Soft Computing Approach for Optimization of siRNA Efficiency Prediction for Post-transcriptional Gene Silencing

Thesis Submitted to

Cochin University of Science and Technology

For the award of the degree of

Doctor of Philosophy

Under

Faculty of Technology

By

Reena Murali

Under the Supervision of

Dr. David Peter S

DEPARTMENT OF COMPUTER SCIENCE

COCHIN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Kochi – 682022

February 2016

DEPARTMENT OF COMPUTER SCIENCE
COCHIN UNIVERSITY OF SCIENCE AND TECHNOLOGY
COCHIN-682022, KERALA, INDIA

CERTIFICATE

This is to certify that the thesis entitled **"Soft Computing Approach for Optimization of siRNA Efficiency Prediction for Post-transcriptional Gene Silencing"** is a bonafide record of the research carried out by Ms. Reena Murali under my supervision and guidance at the Department of Computer Science, in partial fulfillment of the requirements for the Degree of Doctor of Philosophy under the Faculty of Technology, Cochin University of Science and Technology.

Kochi,
15-02-2016

Dr. David Peter S
Supervising Guide
Department of Computer Science
Cochin University of Science and Technology
Kochi-682022, Kerala

DEPARTMENT OF COMPUTER SCIENCE
COCHIN UNIVERSITY OF SCIENCE AND TECHNOLOGY
COCHIN-682022, KERALA, INDIA

CERTIFICATE

This is to certify that all the relevant corrections and modifications suggested by the audience during the pre-synopsis seminar and recommended by the Doctoral Committee of the candidate have been incorporated in the thesis entitled **"Soft Computing Approach for Optimization of siRNA Efficiency Prediction for Post-transcriptional Gene Silencing".**

Kochi,
15-02-2016

Dr. David Peter S
Supervising Guide
Department of Computer Science
Cochin University of Science and Technology
Kochi-682022, Kerala

DECLARATION

I, **Reena Murali**, hereby declare that the thesis titled **"Soft Computing Approach for Optimization of siRNA Efficiency Prediction for Post-transcriptional Gene Silencing",** submitted to Cochin University of Science and Technology under Faculty of Technology is the outcome of the original research done by me under the supervision and guidance of Dr.David Peter S, Professor, Department of Computer Science, Cochin University of Science and Technology. I also declare that this work did not form part of any dissertation submitted for the award of any degree, diploma, associateship, or any other title or recognition from any University or Institution.

Kochi,
15-02-2016

Reena Murali

Dedicated to my Father

Sree R. Muraleedharan

Acknowledgements

This research work has been undertaken with the unconditional support and encouragement of many and so it is my pleasure to convey my deep felt gratitude to all of them.

First and foremost, I express my utmost and profound gratitude to my supervising guide Dr. David Peter S, Professor, Department of Computer Science, Cochin University of Science and Technology for his valuable guidance, encouragement and support throughout the course of work. I am grateful for his constructive comments and careful evaluation of my thesis.

It is my privilege to express my gratitude to Dr. Sumam Mary Idicula, Professor and Head, Department of Computer Science, Cochin University of Science and Technology for providing timely suggestions as well as necessary facilities in the institute, for my research work. I also express my sincere gratitude to Dr. Sheena Mathew of the institute for giving valuable suggestions and encouragement during the course of my study. I thank all the administrative and supporting staff of the institute for their support and help.

I acknowledge with gratitude to Dr. Indiradevi K.P, Principal, Rajiv Gandhi Institute of Technology, who has always been helpful during the entire course of my work. Also I would like to thank my friends Dr. Vinod Chandra S.S and Dr. Vineetha S, for their timely help, sincere suggestions and encouragement during my research.

It would not have been possible to undertake this journey without the support of my family. The love and prayers of my parents towards me have always been a great strength which helped me throughout. Also I would like to express my gratitude towards my husband Dr. Raghunathan Rajesh for his valuable suggestions, advice and mental support. Without his help, the completion of the work would not have been possible. I am deeply indebted to my children Naveen and Nandan, who bore with me in spite of lack of proper attention and care during my research.

I am also grateful to all who have been helpful during the entire course of the work. Above all, I express my gratitude to the Almighty for showering His choicest blessings up on me in my journey through this research.

Reena Murali

ABSTRACT

Post-transcriptional gene silencing by RNA interference is mediated by small interfering RNA called siRNA. This gene silencing mechanism can be exploited therapeutically to a wide variety of disease-associated targets, especially in AIDS, neurodegenerative diseases, cholesterol and cancer on mice with the hope of extending these approaches to treat humans. Over the recent past, a significant amount of work has been undertaken to understand the gene silencing mediated by exogenous siRNA. The design of efficient exogenous siRNA sequences is challenging because of many issues related to siRNA. While designing efficient siRNA, target mRNAs must be selected such that their corresponding siRNAs are likely to be efficient against that target and unlikely to accidentally silence other transcripts due to sequence similarity. So before doing gene silencing by siRNAs, it is essential to analyze their off-target effects in addition to their inhibition efficiency against a particular target. Hence designing exogenous siRNA with good knock-down efficiency and target specificity is an area of concern to be addressed. Some methods have been developed already by considering both inhibition efficiency and off-target possibility of siRNA against a

gene. Out of these methods, only a few have achieved good inhibition efficiency, specificity and sensitivity.

The main focus of this thesis is to develop computational methods to optimize the efficiency of siRNA in terms of "inhibition capacity and off-target possibility" against target mRNAs with improved efficacy, which may be useful in the area of gene silencing and drug design for tumor development. This study aims to investigate the currently available siRNA prediction approaches and to devise a better computational approach to tackle the problem of siRNA efficacy by inhibition capacity and off-target possibility. The strength and limitations of the available approaches are investigated and taken into consideration for making improved solution. Thus the approaches proposed in this study extend some of the good scoring previous state of the art techniques by incorporating machine learning and statistical approaches and thermodynamic features like whole stacking energy to improve the prediction accuracy, inhibition efficiency, sensitivity and specificity. Here, we propose one Support Vector Machine (SVM) model, and two Artificial Neural Network (ANN) models for siRNA efficiency prediction. In SVM model, the classification property is used to classify whether the siRNA is efficient or inefficient in silencing a target gene. The first ANN

model, named siRNA Designer, is used for optimizing the inhibition efficiency of siRNA against target genes. The second ANN model, named Optimized siRNA Designer, OpsiD, produces efficient siRNAs with high inhibition efficiency to degrade target genes with improved sensitivity-specificity, and identifies the off-target knock-down possibility of siRNA against non-target genes. The models are trained and tested against a large data set of siRNA sequences. The validations are conducted using Pearson Correlation Coefficient, Mathews Correlation Coefficient, Receiver Operating Characteristic analysis, Accuracy of prediction, Sensitivity and Specificity.

It is found that the approach, OpsiD, is capable of predicting the inhibition capacity of siRNA against a target mRNA with improved results over the state of the art techniques. Also we are able to understand the influence of whole stacking energy on efficiency of siRNA. The model is further improved by including the ability to identify the "off-target possibility" of predicted siRNA on non-target genes. Thus the proposed model, OpsiD, can predict optimized siRNA by considering both "inhibition efficiency on target genes and off-target possibility on non-target genes", with improved inhibition efficiency, specificity and sensitivity. Since we have taken efforts to optimize the siRNA efficacy in terms of "inhibition efficiency and off

target possibility", we hope that the risk of "off-target effect" while doing gene silencing in various bioinformatics fields can be overcome to a great extent. These findings may provide new insights into cancer diagnosis, prognosis and therapy by gene silencing. The approach may be found useful for designing exogenous siRNA for therapeutic applications and gene silencing techniques in different areas of bioinformatics.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

List of Abbreviations

A	**-**	**Adenine**
Acc	**-**	**Accuracy**
ANN	**-**	**Artificial neural network**
API	**-**	**Application Programme Interface**
AUC	**-**	**Area Under Curve**
bi-siRNA	**-**	**bi-functional siRNA**
C	**-**	**Cytosine**
cDNA	**-**	**complementary DNA**
DNA	**-**	**Deoxyribonucleic Acid**
dsRNA	**-**	**double stranded RNA**
FN	**-**	**False Negative**
FP	**-**	**False Positive**
FPR	**-**	**False Positive Rate**
G	**-**	**Guanine**
G-C	**-**	**Guanine-Cytosine**
HD	**-**	**Huntington Disease**
IDE	**-**	**Integrated Development Environment**
LIBSVM	**-**	**Library for Support Vector Machine**
MCC	**-**	**Mathews Correlation Coefficient**
miRNA	**-**	**micro RNA**
mRNA	**-**	**messenger RNA**
NCBI	**-**	**National Centre for Biotechnology Information**
ncRNA	**-**	**non coding RNA**
nt	**-**	**nucleotide**

OpsiD	**-**	**Optimized siRNA Designer**
PTGS	**-**	**Post Transcriptional Gene Silencing**
RBF	**-**	**Radial Basis function**
RefSeq	**-**	**Reference Sequence**
RISC	**-**	**RNA Induced Silencing Complex**
RNA	**-**	**Ribonucleic acid**
RNAi	**-**	**RNA interference**
ROC	**-**	**Receiver Operating Characteristics**
RProp	**-**	**Resilient Propagation**
rRNA	**-**	**Ribosomal RNA**
SCG	**-**	**Scaled Conjugate Gradient**
siRNA	**-**	**small interfering RNA**
Sn	**-**	**Sensitivity**
Sp	**-**	**Specificity**
SVC	**-**	**Support Vector Classification**
SVM	**-**	**Support vector machine**
SVR	**-**	**Support Vector Regression**
T	**-**	**Thymine**
TGS	**-**	**Transcriptional Gene Silencing**
TN	**-**	**True Negative**
TP	**-**	**True Positive**
TPR	**-**	**True Positive Rate**
tRNA	**-**	**Transfer RNA**
U	**-**	**Uracil**

……..✷♦✷……..

Chapter - 1
Introduction

Contents

1.1 Relevance of siRNA Prediction

RNA interference (RNAi) is a biological process which can control the gene regulation by sequence specific post-transcriptional gene silencing mechanism [1,2]. In functional genomic research, the discovery of RNAi has become much helpful in drug design and therapeutic applications because of its ability to perform gene silencing. It has great potential in future therapeutics as it has the ability to regulate many disease-associated genes. RNAi has been successfully used to target diseases such as AIDS [3], neurodegenerative diseases [4], cholesterol [5] and cancer [6] on

mice with the hope of extending these approaches to treat humans. Post-transcriptional gene silencing by RNAi is mediated by small interfering RNA (siRNA). The siRNA molecules are double stranded nucleic acids approximately 19-21 nucleotide in length that act as the mediators of RNAi. siRNAs interact with their cognate messenger RNAs (mRNA) and subsequently trigger degradation of the target mRNAs in a sequence specific fashion. The consequence of mRNA degradation is a reduction in protein expression or gene silencing. This gene silencing mechanism can be exploited therapeutically to a wide variety of disease-associated targets [7-8], especially in cancer, which is formed because of uncontrolled cell proliferation due to malfunctioning of regular cell division process. Because the RNAi mechanism results in sequence specific mRNA degradation, it has the potential to realize cancer therapy by specifically attacking the cancer cells and minimizing the effect on normal healthy cells. siRNA molecules have the potential to revolutionize cancer therapy by providing highly potent and specific cancer cell killing ability with drastically reduced side effects. Recently, it has been reported that some research area of drug design in cancer therapy is concentrating to artificially inject exogenous siRNA capable of degrading the mRNA responsible for tumour development. Therefore, identification of efficient siRNA capable of degrading target mRNA responsible for tumor development is a key step towards the diagnosis and treatment of cancer. Thus siRNAs are new promising therapeutic agents that are perfectly suited for gene silencing and molecularly targeted

cancer therapy. siRNA can be endogenous or exogenous. The use of exogenous siRNA for performing gene silencing has become an important biological milestone for mRNA target identification and drug design [9-11] in various areas of bioinformatics.

1.2 Issues in Predicting Efficient siRNA

A significant amount of work has been undertaken over the recent past to understand the gene silencing mediated by exogenous siRNA. Many models have been proposed to predict efficient siRNAs against target mRNA. Even though several algorithms and methods have been presented to predict efficiency of siRNA, only a few have achieved an acceptable level of efficacy, due to the following issues related to siRNA.

From the siRNA related studies, it is understood that among all siRNAs that can be generated against a target mRNA, only a few are found successful in causing degradation and the efficiency of such siRNA differs in different target sites of same mRNA. However even those few do not perform equal knock-down effects [12]. Also, it was earlier understood that full complementary siRNA is needed to silence a target gene. But recent studies reveal that siRNA behaves like micro RNA (miRNA) and can suppress protein synthesis even though it is not fully complementary to the target. This shows that mismatches are allowed during target selection by siRNA [13-14].

This may cause a very serious problem of "*off-target effect*" where unintended genes may be suppressed by selected siRNA [15-17].

Like this, there are many challenges in connection with therapies using gene silencing techniques. Most important challenges are target specificity and effectiveness of delivery. These challenges may prevent effective practical applications of exogenous siRNA. The factors of siRNA like specific targeting, efficient delivery system, validated genes and the potent siRNA sequences are all vital important to overcome these barriers. So target specificity and efficient delivery of siRNA molecules for gene silencing is a serious research issue to be addressed. Special care must be given to design efficient methods to deliver and develop specific gene silencing therapeutics using siRNA in a more safe and effective manner.

1.3 How to Address the Predction Issues?

The design of effective siRNA sequences is challenging because the target mRNAs must be selected such that their corresponding siRNAs are likely to be efficient against that target and unlikely to accidentally silence other transcripts due to sequence similarity. Hence to design efficient siRNAs, the ability of knocking down target genes as well as the off-target possibility on any non-target genes are to be considered. So before doing gene silencing by siRNAs, it is essential to analyze their off-target effects in addition to their inhibition efficiency against a particular target. Thus during

efficient exogenous siRNA design, the following points are to be addressed properly.

- How to design siRNA with specific targeting and efficient delivery system such that
  - they are likely to be efficient against that target?
  - they are unlikely to accidentally silence other transcripts due to sequence similarity?

- How to optimize the inhibition efficiency, prediction accuracy and off-target effect of siRNA?
- What are the computational methodologies that can be used for the design?
- How the efficiency of the computational method can be evaluated?

1.4 Research Problem

The issues related to exogenous siRNA predcition must be meaningfully addressed. So designing efficient siRNA against target mRNA or gene, with good knock-down efficiency and target specificity is an area of concern to be addressed. The efficiency of siRNA must be optimized such that they are capable of inhibiting their target mRNA sequences without affecting any other genes. Thus to design siRNAs, two important concepts must be considered: the ability in knocking down target genes and the off- target possibility on any non-target genes. Only a few methods have been developed

by considering "both inhibition efficiency and off-target possibility" of siRNA against a gene.

The main aim of this study is to propose soft computing approach for predicting efficient exogenous siRNAs capable of performing post-transcriptional gene silencing in mammalian cells, with high inhibition capacity on target genes and low off-target possibility on non-target genes. The thesis also focuses on optimizing the efficiency of predicted exogenous siRNA over the state of the art techniques.

1.5 Extending the State of the Art

The techniques emerged to explore the issues related to exogenous siRNA design are classified into two groups, first generation and second generation methods. As the first generation models were not able to achieve the targeted level of efficacy, there was a need to develop techniques to improve the efficiency of predicted siRNA. These second generation models are based on either artificial neural network or linear regression models. Some of the good scoring second generation models like BIOPREDsi [18], DSIR [19], ThermoComposition21 [20], i-Score [21], Scales [22], OptiRNA [23], siDRM [24], RNAxs [25], siRecords [26], E-RNAi [27], MysiRNA-Designer [28], and MysiRNA [29], DISR [30], RNAiAtlas [31], siSPOTR [32] were developed by introducing data mining techniques to improve the efficiency of siRNA with their

experimental inhibition. Among these techniques, we have considered some good scoring methods to integrate with our technique. These methods are BIOPREDsi [18], DSIR [19], ThermoComposition21[20], i-Score [21] and MysiRNA [29].

BIOPREDsi [18], ThermoComposition21 [20], and MysiRNA [29] used the artificial neural network models, while DSIR [19] and i-Score [21] used linear regression models. ThermoComposition21 [20] improved the prediction accuracy by combined position dependent features together with thermodynamic features in single artificial neural network model. The prediction accuracy is improved in DSIR [19] and i-Score [21] using linear regression model. In MysiRNA [29], the prediction accuracy is further improved by artificial neural network model. The approaches proposed in this thesis extend these selected state of the art techniques, by incorporating machine learning and statistical techniques to improve the prediction accuracy and reduce the off-target possibility of siRNA.

1.6 Goals and Objectives

In this study we propose machine learning approach which optimizes the efficacy of predicted siRNA by inhibition efficiency and off-target possibility against target genes, which is built on existing good scoring second generation models.

The main goals of the study are

1. Design exogenous siRNA capable of performing post-transcriptional gene silencing.

2. Identify siRNA with high inhibition capacity against a target mRNA, with minimum off-target silencing.
   `
3. Compare the efficiency of our approach with the state of the art techniques.

The main Objectives of the study are

1. Design efficient siRNAs for any target messenger RNAs (mRNAs) or complementary DNAs (cDNAs).

2. Predict siRNA inhibition efficiency for a given target mRNA using machine learning techniques.

3. Improve the efficiency by including thermodynamic properties of siRNA.

4. Improve the efficacy of siRNA in terms of accuracy of prediction, target specificity, sensitivity and inhibition capacity than those of the existing approaches.

5. Optimize the siRNA efficacy by combined approach of "inhibition capacity and off-target possibility".

1.7 Research Method

The following are the set of machine learning approaches proposed in this study for finding the efficiency of siRNA data. Based on these algorithms, the efficiency of siRNA against target mRNA were modeled and tested.

i. Support Vector Machine (SVM) model is used to classify and observe the efficiency of siRNA against target mRNA.

ii. Two Artificial Neural Network (ANN) models are designed to improve the efficiency of siRNA against target mRNA.

In SVM model, the classification property is used to classify whether the siRNA is efficient or inefficient in silencing a target gene [33]. Out of the two ANN models, first model is named as siRNA Designer and is used to optimize the inhibition efficiency of the predicted siRNA [34-35]. Second model is the optimized siRNA designer, OpsiD, which optimizes the prediction efficacy in terms of inhibition capacity, prediction accuracy, sensitivity-specificity and off-target possibility over the state of the art techniques using feed forward back propagation neural network model. The research method adopted for this study is shown in the block diagram (Fig. 1.1).

