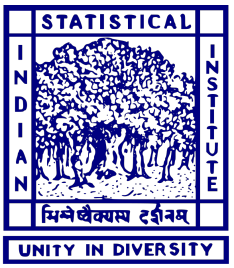


Identifying Drug Resistant miRNAs In Cancer

Dissertation submitted in partial fulfilment for the award of the degree

Master of Technology in Computer Science

by

Akash Dutta

Roll No.: CS2126

M.Tech, 2nd year

Under the supervision of

Prof. Shubhra Sankar Ray

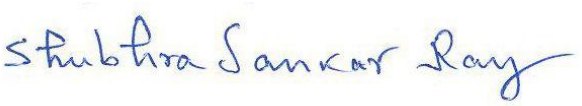
Center for Soft Computing Research,

Indian Statistical Institute, Kolkata

June, 2023

Certificate

I hereby certify that the project titled "*Identifying Drug Resistant miRNAs in Cancer*" has been conducted by Akash Dutta under my supervision and guidance. The thesis represents his original work, which has been carried out through thorough research and investigation. The quality of the thesis meets the standards required for the Master of Technology program in Computer Science, and I highly recommend its submission for evaluation.



Date: 19.06.2023

Prof. Shubhra Sankar Ray
Center for Soft Computing Research,
Indian Statistical Institute, Kolkata

Declaration

I, Akash Dutta, a registered M.Tech student in Computer Science at the Indian Statistical Institute, hereby declare that I have fulfilled all the requirements stated by the institute for the submission of my dissertation.

I declare that the project I have submitted is original and represents the result of my independent investigations and research. It does not contain any form of plagiarism. The work conducted in this project has led to the development of new techniques. Furthermore, I confirm that this work has not been previously submitted to any other university or institution for the purpose of obtaining a degree, diploma, or any other academic award.

I also declare that any text, diagrams, or other materials taken from external sources, such as books, journals, or the internet, have been appropriately acknowledged, referenced, and cited to the best of my knowledge and understanding.

Date: 19.06.2023

Akash Dutta

M.Tech (CS), CS2126

Indian Statistical Institute

Abstract

MicroRNAs (miRNAs) refer to tiny RNA molecules that have a crucial part in regulating drug sensitivity and resistance in cancer. Identifying these miRNAs can significantly enhance the effectiveness of cancer treatment. In this study, a computational method is developed to identify drug resistant miRNAs. Additionally, a comprehensive review of studies focused on identifying those miRNAs is presented. The developed method introduces a scoring system based on expressions of miRNAs in control and resistant groups and involves integration of absolute distance, fold change, and Pearson correlation coefficient in a weighted framework to reduce the average ranking of miRNAs. In the process, the power of the fold change is also varied. Arranging the miRNAs in a descending order based on the score helps in selecting the top ranked miRNAs which helps in classification of the patients. This score offers an effective strategy for identifying miRNAs linked to drug resistance in cancer. Its application may provide valuable insights into potential therapeutic targets, thereby improving the outcomes of cancer treatment.

Acknowledgements

I would like to express my heartfelt gratitude to several individuals who have provided invaluable support and assistance throughout the completion of this report.

First and foremost, I would like to thank my supervisor, Prof. Shubhra Sankar Ray, for his unwavering guidance, expertise, and encouragement. His insightful feedback and constructive criticism have played a pivotal role in shaping the content and direction of this report. I am truly grateful for his mentorship and dedication.

I am also indebted to Mr. Joginder Singh, whose support and assistance have been indispensable. His prompt responses to my queries, valuable suggestions, and commitment to helping me navigate through challenges have been greatly appreciated.

I would like to extend my sincere thanks to my friends and colleagues who have provided unwavering support and encouragement throughout this journey. Their willingness to engage in discussions, share ideas, and provide feedback has been invaluable in enhancing the quality of this report. Their friendship and encouragement have kept me motivated, and I am grateful for their presence in my academic life.

Finally, I would like to express my appreciation to all the individuals who have contributed to this report in various ways but may not be mentioned by name. Their support, whether through discussions, resource sharing, or emotional encouragement, has been incredibly valuable and appreciated.

I am deeply grateful for the collective efforts of all those mentioned above, as well as those who have contributed silently but significantly. Their assistance has played a crucial role in the successful completion of this report.

Thank you.

Akash Dutta

Akash Dutta
Indian Statistical Institute, Kolkata
19.06.23

Contents

List of Tables

Chapter 1

Introduction

Drugs play a crucial role in the treatment of various diseases, ranging from common colds to complex conditions like cancer (1). They are essential for restoring health and well-being. However, it is important to recognize that drugs are foreign substances to the body, and with frequent or prolonged use, the body can develop resistance to them. This resistance mechanism can be both beneficial and detrimental. For instance, in the case of chickenpox, once a person has contracted the disease, their body develops resistance, making it unlikely for them to be affected by it again (2). However, in the context of drug resistance, the body develops resistance to specific drugs, which can hinder the treatment process and allow the disease to progress despite drug administration.

Cancer is a widely known term, and many families have been affected by this severe illness (3). It can originate in various parts of the body and gradually spread through rapid cell division. Unlike normal cells, cancer cells divide rapidly, and both old and new cells continue to grow uncontrollably, forming tumors in some cases (4). In other types of cancer, such as leukemia or blood cancer, tumors may not be present. Unfortunately, the chances of survival from cancer have been relatively low, primarily because it is often detected at later stages when it has already spread extensively throughout the body.

Prior to the formation of cancer cells within the body's tissues, there occur anomalous transformations referred to as hyperplasia and dysplasia (1). Hyperplasia signifies a rise in cell count within an organ or tissue, where these cells appear normal when observed under a microscope. Conversely, dysplasia denotes the presence of abnormal cellular morphology under a microscope, although they have not reached a cancerous state yet.

