa-mir-31	4	3	4	4	3
77	hsa-mir-146a	4	4	4	3	3
78	hsa-mir-758	4	4	4	4	4
79	hsa-mir-10a	4	4	3	4	3
80	hsa-mir-185	4	5	7	6	5
81	hsa-mir-503	4	3	3	3	4
82	hsa-mir-192	3	3	3	2	3
83	hsa-mir-197	3	3	2	3	
84	hsa-mir-222	3	3	3	3	3
85	hsa-mir-143	3	3	6	4	3
86	hsa-mir-23b	3	2	3	3	3
87	hsa-mir-132	3	4	4	2	4
88	hsa-mir-212	3	4	4	2	4
89	hsa-mir-410	3	4	4	2	4
90	hsa-mir-494	2	1	1	1	
91	hsa-mir-335	2	2	1	2	2
92	hsa-mir-148a	2	2	1	1	1
93	hsa-mir-23a	2	2	1	1	
94	hsa-mir-363	2	2	2	2	
95	hsa-mir-193b	2	2	2	2	
96	hsa-mir-186	2	2	2	1	2
97	hsa-mir-18b	2	2	2	1	2
98	hsa-mir-378	2	2	2	2	
99	hsa-mir-378c	2	2	1	2	2
100	hsa-mir-210	1	1	1		
101	hsa-mir-495	1	1	1		
102	hsa-mir-137	1	1	1	1	1
103	hsa-mir-20b	1	1	1	1	

104	hsa-mir-10b	1	1	1	1	
105	hsa-mir-191	1	1	1	1	
106	hsa-mir-190	1	1	1	1	1
107	hsa-mir-424	1	1	1	1	
108	hsa-mir-326	1	1	1	1	1
109	hsa-mir-376c	1	2	2		

References

- [1] S. Ekins, J. Mestres, and B. Testa, “In silico pharmacology for drug discovery: methods for virtual ligand screening and profiling,” *Br. J. Pharmacol.*, vol. 152, no. 1, pp. 9–20, Sep. 2007.
- [2] S. Brogi, T. C. Ramalho, K. Kuca, J. L. Medina-Franco, and M. Valko, “Editorial: In silico Methods for Drug Design and Discovery,” *Front. Chem.*, vol. 8, p. 612, 2020.
- [3] X. Zheng, L. Gan, E. Wang, and J. Wang, “Pocket-based drug design: exploring pocket space,” *AAPS J.*, vol. 15, no. 1, pp. 228–241, Jan. 2013.
- [4] “UniProt: the universal protein knowledgebase in 2021,” *Nucleic Acids Res.*, vol. 49, no. D1, pp. D480–D489, Jan. 2021.
- [5] H. A. Hussein, A. Borrel, C. Geneix, M. Petitjean, L. Regad, and A.-C. Camproux, “PockDrug-Server: a new web server for predicting pocket druggability on holo and apo proteins,” *Nucleic Acids Res.*, vol. 43, no. W1, pp. W436–42, Jul. 2015.
- [6] A. M. Elmoselhy and E. Ramadan, “Utilizing Graph Database for Inferring Domain-Disease Associations,” in *2019 International Conference on Advances in the Emerging Computing Technologies (AECT)*, 2020, pp. 1–6.
- [7] L. Kirton, “Analysis of miRNA and mRNA associated with Epithelial Mesenchymal Transition,” *OpenStax CNX*, vol. Ch 6, pp. 150–153, 2010.
- [8] K. Lanier and L. Williams, “The Origin of Life: Models and Data,” *J. Mol. Evol.*, vol. 84, 2017.
- [9] J. Sergio Hleap and C. Blouin, “The Semantics of the Modular Architecture of Protein Structures,” *Curr. Protein Pept. Sci.*, vol. 17, no. 1, pp. 62–71, 2015.
- [10] R. D. Schaeffer and V. Daggett, “Protein folds and protein folding,” *Protein Eng. Des. & Sel. PEDS*, vol. 24 1–2, pp. 11–19, 2011.
- [11] A. M. Ardekani and M. M. Naeini, “The Role of MicroRNAs in Human Diseases,” *Avicenna J. Med. Biotechnol.*, vol. 2, no. 4, pp. 161–79, Oct. 2010.
- [12] B. Bernardo, F. Charchar, R. Lin, and J. McMullen, “A MicroRNA Guide for Clinicians and Basic Scientists: Background and Experimental Techniques,” *Heart. Lung Circ.*, vol. 21, pp. 131–142, 2011.
- [13] S. Kaur, A. Ali, U. Ahmad, Y. Siahbalaei, A. K. Pandey, and B. Singh, “Role of single nucleotide polymorphisms (SNPs) in common migraine,” *Egypt. J. Neurol. Psychiatry Neurosurg.*, vol. 55, no. 1, p. 47, 2019.
- [14] M. Lu *et al.*, “An analysis of human microRNA and disease associations,” *PLoS One*, vol. 3, no. 10, p. e3420, Oct. 2008.
- [15] Y. Li *et al.*, “HMDD v2.0: a database for experimentally supported human microRNA and