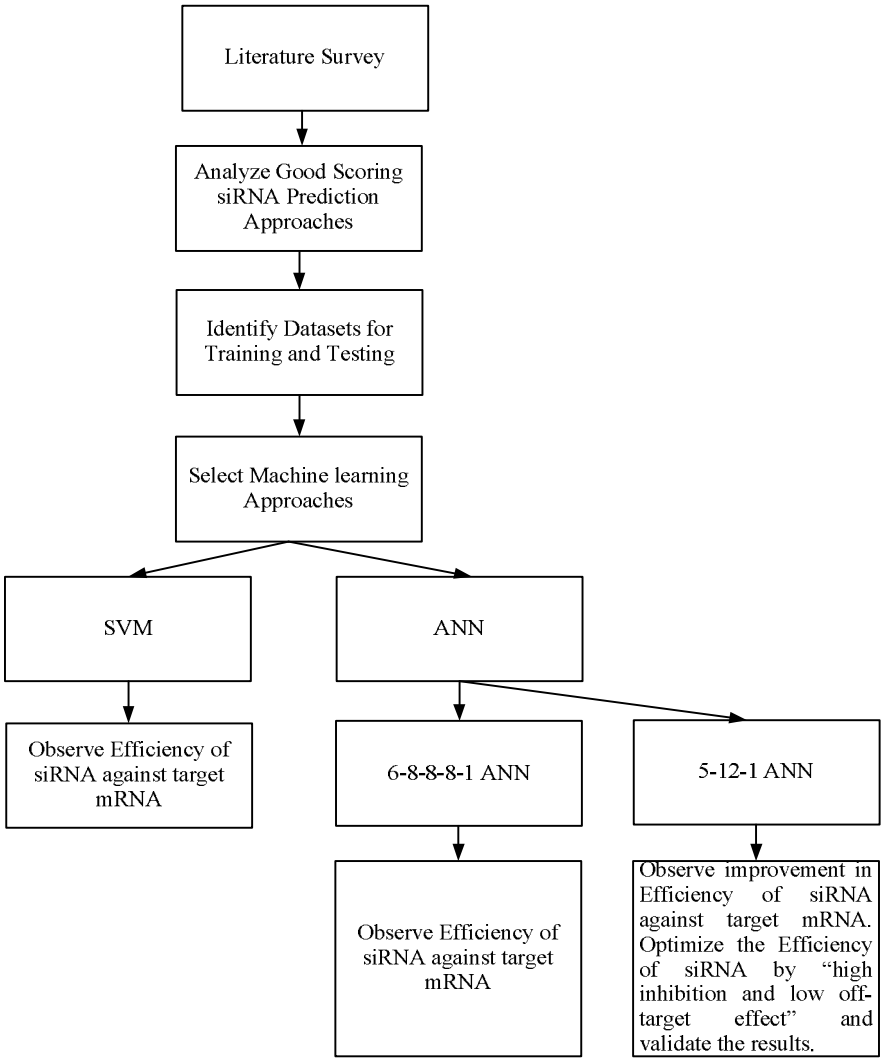


Fig. 1.1: Research Method

1.8 Organization of the Thesis

The layout of the thesis is as follows:

- Chapter 1 describes research problem, goals and objectives of the study.
- Chapter 2 presents biological aspects of gene silencing by RNA interference mechanism and the potential of RNAi in genomics and therapeutics.
- Chapter 3 serves as a brief literature review on relevant work on siRNA efficiency prediction for gene silencing.
- Chapter 4 provides a brief description of materials and methods, machine learning approaches, frame works, training algorithms and validation strategies used in this study.
- Chapter 5 presents the work done for predicting efficiency of siRNA by Support Vector Machines Model.
- Chapter 6 describes the work done for optimizing the inhibition efficiency of predicted siRNA by Artificial Neural Network Model.

- Chapter 7 presents the work done for optimization of predicted siRNA in terms of inhibition efficiency, accuracy of prediction, sensitivity, specificity and off-target possibility.

- Chapter 8 describes the results and discussion. The performance evaluation and comparison with existing approaches are also done in this chapter.

- Chapter 9 summarizes the contributions and some of the limitations as well as future scope of the study.

……..✷♦✷……..

Chapter - 2

Gene Silencing by RNA Interference

Contents

2.1 Introduction

This chapter discusses how gene silencing can be done by RNA interference mechanism. Section 2.2 describes the biological aspects of gene silencing. The mechanism of RNAi and the RNAi pathway are explained the Section 2.3. The next sections, 2.4 and 2.5 present small RNAs mediating RNAi and applications of RNAi respectively. Types of gene silencing like transcriptional gene

silencing and post-transcriptional gene silencing are described in section 2.6. Potential and role of RNAi in genomics and therapeutics, challenges to gene silencing, need and complexities of designing exogenous siRNA are described in sections 2.7, 2.8, 2.9 and 2.10 respectively. Finally, a brief summary of the chapter is presented in section 2.10.

2.2 Biological Aspects of Gene Silencing

Gene is the basic unit of heredity of all living organism which is passed from parents to their offspring [36]. Each gene is a particular segment of DNA with a linear sequence of nucleotides, on a chromosome. They contain chemical information needed for the synthesis of different proteins. A gene determines the characteristics of an individual or a species in the form of protein. Thus genes regulate the operations of organisms and play very important role in differentiating individuals and species. The entire genetic material of an organism is called genome [37]. Genome represents an organism's complete set of DNA, including all its genes and contains entire information needed to build and maintain that particular organism. The genome includes all the genes and the non-coding sequences of the DNA and RNA. DNA carries the essential instructions for building RNA and proteins. Inside the cells of all living things, some molecular mechanisms are constantly reading the information in DNA for building proteins. Thus DNA encodes for the genetic instructions for all living organisms [38-42], and hence DNA is considered as the "blue print of life". A gene is said to be

expressed when a protein is formed due to this molecular mechanism. During gene expression, the information from a gene is used to produce a functional gene product, which may be a protein or a functional RNA.

Genetic codes are set of rules though which the encoding of genetic materials is done. The information in genetic materials is thus translated or encoded into proteins. RNA is a nucleic acid which is responsible for various biological activities like coding of genetic materials into proteins or messenger RNA to amino acids, gene regulation, and expression of genes. Most of the RNAs are single stranded. But there are some special types of RNA with two complementary strands similar to DNA, called double-stranded RNA (dsRNA). An important double-stranded RNA called short interfering RNA or small interferingRNA (siRNA) can trigger RNA interference in eukaryotes, and interferon response in vertebrates [43-46]. RNA can be either coding or non-coding. A non-coding RNA (ncRNA) is a functional RNA molecule that is not translated into a protein [47-50]. Two examples of non-coding RNAs are microRNA (miRNA) and short interfering RNA (siRNA). Coding RNAs play crucial roles in protein synthesis and other cell activities. One important class of coding RNAs is messenger RNA (mRNA). It is a type of RNA that reflects the exact nucleotide sequence of the genetically active DNA. mRNA carries the "message" of the DNA to

the cytoplasm of cells, where protein is made as amino acid sequences specified by the mRNA. Thus mRNA acts as the key intermediary in gene expression by translating the DNA's genetic code into the amino acids that make up proteins. The central dogma of molecular biology describes the flow of genetic information to form proteins [51-52]. It has also been described as "DNA makes RNA and RNA makes protein" [53]. The main steps in Central Dogma are transcription and translation.

2.2.1. Transcription

Transcription is the initial step of gene expression [51-52]. In transcription, a particular segment of DNA is copied into RNA. As a first step, the DNA sequence is read by the enzyme called RNA polymerase. It produces a complementary, anti parallel RNA strand called a primary transcript. The portion of DNA transcribed into an RNA molecule is called a transcription unit and it encodes at least one gene. The transcribed RNA molecule is called mRNA. Fig 2.1 shows the steps during transcription.

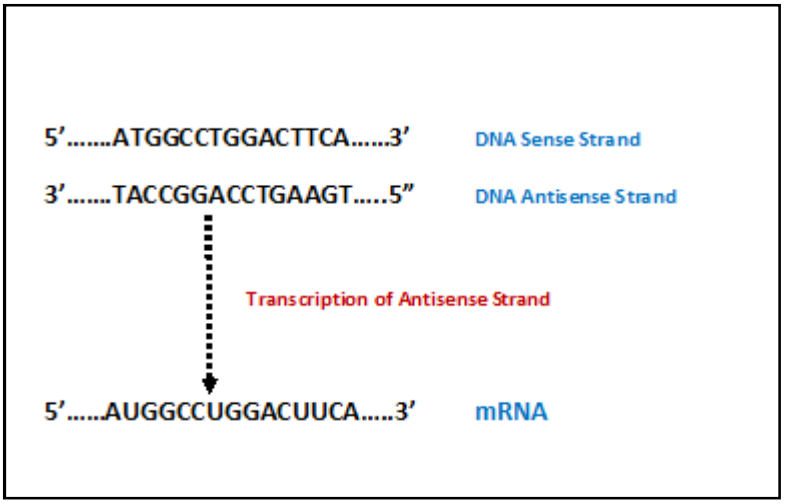


Fig. 2.1: Transcription

2.2.2 Translation

Translation is a process where ribosomes synthesize proteins from the information contained in the mRNA [51-52]. During translation, the ribosome reads a string of three bases on the mRNA (codon) and translates them into one amino acid (Fig. 2.2). Proteins are further processed in various cellular compartments and then transported in and out of the cell to carry out different metabolic functions.

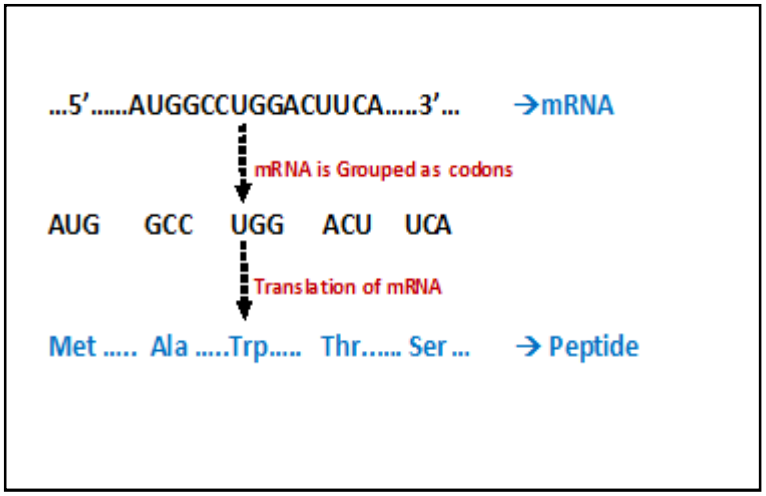


Fig. 2.2: Translation

During gene expression, information encoded in a gene is used for producing the gene products like mRNA and proteins. It covers the entire process from transcription through protein synthesis [53]. In the first step, the DNA on which the gene resides is transcribed to messenger RNA and in second step, it is translated from mRNA to protein. When the protein is synthesized, a gene is said to be "expressed" and the expression level of gene depends on the amount of mRNA it produced. Different cell types in an organism carry out a range of specialized function depends upon the genes that are expressed only in that cell type. Some of the factors affecting gene expression are the age of the person, the type of tissue, the presence of specific chemical signals and heredity.

Gene expression can be controlled by gene regulation [54-55]. Gene regulation is achieved by a process of turning genes on and off. It is the basis of all biological activities like cell growth, cellular differentiation, adaptability and versatility of any organism. Gene regulation controls the appearance of the functional gene product or gene expression. Gene expression is controlled at three levels during the production of an active gene product. First phase is the transcriptional regulation. It mainly takes care of when the gene is transcribed and how much it is transcribed. Second is the translational regulation which controls the amount of proteins synthesized from mRNA. Third phase is post-translational regulation mechanisms which control the level of active gene products. The gene expression can be controlled or altered by making alterations in mRNA or protein. An active mRNA level may be controlled by splicing or by silencing with some of the non-coding RNAs like miRNA (micro RNA), siRNA (short interfering RNA), rRNA (ribosomal RNA) and tRNA (transfer RNA). Also some proteins may undergo self modifications such as folding, enzymatic cleavageand bond formation. These modifications can play crucial roles in the regulation and control of gene expression. Genes can be either up regulated or down regulated. Using down regulation, the expression of a particular gene may be prevented. Gene silencing is done by preventing the expression of a particular gene and there by turning "off" gene expression [54-55].

2.3 Mechanism of RNAi

Earlier, gene knock out was conducted by scientists using antisense, dominant negative or knockout techniques which were time consuming and expensive. The discovery of RNAi and double stranded RNA helped to silence genes very efficiently [2-3, 10-12]. RNAi is an important gene silencing method used in molecular biology over the past few years. The presence of RNAi mechanism was discovered both in plants and animals [56-59]. Short RNAs of length 21-23 nucleotides exist in a double-stranded form, with 2 nucleotide overhangs at each 3′ end [60-62]. They are known as small or short interference RNA. RNAi is a naturally evolved mechanism in insects, nematodes and plants as a result of a developed intrinsic defense against RNA virus [63-67]. This characteristic makes it ideal as the basis for a physiologic approach for both *in vitro* and *in vivo* gene silencing [68-69]. This mechanism has been described in several eukaryotic organisms including human cell lines and primary cells [70-75]. Thus RNAi, is known for co-suppression [76], quelling [77], post-transcriptional gene silencing [78], and plays an important role in cellular anti-viral defenses and silencing mechanisms [79]. The discovery of RNA-mediated gene silencing, changed the view of gene regulation and led to the development of new genetic tools and methods for selective gene silencing, and have opened a way for development of novel therapeutics against various diseases [80].

2.3.1 RNAi Pathway

RNAi targets the protein producing mRNA and controls disease in the transcription phase by generating a non coding RNA called siRNA. The biogenesis of RNAi is divided into 4 steps and shown in Fig. 2.3.

- dsRNA cleavage by Dicer generating siRNAs: When long dsRNA from an external source is introduced into the cell, it is recognized by Dicer. The Dicer is a Ribonuclease III protein which is present in all organisms. The dicer cleaves the dsRNA randomly to generate siRNAs of ~21 to 23 nucleotide in length [81-82]. Each siRNA strand has a 5' phosphate group and a 3' hydroxyl group and has a 2 nucleotide overhang at both ends [82].
- Formation of RISC: The siRNAs created by Dicer initiated cleavage get attached with a nuclease complex called RISC (RNA Induced Silencing Complex). The complex formed is inactive.

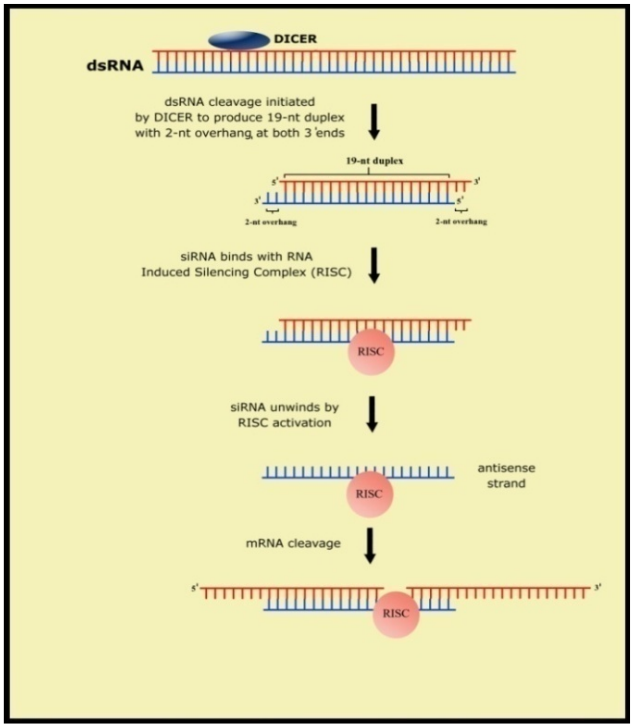


Fig. 2.3: RNAi Pathway

- SiRNA unwinding and RISC activation: Due to RISC activation, siRNA duplexes will unwind and separate into sense and antisense strands. Both the sense and antisense strands of the siRNA are capable of directing RNAi, but specificity depends on the antisense strand. The antisense strand is taken up by RISC. Due to unwinding, siRNA duplex losses one strand that is not bound to the RISC. This single strand RISC complex thus gets activated.

- mRNA targeting and degradation: The activated siRNA-RISC complex will target mRNAs which are complementary with

the siRNA sequence. If the match is perfect, the targeted mRNA is cleaved into smaller fragments which are then degraded [17, 83]. If the match is not perfect the RISC remains stuck to the mRNA, thus translation inhibition occurs.

2.4 Small RNAs of RNAi

Small RNAs are different classes of RNAs which can influence several levels of gene regulation. Here, we are listing two well defined classes of small RNAs: short interfering RNAs and microRNAs.

2.4.1 Short interfering RNA

Short interfering RNA or small interfering RNA (siRNA) is a class of non-coding RNA molecule. siRNAs are short pieces of dsRNAs, which are mediators of RNAi at post-transcriptional level. The structure of siRNA is depicted in Fig. 2.4. This double stranded RNA is composed of a sense and an antisense strand which are paired resulting in a 2 nucleotide 3' overhang at both the ends. siRNA directly induces the RNAi pathway by binding to an almost perfect complementary region of the targeted mRNA transcript and cleaves the mRNA. siRNA plays very crucial role in the RNAi pathway, by degrading the expression of specific genes with complementary nucleotide sequences. siRNAs and their role in post-transcriptional gene silencing (PTGS) in plants were first discovered

by David Baulcombe's group at the Sainsbury Laboratory in Norwich [84]. Later it is reported that synthetic siRNAs could induce RNAi in mammalian cells [60-62,82]. This discovery led to a keen interest in harnessing RNAi for drug development for cancer therapy and various gene silencing applications in biomedical research.

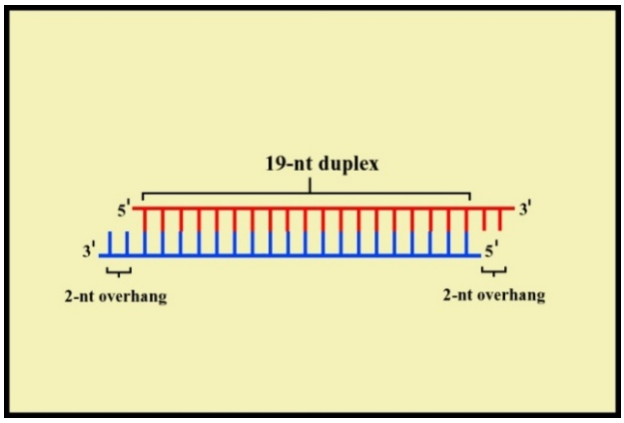


Fig. 2.4: Structure of siRNA

2.4.2 Micro RNA

MicroRNAs (miRNA) were discovered in 1993 by Rosalind Lee, Rhonda Feinbaumand Victor Ambrosduring a study of the gene lin-14 in *C.elegans* development [85]. MicroRNAs are non-coding RNAs that combine to mRNAs and regulate the activities at translational and post-transcriptional level [86]. There are at least 800 miRNAs within the human genome, which may target about 60% of mammalian genes [87-88]. MicroRNAs bind to partially complementary sites in the messenger RNA of other genes and

inhibit the translation of these genes. It has been found that miRNA can effectively regulate biological activities such as cell proliferation, cell differentiation, cell growth, apoptosis, protein synthesis [88-91], and can also act as oncogenes as well as tumor suppressors [92].

2.5 Applications of RNAi

RNAi has become a powerful biological technique for gene function studies and drug discovery [93-94]. It is also becoming increasingly important in developing therapeutic applications for a number of diseases due to its potential for specific targeted silencing [95-96]. Thus RNAi along with siRNA plays important role in gene regulation. Widely used applications of RNAi and siRNA in genomics and therapeutics are

- Selection of possible targets for Tumor therapy [97]
- Gene Therapy [98]
- Better understanding of viral infections [99]
- Gene Silencing [100]

2.6 Gene Silencing by RNAi

When genes are silenced, the expression of those genes is reduced [101]. But when genes are knocked out, they are completely removed from the organism's genome and have no expression at all. It is understood that RNAi is a gene silencing mechanism that

reduces the expression of a gene by at least 70% but do not eliminate completely [101]. The gene regulatory mechanism by RNAi limits the transcript level either by suppressing the transcription or by activating sequence specific mRNA degradation. Based on this, gene silencing is classified as transcriptional gene silencing and post-transcriptional gene silencing.

2.6.1 Transcriptional Gene Silencing

Transcriptional Gene Silencing (TGS) is one type of silencing genes at transcriptional level [2,101]. In this method, due to the effect of silencing, the messenger RNA is not formed and further activities of protein formation are stopped.

2.6.2 Post-transcriptional Gene Silencing

Post-transcriptional gene silencing (PTGS) is another type of silencing genes at post-transcriptional level, means silencing action is done after messenger RNA formation [2,101]. Due to post-transcriptional gene silencing, the targeted messenger RNA is lost or degraded after RNA interference mechanism in gene. Ultimately gene expression will be turned off or gene knock-down may happen.