Chemotherapy serves as the main treatment modality for cancer (3). Nonetheless, its long-term effectiveness is impeded by the emergence of resistance to drugs. This resistance can arise due to various factors, such as genetic alterations, Enhanced expression of the target proteins, modifications in drug targets, impaired drug delivery to target cells, drug expulsion from cells, drug inactivation, enhanced DNA repair mechanisms, impaired apoptosis pathways, modified drug metabolism, and spontaneous mutations induced by drugs, non-mutational changes in gene function triggered by drugs, and drug-induced changes in the structure of chromosomes (5). Recent studies on chemotherapy resistance provide valuable insights into the significant role played by noncoding RNAs in cancer development and their potential contribution to the evolution of drug resistance (3). MicroRNAs (miRNAs) are short noncoding RNAs, they play a vital role in post-transcriptional gene regulation by exerting their influence on mRNA molecules' 3'untranslated region (UTR) in descending order, based on their impact, leading to the downregulation of gene expression. Analyzing the expression patterns of miRNAs can offer valuable insights into the development of chemotherapy resistance. Typically, miRNA expression data obtained from cancer patients are utilized to identify the specific miRNAs responsible for conferring resistance to chemotherapy in that particular type of cancer.

1.1 miRNA in cancer

MicroRNAs (miRNAs) have a significant role in the initiation and progression of cancer, significantly impacting its development and advancement. These small RNA molecules regulate gene expression at the post-transcriptional level (1–4). In the context of cancer, miRNA expression profiles can undergo alterations, disrupting essential cellular processes such as cell apoptosis, differentiation, and proliferation.

The influence of miRNAs on cancer is multifaceted. Some miRNAs act as suppressors of tumor growth by restraining the expression of oncogenes or genes that promote tumor growth. Conversely, certain miRNAs can function as oncogenes by suppressing the genes responsible for maintaining normal cellular functions (1).

The dysregulated expression of miRNAs in cancer can arise from various mechanisms (5). It can be a result of genetic alterations affecting the miRNA genes themselves or the regulatory elements controlling their expression, including amplifications, deletions, or mutations. Epigenetic modifications such as histone modifications or DNA methylation can also influence the patterns of miRNA expression in cancer cells.

The irregular expression of specific miRNAs in cancer has been linked to multiple aspects of the disease, such as the initiation of tumors, their progression, metastasis, and the response to therapy. Therefore, miRNAs have emerged as potential diagnostic biomarkers for cancer detection and prognostic indicators for patient outcomes. Moreover, miRNAs hold promise as therapeutic targets or agents themselves, with ongoing research focused on developing miRNA-based therapies for cancer treatment.

It's important to note that the field of miRNA research in cancer is rapidly evolving, and discoveries continue to uncover the complex roles of miRNAs in tumor biology.

1.2 Foundational Concepts and Terminology

a. Drug Resistance: Resistance occurs when the body builds a defense mechanism against a drug due to its frequent or prolonged use, resulting in the advancement of the disease despite the drug's administration (1).

b. Computational techniques: Computational techniques (6) rooted in computer science and mathematical models aid in the analysis and resolution of scientific problems.

c. Bioinformatics: Bioinformatics is an interdisciplinary field that merges biology, computer science, and statistics to examine and make sense of biological data. It involves the application of computational methods, algorithms, and databases to understand biological processes, genetic information, and molecular interactions (6).

d. Chemotherapy: Chemotherapy (5) is a common treatment approach used in the management of cancer and certain other diseases. It entails the utilization of pharmaceutical substances formulated to eliminate or impede the proliferation of cancer cells. Chemotherapy drugs can be utilized or employed orally, intravenously, or through other routes depending on the specific treatment plan.

e. Cancer: Cancer is a multifaceted and diverse collection of illnesses marked by the unregulated proliferation and dissemination of anomalous cells throughout the body. It can affect almost any part of the body and can arise from various types of cells (7).

1.3 Motivation

The research work is highly motivated by the profound impact of cancer on human life. Cancer is a devastating disease that affects countless individuals and their loved ones. Extensive research efforts are underway worldwide to enhance cancer treatment and ultimately find a cure. However, one significant challenge in the treatment process is the development of drug resistance, which hampers the effectiveness of chemotherapy and reduces patients' lifespan.

The computational technique developed in this research aims to address this critical issue by predicting the specific microRNAs (miRNAs) responsible for drug resistance mechanisms in cancer. Identifying these miRNAs holds the potential to improve cancer treatment procedures and facilitate the development of more effective drugs (1). The significance of this research lies in the potential positive impact it can have on the lives of cancer patients, providing them with better treatment options and extending their survival.

The diligent efforts and hard work invested in this research are justified by the potential outcomes it can bring about. By unraveling the miRNAs involved in drug resistance, this research has the potential to contribute significantly to advancements in cancer treatment and provide hope for patients and their families. Ultimately, the ultimate goal is to alleviate the suffering caused by cancer and enhance the quality of life for those affected by this devastating disease.

1.4 Objective

This study aims to determine the specific miRNAs related to drug resistance in cancer patients affected by different types of cancer. The levels of expression of these miRNAs are analyzed to achieve this objective. The research collects miRNA expression data from two distinct patient groups: the control group and the resistant group. The study focuses on seven different types of cancer data to identify the relevant miRNAs involved.

1.5 Problem Statement

Cancer is a devastating disease that claims the lives of numerous individuals, often resulting in a painful demise. Unfortunately, there is currently no definitive cure for advanced-stage

cancer. However, certain drugs, such as cisplatin and methotrexate, are widely utilized and renowned for their effectiveness in cancer treatment, particularly in chemotherapy. Nevertheless, the repeated use of these drugs can prompt the development of drug resistance within the body of cancer patients, leading to disease progression despite undergoing chemotherapy.