- disease associations,” *Nucleic Acids Res.*, vol. 42, no. D1, pp. D1070–D1074, Jan. 2014.
- [16] A. Vashist, C. A. Kulikowski, and I. B. Muchnik, “MULTIPARTITE GRAPH CLUSTERING FOR STRUCTURED DATASETS AND AUTOMATING ORTHOLOG EXTRACTION.”
 - [17] I. Robinson, J. Webber, and E. Eifrem, *Graph databases*. “ O’Reilly Media, Inc.,” 2013.
 - [18] C. T. Have and L. J. Jensen, “Are graph databases ready for bioinformatics?,” *Bioinformatics*, vol. 29, no. 24, p. 3107, 2013.
 - [19] O. Mason and M. Verwoerd, “Graph theory and networks in biology,” *IET Syst. Biol.*, vol. 1, no. 2, pp. 89–119, 2007.
 - [20] Y. Park, M. Shankar, B.-H. Park, and J. Ghosh, “Graph databases for large-scale healthcare systems: A framework for efficient data management and data services,” in *2014 IEEE 30th International Conference on Data Engineering Workshops*, 2014, pp. 12–19.
 - [21] L. K. Branting, F. Reeder, J. Gold, and T. Champney, “Graph analytics for healthcare fraud risk estimation,” in *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2016, pp. 845–851.
 - [22] G. Drakopoulos and A. Kanavos, “Tensor-based document retrieval over Neo4j with an application to PubMed mining,” in *2016 7th International Conference on Information, Intelligence, Systems & Applications (IISA)*, 2016, pp. 1–6.
 - [23] P. J. A. Cock *et al.*, “Biopython: freely available Python tools for computational molecular biology and bioinformatics,” *Bioinformatics*, vol. 25, no. 11, pp. 1422–1423, 2009.
 - [24] A. Lysenko, I. A. Roznov\ua\ct, M. Saqi, A. Mazein, C. J. Rawlings, and C. Auffray, “Representing and querying disease networks using graph databases,” *BioData Min.*, vol. 9, no. 1, p. 23, 2016.
 - [25] J. Piñero *et al.*, “DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants,” *Nucleic Acids Res.*, p. gkw943, 2016.
 - [26] H. Hermjakob *et al.*, “IntAct: an open source molecular interaction database,” *Nucleic Acids Res.*, vol. 32, no. suppl_1, pp. D452--D455, 2004.
 - [27] W. Zhang, M. P. Coba, and F. Sun, “Inference of domain-disease associations from domain-protein, protein-disease and disease-disease relationships,” in *BMC systems biology*, 2016, vol. 10, no. 1, p. S4.
 - [28] K. Piletič and T. Kunej, “MicroRNA epigenetic signatures in human disease,” *Arch. Toxicol.*, vol. 90, no. 10, pp. 2405–2419, 2016.
 - [29] S. Swarbrick, N. Wragg, S. Ghosh, and A. Stolzing, “Systematic Review of miRNA as Biomarkers in Alzheimer’s Disease,” *Mol. Neurobiol.*, vol. 56, no. 9, pp. 6156–6167, 2019.
 - [30] L.-H. H. Peng, C.-N. N. Sun, N.-N. N. Guan, J.-Q. Q. Li, and X. Chen, “HNMDA: heterogeneous network-based miRNA–disease association prediction,” *Mol. Genet. Genomics*, vol. 293, no. 4, pp. 983–995, Aug. 2018.