2.7 Potential of RNAi in Genomics and Therapeutics

The use of RNAi has led to the development of a new technology called siRNA mediated gene silencing. It is used for gene therapy applications in medical research, especially in cancer

therapeutics. Gene specific silencing has allowed systematic approach of designing new drugs, and for enhancing the effect of already existing drugs. RNAi could enable gene silencing with high specificity and improved efficiency than with any other techniques. Instead of transfecting big dsRNA molecules in to the cells, chemically engineered siRNA's enable targeting the specific genes. In principle any gene may be knocked-down by a synthetic siRNA with exact complementary sequence. Hence in the post-genomic era, siRNA is considered as an important tool for validating gene function and drug targeting. The gene silencing capacity of RNAi has been used in cell cultures and in animal models that encourage siRNA based reagents for clinical usage to treat cancer [6] as well as other diseases such as neurodegenerative disorders, cholesterol and viral diseases [4,102-103].

Cancer treatment will be successful if it is able to do complete removal of the tumor without making damage to any other parts of the body. This shall be achieved by doing surgery, to a certain level. But surgery is not as effective if the disease has already spread to other locations of the body. Chemotherapy is sometimes toxic to healthy tissues as it is not specific to cancer cells. Radiation also damage normal cells and tissues. By considering all these limitations of the existing cancer therapy techniques, it is very essential to develop novel target specific therapeutics for the effective treatment of cancer. Recently it is understood that RNAi can be successfully

used in cancer therapies. Nowadays there are lot of insights and promises for using siRNAs as drugs targeted only into the cancer cells. Genes associated with several cancers can be silenced by RNA interference. For example, in *in-vitro* studies of one type of leukemia, it is shown that siRNA could damage the fusion protein, which prevents the drug from binding to the cancer cells [104]. Cleaving or damaging the fusion protein will reduce the amount of transformed cells that spread throughout the body. This is done by increasing the sensitivity of the cells to the drug [104]. RNA interference can be used to target particular mutants. For example, siRNAs were able to bind specifically to tumor suppressor p53 molecules containing a single point mutation and destroy it [105].

Researchers have used siRNAs to selectively regulate the expression of cancer related genes. siRNA molecules are used to target the uncontrolled production of cancer cells, proliferation of breast cancer [106]. Also it is understood that siRNAs can be used to reduce protein formation and can thereby increase the sensitivity of the cancer cells towards chemotherapy treatments [107-108]. *In-vivo* studies are being utilized to study the potential use of siRNA molecules in cancer therapeutics [108]. RNAi has already been used to target particular genes in several serious viral diseases like hepatitis and human immunodeficiency virus (HIV) [102-103]. Especially, siRNA was used to silence the primary HIV receptor named chemokine receptor-5 [109] to prevent the virus from entering

the human peripheral blood lymphocytes and the primary hematopoietic stem cells [109-110]. Gene silencing techniques using RNAi have also been successfully used to target other viruses, such as hepatitis B and C, human Papilloma virus, West Nile Virus and so on. In hepatitis B, siRNA silencing technique was successfully used to target the hepatitis B virus and could effectively decrease the number of viral components [111]. Also, siRNA techniques used in hepatitis C were able to reduce the quantum of the virus in the cell by 98% [112-113]. From recent studies it is understood that siRNA may also be used for diseases like cystic fibrosis and chronic obstructive pulmonary disease, asthma and Huntington's disease (HD) [113-116].

RNAi has great potential in future therapeutics since it has the potential to regulate disease related genes. So any disease caused by abnormal enhancement activity of one or more genes could be regulated by RNAi-based therapies [117]. Over the past several years, a number of RNAi-based preclinical and clinical trials have grown to understand brain and skin diseases, viral infections, respiratory disorders, cancer and metabolic diseases [118]. Till date, RNAi therapies in clinical trials have targeted approximately 14 different diseases [118]. Many of the siRNA therapies are at preclinical stage. The methods for delivering siRNA drugs had been improved to maximize the specificity of siRNA and to minimize the toxicity and degradation effects that compromise drug efficacy [65]. Three clinical trials have used ex-vivo delivery of the siRNA

therapeutics. In this method, cells were collected from patients and treated with siRNAs and re-infused back into the patient [119]. One of the three clinical trials involves use of an anti-tumor bifunctional siRNA (bi-siRNA) for treatment of metastatic melanoma, a form of cancer that originates in melanocytes. The idea used in cancer treatment with RNAi is that cancer cells will be killed through the actions of the patient's own immune system.

2.8 Challenges to Gene Silencing Therapeutics

siRNA is the mediator of RNAi and can do efficient and specific gene silencing. Thus it is extremely promising for various therapeutic applications. But there are many barriers for making effective practical applications of siRNA. siRNA can be transfected directly into the cells or organs. But stability in the blood stream, the duration of the effect and the delivery techniques are still quite big questions before RNAi-based therapy can be used. siRNA stability and targeting may be highly influenced to degradation by various enzymes found in tissues. The life of siRNAs in serum may range from minutes to an hour [120]. Because of this survival problem of siRNA, target site accumulation for therapeutical applications is a major challenge [121]. There are many other challenges in connection with therapies using gene silencing techniques. Most important challenges are target specificity and effectiveness of delivery. For example, in case of neurodegenerative disorders, gene silencing particles must be directly delivered to the brain. The brain-

blood barrier may block to deliver the gene silencing molecules exactly into the brain. This untargeted delivery may happen either by preventing the passage of the molecules that are injected or by absorbing into the blood [116,122]. Thus, it is found that gene silencing molecules must be either injected directly or using implant pumps which push them into the brain [116]. Once inside the brain, the molecules must move inside the targeted cells. This method of delivery may also make some problems as it can induce an immune response against the gene silencing molecules [116]. In addition to targeted delivery problem, target specificity is also an issue while doing gene silencing. There is a possibility of siRNA molecules to bind with the wrong mRNA molecule which may lead to undesirable results [116].

Recent studies have revealed that siRNA treatment can result in off-target gene silencing, means silencing genes other than the intended targets [17]. Off-target silencing may lead to mutation of gene expression and cell transformation in undesirable form. Most off-target silencing is resulting because of sequence similarity with six to seven nucleotides in the "seed region" of the siRNA sequence [17,123-124]. So by doing careful selection of guide strand of siRNA, the probability of matching with undesired targets can be avoided to some extent. Off-target silencing is an important issue to be addressed upon while doing siRNA-based therapeutics. For the

potential and efficiency of siRNA for therapeutic applications without doing "off-target" silencing must also be heavily tested before application. So target specificity and efficient delivery of siRNA molecules for gene silencing is a serious research issue to be addressed. Special care must be given to design efficient methods to deliver and develop specific gene silencing therapeutics using siRNA in a more safe and effective manner.

2.9 Need of exogenous siRNA design

Even though dsRNA had shown to induce gene-specific silencing capacity in early mouse embryos [2, 125], the attempts to use dsRNA in mammalian systems were not conclusive. In these experiments the application of long dsRNAs generated an overall decrease in mRNA eventually leading to apoptosis, instead of triggering RNAi and also created a response mediated by dsRNA dependent protein kinase [126]. Later it is understood that this type of non-specific response can be bypassed by using chemically synthesized 19 to 22 nucleotide siRNAs [62, 127-129]. By the transfection of these chemically synthesized siRNAs, strong and sequence specific silencing of gene expression in various mammalian cells could be done very effectively. Because of this potential of RNAi-based technologies [130] in therapeutic applications, the use of exogenous siRNA technology has become widespread to study mammalian gene function including clinically relevant genes.

2.10 Complexity in siRNA Design

The design of effective siRNA sequences is a challenging work because the target mRNAs must be selected such that their corresponding siRNAs are likely to be efficient against that target and unlikely to accidentally silence other transcripts due to sequence similarity [12-14]. So it is desirable to consider two important concepts while designing exogenous siRNAs: the ability in knocking down target genes and the off target possibility on any non target genes. Hence before doing gene silencing by siRNAs, it is essential to analyze their off target effects in addition to their inhibition efficacy against a particular target [15-17]. Many barriers prevent practical applications of siRNA. Concepts of siRNA like specific targeting, efficient delivery system, validated genes and the potent siRNA sequences are all vital important to overcome these barriers.

Although reasonable progress has been made in analyzing how the RNAi and siRNA mediates gene silencing, the design of potent siRNAs remains still challenging [15-17]. While considering to optimize the efficiency of siRNA, the above mentioned complexities may lead to the following questions.

- How to identify and validate target genes to design potent siRNA?

- How to design siRNA with good inhibition efficiency?
- How to select functional siRNA sequences with good inhibition efficiency?
- How to eliminate near perfect matched off target genes?

2.11 Summary

The role of siRNA in post-transcriptional gene silencing, the need, complexity, and challenges of designing exogenous siRNA for therapeutical applications are briefly described in this chapter. In this thesis, we try to address some of the complexities of siRNA design so that exogenous siRNA can be designed effectively with specific targeting and efficient delivery system, which may be helpful for effective gene silencing.

……..✷ ♦ ✷……..

Chapter - 3

Study of siRNA Design Approaches

Contents

3.1 Introduction

This chapter aims to provide the literature review of the existing siRNA design approaches. Section 3.2 deals with first generation methods for siRNA design by briefly describing the rules used to design siRNAs. Section 3.3 deals with second generation methods and the machine learning models used in these approaches. Section 3.4 explains the study of selected 23 good scoring siRNA design methods by making a comparison of them. This section is concluded by describing the important approaches selected to integrate in our study. Finally in section 3.5, a brief summary of the chapter is presented.

3.2 First Generation Methods

Several techniques have emerged in the past few years to explore the difficulties in designing exogenous siRNAs. Most studies suggest that positional features like presence or absence of specific nucleotides in certain positions within the siRNA, thermodynamic properties like 'whole stacking energy' and secondary structures of siRNAs are important in predicting efficacy [81,131-135]. These methods are classified into two groups, first generation and second generation methods. First generation methods follow certain rules and regulations for designing siRNA. The following section describes some important siRNA prediction rules followed by first generation methods. These studies reveal that position specific features (presence or absence of specific nucleotides in certain positions within the siRNA), thermodynamic properties and secondary structures of the target site are important in determining the regulatory efficiency of siRNA.

3.2.1. Rules for Designing siRNA

3.2.1.1 Tuschl Rules

Tuschl Rule is the first technique for designing effective and efficient siRNAs and is developed by Elbashir et al. [62]. They recommended that synthesizing siRNA duplexes with a 23 nucleotide sense strand and a 21 nucleotide antisense strand, paired with 2 nucleotide 3' overhang on both ends, mediates the efficiency of target mRNA cleavage. The important rules in this design are summarized below:

- The target region starts 50 to 100 nucleotides downstream of the start codon of a given transcript.
- First search for 23-nt sequence motif AA(N19)TT13
- After it, search for 23-nt sequence motif NA(N21) and convert the 3' end of the sense siRNA to TT
- Finally search for NAR(N17)YNN, where R 2 {A, G} and Y 2 {C, T}
- Target sequence should have a Guanine-Cytosine (G-C) content of around 50%.

3.2.1.2 Amarzguioui Rules

Amarzguioui and Prydz [81] designed the following siRNA design rules, based on their study of 46 siRNAs with a knockdown rate of more than 70%. The rules were tested on another 34 independent siRNAs.

- Strong binding of 5'sense strand
- Weak binding of 3'sense strand
- asymmetry in the stability of the duplex ends
- Presence of G/C at position 1
- Presence of A at position 6
- Absence of U at position 6
- Absence of U at position 1
- Absence of G at position 19
- Presence of A/U at position 19

3.2.1.3 Reynolds Rules

180 siRNAs are analyzed by Reynolds et al. [133]. Based on their regulation efficiency, they divided the siRNAs into different groups and tested whether siRNAs with high functionality have any similarities in their sequence. Based on their analysis, they proposed some rules of how to design highly potent siRNAs. They assigned a score to each siRNA based on the number of rules satisfied. Each siRNA exceeding a specific threshold is predicted to be functional.

- GC content has to be between 30% and 52%
- Presence of nucleotide A at position 3 and 19
- Presence of U at position 10
- Absence of G or C at position 19
- Absence of G at position 13
- Presence of A/U in positions 15 through 19

3.2.1.4 Ui-Tei Rules

These rules are established by Ui-Tei et al. [134] based on 62 siRNA from 5 genes. They analyzed 62 targets in mammalian and Drosophila cells and came up with a conclusion that four features of siRNA listed below should simultaneously satisfy to cause efficient silencing. These rules were found to be applicable to mammalian cells.

- A/U at the first nucleotide of the 5’ end of the antisense strand

- G/C at the first nucleotide of the 5' end of the sense strand
- At least five A/U nt in the 5' terminals first-third of the antisense strand
- No 'GC' stretch of more than 9 nt in length

3.2.1.5 Chalk Rules

Many rules established in recent studies are reviewed on a dataset of 398 siRNAs of known efficiency from 92 genes. This prediction algorithm by Chalk et al. [135] incorporates the thermodynamic properties of the siRNA. The rules are

- Total hairpin energy < 1 kcal/mol
- 5' end binding energy < 9 kcal/mol in the antisense strand
- 5' end binding energy in the range 5-9 kcal/mol exclusive in the sense strand
- G/C Content between 36% and 53%
- Middle area (7-12) binding energy < 13 kcal/mol
- Energy difference between antisense and sense 5' energies <0 kcal/mol
- Energy difference between antisense and sense 5' energies within -1 kcal/mol and 0 kcal/mol

3.2.1.6 Khvorova Rules

These rules are proposed by Khovorova et al. [136] based on 180 siRNAs from one gene. The main aim is to study the internal stability of miRNAs and siRNAs, whose functional duplexes display a lower internal stability at the 5' end antisense strand than nonfunctional duplexes. They could establish that the thermodynamic properties play a critical role in duplex unwinding and strand retention by RISC. The rules are as follows:

- Low stability 3' (sense strand) > -8.5 [kcal/mol]
- Low stability 6-11 (sense strand) > -7 [kcal/mol]
- High stability 5' (sense strand) < -9 [kcal/mol]

3.2.1.7 Takasaki Rules

Takasaki et al. [137] conducted a research on 249 siRNA from one gene. The rules are as follows:

- No A/U at position 1 (sense strand)
- G at position 1 (sense strand)
- A at position 6 (sense strand)
- G at position 7 (sense strand)
- No U at position 7 (sense strand)
- A at position 8 (sense strand)
- No G at position 8 (sense strand)
- No G at position 9 (sense strand)
- U at position 9 (sense strand)

- U at position 15 (sense strand)
- No G at position 19 (sense strand)

3.2.1.8 Hohjoh Rules

These rules are proposed by Hohjoh in 2004 [138]. It is shown that newly designed siRNA duplexes, called "forksiRNA duplexes", can enhance RNAi activity over conventional siRNA duplexes in cultured mammalian cells.

- "Fork-siRNA" mismatch at the 3' sense strand-siRNA
- G/C at position 1 (sense strand)
- A/U at position 19 (sense strand)
- A/U at position 8 (sense strand)

3.2.1.9 Hsieh Rules

Study by Hseih et al. [139] involved about 148 siRNAs from 30 genes and proposed the following rules:

- No C at position 6 (sense strand)
- C/G at position 11 (sense strand)
- A at position 13 (sense strand)
- G at position 16 (sense strand)
- U at position 19 (sense strand)
- No G at position 19 (sense strand)

3.3 Second Generation Methods

Many of the first generation models were not achieving good inhibition efficiency against target genes. Also most of them were not considering many important aspects of siRNA like sensitivity, specificity, off-target possibility while designing siRNA. So there was a need to develop techniques to improve the efficacy of predicted siRNA. These methods are called second generation models which are mostly based on either support vector machine models, linear regression models or artificial neural network models.

3.3.1 Machine Learning Models

3.3.1.1 Support Vector Machines

Support Vector Machine Model (SVM) can be applied to a labeled data to perform classification or regression and can handle multiple continuous and categorical variables [33]. SVM is widely used for applications in bioinformatics [140], text classification [141], pattern recognition [142]. SVM performs classification by constructing hyper-planes in a multidimensional space that separates cases of different class labels. So it is considered to be working based on the concept of decision planes. From each output class, it identifies a subset of the training data. This subset is called support vectors. The model will learn the data using the support vectors. When a new data point needs to be classified, it uses the support vectors to make the classification. In the case of linear classifier, as

shown in Fig.3.1, it separates the data points or objects into their respective groups with a line. If classification is done based on drawing separating lines to distinguish between objects of different class memberships, it is known as hyper-plane classifiers (Fig.3.2). Then SVM finds a separating hyper-plane which has the maximum margin between the training examples and the class boundary. In principle, SVM can be viewed as the maximum margin classifier defined in terms of the support vector approach (Fig.3.3). Maximizing this margin will result in minimizing the maximum loss [143].

3.3.1.1.1 Kernel Functions

SVM will project the training data into a higher dimensional space through a kernel function. Different kernels like linear, polynomial, radial basis function (RBF) and sigmoid can be used in SVM models. Because of the localized and finite responses across the entire range of the real x-axis, RBF is the most popular choice of kernel types used in Support Vector Machines.

$$K(X_i, X_j) = \begin{cases} X_i \cdot X_j \\ (\gamma X_i . X_j + C)^d \\ \exp(-\gamma |X_i - X_j|^2) \\ \tanh(\gamma X_i \cdot X_j + C) \end{cases} \qquad (3.1)$$

Here $K(X_i, X_j) = \phi(X_i) \cdot \phi(X_j)$ means the kernel function represents a dot product of input data points mapped into the higher dimensional feature space by transformation ϕ. The SVM uses a nonlinear mapping function ϕ, that maps the data to a higher dimension, here a separating hyperplane can always be found. Each data point X_k is mapped implicitly to $Y_k = \phi(X_i)$

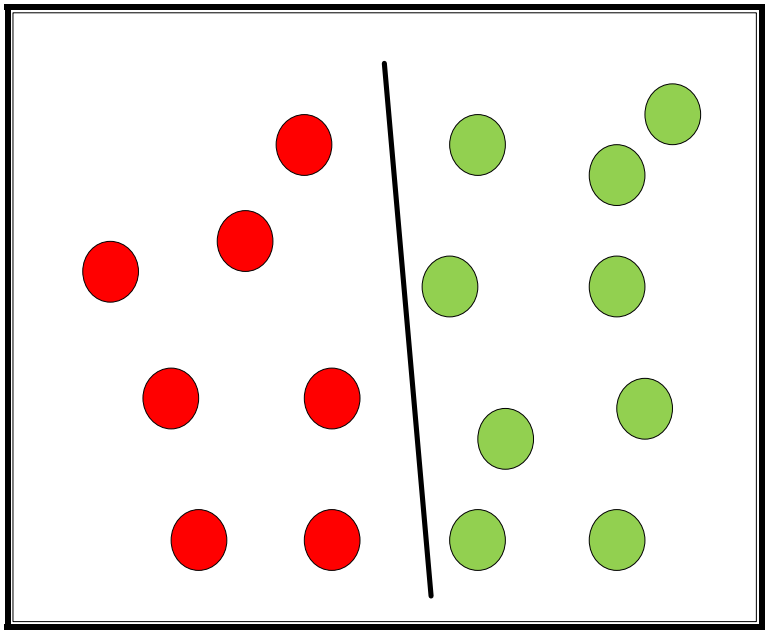


Fig.3.1: Linear classifier

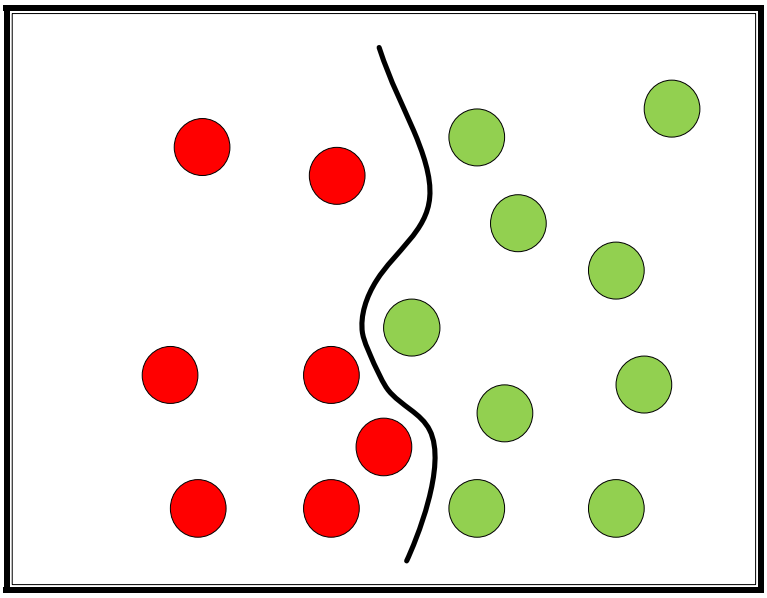


Fig.3.2: Hyper-plane classifier

3.3.1.1.2 Classification of SVM

SVM constructs the optimal hyperplane by iterative training algorithm and minimize an error function. SVM models can be classified into four groups based on the error function: Classification SVM Type 1, Classification SVM Type 2, Regression SVM Type 1 and Regression SVM Type 2.