Recent research has illuminated the involvement of microRNAs in the mechanism of drug resistance. microRNAs play a crucial role in gene regulation by binding to the 3'untranslated region (UTR) of target messenger RNA (mRNA) of genes. This interaction can result in mRNA cleavage or translation inhibition, ultimately affecting gene expression (1). Recent discoveries have highlighted the contribution of miRNAs to the emergence of resistance to drugs for cancer treatment patients.

In this study, instead of conducting wet lab experiments to collect miRNA expression data, Gene Expression Omnibus (GEO) data archives have been utilized. Specifically, datasets from seven different types of cancer have been collected. The primary objective is to develop a computational approach that utilizes the expression values of miRNAs to identify those responsible for the mechanism of drug resistance in cancer patients. By leveraging existing data, the aim is to gain insights into the specific miRNAs associated with drug resistance, without the need for additional experimental procedures.

Chapter 2

Literature Survey

2.1 REVIEW OF EXISTING STUDIES

2.1.1 Biochemical Studies

The study conducted by Miller et al. (8) holds great significance as it predates recent research on the involvement of MicroRNA (miRNA) in drug resistance. Miller's study specifically explored the role of MicroRNA-221/222 in the development of resistance to Tamoxifen, a commonly used drug for the therapeutic approach for breast cancer with estrogen receptor positivity. The effects of Tamoxifen on MCF-7 cell lines were assessed using microarray analysis, which identified gene expression changes with a fold change greater than 1.8. The study revealed the up-regulation of three promising miRNAs (miR-221,222 and 181) and the down-regulation of three other miRNAs (miR-21,342 and 489) in both Tamoxifen-sensitive and Tamoxifen-resistant MCF-7 cell lines. Moreover, introducing miR-221/222 into parental MCF-7 cells resulted in their resistance to Tamoxifen. This research was the first to establish a link between the expression of miR-221/222, the overexpression of HER2/neu in breast cancer tissues obtained directly from patients, and inherent resistance to Tamoxifen therapy.

Bockhorn et al. (9) conducted research on the role of MicroRNA-30c in chemotherapy for breast tumors. The study aimed to understand how this specific miRNA contributes to chemo-resistance by targeting TWF1 directly. To investigate the relationship between miRNA targets, epithelial-to-mesenchymal transition (EMT), and drug resistance in breast cancer cells, the author utilized microarray analyses, miRNA target prediction algorithms, and experimental validation methods. Initially, gene and miRNA expression analyses were

performed on frozen primary breast tumor tissues (University of Chicago [UC] tumor set, n=46) and adjacent normal tissues (n=5). Out of the 757 expressed miRNAs, 152 miRNAs meeting specific criteria underwent unsupervised hierarchical clustering, resulting in two distinct expression patterns known as Cluster A and Cluster B. To examine the involvement of miR-30c in breast tumor progression and resistance to chemotherapy, the study assessed the impact of miR-30c on chemotherapy-induced cytotoxicity in tumor cells. The findings revealed a consistent and significant inverse association between miR-30c expression levels and cell survival across six different cell lines (T47D, MCF-7, MDA-MB-231, BT-20, HCC-70, and HCC-38) when treated with paclitaxel and doxorubicin. This association was particularly evident at higher concentrations of paclitaxel (50 nM) and doxorubicin (500 nM).

Kovalchuk et al. (1) conducted a study highlighting the role of microRNA-451 in the resistance of MCF-7 breast cancer cells to the chemotherapy drug doxorubicin. According to Kovalchuk, breast cancer patients often develop drug resistance during treatment, leading to relapse and a worsened prognosis. The study utilized microarray analysis to examine miRNA expression profiles and identify miRNA deregulations. Hierarchical clustering was performed on the miRNA genes that exhibit distinct expression patterns using ANOVA, resulting in the identification of 137 miRNA genes that exhibited significant differential expression ($p < 0.05$). Among these genes, microRNA-451 showed the most promising association with drug resistance.

Palagani et al. (10) conducted research on multiple myeloma cells and published his findings, suggesting that ectopic transcription of MicroRNA-150-5p sensitizes the response to glucocorticoid therapy. The research involved comparing the mRNA and microRNA transcription profiles of glucocorticoid (GC)-sensitive MM1S cells with GC-resistant MM1R cells. The analysis of the transcriptome unveiled that GCs regulated the expression of numerous genes responsible for cell cycle regulation, cellular organization, cell death, and immunological diseases specifically in MM1S cells, while MM1R cells remained unaffected. Among the validated microRNAs (mir-26b,125a-5p,146-5p,150-5p, and 184), mir-150-5p exhibited the most consistent response to GC regulation. Interestingly, the transfection of mir-150-5p in combination with low doses of GC in MM1S cells increased the sensitivity to the response of the treatment, while the contrary effects were observed when a specific antagomir targeting mir-150-5p was utilized.

Kurokawa et al. (11) conducted a study focusing on colon cancer cells and published the findings regarding the involvement of miR-19b and its target mRNAs in the development of resistance to 5-fluorouracil (5-FU), a commonly used chemotherapy drug. Two variants

of colon cancer cells resistant to 5-fluorouracil (5-FU) were created from the DLD-1 and KM12C cell lines. A microarray containing 723 microRNAs was employed to analyze the expression levels of microRNAs, and the results were confirmed using qRT-PCR. In order to identify potential direct targets of the microRNAs, an mRNA immunoprecipitation (RIP)-Chip technique and a pathway analysis tool were employed. The study demonstrated the up-regulation of miR-19b in response to 5-FU and identified potential targets of miR-19b involved in cell cycle regulation during 5-FU treatment.