- [31] M. Chen, Y. Zhang, A. Li, Z. Li, W. Liu, and Z. Chen, “Bipartite Heterogeneous Network Method Based on Co-neighbor for MiRNA-Disease Association Prediction,” *Front. Genet.*, vol. 10, no. APR, p. 385, Apr. 2019.
- [32] X. Chen, L.-Y. Y. Wang, and L. Huang, “NDAMDA: Network distance analysis for MiRNA-disease association prediction,” *J. Cell. Mol. Med.*, vol. 22, no. 5, pp. 2884–2895, May 2018.
- [33] B. Shao, B. Liu, and C. Yan, “SACMDA: MiRNA-Disease Association Prediction with Short Acyclic Connections in Heterogeneous Graph,” *Neuroinformatics*, vol. 16, no. 3–4, pp. 373–382, Oct. 2018.
- [34] J. Mistry *et al.*, “Pfam: The protein families database in 2021,” *Nucleic Acids Res.*, vol. 49, no. D1, pp. D412–D419, Jan. 2021.
- [35] K. L. Howe *et al.*, “Ensembl 2021.,” *Nucleic Acids Res.*, vol. 49, no. D1, pp. D884–D891, 2021.
- [36] H.-Y. Huang *et al.*, “miRTarBase 2020: updates to the experimentally validated microRNA–target interaction database,” *Nucleic Acids Res.*, Oct. 2019.
- [37] K. Tomczak, P. Czerwińska, and M. Wiznerowicz, “The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge,” *Contemp. Oncol. (Poznan, Poland)*, vol. 19, no. 1A, pp. A68–77, 2015.
- [38] V. Agarwal, G. W. Bell, J.-W. Nam, and D. P. Bartel, “Predicting effective microRNA target sites in mammalian mRNAs,” *Elife*, vol. 4, Aug. 2015.
- [39] W. M. C. da Silva, P. Wer celens, M. E. M. T. Walter, M. Holanda, and M. Br`vigo, “Graph Databases in Molecular Biology,” in *Brazilian Symposium on Bioinformatics*, 2018, pp. 50–57.
- [40] A. Patel and J. Dharwa, “Graph Data: The Next Frontier in Big Data Modeling for Various Domains,” *Indian J. Sci. Technol.*, vol. 10, no. 21, pp. 1–7, 2017.
- [41] W. Wang, W. Zhang, R. Jiang, and Y. Luan, “Prioritisation of associations between protein domains and complex diseases using domain--domain interaction networks,” *IET Syst. Biol.*, vol. 4, no. 3, pp. 212–222, 2010.
- [42] G. M. Qin, R. Y. Li, and X. M. Zhao, “Identifying Disease Associated miRNAs Based on Protein Domains,” *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 13, no. 6, pp. 1027–1035, Nov. 2016.
- [43] J. P. Overington, B. Al-Lazikani, and A. L. Hopkins, “How many drug targets are there?,” *Nat. Rev. Drug Discov.*, vol. 5, no. 12, p. 993, 2006.
- [44] R. Apweiler *et al.*, “UniProt: the universal protein knowledgebase,” *Nucleic Acids Res.*, vol. 32, no. suppl_1, pp. D115--D119, 2004.
- [45] R. J. Kinsella *et al.*, “Ensembl BioMarts: a hub for data retrieval across taxonomic space,” *Database (Oxford)*, vol. 2011, p. bar030, 2011.
- [46] R. R. Vallabhajosyula, D. Chakravarti, S. Lutfeali, A. Ray, and A. Raval, “Identifying hubs