(i) Classification Type I: In this, training involves the minimization of the error function as

$$\frac{1}{2}w^T + C \sum_{i=1}^{N} \zeta_i \qquad (3.2)$$

Subject to the the constraints:

$$y_i(w^T\phi(x_i) + b) \geq 1 - \zeta_i$$

$$\zeta_i \geq 0, i = 1, \ldots, N$$

Where C is the capacity constant,

w is the vector of coefficients,

b is a constant,

ζ_i represents parameters for handling nonseparable inputs

Index i labels the N training cases

y $\epsilon \pm 1$ represents the class labels

x_i represents the independent variables

ϕ is kernel function used to transform data from the input to the feature space

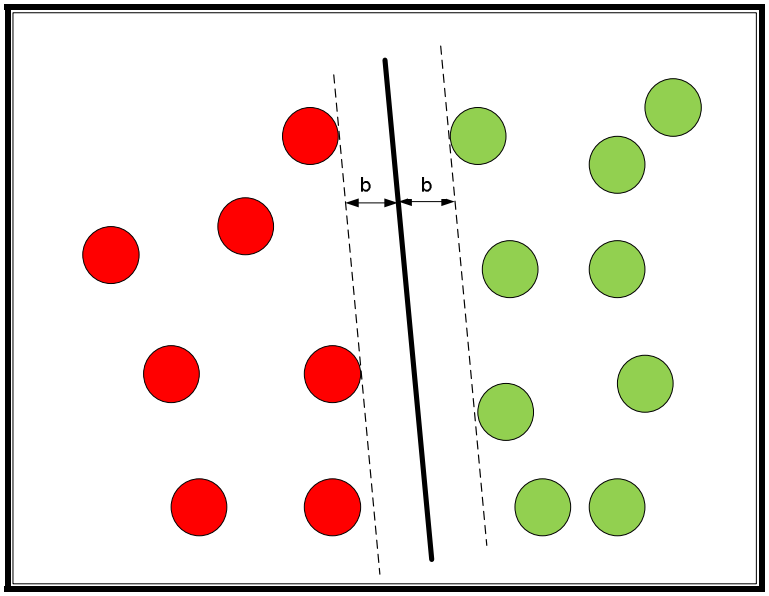


Fig. 3.3: Maximum margin hyper-plane for a two class problem

(ii) Classification Type II: This model minimizes the error function as:

$$\frac{1}{2}w^T - \nu\rho + \frac{1}{N}\sum_{i=1}^{N}\zeta_i \quad (3.3)$$

Subject to the the constraints:

$$y_i(w^T\phi(x_i) + b) \geq \rho - \zeta_i$$

$$\zeta_i \geq 0, i = 1, \ldots, N$$

$$\rho \geq 0$$

(iii) Regression Type I : In this type of SVM, the error function is given by

$$\frac{1}{2}w^T w + C\sum_{i=1}^{N}\zeta_i + C\sum_{i=1}^{N}\dot{\zeta}_i \quad (3.4)$$

Subject to the constraints:

$$w^T\phi(x_i) + b - y_i \leq \varepsilon - \dot{\zeta}_i$$

$$y_i - w^T\phi(x_i) - b_i \leq \varepsilon + \dot{\zeta}_i$$

$$\zeta_i, \dot{\zeta}_i \geq 0, i = 1, \ldots, N$$

(iv) Regression Type II: For this SVM model, the error function is given by

$$\frac{1}{2}w^T w - C(\nu\varepsilon + \frac{1}{N}\sum_{i=1}^{N}(\zeta_i + \dot{\zeta}_i) \quad (3.5)$$

In which minimization of error is subject to

$$w^T\phi(x_i) + b - y_i \leq \varepsilon + \dot{\zeta}_i$$

$$y_i - w^T\phi(x_i) - b_i \leq \varepsilon + \dot{\zeta}_i$$

$$\zeta_i, \dot{\zeta}_i \geq 0, i = 1, \ldots, N$$

3.3.1.2 Artificial Neural Network

Artificial Neural Networks (ANNs) [34-35] are considered as neural network models in artificial intelligence, and is represented as a function $f : X \rightarrow Y$ or a distribution over X or both X and Y. In an ANN, the basic units called perceptrons or neurons are interconnected between different layers of the system. An ANN is defined by three types of parameters: interconnection pattern between the different layers of neurons, learning process for updating the weights of the interconnections, activation function that converts a neuron's weighted input to its output activation. Perceptrons are organized in different ways to form the neural network network's structure. Each perceptrons receives the input which can be an independent raw data or the output of other perceptrons. After processing, the input perceptron delivers a single output, which can be the final result or inputs to other perceptrons. A graphical representation of a perceptron is shown Fig.3.4. A perceptron takes a vector of real-valued inputs, calculates a linear combination of these inputs, then outputs

- a 1 if the result is greater than some threshold
- –1 otherwise.

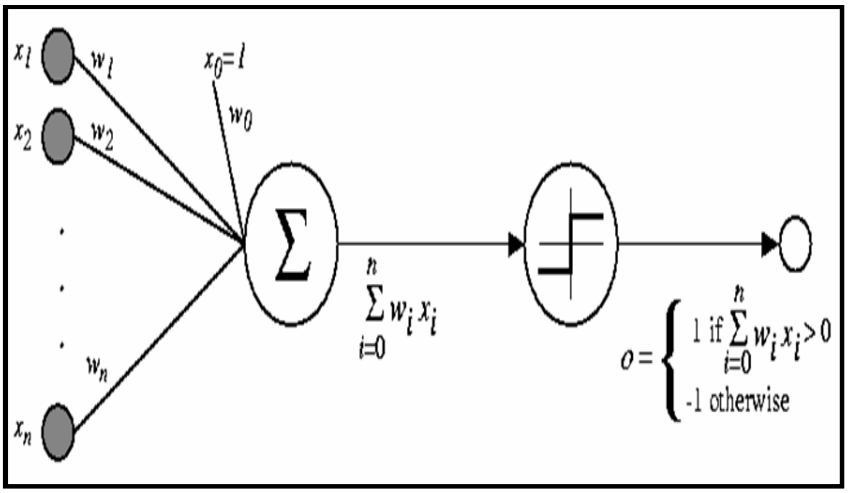


Fig.3.4: Graphical Representation of a Perceptron

(Source of Figure: www.cse-wiki.unl.edu)

3.3.1.2.1 The Network Architecture

Each ANN is composed of a collection of perceptrons grouped in layers. A typical structure of a multi layer neural network is shown in Fig.3.5.

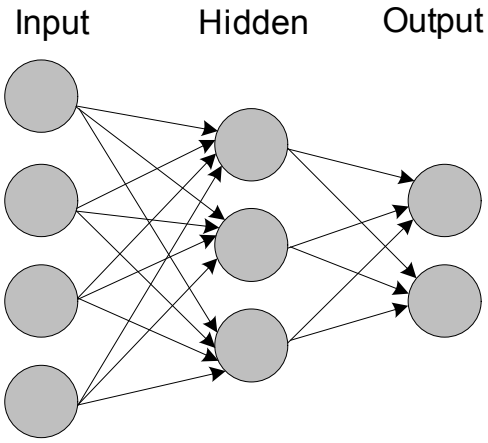


Fig.3.5: A three layer neural network

Based on the pattern of connections between the neurons or perceptrons and the propagation of data, the neural network models are classified into feed forward and feedback networks.

(i) **Feed forward Networks:** The data flow from input to output unit is strictly feed-forward. The processing of data can be extended over multiple layers of the network. But feedback connections extending the output units to previous layers or to input units are not at all present anywhere in the network. Basic structure of a feed forward neural network is shown in Fig.3.6 (a).

(ii) **Feedback Networks:** Feedback networks are also known as recurrent networks, which contain feedback connections. In recurrent neural network, the connections between units form a directed cycle. Basic structure off a feed forward neural network is shown in Fig.3.6 (b).

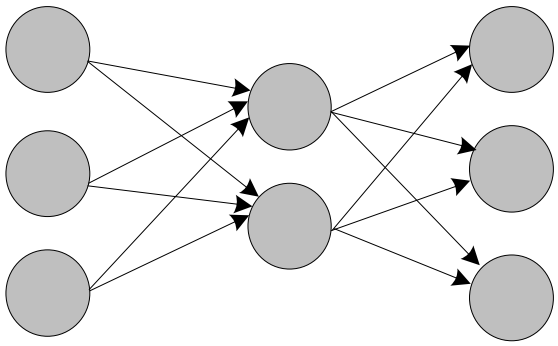


Fig. 3.6: (a) Feed Forward Neural Network

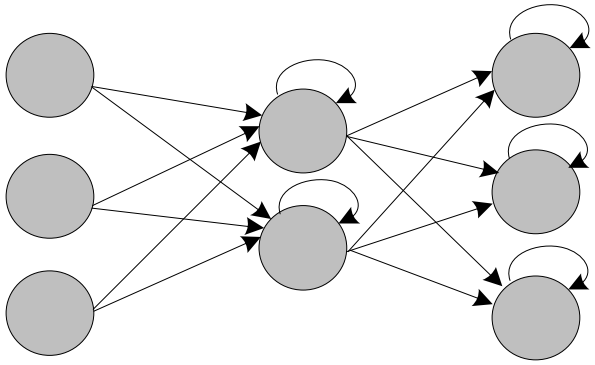


Fig. 3.6: (b) Feedback Neural Network

3.4 siRNA Design Approaches

3.4.1 Study of siRNA Design Methods

Using second generation siRNA models, the complexities in designing efficient siRNA have been addressed to a great extent. But even though many models have emerged, only a few have achieved accepted level of inhibition efficiency (minimum 0.70), accuracy (minimum 0.70), sensitivity (minimum 0.60) and specificity (minimum 0.80). We studied 23 available good scoring siRNA design approaches to understand the efficacy level of each of them. Table 3.1 describes some of the important good scoring siRNA design methods and their characteristics.

Table.3.1. Good Scoring siRNA prediction Methods

Sl.No	siRNA Design Methods	Comments
1	Sfold [144]	Target accessibility prediction and RNA duplex thermodynamics for rational siRNA design.
2	DEQOR [145]	Predicts the probability that an mRNA fragment will cross-react with other genes in the cell
3	BIOPREDsi [18]	Designed a genome wide siRNA library using ANN Network Model
4	siVirus [146]	A web-based system that provides efficient siRNA design for antiviral RNA interference.
5	ThermoComposition21 [20]	Improved the inhibition efficiency of siRNA usig ANN Model, used thermodynamic features.
6	siRNArules [147]	An open-source JAVA program predicting active siRNAs.

Sl.No	siRNA Design Methods	Comments
7	DSIR [19]	Linear Regression Models, A model for siRNA and shRNA target design.
8	iScore [21]	Designed an accurate and interpretable model for siRNA efficacy prediction using linear Regression.
9	Scales [22]	Linear Regression Model
10	OptiRNA [23]	A prediction server for ranking siRNA target sites.
11	RNAxs [25]	Design potent siRNAs to knock down gene of interest.
13	OligoWalk [148]	The web server generates a list of siRNA candidate sequences, ranked by the probability of being efficient siRNA (silencing efficacy greater than 70%).

Sl.No	siRNA Design Methods	Comments
14	AsiDesigner [149]	Exon-based siRNA design server considering alternative splicing
15	BiLTR [150]	A generic framework to enhance siRNA knockdown efficacy prediction.
16	siDRM [24]	An implementation of the DRM rule sets for selecting effective siRNAs.
17	siRecords [26]	A database of siRNAs experimentally tested by researchers with consistent efficacy ratings.
18	E-RNAi [27]	A model for the design and evaluation of RNAi reagents for a variety of species.
19	MysiRNA-Designer [28]	Integrates several factors in an automated work-flow considering mRNA transcripts variations, siRNA and mRNA target accessibility, and both near-perfect and partial off-target matches.

Sl.No	siRNA Design Methods	Comments
20	MysiRNA [29]	Including positional preferences, target accessibility and other thermodynamic features. Used ANN model to predict siRNA inhibition activity
21	DISR [30]	A new version of DSIR incorporating new findings, as well as the list of validated siRNA against the tested cancer genes
22	RNAiAtlas [31]	Provides a siRNA oligonucleotide data from different sources and companies, and visualize interactions between siRNA and predicted off-target.
23	siSPOTR [32]	Allows to determine the off-targeting potential of already designed siRNAs.

3.4.2 Methods Selected for Our Work

We studied 23 siRNA prediction models described in Table 3.1 and tried to find out the prediction accuracy by independent models a well as with combinations of these approaches. The prediction accuracy of various combinations is tested against data sets. Finally it is noticed that accuracy is reaching closer to the original experimental values with a combination of five scoring algorithms: BIOPREDsi [18], DSIR [19], ThermoComposition21

[20], i-Score [21], MysiRNA [29]. So we selected these five state of the art techniques to be integrated in our work to improve the efficiency further. Out of these methods, MysiRNA model [29] is showing good results in terms of inhibition efficiency. Also in their model MysiRNA-Designer [28], they have tried to address the effect of off-target possibilities. All these selected algorithms have been developed by introducing data mining techniques to improve the efficiency of siRNA with their experimental inhibition. BIOPREDsi, ThermoComposition21, MysiRNA-Designer package and MysiRNA used the artificial neural network models, while DSIR, i-Score and Scales used linear regression models. The ThermoComposition21 improved the prediction accuracy by combined position dependent features together with thermodynamic features in single artificial neural network model. The prediction accuracy is improved in DSIR, i-Score and Scales using linear regression model. Further the MysiRNA-Designer package and MysiRNA much improved the prediction accuracy by artificial neural network model. These methods are described below.

3.4.2.1 BIOPREDsi and s-Biopredsi

Huesken et al. [18] designed a genome wide siRNA library to overcome the burden of shortage of interfering short hairpin RNAs for conducting gene knock-down experiments. They used the Stuttgart Neural Net Simulator to train algorithms. The experiments were conducted on a data set of 2,182 randomly selected siRNAs

targeted to 34 mRNA species. These were assayed through a high-throughput fluorescent reporter gene system. This algorithm is known as BIOPREDsi, in which they could predict the inhibition activity of a test data set of 249 siRNAs with Pearson coefficient R= 0.66. They have done the experiments on both 21nucleotide and 19 nucleotide sequences and identified that neural networks trained on a complementary 21 nucleotide sequences were superior to those on 19 nucleotide sequences.

Since we were not able to access most of data from the original BIOPREDsi model [18], we used the simulated-Biopredsi (s-Biopredsi), as in Ichihara's work [21], rather than the original BIOPREDsi. In [21], Ichihara et.al could prove the correspondence between s-Biopredsi and BIOPREDsi by achieving a Pearson correlation coefficient of 1 and identical receiver operating characteristics in ROC analysis.

3.4.2.2 DSIR

Vert et.al. [19] proposed a simple linear model, DSIR, by combining the basic features of siRNA sequences for siRNA inhibition efficacy prediction. It performs well in terms of prediction accuracy. They have used large data set of 2431 randomly selected siRNAs targeting 34 different mRNAs identified by Huesken et al. [18]. They have divided the entire 2431 siRNAs into a training set of 2182 sequences and a test set of 249 sequences. Each siRNA sequence was converted to a vector of features using PYTHON. In

conclusion, they have developed an accurate and interpretable model for siRNA efficacy prediction which performs at least as well as the current state of art.

3.4.2.3 ThermoComposition21

Shabalina et al.[20] developed a model by considering the important thermodynamic properties of siRNA. They collected a heterogeneous set of 653 siRNAs as training data set from various literatures. They have used this training set to fix siRNA features and optimize computational models. They have improved the inhibition efficiency of siRNA molecules by performing thermodynamic and correlation analysis of the training data set. Using a neural network model, they could prove the efficiency of the model against the efficiency prediction at different concentrations. The main advantage of this model over other is the less number of parameters. Because of this advantage, this model requires a very small training data set to get consistent results.

3.4.2.4 i-Score

Ichihara et al. [21] developed an algorithm to predict efficient siRNAs with their inhibitory-Score (i-Score). They have applied a linear regression model to 2431 siRNAs. The only parameter used in this algorithm is the nucleotide preferences at each position. For testing they have used a dataset consisting of 419 siRNAs. With this validation data set, they could predict the accuracy of prediction as well as those of BIOPREDsi[18], ThermoComposition21 [20] and

DSIR [19], in which they employed neural network model or linear regression model. Also they could establish relationship between whole stacking energy and prediction accuracy of siRNA. They could identify that exclusion of siRNAs with a threshold of whole stacking energy, will improve the prediction accuracy.

3.4.2.5 MysiRNA

Mysara et al. [28-29] designed a model called MysiRNA. They identified that many factors including positional preferences, target accessibility and other thermodynamic features will affect the functionality of siRNA. They could develop a model which optimizes the selection of target siRNAs by identifying siRNAs having high experimental inhibition. This uses an artificial neural network model to predict siRNA inhibition activity. This is mainly built on two previous models (Thermo Composition21 [20] and i-Score [21]) together with whole stacking energy (ΔG) in a multi layer artificial neural network. Comparatively, this model results in good siRNA efficiency in terms of specificity, and sensitivity. They have also addressed the off target possibility of siRNA.

3.5 Summary

A comparative study of various good scoring siRNA design approaches is done and the results are analyzed for finding their efficacy. After analyzing the efficacy in terms of inhibition efficiency and off-target possibility of each model in Table 3.1, it is understood

that only a few mechanism were developed for addressing "both inhibition efficiency and off-target effect" of predicted siRNA against genes. Our aim is to develop an approach which optimizes the accuracy of predicted siRNA against target genes by taking care of both inhibition efficiency and off-target effect. The approaches proposed in this thesis extend five previous state of the art techniques named BIOPREDsi [18], DSIR [19], ThermoComposition21 [20], i-Score [21], and MysiRNA [29], by incorporating machine learning and statistical techniques to improve the prediction accuracy and reduce the off-target possibility of siRNA. Out of these methods, MysiRNA model is showing the best results in terms of inhibition efficiency and off-target possibility prediction. In this study, we try to further improve and optimize the predictive ability of siRNA in terms of inhibition efficiency, sensitivity and specificity, accuracy of prediction, off-target identification, by combining the selected state of the art siRNA design techniques.

……..✷♦✷……..

Chapter - 4

Materials and Methods

Contents

4.1 Introduction

This chapter provides the materials and methods used in this research work. The data sets used in this study are presented in section 4.2. The sections 4.3, 4.4 and 4.5 describe the important aspects like siRNA efficiency, need of target specificity while designing exogenous siRNA, and whole stacking energy used during siRNA prediction. The actual machine learning approaches, machine learning frameworks and training algorithms to predict the efficiency of siRNA against target messenger RNAs are described in section

4.6, 4.7 and 4.8 respectively. Section 4.9 describes various validation strategies used for evaluating the performance of the proposed approaches. Finally in section 4.10, a brief summary of the chapter is discussed.