The regulatory role of the cluster of miRNAs consisting of miR-134,487b and 655 in transforming growth factor-b-induced epithelial-mesenchymal transition (EMT) and drug resistance to gefitinib by targeting MAGI2 in lung adenocarcinoma cells was reported in one study. The research aimed to explore the microRNA (miRNA)-related mechanisms involved in EMT and the acquisition of resistance to EGFR tyrosine kinase inhibitor (EGFR-TKI) in non-small cell lung cancer (NSCLC). The expression profiles of miRNAs were examined in four human adenocarcinoma cell lines, with or without epithelial-mesenchymal transition (EMT), before and after exposure to transforming growth factor b1 (TGF-b1). Through miRNA array and qRT-PCR analysis, it was discovered that TGF-b1 significantly upregulated the expression of miR-134,487b and 655, which are clustered together on chromosome 14q32, in lung adenocarcinoma cells undergoing EMT. The study demonstrated that the miR cluster played a crucial part in TGF-b1-induced EMT and influenced resistance to gefitinib by directly targeting the MAGI2 gene. The restain of MAGI2 subsequently resulted in the destabilization of PTEN in lung cancer cells.

Wolfson et al. (5) conducted a study focusing on the involvement of miR-140 in early-stage of breast cancer associated with stem cells. The author observed that genetic and epigenetic changes commonly observed in ductal carcinoma in situ (DCIS) are frequently shared with invasive ductal carcinoma (IDC), and both mRNA and miRNA expression profiles undergo significant alterations. As the tumor grade increases, there is a progressive downregulation of miRNA-140. The study revealed that miR-140 plays a crucial role in suppressing tumors within the Wnt, SOX2, and SOX9 stem cell regulator pathways. The downregulation of miR-140 removes the blocking of these pathways, leading to an increased population of cancer stem cells and the progression of breast cancer.

R. Hummel et al. (7) conducted a study on microRNA signatures observed in chemotherapy-resistant esophageal cancer cell lines. The study utilized an in-vitro model to investigate acquired chemotherapy resistance in esophageal adenocarcinoma (EAC) and squamous cell carcinoma (ESCC) cells. MicroRNA expression profiles were compared between

chemotherapy-resistant variants (cisplatin or 5-fluorouracil) and microarray and quantitative real-time polymerase chain reaction (PCR) techniques were employed to compare chemotherapy-sensitive controls with chemotherapy-resistant samples. Distinct miRNA signatures were observed in the chemotherapy-resistant sublines, and these signatures varied between cisplatin and 5-FU-resistant cells within the same tumor cell line, as well as between EAC and ESCC cells resistant to the same chemotherapy agent. The findings of the study provide evidence supporting the hypothesis that microRNA expression is involved in chemotherapy resistance in esophageal cancer.

2.1.2 Computational Approaches

A computational strategy to reconstruct mutational pathways of drug resistance by utilizing a combination of phylogenetic and Markov models was developed in (6). The objective of the research was to gain a comprehensive understanding of the emergence of drug resistance. The approach consisted of two main phases. Firstly, a model is constructed to simulate the evolutionary process of the virus within the human host. Secondly, a Markov model was developed to estimate the probabilities of transitioning between various genotypes. The resulting model offered a more comprehensive representation of the evolutionary trajectory over time compared to previous models.

A study, demonstrating the potential and effectiveness of a miRNA-based fuzzy mechanism called fuzzy cognitive map, which combines neural network and fuzzy logic principles was presented in (12). The research aimed to investigate the apoptosis/proliferation control mediated by the miRNA-17-92 cluster/E2F1/cMYC circuitry. By experimentally validating the concept of fuzzy control, the author proposed that it could revolutionize the analysis and modeling of gene expression, potentially influencing the design of therapeutic interventions based on miRNA silencing. The study highlighted the capability of the fuzzy cognitive map (FCM) to represent models of various complexities, consider linear and non-linear relationships, and enable causal propagation.

Another research developed a method for improving HIV-1 drug resistance prediction by utilizing cross-resistance information through multilabel classification. The researcher performed experiments on a large dataset consisting of over six hundred reverse transcriptase sequences, along with corresponding resistance data for six nucleoside analogs. Through the use of multilabel classification models and considering cross-resistance patterns, significant improvements in the overall accuracy of drug prediction were achieved compared to

using individual binary classifiers without additional data. Furthermore, the study identified specific sequence patterns in the reverse transcriptase sequences that could determine the optimal sequence of classifiers within the classifier chains.

Two methods, namely EDWFC and HCEDFCR, to identify miRNAs were introduced in (2) that contribute to drug resistance in cancer. The study found that the identified miRNAs consistently ranked among the top positions in the datasets, highlighting their significance. Additionally, the miRNAs that were previously reported as significant in biochemical studies were also among the top-ranked miRNAs. The EDWFC method can be applied to similar miRNA datasets, with a focus on analyzing the top 20 miRNAs through biochemical testing to identify the specific miRNAs related to drug resistance. The HCEDFCR method demonstrated that the top clusters contain miRNAs linked to drug resistance across most datasets, successfully predicting the most influential miRNAs in certain types of cancer. This suggests that HCEDFCR can effectively detect the primary miRNAs responsible for drug resistance in cancer patients. Conducting biochemical testing on the top miRNAs from the identified clusters can provide valuable insights into their role in drug resistance.

L. Luz Gomes (4) developed the application of Self-Organizing Maps (SOM) to identify miRNA expression profiles in Gastric Cancer. The SOM network was constructed based on the observed differences in miRNA expression between healthy and cancerous gastric tissues, with a particular focus on miRNAs that exhibited either under-expression or over-expression. By analyzing 514 miRNAs in gastric cancer samples, the neural network identified a specific miRNA signature comprising nine miRNAs: mir-21,29a,29c,148a,141,hsa-let-7b,mir-31,451 and 192. These miRNAs displayed significant values ($p - value < 0.01$ *and* *fold change* > 5) and facilitated the clustering of samples into two distinct groups: healthy tissue and gastric cancer tissue. In a clustering approach called PatternClus for gene expression data, regulation-based clustering was incorporated and subclusters were identified using an order-preserving ranking method. The method was evaluated using real-life datasets, and the results were reported to be satisfactory. A comparison with popular clustering algorithms such as k-means and hierarchical clustering showed that PatternClus outperformed them in terms of cluster validation measured by z-score.