- in protein interaction networks.,” *PLoS One*, vol. 4, no. 4, p. e5344, 2009.
- [47] R. L. Grossman *et al.*, “Toward a Shared Vision for Cancer Genomic Data.,” *N. Engl. J. Med.*, vol. 375, no. 12, pp. 1109–1112, Sep. 2016.
 - [48] M. Ahmed, H. Nguyen, T. Lai, and D. R. Kim, “miRCancerdb: a database for correlation analysis between microRNA and gene expression in cancer,” *BMC Res. Notes*, vol. 11, no. 1, p. 103, 2018.
 - [49] F. Zhou, S. Mahler, and H. Toivonen, “Simplification of networks by edge pruning,” in *Bisociative knowledge discovery*, Springer, 2012, pp. 179–198.
 - [50] Q. Luo *et al.*, “MiR-15a is underexpressed and inhibits the cell cycle by targeting CCNE1 in breast cancer,” *Int J Oncol*, vol. 43, no. 4, pp. 1212–1218, 2013.
 - [51] Y. Li and K. V Kowdley, “MicroRNAs in common human diseases.,” *Genomics. Proteomics Bioinformatics*, vol. 10, no. 5, pp. 246–253, Oct. 2012.
 - [52] S. Shyamasundar, J. P. Lim, and B. H. Bay, “miR-93 inhibits the invasive potential of triple-negative breast cancer cells in vitro via protein kinase WNK1.,” *Int. J. Oncol.*, vol. 49, no. 6, pp. 2629–2636, Dec. 2016.
 - [53] G. An *et al.*, “Effects of miR-93 on epithelial-to-mesenchymal transition and vasculogenic mimicry in triple-negative breast cancer cells.,” *Mol. Med. Rep.*, vol. 23, no. 1, Jan. 2021.
 - [54] Y. SHINDEN *et al.*, “Diminished Expression of MiR-15a Is an Independent Prognostic Marker for Breast Cancer Cases,” *Anticancer Res.*, vol. 35, no. 1, pp. 123 LP – 127, Jan. 2015.
 - [55] W. Dai *et al.*, “miR-148b-3p, miR-190b, and miR-429 Regulate Cell Progression and Act as Potential Biomarkers for Breast Cancer,” *jbc*, vol. 22, no. 2, pp. 219–236, Apr. 2019.
 - [56] D.-D. Xiong *et al.*, “A nine-miRNA signature as a potential diagnostic marker for breast carcinoma: An integrated study of 1,110 cases.,” *Oncol. Rep.*, vol. 37, no. 6, pp. 3297–3304, Jun. 2017.
 - [57] L. Zhao, Y. Zhao, Y. He, Q. Li, and Y. Mao, “The functional pathway analysis and clinical significance of miR-20a and its related lncRNAs in breast cancer,” *Cell. Signal.*, vol. 51, pp. 152–165, 2018.
 - [58] U. Berber, I. Yilmaz, G. Narli, A. Haholu, Z. Kucukodaci, and D. Demirel, “miR-205 and miR-200c: Predictive Micro RNAs for Lymph Node Metastasis in Triple Negative Breast Cancer,” *jbc*, vol. 17, no. 2, pp. 143–148, Jun. 2014.
 - [59] Z. Huang, Y. Xu, M. Wan, X. Zeng, and J. Wu, “miR-340: A multifunctional role in human malignant diseases.,” *Int. J. Biol. Sci.*, vol. 17, no. 1, pp. 236–246, 2021.
 - [60] Z. Shi *et al.*, “MiR-340 Inhibits Triple-Negative Breast Cancer Progression by Reversing EZH2 Mediated miRNAs Dysregulated Expressions.,” *J. Cancer*, vol. 8, no. 15, pp. 3037–3048, 2017.
 - [61] L.-K. Hou *et al.*, “miR-340 and ZEB1 negative feedback loop regulates TGF- β - mediated breast cancer progression.,” *Oncotarget*, vol. 7, no. 18, pp. 26016–26026, May 2016.