4.2 Data Sets

The neural network models used in this study are trained using the experimental inhibition capacity values of the siRNAs in Huesken data set [18], Data Set 1. This data set contains a total of 2431 siRNAs derived from 34 genes, and their corresponding experimental inhibition capacity values prepared by Huesken et al [18]. This is known as Huesken data set, which is the reliable training data set used by most of the siRNA design approaches. We used entire Data Set 1 for training our neural network model. For testing our neural network, we used two more data sets: Data Set 2 and Data Set 3. Data Set 2, which is mutually exclusive from Data Set 1, consists of 419 siRNAs taken from various sources such as Reynolds et al. [133], Ui-Tei et al. [134], Vickers et al. [157], Khvorova et al. [136] and Harborth et al. [161]. This data set was compiled by Ichihara et al. [21] for their i-Score designer model and is used by Mysara et al. [29] for testing their MysiRNA model. Data Set 3 is used for evaluating the sensitivity and specificity of our model. Data Set 3, which is entirely different from Data Set 1 and Data Set 2, consists of 476 siRNAs presented by Mysara et al. [29]. These 476

siRNAs were originally taken from a larger data set of 18,593 siRNAs introduced by Fellman et al. [162]. The details of data sets used in the study are shown in Table 4.1. We have particularly selected these data sets for training and testing, since we can make an easy comparison of results with previous state of art techniques. In addition to these data sets, we have maintained our own data set (Data Set 4) containing 743 siRNAs collected manually from RefSeq [163] for working with the model. Also the model can directly take any mRNA or cDNA sequence from RefSeq [163] and will automatically create the siRNA sequences corresponding to the mRNA or cDNA and continue with efficiency prediction. The sample cDNA sequences used for designing siRNA are shown in Appendix 1.

Table 4.1: Data Sets used for Training and Testing ANN models

Train/Test Data Set	**Name of Data Set**	**No of Genes**	**No of siRNA used**	**siRNA with 50% to 70 % inhibition**	**siRNA with 70% to 90 % inhibition**	**siRNA with > 90 % inhibition**
Train Data Set	Data Set 1	34	2431	778	853	369
Test Data Set	Data Set 2	12	419	60	117	96
Test Data Set	Data Set3	9	476	70	53	127

4.3 siRNA Efficiency

The goal of siRNA efficacy prediction is to help in designing siRNA sequences that are highly efficient against their target mRNA sequences. Reynolds et al. [133] observed in their siRNA knock-down experiments that properties of the target mRNA did not affect knockdown and efficacy seems to be solely based on properties of the siRNA. Gene silencing related studies indicate that out of the possible siRNAs that can be synthesized against a particular target, only a few are found successful in causing any degradation [12,151]. Among those successful siRNAs, all do not result with equal knockdown effects [12]. Also the efficacy of same siRNA may be different among different target sites for the same mRNA. Some studies reveal that stability factors like secondary structure and thermodynamic properties of the siRNA are also important determinants of functionality [131-132]. So for performing effective gene silencing, it is important to select effective siRNA sequences with good inhibition capacity, i.e., siRNAs that are highly functional in causing a certain percentage of the target mRNA sequence to degrade. In most studies, siRNAs causing knockdown of more than 70% of the target mRNA are considered highly efficient but the threshold varies depending on the level of silencing required [12,131-132, 151-152].

4.4 siRNA Specificity

In addition to inhibition efficiency, another important factor to be considered while siRNA design is the specificity of the siRNA [16,62,153]. siRNA mediated gene silencing is generally believed to be highly sequence specific. Sometimes siRNAs may tolerate mismatches with the target mRNA, but knockdown of genes other than the intended target could make serious consequences. Gene expression profiling in cultured human cells demonstrated silencing of non-targeted genes. Even though eleven complementary matches out of the 19 nucleotides of siRNA was enough to cause silencing [16], in some cases even a single base mismatch between the siRNA and its mRNA target abolished gene silencing [62]. This indicates that siRNA may cross-react with targets of limited sequence similarity. While maximum degradation of target mRNA is required, silencing of non-target mRNA should be avoided. Therefore, due consideration must be given to the implications arising from siRNA specificity in design algorithms. This can be achieved by selecting target mRNA such that their corresponding siRNAs are likely to be efficient against that target and unlikely to accidentally silence other transcripts due to sequence similarity. So to design siRNAs, two important concepts must be considered: the ability in knocking down

target genes and the off target possibility on any non target genes [16,62].

4.5 Whole Stacking Energy

We are using an important thermodynamic property of siRNA called whole stacking energy (ΔG) as one of the input parameters to our approaches since it reflects the stability of siRNA duplexes and shows good correlation with inhibition efficiency [21,23,154]. We have used nearest neighbour model [29,155-156] to calculate the whole stacking energy of siRNA strand. The method used is same as that of iScore designer [152] and MysiRNA [29]. For calculating ΔG, the sum of the ΔG values in kcal/mol contributed by each nearest neighbour pair in the siRNA sequence is found out as shown in Table 4.2.

$$\text{Whole } \Delta G = \sum_{i=1}^{n-1} \Delta G_{37}(\text{Seq}[i]\text{Seq}[i+1]) \qquad (7.1)$$

For example, if the siRNA sequence is AGACUA,

Whole ΔG = ΔG(AG) + ΔG(GA) + ΔG(AC) + ΔG(CU) + ΔG(UA) = -2.1 + -2.4 + -2.2 + -2.1 + -1.3
= -10.1 kcal/mol.

Table 4.2: ΔG values of nearest neighbor pairs

Nearest Neighbour Pair	ΔG_{37} (kcal / mol)	Nearest Neighbour Pair	ΔG_{37} (kcal / mol)
AA	-0.9	GA	-2.4
AU	-1.1	GU	-2.2
AG	-2.1	GG	-3.3
AC	-2.2	GC	-3.4
UA	-1.3	CA	-2.1
UU	-0.9	CU	-2.1
UG	-2.1	CG	-2.4
UC	-2.4	CC	-3.3

4.6 Machine Learning Approaches

The following are the set of machine learning approaches proposed in this study for finding the efficiency of siRNA data. Based on these algorithms the efficiency of siRNA against target mRNA are modeled and tested.

i. Support Vector Machine (SVM) model is used to classify and observe the efficiency of siRNA against target mRNA [33].

ii. Two Artificial Neural Network (ANN) models are designed to find the efficiency of siRNA against target mRNA [34-35].

4.7 Machine Learning Frameworks

In this study, we are using three machine learning approaches named LIBSVM [158], Neuroph Studio [159] and Encog Workbench IDE [160].

4.7.1 LIBSVM

LIBSVM [158] is used as a library for Support Vector Machines. The practical use of LIBSVM involves mainly two steps. In the first step, training is done with a known data set to obtain a model. In the second step, it will predict information of a testing data set using the developed model. LIBSVM supports various SVM formulations for classification, regression, and distribution estimation. It supports multi class classification like One-class SVM, SVC (Support Vector Classification for two-class and multi-class) and SVR (Support Vector Regression) [33,158]. We have used LIBSVM for working with our SVM model.

4.7.2 Neuroph Studio

One of our artificial neural network models is designed by using Neuroph studio [159]. The Neuroph library for Java is used to create neural network model. Neuroph is a lightweight Java neural network framework to develop common neural network architectures. The Neuroph Studio IDE provided by Neuroph is used to easily design and test the model. The IDE provides an easy-to-use graphical interface to design various neural network configurations, and to train

or test the network using various neural network training algorithms.

Website: http://www.neuroph.sourceforge.net

4.7.3 Encog Workbench IDE

The Encog Workbench IDE [160] is used for creating our second neural network model, i.e., optimized siRNA designer. The Encog machine learning framework for Java is used to create and use the siRNA designer neural network model. Encog is an advanced, lightweight Java machine learning framework which can be used to develop common neural network and other machine learning models like Support Vector Machines, Genetic Algorithms, Bayesian Networks, Hidden Markov Models. The Encog Workbench IDE is used to easily design and test the model. The IDE provides an easy-to-use graphical interface to design various neural network configurations, and to train or test the network using various neural network training algorithms. In addition to Java, the Encog framework is also available for .NET and C/C++.

Website: http://www.heatonresearch.com/encog

4.8 Training Algorithms

The back propagation algorithms [164-165] are used for training our neural networks. For training with back propagation, the input patterns should be known apriory. Then the algorithm can be used for training a given feed-forward multilayer neural network with

a known set of input patterns with the classifications. For each of the sample input presented to the network, it examines the output response. Then the network will compare the output response to the known desired output and the error value is calculated accordingly. The connection weights are adjusted based on the error. The set of sample patterns are repeatedly presented to the network until the error value is minimized. We have used two back propagation algorithms namely Resilient Propagation (RProp) [166-167] and Scaled Conjugate Gradient (SCG) [168-172] in our neural network methods.

4.8.1 Resilient Propagation

Resilient back propagation (Rprop) [166-167] is an algorithm which is used for training a neural network which is same as that of a regular back propagation algorithm. Training with Rprop is faster than back propagation and Rprop doesn't require specifying any free parameter values for learning rate. But the main disadvantage of Rprop algorithm is that it is more complex to implement than back propagation. The Rprop algorithm has two significant differences with the back propagation algorithm. First, Rprop uses only the sign of the gradient instead of magnitude to determine weight delta. Second, Rprop maintains separate weight deltas for each weight and bias, and adapts these deltas during training, instead of using a single learning rate for all weights and biases.

4.8.2 Scaled Conjugate Gradient

Many adaptive learning algorithms for feed forward neural networks have been introduced [164]. But most of them are based on Gradient Descent algorithm and have poor convergence rate. For example standard back propagation algorithm [165] often behaves badly on large scale problems. But Conjugate Gradient Methods are one class of optimization methods that are able to handle larger scale problems very effectively [166-169]. Several Conjugate Gradient algorithms have been introduced as learning algorithms in neural networks [170-171]. Finally Scaled Conjugate Gradient (SCG) [173] a supervised learning algorithm is introduced for improving the requirements of feed forward neural networks with good convergence rate. SCG is based on optimization techniques in numerical analysis.

4.9 Validation Strategies

After training and testing of the neural network model, validation of results are done thorough Pearson Correlation analysis, followed by Accuracy of Prediction, Sensitivity and Specificity, Matthews Correlation Coefficient and Receiver Operating Characteristics analysis [174-175]. These validation strategies are explained briefly.

4.9.1 Pearson Correlation Coefficient

Pearson correlation coefficient (R) [174-175] is a measure of the linear dependence between two variables X and Y. The correlation values range between -1 to 1, where values closer to -1 indicates negative correlation and values closer to +1 indicates strong positive correlation and those tending towards 0 indicates no correlation. The interpretation of the Pearson correlation coefficient is as follows.

R = +0.70 or above indicates very strong positive correlation
R = +0.40 to +0.69 indicates strong positive correlation
R = 0 indicates no correlation
R = -0.40 to -0.69 indicates strong negative correlation
R = -0.70 or above indicates very strong negative correlation

In our approaches, Pearson correlation is calculated to find the accuracy of our results with original experimental values. For this, correlation between the predicted siRNA inhibition efficiency by our model against original experimentally proven siRNA inhibition efficiency is observed and analyzed.

4.9.2 Sensitivity, Specificity, Accuracy

A diagnostic test may be highly specific without being sensitive, or it may be highly sensitive without being specific. But both factors are equally important. A diagnostic test is considered as "good" if the test has both high sensitivity and specificity. The sensitivity, specificity and accuracy are described in terms of true positive (TP), true negative (TN), false negative (FN), and false positive (FP). If a disease is proven present in a patient and the given diagnostic test also indicates the presence of disease, the result of the diagnostic test is considered true positive. Similarly, if a disease is proven absent in a patient and the diagnostic test suggests the disease is absent as well, the test result is true negative (TN). From Table 4.3 it is understood that both true positive and true negative suggest a consistent result between the diagnostic test and the proven condition. However, if the diagnostic test indicates the presence of disease in a patient who actually has no such disease, the test result is false positive (FP). Similarly, if the result of the diagnosis test suggests that the disease is absent for a patient with disease for sure, the test result is false negative (FN). Both false positive and false negative indicate that the test results are opposite to the actual condition.

Table 4.3: Template for diagnostic test results

Diagnostic Test Result	**Existence of Disease as determined by the standard of truth**		
	Positive	**Negative**	**Row Total**
Positive	TP	FP	TP+FP
Negative	FN	TN	FN+TN
Column total	TP+FN	FP+TN	N=TP+TN+FP+FN

Sensitivity: Sensitivity [174-175] is the proportion of true positives that are correctly identified by a diagnostic test. It shows the goodness of the test while detecting a disease. The numerical values of sensitivity represents the probability of a diagnostic test identifies patients who do in fact have the disease. As the numerical value of sensitivity is higher, the possibility of diagnostic test returns false-positive results is less. For example, if sensitivity = 95%, it means: when we conduct a diagnostic test on a patient with certain disease, there is 95% of chance, this patient will be identified as positive. A test with high sensitivity may capture all possible positive conditions without missing anyone. So a test showing high sensitivity is often used to screen for disease.

Sensitivity is defined as

Sn = Number of true positive assessment / Number of all positive assessment

$$\text{Sn} = \frac{\text{TP}}{\text{TP+FN}} \qquad (4.1)$$

Specificity: Specificity [174-175] is the proportion of the true negatives correctly identified by a diagnostic test. It indicates how good the test is at identifying normal (negative) condition. The numerical value of specificity represents the probability of a test diagnoses a particular disease without giving false-positive results. For example, the specificity of a test is 95% means: when we conduct a diagnostic test on a patient without certain disease, there is 95% chance of this patient to be identified as negative. Specificity is defined as

Specificity = Number of true negative assessment/Number of all negative assessment

$$\text{Sp} = \frac{\text{TN}}{\text{TN+FP}} \qquad (4.2)$$

Along with sensitivity and specificity, the measures like Positive Predictive Value (PPV), Negative Predictive Value (NPV), False Positive Rate (FPR), False Negative Rate (FNR), False Discovery Rate (FDR), F-Score (F) are also be used for describing the performance of diagnostics.

$$\text{PPV} = \frac{\text{TP}}{\text{TP+FP}} \tag{4.3}$$

$$\text{NPV} = \frac{\text{TN}}{\text{TN+FN}} \tag{4.4}$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP+TN}} \tag{4.5}$$

$$\text{FNR} = \frac{\text{FN}}{\text{TP+FN}} \tag{4.6}$$

$$\text{FDR} = \frac{\text{FP}}{\text{TP+FP}} \tag{4.7}$$

$$\text{F} = \frac{2\text{TP}}{2\text{TP+FP+FN}} \tag{4.8}$$

Accuracy: Accuracy measures [174-175] the degree of veracity of a diagnostic test on a condition. The numerical value of accuracy represents the proportion of true positive results (both true positive and true negative) in the selected population. An accuracy of 95% means the test result is accurate, regardless positive or negative. However, the equation of accuracy implies that even if both sensitivity and specificity are high, it does not suggest that the accuracy of the test is equally high as well. Apart from sensitivity and specificity, accuracy of the test is also used as a measure to determine how common the disease in a selected population. A diagnosis for

rare conditions in a particular population might result in high sensitivity and specificity but with low accuracy.

Accuracy is defined as

Accuracy = Number of correct assessments/Number of all assessments

$$\text{Acc} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{TN}+\text{FN}} \tag{4.9}$$

4.9.3 Matthews Correlation Coefficient

Matthews Correlation Coefficient (MCC) [174-175] is typically used in machine learning as a metric for assessing the quality of the predicted value to the observed value. Or it is a measure of quality of prediction which is the correlation coefficient between the observed and predicted binary classifications by considering false positive (FP), false negative (FN), true positive (TP) and true negative (TN). An MCC with -1 indicates negative correlation, 0 indicates no correlation (random selection) and +1 indicates positive (perfect) correlation.

Mathews correlation coefficient,

$$\text{MCC} = \frac{(\text{TP}\times\text{TN}) - (\text{FN}\times\text{FP})}{\sqrt{(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TP}+\text{FP})(\text{TN}+\text{FN})}} \tag{4.10}$$

4.9.4 Receiver Operating Characteristic

Receiver Operating Characteristic (ROC) [174-175] is one among the robust tool used for diagnostic tests. ROC plots sensitivity on Y axis against '1- specificity' on X axis. Using ROC analysis, we can calculate the Area under Curve (AUC) as a measure of performance. The AUC can also be realized as the average sensitivity over entire range of all possible specificities, or the average specificity over entire range of all possible sensitivities. An AUC of 1 identifies perfect classification and an AUC of 0.5 identifies random classification.

For siRNA efficacy prediction, it is desirable to have low false positive rates. For RNAi studies, where functional siRNAs are required, it is very important that siRNAs having low efficacy must not predicted to be functional. But, misclassifying siRNAs with high efficiency rates as nonfunctional is of much lesser consequence. ROC curve is a plot of a test's sensitivity versus (1-specificity). ROC curves are useful in comparing classifiers based on true positive and false positive rates. For a given diagnostic test, the true positive rate (TPR) against false positive rate (FPR) can be measured, where

$$\mathrm{TPR} = \frac{\mathrm{TP}}{(\mathrm{TP}+\mathrm{FN})} \tag{4.11}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{(\mathrm{FP}+\mathrm{TN})} \tag{4.12}$$

From the above equations it can be noted that, TPR is equivalent to sensitivity and FPR is equivalent to (1 – specificity). All possible combinations of TPR and FPR compose a ROC space. A single point in the ROC space is determined by one TPR and one FPR. The position of a point in the ROC space indicates the tradeoff between sensitivity and specificity. An increase in sensitivity is accompanied by a decrease in specificity. Thus the location of the point in the ROC space depicts whether the diagnostic classification is good or not. If a point determined by both TPR and FPF gives coordinates (0,1), we can say that this point falls on the upper left corner of the ROC space. This ideal point indicates the diagnostic test has a sensitivity of 100% and specificity of 100%. It is also called perfect classification. Diagnostic test with 50% sensitivity and 50% specificity can be visualized on the diagonal determined by coordinate (0, 0) and coordinates (1, 0). If a point predicted by a diagnostic test fall into the area above the diagonal, it represents a good diagnostic prediction, otherwise a bad classification. A graphic representation is shown in Fig. 4.1. It shows that shadow area represents better diagnostic classification.

The interpretation of ROC curve is similar to a single point in the ROC space. If the point on the ROC curve is closer to the ideal coordinate, the test result will be more accurate. If the point on the ROC curve is closer to the diagonal, the test result is less accurate.

The properties of ROC are as follows,

- As the curve approach the ideal point faster, the test results are more useful;
- The slope of the tangent line to a cut-point indicates the ratio of the probability of identifying true positive over true negative. If the ratio is greater than 1, true positive results will be identified and if the ratio is less than 1, disease likelihood is decreased.

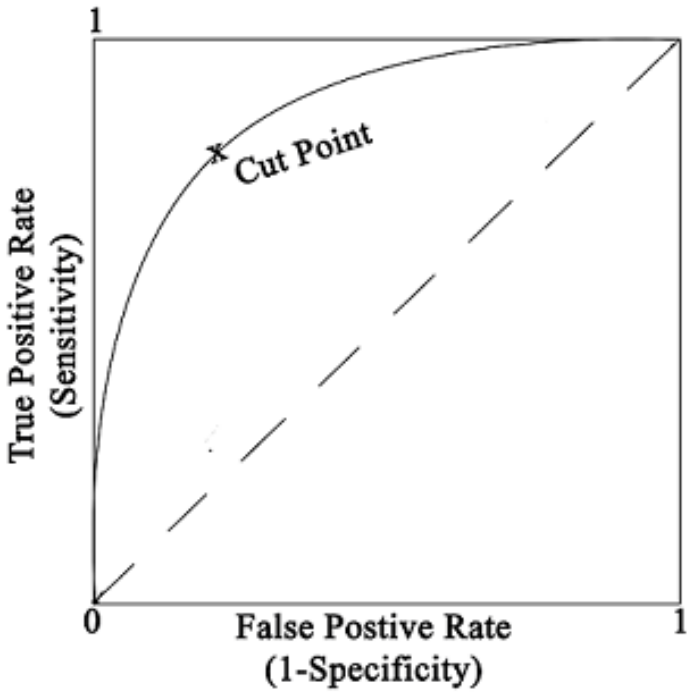


Fig. 4.1: ROC Curve

- AUC, the area under ROC curve is a measure of accuracy of a diagnostic test. Accuracy classification by AUC for a diagnostic

test is shown in Table 4.4. As the area is larger, the more accurate the diagnostic test is.