Chapter 3

Datasets

3.1 Dataset Collection

Human miRNA expression data is obtained from the GEO website, also known as the Gene Expression Omnibus `www.ncbi.nlm.nih.gov/geo/`. The GEO serves as a public repository for genomics data, including sequence or array-based data submitted by researchers after their studies have been published. This data is sourced from actual patients and can be freely downloaded by users for their experiments. Essentially, it functions as an accessible archive of publicly available gene-related data. Each dataset in GEO is assigned a unique identifier called a GEO accession number (e.g., GSE28549), which provides information about the specific dataset, including the type of data it contains and the title of the associated research paper. By using the GEO accession number, researchers can retrieve the desired dataset through the GEO Accession viewer. To determine if a dataset is relevant to our research, we can examine the title of the associated paper and assess the similarities between the work described in that paper and our current research objectives.

This study involves gathering human miRNA expression data from patients diagnosed with various types of cancer, specifically Lung cancer, Breast cancer, Lymphoblastic Leukemia, Esophageal cancer, and Colon cancer. Consequently, data on miRNA expression is collected from individuals affected by these distinct cancer types.

3.2 Datasets

For each cancer type, the patients are categorized into two groups: the control group and the resistant group. The control group consists of patients with cancer who have not received chemotherapy treatment yet, while the resistant group comprises patients who have shown resistance to chemotherapy, with their disease progressing despite the administration of drugs. Datasets with a small number of patients are excluded, and only datasets with a minimum of four patients in each group (control and resistant) are considered for the study.

The datasets available in GEO, are accompanied by documents that provide information about the data acquisition process, including the number of patients involved, patient categorization for the experiment, the specific cancer treatment drug used, and the duration of treatment. It is crucial to have a clear understanding of the dataset before proceeding with the main work, particularly when it pertains to a critical issue like cancer. Additionally, each miRNA in the dataset is assigned an identification (ID), and it is important to determine which microRNA is represented by each ID. This information can be obtained by referring to the document that contains the platform details, where the miRNAs corresponding to their respective IDs can be found.

This study utilizes seven distinct cancer datasets sourced from the publicly available Gene Expression Omnibus (GEO) database. These datasets cover a range of cancer types, namely breast cancer, colon cancer treated with fluorouracil (colon-FU), colon cancer with methotrexate (colon-M), esophageal cancer with cisplatin (esophageal cis), esophageal cancer with fluorouracil (esophageal FU), and lung cancer. The primary focus of these datasets is to analyze the expression levels of microRNAs (miRNAs) in both control (untreated) samples and samples obtained from patients who have developed drug resistance.

The selected datasets have been previously studied, and the miRNAs related to drug resistance have been identified. The GEO repository provides readily available expression data for both control and drug-resistant samples.

The breast cancer dataset consists of twelve samples, with six samples representing control cases and 6 samples representing drug-resistant cases. The dataset includes information on the expression levels of 654 miRNAs.

For colon-FU cancer, the dataset includes 8 samples, with 4 control samples and 4 drug-resistant samples. It comprises expression data for 723 miRNAs.

Similarly, the dataset for colon-M cancer consists of six samples, with three control samples and three drug-resistant samples. It includes expression data for 723 miRNAs.

Both esophageal cancer datasets consist of twelve samples, with six control samples and six drug-resistant samples. These datasets encompass expression data for 847 miRNAs.

Finally, the lung cancer dataset comprises eight samples, with four control samples and four drug-resistant samples, along with expression data for 377 miRNAs.

Table 3.1: miRNA expressions for Resistant group in Breast Cancer

	0	**1**	**2**	**3**	**4**	...	**651**	**652**	**653**
0	11.43	10.75	6.72	8.88	3.63	...	4.85	6.06	5.45
1	11.37	10.73	6.98	9.04	3.73	...	4.98	6.05	5.32
2	11.27	10.84	6.85	9.02	3.61	...	4.95	6.13	5.62
3	11.19	10.84	6.80	9.26	3.70	...	5.05	6.03	5.31
4	11.16	10.88	6.76	9.23	3.57	...	4.58	6.02	5.45
5	11.17	10.97	6.77	9.30	3.89	...	5.13	5.95	5.50

Table 3.2: miRNA expressions for Control group in Breast Cancer

	0	**1**	**2**	**3**	**4**	...	**651**	**652**	**653**
0	7.8	6.9	4.1	7.0	4.8	...	5.2	4.7	5.2
1	7.8	6.9	4.1	7.0	4.9	...	4.9	4.7	5.3
2	7.9	6.8	4.3	6.9	4.9	...	4.9	4.5	5.2
3	7.8	6.8	5.7	6.4	4.7	...	5.1	6.5	4.9
4	7.7	6.8	5.5	6.5	4.6	...	4.8	6.5	5.2
5	7.5	6.8	5.4	6.3	4.6	...	4.8	6.4	5.0

Chapter 4

Proposed Approach

As mentioned earlier, the primary objective of this study is to identify the specific microRNAs (miRNAs) related to drug resistance in various cancer types. In this regard, a new score called, ADFCPCC, is developed where absolute distance(AD), fold change(FC), and Pearson correlation coefficient(PCC) are integrated into a weighted framework. To accomplish this, the dataset was divided into two groups: control and resistant. The control group comprises miRNA expression data from cancer patients who have not yet undergone chemotherapy. On the other hand, the resistant group consists of miRNA expression data from patients who have developed resistance to the chemotherapy drugs they are receiving as part of their treatment. The main steps of ADFCPCC are as follows:

i. Load the dataset of control and resistant groups and some known data of both groups responsible for a single type of cancer (e.g., Breast cancer).