- [62] P. L. Farré *et al.*, “MiR-106b-5p: A Master Regulator of Potential Biomarkers for Breast Cancer Aggressiveness and Prognosis,” *International Journal of Molecular Sciences*, vol. 22, no. 20, 2021.
- [63] Q. Zheng *et al.*, “miR-200b inhibits proliferation and metastasis of breast cancer by targeting fucosyltransferase IV and α 1,3-fucosylated glycans,” *Oncogenesis*, vol. 6, no. 7, pp. e358–e358, 2017.
- [64] I. K. Guttilla and B. A. White, “Coordinate regulation of FOXO1 by miR-27a, miR-96, and miR-182 in breast cancer cells,” *J. Biol. Chem.*, vol. 284, no. 35, pp. 23204–23216, Aug. 2009.
- [65] W. Li *et al.*, “miR-29c plays a suppressive role in breast cancer by targeting the TIMP3/STAT1/FOXO1 pathway,” *Clin. Epigenetics*, vol. 10, no. 1, p. 64, 2018.
- [66] H. Xia, J. Long, R. Zhang, X. Yang, and Z. Ma, “MiR-32 contributed to cell proliferation of human breast cancer cells by suppressing of PHLPP2 expression,” *Biomed. Pharmacother.*, vol. 75, pp. 105–110, Oct. 2015.
- [67] S. Ahn, A. Kwon, Y. H. Huh, S. Rhee, and W. K. Song, “Tumor-derived miR-130b-3p induces cancer-associated fibroblast activation by targeting SPIN90 in luminal A breast cancer,” *Oncogenesis*, vol. 11, no. 1, p. 47, 2022.
- [68] H.-D. Zhang, L.-H. Jiang, D.-W. Sun, J. Li, and J.-H. Tang, “miR-30a inhibits the biological function of breast cancer cells by targeting Notch1,” *Int. J. Mol. Med.*, vol. 40, no. 4, pp. 1235–1242, Oct. 2017.
- [69] C. Gao *et al.*, “The construction and analysis of ceRNA networks in invasive breast cancer: a study based on The Cancer Genome Atlas,” *Cancer Manag. Res.*, vol. 11, pp. 1–11, 2019.
- [70] Z.-G. Cao *et al.*, “High expression of microRNA-454 is associated with poor prognosis in triple-negative breast cancer,” *Oncotarget*, vol. 7, no. 40, pp. 64900–64909, Oct. 2016.
- [71] L.-X. Yan *et al.*, “MicroRNA miR-21 overexpression in human breast cancer is associated with advanced clinical stage, lymph node metastasis and patient poor prognosis,” *RNA*, vol. 14, no. 11, pp. 2348–2360, Nov. 2008.
- [72] M. Blum *et al.*, “The InterPro protein families and domains database: 20 years on,” *Nucleic Acids Res.*, vol. 49, no. D1, pp. D344–D354, Jan. 2021.
- [73] D. W. Huang, B. T. Sherman, and R. A. Lempicki, “Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources,” *Nat. Protoc.*, vol. 4, no. 1, pp. 44–57, 2009.

Vitae

Publications:

- A. M. Elmoselhy and E. Ramadan, "Utilizing Graph Database for Inferring Domain-Disease Associations," *2019 International Conference on Advances in the Emerging Computing Technologies (AECT)*, 2020, pp. 1-6, doi: 10.1109/AECT47998.2020.9194219.

Name :Ahmed Mohamed Mahmoud Elmoselhy |

Nationality :Egyptian |

Date of Birth :8/23/1993|

Email :ahmedmelmoselhy@gmail.com|

Address :Alexandria, Egypt|

Postal Code :21519|

Academic Background :Bachelor of Computer Science |

---END OF TEXT---