AUC of ROC curve is measured by the following equation,

$$\text{AUC} = \int_0^1 \text{ROC}\,(t)dt \qquad (4.7)$$

Where t = (1 – specificity) and ROC (t) is sensitivity.

Table 4.4: Accuracy classification by AUC for a diagnostic test

Range of AUC	**Classification**
AUC between 0.9 and 1.0	Excellent
AUC between 0.8 and 0.9	Good
AUC between 0.7 and 0.8	Worthless
AUC between 0.6 and 0.7	Not good

4.10 Summary

The materials and methods like machine learning approaches (SVM [33] and ANN [34-35]), machine learning frameworks (LIBSVM [149], Neuroph Studio [159], Encog Workbench IDE

[160]), machine learning algorithms [166-172] (Rprop and SCG), Data Sets for training and testing, validation strategies like Pearson Correlation Coefficient, MCC, ROC, Sensitivity, Specificity and Accuracy of Prediction [174-175] for evaluating the performance of the proposed approaches are briefly explained in this chapter.

……..✷ ♦ ✷……..

Chapter - 5

siRNA Efficiency Prediction by Support Vector Machine Model

Contents

5.1 Introduction

As the first step of our study, we selected Support Vector Machine model, to start predicting efficiency of siRNA against target mRNA or cDNA sequences. This chapter describes the method for predicting siRNA efficiency using one of the machine learning approaches called Support Vector Machine. Section 5.2 describes the input parameters selected for the SVM model. Section 5.3 briefly presents how training and testing is done with SVM. In section 5.4, various steps during training and testing phase are explained. Finally, an overview of the chapter is discussed in section 5.5. Using this model, we try to classify a given siRNA as efficient or inefficient to silence a target mRNA sequence. Also we filtered the results to find the influence of melting temperature, one of the

important thermodynamic properties of siRNA on the inhibition efficiency.

5.2 Input parameters

The model is initially trained with input parameters like positional features, percentage of G-C content, and some thermodynamic properties of siRNA. Since thermodynamic properties are important stability factors of siRNAs, we finalized the input parameters as four thermodynamic properties of siRNA: whole stacking energy (ΔG), enthalpy (ΔH), entropy (ΔS) and melting temperature (Tm). These values of siRNAs are calculated according to the nearest neighbor model [155-156]. NetBeans IDE 6.9.1, the open-source Integrated Development Environment with Glass Fish Application Server is used to develop the model. Apache Derby Network Server is used for the implementation of servlets and JSP. LIBSVM [158], the publicly available SVM program written in Java is used by us for solving the classification problem.

5.3 Training and Testing with SVM

The training data set is prepared as follows. Based on the gene silencing activity, we have collected 653 siRNAs from the published data from RefSeq [163]. Based on the reported gene silencing activity, we have filtered the siRNAs into two categories. The siRNAs with greater than or equal to 60 percentage gene silencing activity is considered as efficient and siRNAs with less

than or equal to 30 percentage gene silencing activity is considered as inefficient. Thus manually we could separate 359 siRNAs out of 653. For each siRNA, the input parameters are calculated and training is done by SVM. The output is then scaled to keep features with large numerical scales with small numerical scales, to the range [0, 1]. Testing is done with target mRNA sequences. The input is taken as mRNA sequence or cDNA sequence. The flowcharts shown in Fig.5.1 and Fig.5.2 describe the training and testing phase of SVM. SVM finally lists all possible efficient and inefficient siRNAs for a specified mRNA or cDNA sequence. The predicted siRNAs by our model are compared and analyzed with existing available siRNA target finder models and the results are verified.

5.4 Steps in Training and Testing

5.4.1 Files used:

1. **Svm_trainingset.csv**: Contains the set of siRNAs and their known efficiency which are used to train the SVM.
2. **Train**: Contains 4 parameters and efficiency of training siRNAs.
3. **Train.scale**: Scaled values of parameters in the file "Train".
4. **Train.model**: Stores the trained SVM.
5. **Test**: Stores the parameters of siRNAs generated from the user given mRNAs.
6. **Test.scale**: Contains the scaled version of the file "Test".

7. **Scalingfactors**: Stores the scaling factors used to scale the values.

5.4.2 Training Phase:

⇨ Read siRNAs from svm_trainingset.csv file, calculate their parameters and write to the file "Train".

⇨ Scale the parameters in the file "Train" and write the scaled values to file "Train.scale".

⇨ Input the train.scale file to the SVM and write the trained SVM to the file "Train.model".

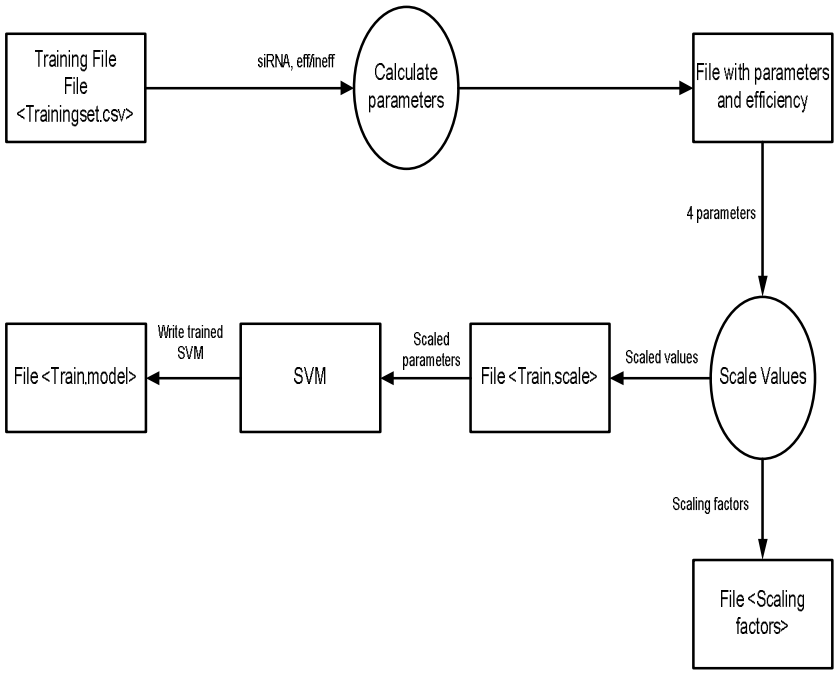


Fig 5.1: Flow Chart for Training Phase

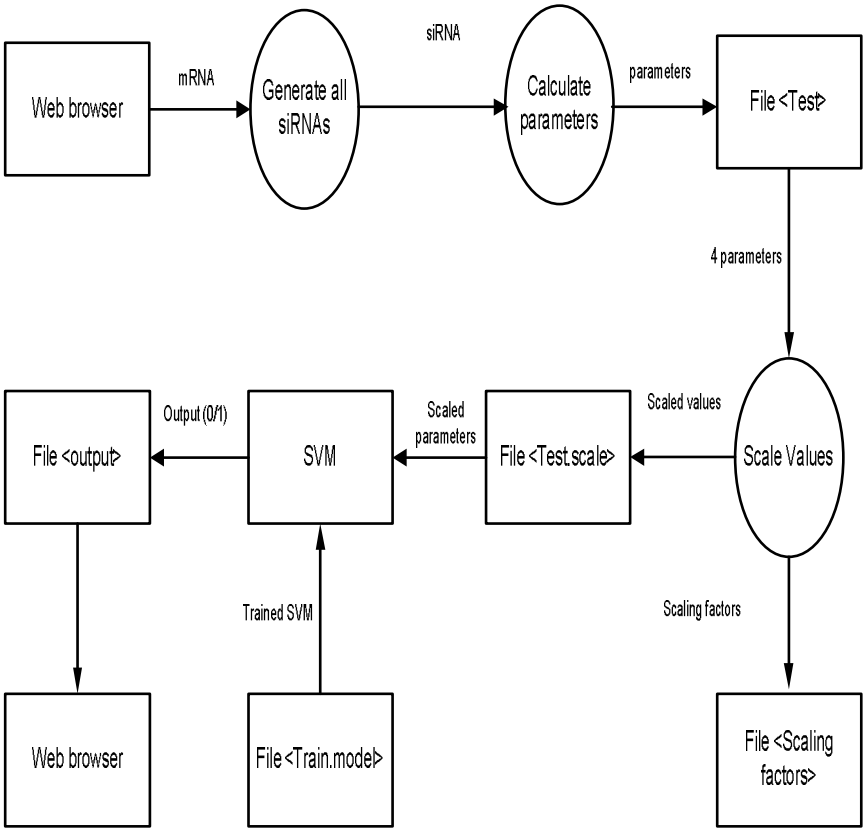


Fig 5.2: Flow Chart for Testing Phase

5.4.3 Testing Phase:

- ⇨ Read the mRNA submitted by the user from the web browser.
- ⇨ Generate all possible siRNAs from the given mRNA.
- ⇨ Calculate the 4 parameters of all these siRNAs and write to the file “Test”.

- ⇨ Scale the values from "Test" and write to the file "Test.scale". The scaling factors are obtained from the file "Scalingfactors".
- ⇨ Feed the scaled parameters of each siRNA from the file "Test.scale" to the SVM and obtained the output 0 or 1. If the output is 1, the corresponding siRNA is efficient otherwise inefficient. SVM predicts siRNAs one by one.

5.5 Summary

Using this SVM model, we are able to achieve the first objective of our study, i.e., designing efficient siRNAs for any target mRNA or cDNA. In efficiency prediction using SVM, we are classifying the siRNA into efficient or inefficient: means whether siRNA is able to silence a target mRNA sequence or a gene. The predicted efficiency is verified with existing siRNA design approaches. Also we are able to notice a relationship between the thermodynamic property and inhibition efficiency of siRNA using this model. Results and discussion of SVM model is further elaborated in chapter 8.

In this model, since we considered the classification property of SVM, we are able to only identify whether the predicted siRNA is efficient or inefficient against target genes. But in the optimized siRNA prediction model, we are supposed to find the percentage of

inhibition efficiency of each predicted siRNA too. When we analyzed the available siRNA prediction approaches, it is understood that most of them are using artificial neural network model for finding the efficiency of siRNA. So we have moved further to design artificial neural network machine learning models to identify as well as predict the inhibition efficiency of each siRNA against target genes. These models are described in Chapter 6 and Chapter 7.

……..✷♦✷……..

Chapter - 6

siRNA Efficiency Prediction by Artificial Neural Network Model

Contents

6.1 Introduction

This chapter describes the method used for predicting siRNA inhibition efficiency using one of the machine learning approaches called Artificial Neural Network. An approach named *siRNA Designer*, is built through a multi layer perceptron feed forward neural network (6-8-8-8-1 ANN) based on five previous second generation models BIOPREDsi, DSIR, ThermoComposition21, i-Score and MysiRNA together with one of the important thermodynamic property of siRNA called whole stacking energy (ΔG), to predict the efficiency of siRNA. The chapter is divided into four sections. Section 6.2 describes the architecture of the 6-8-8-8-1 neural network model. The work flow of the model is described in section 6.3. The chapter ends with section 6.4, which gives a summary of this chapter. Using this model, we try to find the

percentage of inhibition efficiency of each siRNA generated against a target mRNA or cDNA sequence and to optimize the result in terms of inhibition efficiency.

6.2 Neural Network Architecture

A multi-layer perceptron, feed-forward neural network trained using the Resilient Propagation algorithm [166] is used for computing the final score. The neural network which we use is a 6-8-8-8-1 ANN, which has 6 neurons in the input layer (x1 to x6); three hidden layers of 8 neurons each and 1 neuron in the output layer (y) (Fig. 6.1). The neural network is built and trained using Neuroph Studio [159] and integrated into our siRNA designer model. The Neuroph library for Java is used to create and use the siRNA designer neural network model.

Neuroph is a lightweight Java neural network framework to develop common neural network architectures. The Neuroph Studio IDE [159] provided by Neuroph is used to easily design and test the model. The IDE provides an easy-to-use graphical interface to design various neural network configurations, and to train or test the network using various neural network training algorithms. It is available under version 2.0 of the Apache License [176]. Apache License is free open source software from the Apache Software Foundation. The Apache library is used to read and write Microsoft Excel files, such as i-Score designer Excel file. It provides a set of

Java APIs for creating and manipulating Microsoft Office Documents. This model is named as siRNA Designer.

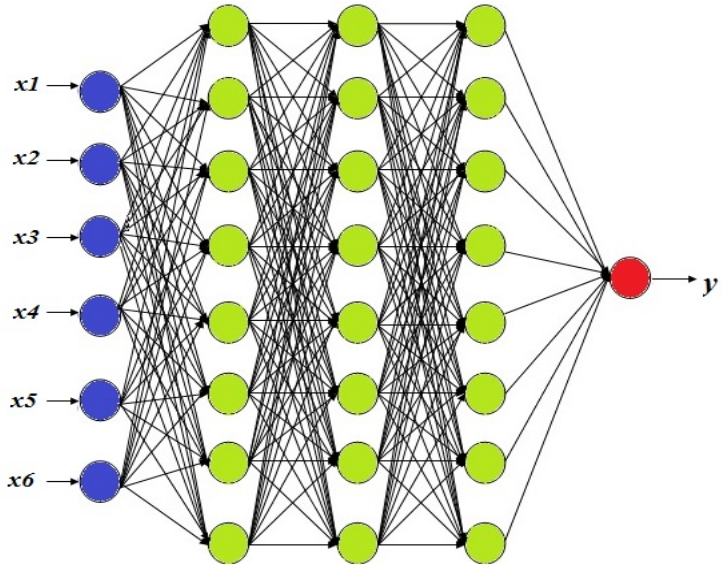


Fig. 6.1: 6-8-8-8-1Neural Network Model

6.2.1 Input Parameter Selection

The existing siRNA design approaches use different features and weights in their model design. We made an attempt to combine these features for improving the design. For this, we considered the features of many good scoring siRNA design models to get better prediction value. After several iterations and trials, we are able to obtain a combination of 5 approaches with a very good prediction power. These models are BIOPREDsi [18], ThermoComposition21 [20], i-Score [21], DSIR [19] and MysiRNA [29]. Since most of the data from the original BIOPREDsi model [18] are not available

directly, we used the simulated-Biopredsi (s-Biopredsi), as in Ichihara's work [21], rather than the original BIOPREDsi [18]. In the next step, we included one more parameter, whole stacking energy (ΔG), to find the effect of ΔG oin inhibition efficiency.

6.2.2 Normalization of Input and Output

The input values given to the neural network, i.e. the 6 metrics described above, are normalized using the z-normalization method [177]. That is, the normalized input values are given by:

$$x_i^{'} = \frac{x_i - \mu_x}{\sigma_x}$$

Where $x_i^{'}$ is the normalized value of the metric for the i^{th} siRNA

x_i is the actual value

μ_x is the mean value of the metric x for the entire set of siRNAs

σ_x is the standard deviation

The mean and standard deviation values obtained for the training data set are used for normalizing the input values. The neural network gives a single output value in the range [0, 1], which is multiplied by 100 to give the final score which is displayed for each siRNA.

6.3 siRNA Designer Workflow

The workflow of the approach is shown in Fig. 6.2.

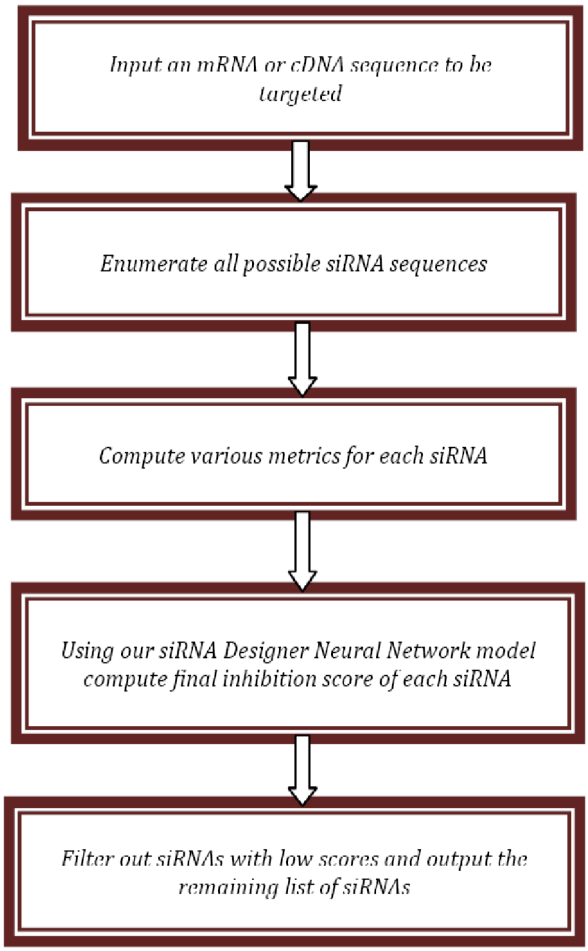


Fig. 6.2: Workflow of 6-8-8-8-1 Model

6.4 Summary

In this work, an approach named siRNA Designer (6-8-8-8-1 ANN Model) is built based on artificial neural network model to predict siRNA inhibition activity based on five good scoring state of the art models, BIOPREDsi, DSIR, ThermoComposition21, i-Score and MysiRNA together with whole stacking. Using this 6-8-8-8-1 ANN model, we are able to achieve the second objective of our study, i.e., predicting the percentage of inhibition efficiency of each predicted siRNA against a target mRNA or cDNA sequence. By maintaining a cut-off in inhibition efficiency (normally cut-off will be 70%-80% depending on the amount of silencing needed), one can select efficient siRNAs which are capable of inhibiting target mRNAs. The result and discussion of this ANN model is elaborated in section 8.3 of chapter 8.

It is found that this model results in very good performance in terms of inhibition efficiency. Using this model even though we are able to optimize the efficiency of siRNA in terms of inhibition capacity, we are not able to address the issues like sensitivity and specificity of siRNA properly. Hence another ANN model is designed for optimizing the siRNA efficiency in terms of inhibition capacity, sensitivity, specificity, accuracy of prediction and off-target possibility. With this new 5-12-1 artificial neural network model, we

try to improve and optimize the inhibition efficiency of siRNA on target genes and off target possibility on non- target genes, and is elaborated in Chapter 7.

……..✷♦✷……..

Chapter - 7

Optimized siRNA Prediction by Artificial Neural Network Model

Contents

7.1 Introduction

This chapter describes the method used for optimization of siRNA efficiency prediction using Artificial Neural Network model. An approach named *Optimized siRNA Designer (OpsiD)* is built through a multi layer perceptron feed forward neural network (5-12-1 ANN) based on four previous second generation models DSIR, ThermoComposition21, i-Score and MysiRNA together with whole stacking energy (ΔG) to predict the efficiency of siRNA. The chapter is divided into five sections. Section 7.2 describes the architecture of 5-12-1 neural network model. In section 7.3, the work flow of the model is presented by briefly explaining the input and output parameters, frame works used for designing and the

prerequisites. The working model and off-target possibility prediction are explained in section 7.4. Finally a brief summary about the approach is given in section 7.5. Using this model, we try to optimize the efficiency of predicted siRNA in terms of inhibition efficiency, off-target possibility, sensitivity, specificity and accuracy of prediction.