ii. Calculate the mean of control and resistant expressions for each miRNA for both the known and unknown data.

$$mc_x = \frac{1}{n}\sum_{i=0}^{n-1} c_i^x \tag{4.1}$$

$$mr_x = \frac{1}{n}\sum_{i=0}^{n-1} r_i^x \tag{4.2}$$

iii. Compute absolute distance and fold change values by using the means of control and resistant expressions of each miRNA.

$$D_x = |mc_x - mr_x| \tag{4.3}$$

$$\begin{aligned} F_x &= \frac{mc_x - mr_x}{mc_x}, \text{ if } mc_x - mr_x \geq 0 \\ &= \frac{mr_x - mc_x}{mr_x}, \text{ if } mr_x - mc_x \geq 0 \end{aligned} \tag{4.4}$$

iv. Compute the Pearson correlation coefficient of each miRNA between control and resistant data in both known and unknown data and scale them in [0,1].

$$corr_{A,B} = \frac{\sum (A_i - \bar{A})(B_i - \bar{B})}{\sqrt{\sum (A_i - \bar{A})^2 \sum (B_i - \bar{B})^2}} \tag{4.5}$$

v. Then absolute distance(AD), fold change(FC), and Pearson correlation coefficient(PCC) are integrated into a weighted framework, and the power of fold change and the weights are adjusted to minimize the average ranking of known miRNAs.

$$l_x = \frac{a * D_x + b * abs(F_x)^p + c * corr_x}{a + b + c} \tag{4.6}$$

vi. Arrange the miRNAs in higher to lower order based on the measured value mentioned above.

vii. Choose a subset of miRNAs from the ordered list, focusing on the top-ranked ones, and perform biological validation to confirm their significance and functional relevance.

Chapter 5

Experimental Results

This method generates a list of ranked miRNAs.We assess the effectiveness of this method by examining their ability to identify the miRNAs associated with drug resistance in the top positions of the list. To validate the results of our computational methods, we compare the identified miRNAs with those obtained from biochemical experiments as known drug-resistant miRNAs.

5.1 Self Evaluation

Our method was evaluated using existing biochemical studies. Table 5.1 presents the highest ranks of miRNAs achieved by this method for different datasets. The table reveals that the responsible miRNA for the colon treated with fluorouracil is ranked at the top of the list.

The results clearly indicate that in all examined cases, the miRNAs identified as responsible in the corresponding biological studies are among the top 30 in the ranked list. Interestingly, in Fluorouracil treatments in colon cancer (as shown in Table 5.1), the responsible miRNAs achieve the highest rank. Similarly, in Lymphoblastic Leukemia cancer, the responsible miRNA attains the 4th highest rank. The only exception is Esophageal_CIS cancer data, where the only responsible miRNAs hold the 32nd and 33rd ranks.

TABLE 5.1: Results Table

Cancer Type	Known miRNA names	Ranks
Breast	hsa-miR-21	10
	30c	11
	221	13
	451	14
	320a	17
Colon_FU	21	1
	15b	9
	24	13
	107	18
	let-7b	19
	26b	20
Colon_M	193b	12
	210	13
	365	16
	455-3p	18
	27b	19
Esophageal_FU	192_st	26
Esophageal_CIS	532-5p_st	32
	638_st	33
Lung	135b	6
Lymphoblastic Leukemia	100	4
	125b	5
	223	10
	125a	14

5.2 Comparison with Related Methods

We assess the effectiveness of our proposed method by conducting a comparative analysis with several established feature selection methods commonly employed in miRNA or gene expression analysis in cancer. These methods, including EDWFC(2), SPEM(3), 1D-CNN(13), and MRMR(14), encompass a combination of classical approaches with well-established track records and recent advancements in miRNA or gene selection based on expression values.

To evaluate the performance of these approaches, we rank all the miRNAs using each respective method and select the top 5% miRNAs from the resulting ranked lists. To compare the selected miRNAs in terms of their accuracy in distinguishing between control and drug-resistant cancer patients, we utilize support vector machine (SVM) as the primary

classifier. Additionally, we conduct a similar comparison using random forest as an alternative classifier, in addition to SVM. Finally, we assess the performance of these methods by measuring different metrics using the top 5% miRNAs selected by each approach.

Our method using SVM consistently outperforms or shows similar results to other methods in most cases, as indicated by the bold markings in Table 5.2. The exceptions are observed using i) Colon_FU data where the best result is (0.81) in terms of sensitivity is obtained by EDWFC method. The second best result (0.80) is obtained using our ADFCPCC, ii) Colon_M data by SPEM & MRMR, and iii) Lymphoblastic Leukemia data using 1D CNN, where our method slightly falls behind in terms of sensitivity (0.80 vs. 0.83), accuracy (88.5 vs. 91.37), F1 score (0.87 vs. 0.91), and MCC (0.77 vs. 0.80). In summary, the ADFCPCC method yields the best results in 169 cases out of 175 (7 datasets x 5 methods x 5 criteria). In a similar way, using random forest classifier(see Table 5.3), the ADFCPCC method exhibits superior performance in 167 cases out of 175.