7.2 Neural Network Architecture

A multi-layer perceptron feed-forward neural network is modeled for finding optimized siRNAs with improved efficacy against target mRNA in terms if inhibition capacity, sensitivity, specificity, accuracy of prediction and off-target possibility than existing state of the art techniques. For selecting the optimized neural network model, we tested various configurations of feed-forward neural networks such as 4-8-8-1, 5-7-7-1, 5-8-1, 5-8-8-1, 5-10- 1, 5-12-1, 6-7-7-1, 6-8-8-1, 6-8-8-8-1, 6-10-1 and 6-12-1 using the proven good scoring models like BIOPREDsi [18], ThermoComposition21 [20], i-Score [21], DSIR [19] and MysiRNA [29] together with whole stacking energy (ΔG) as inputs parameters. Since we were not able to access most of data from the original BIOPREDsi model [18], we used the simulated-Biopredsi (s-Biopredsi), as in Ichihara's work [21], rather than the original BIOPREDsi. Initially we had considered s-Biopredsi as an input metric, but after some experimentation, it is found that the inclusion of s-Biopredsi

decreased the accuracy of the neural network model for various data sets. We finally chose a configuration of 5-12-1 ANN. For calculating the final score of each siRNA, we computed four different metrics for the siRNA strand's inhibition capacity taken from earlier works and used as the input values of our neural network. These final methods considered for combining in our models are ThermoComposition21, i-Score, DSIR and MysiRNA. Along with these, whole stacking energy (ΔG) of each siRNA strand is taken as fifth input metric. The 5-12-1 neural network shown in Fig.7.1 consists of an input layer with 5 neurons (x1 to x5), a single hidden layer with 12 neurons and output layer with 1 neuron (y1). A number of neural network training algorithms such as the classic Back Propagation, Resilient Propagation and Scaled Conjugate Gradient are tried out [166-173]. Varying number of training iterations are also tried out depending on the network configuration and the training algorithm. The training is started from a randomized state and is done for 1, 36,000 iterations. The neural network is built and trained using the Encog Workbench IDE [160] and later integrated into our siRNA designer model. Scaled Conjugate Gradient training algorithm provided by Encog [172-173] is used for computing the final score of each siRNA. The Scaled Conjugate Gradient algorithm is based upon a class of optimization techniques well known in numerical analysis as the Conjugate Gradient Methods. This model is named as Optimized siRNA Designer (OpsiD).

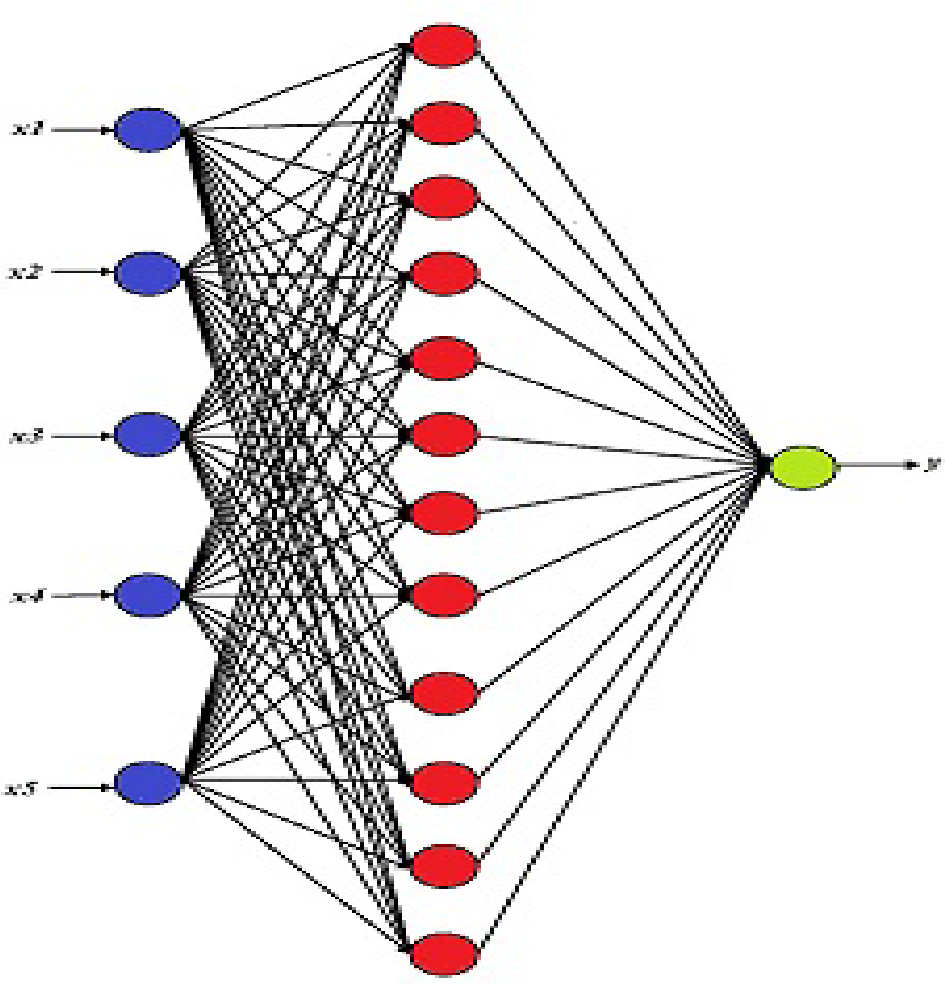


Fig. 7.1: 5-12-1 Neural Network Model.

7.3 Optimized siRNA Designer (OpsiD) Workflow

7.3.1 Input Parameter Selection

For computing the final score of each siRNA, we considered five different metrics: whole stacking energy (ΔG), DSIR score [19],

ThermoComposition21 score [20], i-Score prediction value [21] and MysiRNA score [29]. All these must be downloaded from their respective sources before using our model, OpsiD. In our experiments, these metrics are found to work well and give good results. Along with these values, whole stacking energy is also calculated as described in section 4.5, for each siRNA strand. These five values are used as the input of our neural network.

7.3.2 Normalization of Input and Output

The input values given to the neural network, i.e. the 5 metrics described above, are normalized using the range normalization method, also known as min-max normalization. That is, the normalized input values are given by:

$$A_i^N = \frac{A_i - \min_A}{\max_A - \min_A}(\max_R - \min_R) + \min_R \qquad (7.1)$$

Where $\boldsymbol{A_i^N}$ is the normalized value of the metric A for the i^{th} tuple of the training data, $\boldsymbol{A_i}$ is the actual value, $\boldsymbol{max_A}$ is the maximum value of the metric A for the entire training set of siRNAs and $\boldsymbol{min_A}$ is the minimum value. The values are normalized to the range [$\mathbf{min_R}$, $\mathbf{max_R}$].

The minimum and maximum values of each input metric are used for normalizing the input values. The input values for the neural

network are normalized to the range [-1, 1] before being given as input. The experimental inhibition values from the training data set were also normalized to the range [0, 1] before training the neural network. The neural network gives a single output value in the range [0, 1], which is then multiplied by 100 to give the final score which is displayed for each siRNA.

7.3.3 Frameworks

7.3.3.1 NCBI BLAST

The NCBI BLAST tool (blastall) [178] is used to filter out siRNAs with high off-target effect by running a BLAST search on the NCBI RefSeq database. The blastall tool is bundled along with OpsiD, but the NCBI RefSeq database must be downloaded separately.

Website: ftp://ftp.ncbi.nlm. nih.gov/blast/executables/ release/ LATEST/.

7.3.3.2 Encog Workbench IDE

The Encog machine learning framework for Java is used to create and use the siRNA designer neural network model. The IDE [160] provides an easy-to-use graphical interface to design various neural network configurations, and also to train as well as test the neural network using various neural network training algorithms.

Website: http://www.heatonresearch .com/encog

7.3.3.3 Apache POI

The Apache POI library [176] is used to read and write Microsoft Excel files, such as the i-Score designer Excel file. The Apache POI library is an open-source library developed by the Apache Software Foundation which provides a set of Java APIs for creating and manipulating Microsoft Office Documents (both the new OOXML formats and the old OLE2 Compound Document Format). It is available under version 2.0 of the Apache License. **Website:** http://poi.apache.org/

7.3.4 Prerequisites

The siRNA designer software (OpsiD) requires the following software and files to be downloaded and installed:

i. Java 7
ii. Any Perl Distribution
iii. i-Score Designer Excel file [21]
iv. ThermoComposition21 [20]
v. MysiRNA designer and model file [29]
vi. NCBI BLAST tool (blastall)[a] [178]
vii. NCBI RefSeq RNA BLAST database [163][a]
viii. Encog Workbench[b] [160]

[a] Required only for BLAST search filtering (optional)

[b] Required only for testing the neural network model using the supplied test data

7.4 Working Model

7.4.1 Input

The approach takes an mRNA or cDNA gene sequence as input. The nucleotides may be specified using the uppercase letters A, T, G, C and U or the corresponding lowercase letters, and any spaces or newlines within the sequence are ignored. The user can also enter the RefSeq number of the gene if it has been taken from the NCBI RefSeq RNA database [163]. This is only required if the user selects to perform BLAST search.

7.4.2 Processing

The siRNA designer model first enumerates each possible 19 nucleotide siRNA sequences from the input nucleotide sequence. Then, it computes the parameters for each siRNA strand such as G-C content percentage and whole stacking enenrgy. For calculating whole stacking enenrgy, we use the nearest-neighbor model from Sugimoto et al. [156], which is also used by other models such as i-Score and MysiRNA. If the user has selected the option to filter siRNAs based on G-C content, the software removes all siRNAs which do not have GC content % between the minimum and maximum values specified by the user. Then, it gives each siRNA sequence as input to various pre-existing second generation siRNA designer models such as i-Score, ThermoComposition21, DSIR and

MysiRNA to get their scores for the strand's inhibition capacity. The siRNA strand's i-Score, ThermoComposition21, DSIR, MysiRNA scores and the initially calculated whole delta G value are given as input to the neural network model. The model gives an output value in the range [0, 1] which is then multiplied by 100 to give the displayed final score (in the range 0 – 100) for the siRNA strand. If the user has selected the option to filter siRNAs based on their score, the software will remove all siRNAs whose final scores lie below a certain threshold value specified.

7.4.3 Off-Target Possibility Prediction

If the user wants to filter out siRNA with high off-target effect, the BLAST option must be selected through the interface. The BLAST score outputs for a particular siRNA's maximal match obtained for that sequence against some other mRNA subsequence in the selected gene database. The use of siRNAs with high BLAST score may lead to off-target effect. Thus the user must make the trade-off between the "goodness" of siRNA with respect to inhibition capacity, and its similarity to other mRNA fragments. The work flow of the model, OpsiD is shown in Fig. 7.2.

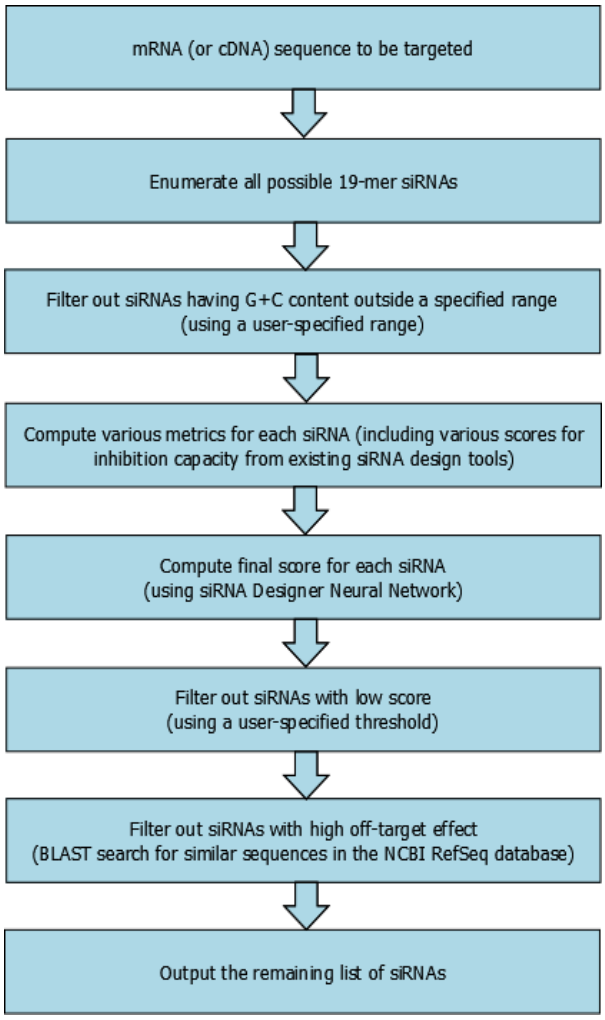


Fig. 7.2: Workflow of OpsiD

7.5 Summary

In this work, a 5-12-1 artificial neural network model named Optimized siRNA Designer (OpsiD), is designed to achieve the goals and objectives of our study. Using this model, we are able to optimize the siRNA efficacy in terms of inhibition efficiency, sensitivity-

specificity, accuracy of prediction and off-target possibility. It is mainly built on four previous second generation models DSIR, ThermoComposition21, i-Score and MysiRNA together with an important thermodynamic property of siRNA called whole stacking energy (ΔG). Using OpsiD, we are able to observe the percentage of inhibition efficiency of each predicted siRNA against a target mRNA or cDNA sequence and able to address some of the issues like sensitivity and specificity.OpsiD also provides the choice for detecting the off-target possibility of each siRNA on any unintended genes during silencing. In OpsiD, the off-target possibility of each predicted siRNA can be observed by BLAST search option provided by the model. Using this method, the risk of "off-target" effect on non-target genes can be easily understood early. Thus, OpsiD provides the chance of identifying optimized siRNA with high inhibition capacity on target genes and low off-target possibility on non-target genes. The results and discussion of OpsiD is elaborated in section 8.4 of chapter 8.

……..✷♦✷……..

Chapter - 8
Results and Discussion

Contents

8.1 Introduction

In this study, one SVM model and two ANN models for predicting efficiency of siRNA against target genes are proposed. The results and discussion of these models are presented in this chapter. The chapter is divided into five sections. Section 8.2 describes the

results and discussion of the SVM model. Section 8.3 and section 8.4 present the results and discussion of siRNA Designer with 6-8-8-8-1 ANN Model and Optimized siRNA Designer, OpsiD, with 5-12-1 ANN Model respectively. The performance evaluation of these models and comparison with existing approaches are also done in these sections. Finally, a summary about the chapter is given in section 8.5.

8.2 SVM Model

8.2.1 Results

In SVM model, as a first attempt of the study, we considered only the classification property for analyzing the efficiency of an siRNA to silence a target mRNA. Because of this, we can only test the correctness of the results, but can't evaluate the performance using the validation strategies shown in chapter 4. So we are not able to present any performance evaluation for SVM model. The sample input interface and output for SVM model is shown in Fig 8.1 and 8.2.

8.2.2. Discussion

In efficiency prediction using SVM, we are classifying the siRNA into efficient and inefficient, i.e., we are able to identify and predict whether an siRNA is efficient or inefficient to silence a target mRNA sequence or a gene. Also, since we have used some thermodynamic features of siRNA as input parameters, we tried to find the relationship among these parameters with the inhibition efficiency of siRNA. From the results it is observed that most of the efficient siRNA stands have G-C content between 50-75 percentage, melting temperature between 60 to 75 and delta G between -30.0 to -38.0. From these results, we can come to the conclusion that the efficiency of siRNA is strongly connected with thermodynamic properties like melting temperature and delta G. So we have decided to include thermodynamic propertiy of siRNA as parameters in our neural network models.

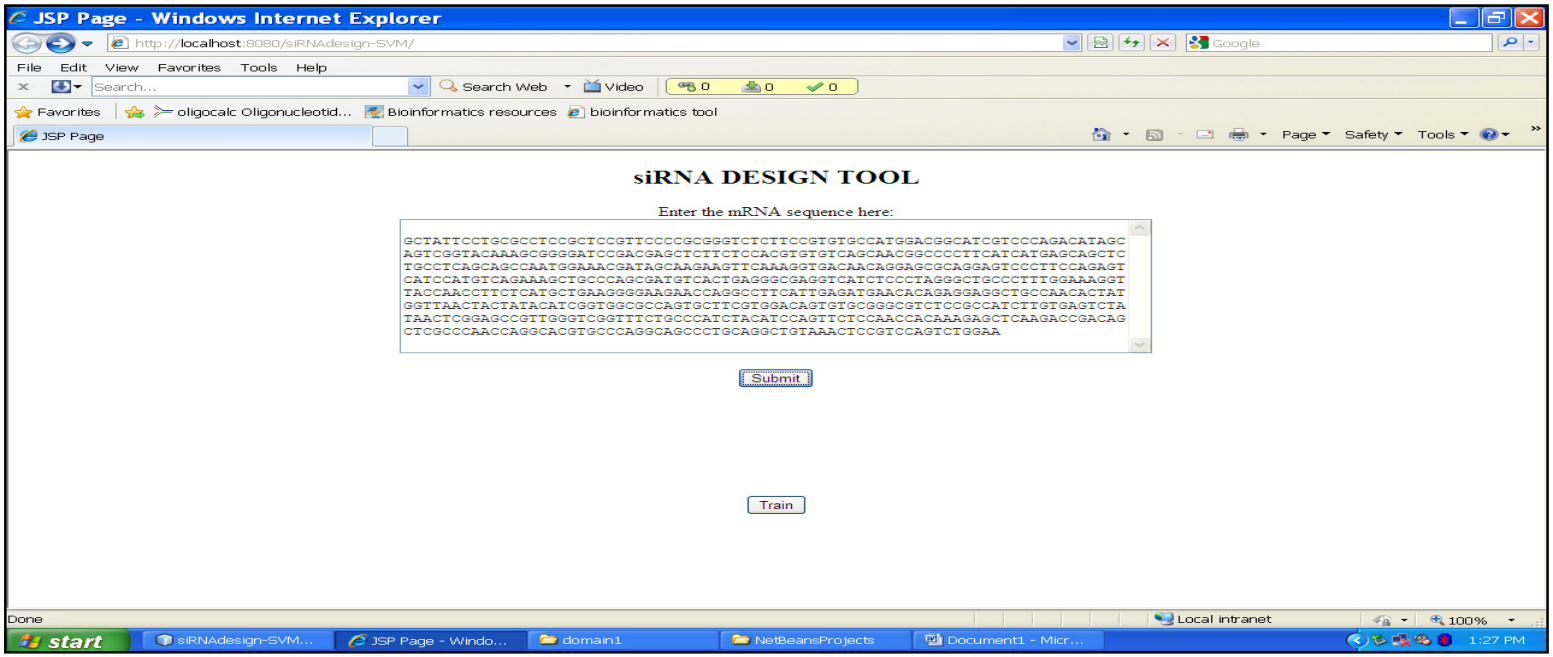


Fig 8.1: Input interface of the SVM Model

design output

Target sequense	Antisense sequense	target position	GC content	Tm	deltaG	Efficacy
GGGGATCCGACGAGCT	AGCUCGUCGGAUCCCCGCU	85	61.904762	70.62809	-46.399998	EFFICIENT
CCCCTTCATCATGAGC	GCUCAUGAUGAAGGGGCCG	125	57.142857	66.727425	-43.2	EFFICIENT
CAAAGGTGACAACAGG	CCUGUUGUCACCUUUGAAC	182	42.857143	59.193245	-36.4	EFFICIENT
TGACAACAGGAGCGCA	UGCGCUCCUGUUGUCACCU	188	52.38095	67.29598	-42.800003	Inefficient
GAGCGCAGGAGTCCCT	AGGGACUCCUGCGCUCCUG	197	61.904762	71.35027	-46.499996	EFFICIENT
TGCCCAGCGATGTCAC	GUGACAUCGCUGGGCAGCU	238	57.142857	68.3326	-44.2	EFFICIENT
TCTCATGCTGAAGGGG	CCCCUUCAGCAUGAGAAGG	305	52.38095	64.98411	-41.399998	EFFICIENT
GAAGAACCAGGCCTTC	GAAGGCCUGGUUCUUCCCC	320	57.142857	68.51952	-43.699997	EFFICIENT
CCAGGCCTTCATTGAG	CUCAAUGAAGGCCUGGUUC	326	47.61905	62.01272	-39.1	Inefficient
GGCCTTCATTGAGATG	CAUCUCAAUGAAGGCCUGG	329	47.61905	61.64982	-39.199997	EFFICIENT
AGAGGAGGCTGCCAAC	GUUGGCAGCCUCCUCUGUG	350	57.142857	68.23084	-43.6	EFFICIENT
CTATACATCGGTGGCG	CGCCACCGAUGUAUAGUAG	380	47.61905	61.689503	-38.3	Inefficient
GGAGCCGTTGGGTCGG	CCGACCCAACGGCUCCGAG	452	66.666664	71.265144	-45.899998	Inefficient
CAAAGAGCTCAAGACC	GGUCUUGAGCUCUUUGUGG	500	47.61905	62.213776	-38.999996	EFFICIENT
GCTCAAGACCGACAGC	GCUGUCGGUCUUGAGCUCU	506	52.38095	65.54435	-42.1	Inefficient
CGACAGCTCGCCCAAC	GUUGGGCGAGCUGUCGGUC	515	61.904762	69.27297	-44.600002	EFFICIENT
GGCACGTGCCCAGGCA	UGCCUGGGCACGUGCCUGG	533	66.666664	74.553276	-48.200005	EFFICIENT

Fig 8.2: siRNA efficiency Prediction by SVM Model

8.3 siRNA Designer Approach

8.3.1 Results

The results and discussion of siRNA Designer with 6-8-8-8-1 ANN Model is described in this section. Predicted inhibition capacity of each siRNA for a targeted mRNA has been observed with the model. The sample screen shot depicting user interface and output are shown Fig. 8.3 and Fig. 8.4 respectively.