TABLE 5.2: COMPARING ADFCPCC WITH DIFFERENT METHODS USING SVM CLASSIFIER

Cancer Type	Method	Sensitivity	Specificity	Accuracy	F1-score	MCC
Breast	**ADFCPCC**	**1**	**1**	**100**	**1**	**1**
	EDWFC	1	1	100	1	1
	SPEM	1	1	100	1	1
	CNN	1	1	100	1	1
	MRMR	1	1	100	1	1
Colon_FU	**ADFCPCC**	**0.8**	**0.86**	**83.3**	**0.8**	**0.66**
	EDWFC	0.81	0.71	76.19	0.77	0.52
	SPEM	0.79	0.56	62.50	0.66	0.38
	CNN	0.72	0.80	75.0	0.76	0.51
	MRMR	0.24	0.57	40.48	0.34	-0.20
Colon_M	**ADFCPCC**	**0.67**	**1**	**83.3**	**0.8**	**0.71**
	EDWFC	0.63	0.81	67.86	0.74	0.40
	SPEM	0.84	0.58	63.80	0.68	0.41
	CNN	0.60	0.73	70.83	0.67	0.37
	MRMR	0.71	0.71	71.42	0.71	0.43
Esophageal_FU	**ADFCPCC**	**1**	**1**	**100**	**1**	**1**
	EDWFC	0.91	0.88	89.58	0.89	0.79
	SPEM	0.83	0.92	87.50	0.87	0.73
	CNN	0.89	0.73	83.44	0.60	0.76
	MRMR	0.88	0.90	89.28	0.89	0.79
Esophageal_CIS	**ADFCPCC**	**1**	**1**	**100**	**1**	**1**
	EDWFC	0.91	0.88	89.58	0.89	0.79
	SPEM	0.83	0.92	87.50	0.87	0.73
	CNN	0.91	0.69	80.64	0.58	0.71
	MRMR	0.79	0.87	83.33	0.83	0.67
Lung	**ADFCPCC**	**0.8**	**0.86**	**83.3**	**0.8**	**0.66**
	EDWFC	0.75	0.75	75.00	0.75	0.50
	SPEM	0.61	0.72	66.57	0.59	0.32
	CNN	0.66	0.71	68.75	0.70	0.40
	MRMR	0.42	0.33	37.50	0.37	-0.25
Lymphoblastic Leukemia	**ADFCPCC**	**0.8**	**0.95**	**88.5**	**0.87**	**0.77**
	EDWFC	0.75	0.71	72.99	0.72	0.46
	SPEM	0.74	0.69	70.67	0.66	0.34
	CNN	0.83	0.92	91.37	0.91	0.8
	MRMR	0.63	0.63	63.21	0.63	0.26

TABLE 5.3: COMPARING ADFCPCC WITH DIFFERENT METHODS USING RANDOM FOREST CLASSIFIER

Cancer Type	Method	Sensitivity	Specificity	Accuracy	F1-score	MCC
Breast	**ADFCPCC**	**1**	**1**	**100**	**1**	**1**
	EDWFC	1	1	100	1	1
	SPEM	0.98	0.95	96.61	0.98	0.95
	CNN	1	1	100	1	1
	MRMR	0.93	0.97	95.00	0.95	0.90
Colon_FU	**ADFCPCC**	**0.8**	**0.71**	**75**	**0.73**	**0.51**
	EDWFC	0.74	0.79	71.88	0.74	0.44
	SPEM	0.79	0.56	62.50	0.66	0.38
	CNN	0.80	0.66	70.32	0.74	0.47
	MRMR	0.53	0.35	43.75	0.48	-0.13
Colon_M	**ADFCPCC**	**0.67**	**1**	**83.3**	**0.8**	**0.71**
	EDWFC	0.71	0.63	66.67	0.68	0.35
	SPEM	0.66	0.53	60	0.62	0.18
	CNN	0.73	0.66	68.75	0.65	0.38
	MRMR	0.70	0.74	72.22	0.72	0.45
Esophageal_FU	**ADFCPCC**	**0.83**	**1**	**91.66**	**0.91**	**0.84**
	EDWFC	0.90	0.89	89.82	0.90	0.80
	SPEM	0.90	0.88	88.88	0.89	0.77
	CNN	0.88	0.72	80.56	0.82	0.63
	MRMR	0.83	0.80	81.67	0.82	0.64
Esophageal_CIS	**ADFCPCC**	**1**	**1**	**100**	**1**	**1**
	EDWFC	0.91	0.89	89.84	0.90	0.8
	SPEM	0.9	0.88	88.89	0.89	0.77
	CNN	0.91	0.74	83.33	0.84	0.66
	MRMR	0.86	0.82	85	0.85	0.7
Lung	**ADFCPCC**	**0.8**	**1**	**91.6**	**0.89**	**0.84**
	EDWFC	0.75	0.75	75.00	0.75	0.50
	SPEM	0.66	0.68	67.50	0.68	0.37
	CNN	0.62	0.73	65.63	0.69	0.33
	MRMR	0.53	0.49	50.71	0.52	0.02
Lymphoblastic Leukemia	**ADFCPCC**	**0.68**	**0.89**	**79.3**	**0.75**	**0.59**
	EDWFC	0.65	0.70	67.68	0.67	0.36
	SPEM	0.67	0.64	66.20	0.67	0.32
	CNN	0.81	0.87	83.62	0.84	0.65
	MRMR	0.49	0.51	50.68	0.50	0.02

Chapter 6

Conclusion and Future Works

6.1 Future Works

Future work for this thesis could include the following:

1. Development of a predictive model: Further refine and improve the predictive model for identifying drug-resistant miRNAs in cancer. Explore different machine learning algorithms and techniques to enhance the accuracy and reliability of predictions.

2. Integration of multi-omics data: Incorporate additional data sources such as gene expression profiles, genomic variations, and clinical information to enhance the predictive power of the model. Integration of multi-omics data can provide a more comprehensive understanding of drug resistance mechanisms in cancer.

3. Validation of predictions: Validate the predictions made by the model using independent datasets or experimental studies. Conduct functional experiments to confirm the role of identified drug-resistant miRNAs in cancer drug resistance.

4. Exploration of underlying mechanisms: Investigate the biological mechanisms and pathways associated with drug-resistant miRNAs. Explore the interactions between miRNAs, target genes, and drug targets to gain insights into the molecular basis of drug resistance.

5. Application in personalized medicine: Explore the potential of using the predicted drug-resistant miRNAs as biomarkers for personalized medicine approaches. Assess the clinical utility of these miRNAs in predicting treatment response and guiding therapeutic decisions.

6. Comparative analysis: Compare the effectiveness of the created model with existing methods for predicting drug-resistant miRNAs in cancer. Evaluate its strengths and limitations, and identify areas for further improvement.

7. Data sharing and collaboration: Share the developed dataset, model, and findings with the scientific community to facilitate collaboration and further advancements in the field of predicting drug-resistant miRNAs in cancer.

By addressing these future directions, the paper can contribute to the ongoing efforts in understanding and combating drug resistance in cancer, ultimately leading to improved treatment strategies and patient outcomes.

6.2 Conclusion

The developed method presents a comprehensive approach for identifying miRNAs that are associated with drug resistance in various types of cancer. The work emphasizes the importance of integrating various similarity measures through weights using existing biological knowledge for selecting drug resistant miRNAs in cancer. The weights are varied in a way to maximize the classification accuracy of cancer patients for the selected genes. The developed predictive model can more accurately identify miRNAs responsible for drug resistance than related methods through the analysis of large-scale miRNA expression datasets and integration of relevant clinical information. The findings highlight the potential of miRNAs as predictive biomarkers for treatment response and patient classification in cancer.

The research may help in developing personalized medicine for targeted therapies. Identification of specific miRNAs, associated with drug resistance, should help clinicians and researchers to make informed decisions regarding treatment strategies and potentially overcome resistance mechanisms. The model can serve as a valuable resource for future studies and contributes towards expanding the pool of information and broader understanding of miRNA-mediated drug resistance. We hope that these findings will ultimately lead to improved patient outcomes and more effective cancer treatments.

Bibliography

[1] O. Kovalchuk, J. Filkowski, J. Meservy, Y. Ilnytskyy, V. P. Tryndyak, C. F. Vasyl, and I. P. Pogribny, "Involvement of microrna-451 in resistance of the mcf-7 breast cancer cells to chemotherapeutic drug doxorubicin," *Molecular cancer therapeutics*, vol. 7, no. 7, pp. 2152–2159, 2008.

[2] A. Kundu, J. Singh, J. Kumar Pal, and S. Sankar Ray, "Predicting drug-resistant mirnas in cancer," *Netw Model Anal Health Inform Bioinforma 12, 6*, December 2022.

[3] J. K. Pal, S. S. Ray, and S. K. Pal, "Identifying drug resistant mirnas using entropy based ranking," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 18, no. 3, pp. 973–984, 2019.

[4] L. L. Gomes, F. C. Moreira, I. G. Homay, S. Santos, P. Assumpcao, A. L. Santana, and A. Ribeiro-dos Santos, "Identification of mirnas expression profile in gastric cancer using self-organizing maps (som)," *Bioinformation*, vol. 10, no. 5, pp. 246–250, 2014.

[5] B. Wolfson, G. Eades, and Q. Zhou, "Roles of microrna-140 in stem cell-associated early stage breast cancer," *World Journal of Stem Cells*, vol. 6, no. 5, pp. 591–597, 2014.

[6] P. Buendia, B. Cadwallader, and V. DeGruttola, "A phylogenetic and markov model approach for the reconstruction of mutational pathways of drug resistance," *Bioinformatics*, vol. 25, no. 19, pp. 2522–2529, 2009.

[7] R. Hummel, C. Sie, D. I. Watson, T. Wang, A. Ansar, M. Z. Michael, M. Van der Hoek, J. Haier, and D. J. Hussey, "Microrna signatures in chemotherapy resistant esophageal cancer cell lines," *World J Gastroenterol*, vol. 20, no. 40, pp. 14904–14912, 2014.

[8] T. E. Miller, K. Ghoshal, B. Ramaswamy, S. Roy, J. Datta, C. L. Shapiro, S. Jacob, and S. Majumder, "Microrna-221/222 confers tamoxifen resistance in breast cancer

by targeting p27kip1*," *Journal of Biological Chemistry*, vol. 283, no. 44, pp. 29897–29903, 2008.

[9] J. Bockhorn *et al.*, "Microrna-30c inhibits human breast tumour chemotherapy resistance by regulating twf1 and il-11," *Nature Communications*, vol. 4, no. 1, pp. 1–14, 20XX.

[10] A. Palagani, K. Op de Beeck, S. Naulaerts, J. Diddens, C. S. Chirumamilla, G. Van Camp, K. Laukens, K. Heyninck, S. Gerlo, P. Mestdagh, J. Vandesompele, and W. Vanden Berghe, "Ectopic microrna-150-5p transcription sensitizes glucocorticoid therapy response in mm1s multiple myeloma cells but fails to overcome hormone therapy resistance in mm1r cells," *PLOS ONE*, 2014.

[11] K. Kurokawa, T. Tanahashi, T. Iima, Y. Yamamoto, Y. Akaike, K. Nishida, K. Masuda, Y. Kuwano, Y. Murakami, M. Fukushima, and K. Rokutan, "Role of mir-19b and its target mrnas in 5-fluorouracil resistance in colon cancer cells," *Journal of Gastroenterology*, vol. 47, pp. 883–895, August 2012.

[12] A. Ripoli, G. Rainaldi, M. Rizzo, A. Mercatanti, and L. Pitto, "The fuzzy logic of microrna regulation: A key to control cell complexity," *PMC Current Genomics*, vol. 11, no. 5, pp. 350–353, 2010.

[13] M. Mostavi, Y.-C. Chiu, Y. Huang, and Y. Chen, "Convolutional neural network models for cancer type prediction based on gene expression," *BMC medical genomics*, vol. 13, no. 5, pp. 1–13, 2020.

[14] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[15] P. A. Mundra and J. C. Rajapakse, "Svm-rfe with mrmr filter for gene selection," *IEEE Transactions on NanoBioscience*, vol. 9, no. 1, pp. 31–37, 2010.