8.3.2 Performance Evaluation

For 6-8-8-8-1 ANN, predicted inhibition capacity of each siRNA for a targeted mRNA is found out and the performance evaluation in terms of Pearson Correlation is done. Pearson correlation gives the correlation between the inhibition efficiency of our predicted model with the original experimental inhibition efficiency. We achieved a good Pearson correlation coefficient of R= 0.727 for Data Set 1. This R value shows that the predicted value of inhibition by our model is closer to the original experimental values. The inhibition activity of our model is plotted with original experimental inhibition for Data Set 1 and is shown in Fig. 8.5.

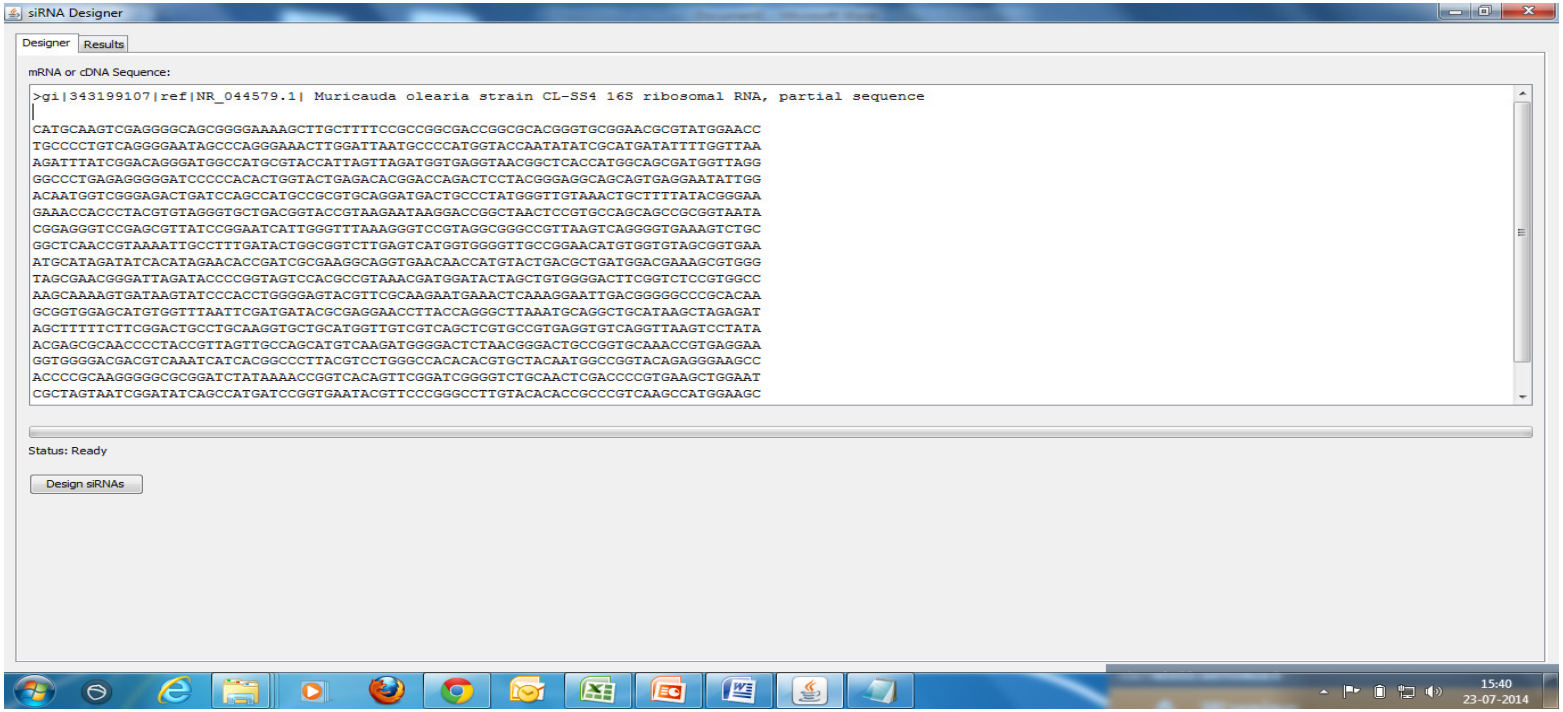
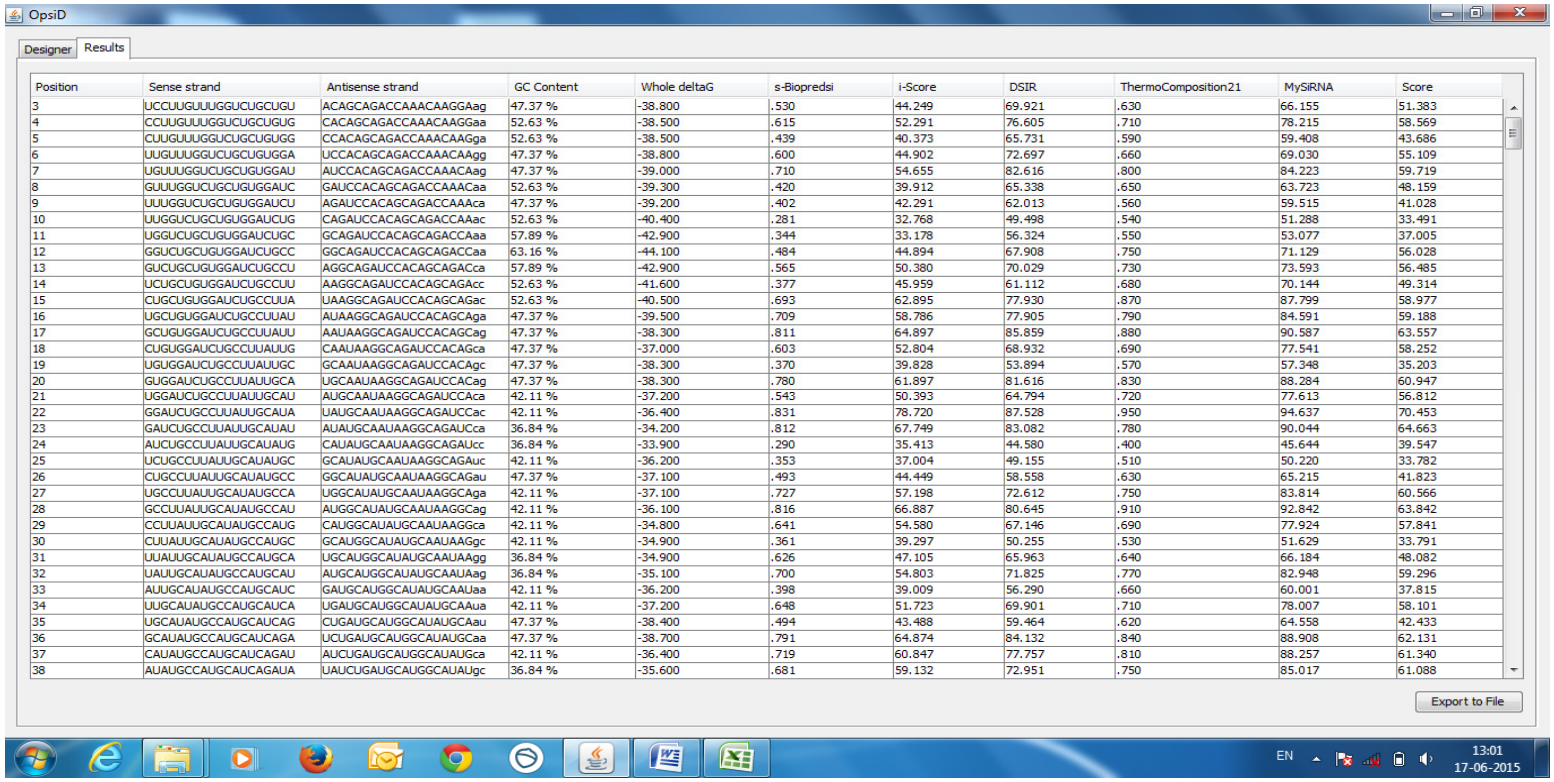


Fig. 8.3: Sample Screen Shot showing the user interface



Position	Sense strand	Antisense strand	GC Content	Whole deltaG	s-Biopredsi	i-Score	DSIR	ThermoComposition21	MySiRNA	Score
3	UCCUUGUUUGGUCUGCUGU	ACAGCAGACCAAACAAGGAag	47.37 %	-38.800	.530	44.249	69.921	.630	66.155	51.383
4	CCUUGUUUGGUCUGCUGUG	CACAGCAGACCAAACAAGGaa	52.63 %	-38.500	.615	52.291	76.605	.710	78.215	58.569
5	CUUGUUUGGUCUGCUGUGG	CCACAGCAGACCAAACAAGga	52.63 %	-38.500	.439	40.373	65.731	.590	59.408	43.686
6	UUGUUUGGUCUGCUGUGGA	UCCACAGCAGACCAAACAAgg	47.37 %	-38.800	.600	44.902	72.697	.660	69.030	55.109
7	UGUUUGGUCUGCUGUGGAU	AUCCACAGCAGACCAAACAag	47.37 %	-39.000	.710	54.655	82.616	.800	84.223	59.719
8	GUUUGGUCUGCUGUGGAUC	GAUCCACAGCAGACCAAACaa	52.63 %	-39.300	.420	39.912	65.338	.650	63.723	48.159
9	UUUGGUCUGCUGUGGAUCU	AGAUCCACAGCAGACCAAAca	47.37 %	-39.200	.402	42.291	62.013	.560	59.515	41.028
10	UUGGUCUGCUGUGGAUCUG	CAGAUCCACAGCAGACCAAac	52.63 %	-40.400	.281	32.768	49.498	.540	51.288	33.491
11	UGGUCUGCUGUGGAUCUGC	GCAGAUCCACAGCAGACCAaa	57.89 %	-42.900	.344	33.178	56.324	.550	53.077	37.005
12	GGUCUGCUGUGGAUCUGCC	GGCAGAUCCACAGCAGACCaa	63.16 %	-44.100	.484	44.894	67.908	.750	71.129	56.028
13	GUCUGCUGUGGAUCUGCCU	AGGCAGAUCCACAGCAGACca	57.89 %	-42.900	.565	50.380	70.029	.730	73.593	56.485
14	UCUGCUGUGGAUCUGCCUU	AAGGCAGAUCCACAGCAGAcc	52.63 %	-41.600	.377	45.959	61.112	.680	70.144	49.314
15	CUGCUGUGGAUCUGCCUUA	UAAGGCAGAUCCACAGCAGac	52.63 %	-40.500	.693	62.895	77.930	.870	87.799	58.977
16	UGCUGUGGAUCUGCCUUAU	AUAAGGCAGAUCCACAGCAga	47.37 %	-39.500	.709	58.786	77.905	.790	84.591	59.188
17	GCUGUGGAUCUGCCUUAUU	AAUAAGGCAGAUCCACAGCag	47.37 %	-38.300	.811	64.897	85.859	.880	90.587	63.557
18	CUGUGGAUCUGCCUUAUUG	CAAUAAGGCAGAUCCACAGca	47.37 %	-37.000	.603	52.804	68.932	.690	77.541	58.252
19	UGUGGAUCUGCCUUAUUGC	GCAAUAAGGCAGAUCCACAgc	47.37 %	-38.300	.370	39.828	53.894	.570	57.348	35.203
20	UGGAUCUGCCUUAUUGCA	UGCAAUAAGGCAGAUCCACag	47.37 %	-38.300	.780	61.897	81.616	.830	88.284	60.947
21	UGGAUCUGCCUUAUUGCAU	AUGCAAUAAGGCAGAUCCAca	42.11 %	-37.200	.543	50.393	64.794	.720	77.613	56.812
22	GGAUCUGCCUUAUUGCAUA	UAUGCAAUAAGGCAGAUCCac	42.11 %	-36.400	.831	78.720	87.528	.950	94.637	70.453
23	GAUCUGCCUUAUUGCAUAU	AUAUGCAAUAAGGCAGAUCca	36.84 %	-34.200	.812	67.749	83.082	.780	90.044	64.663
24	AUCUGCCUUAUUGCAUAUG	CAUAUGCAAUAAGGCAGAUcc	36.84 %	-33.900	.290	35.413	44.580	.400	45.644	39.547
25	UCUGCCUUAUUGCAUAUGC	GCAUAUGCAAUAAGGCAGAuc	42.11 %	-36.200	.353	37.004	49.155	.510	50.220	33.782
26	CUGCCUUAUUGCAUAUGCC	GGCAUAUGCAAUAAGGCAGau	47.37 %	-37.100	.493	44.449	58.558	.630	65.215	41.823
27	UGCCUUAUUGCAUAUGCCA	UGGCAUAUGCAAUAAGGCAga	42.11 %	-37.100	.727	57.198	72.612	.750	83.814	60.566
28	GCCUUAUUGCAUAUGCCAU	AUGGCAUAUGCAAUAAGGCag	42.11 %	-36.100	.816	66.887	80.645	.910	92.842	63.842
29	CCUUAUUGCAUAUGCCAUG	CAUGGCAUAUGCAAUAAGGca	42.11 %	-34.800	.641	54.580	67.146	.690	77.924	57.841
30	CUUAUUGCAUAUGCCAUGC	GCAUGGCAUAUGCAAUAAGgc	42.11 %	-34.900	.361	39.297	50.255	.530	51.629	33.791
31	UUAUUGCAUAUGCCAUGCA	UGCAUGGCAUAUGCAAUAAgg	36.84 %	-34.900	.626	47.105	65.963	.640	66.184	48.082
32	UAUUGCAUAUGCCAUGCAU	AUGCAUGGCAUAUGCAAUAag	36.84 %	-35.100	.700	54.803	71.825	.770	82.948	59.296
33	AUUGCAUAUGCCAUGCAUC	GAUGCAUGGCAUAUGCAAUaa	42.11 %	-36.200	.398	39.009	56.290	.660	60.001	37.815
34	UUGCAUAUGCCAUGCAUCA	UGAUGCAUGGCAUAUGCAAua	42.11 %	-37.200	.648	51.723	69.901	.710	78.007	58.101
35	UGCAUAUGCCAUGCAUCAG	CUGAUGCAUGGCAUAUGCAau	47.37 %	-38.400	.494	43.488	59.464	.620	64.558	42.433
36	GCAUAUGCCAUGCAUCAGA	UCUGAUGCAUGGCAUAUGCaa	47.37 %	-38.700	.791	64.874	84.132	.840	88.908	62.131
37	CAUAUGCCAUGCAUCAGAU	AUCUGAUGCAUGGCAUAUGca	42.11 %	-36.400	.719	60.847	77.757	.810	88.257	61.340
38	AUAUGCCAUGCAUCAGAUA	UAUCUGAUGCAUGGCAUAUgc	36.84 %	-35.600	.681	59.132	72.951	.750	85.017	61.088

Fig 8.4: siRNA efficiency Prediction by siRNA Designer with 6-8-8-8-1 ANN Model

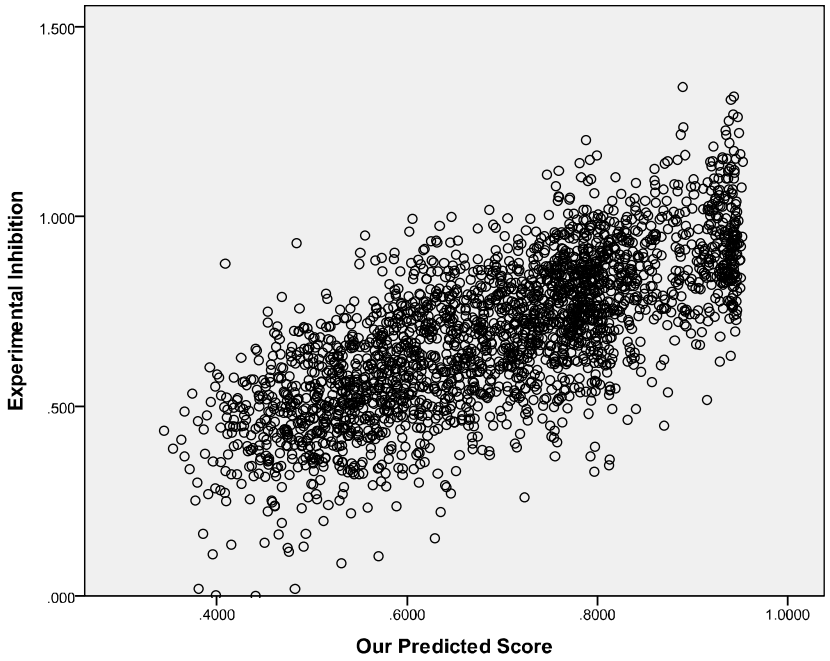


Fig. 8.5: Distribution between experimental inhibition and predicted inhibition for Data Set 1 by siRNA Designer with 6-8-8-8-1 model (R=0.727)

8.3.3 Effect of ΔG on Performance

The results are further analyzed to study the influence of whole stacking energy on inhibition efficiency of siRNA. The Pearson correlation coefficient is calculated for experimental versus predicted inhibition efficiency for the data sets, Data Set 1and Data Set 2, at various thresholds of whole stacking energies. It is noticed that the inhibition efficiency of our model is much closer to the original experimental values when threshold of whole tacking energy is >= -32.5 kcal/mol. We are able to achieve an improved correlation coefficient of R=0.753 when whole tacking energy is >= -32.5

kcal/mol, which shows improvement in the performance compared to our previous results. From this it is understood that while designing exogenous siRNA for gene silencing, whole stacking energy of each designed siRNA can also be analyzed for selecting efficient siRNAs with better inhibition capacity. Sample predicted inhibition values of siRNA for Data Set1 at whole tacking energy >= -32.5 kcal/mol is shown in Appendix 2. Fig 8.6 is the scatter plot showing distribution between experimental inhibition and predicted inhibition for Data Set 1 at ΔG >=-32.5 kcal/mol.

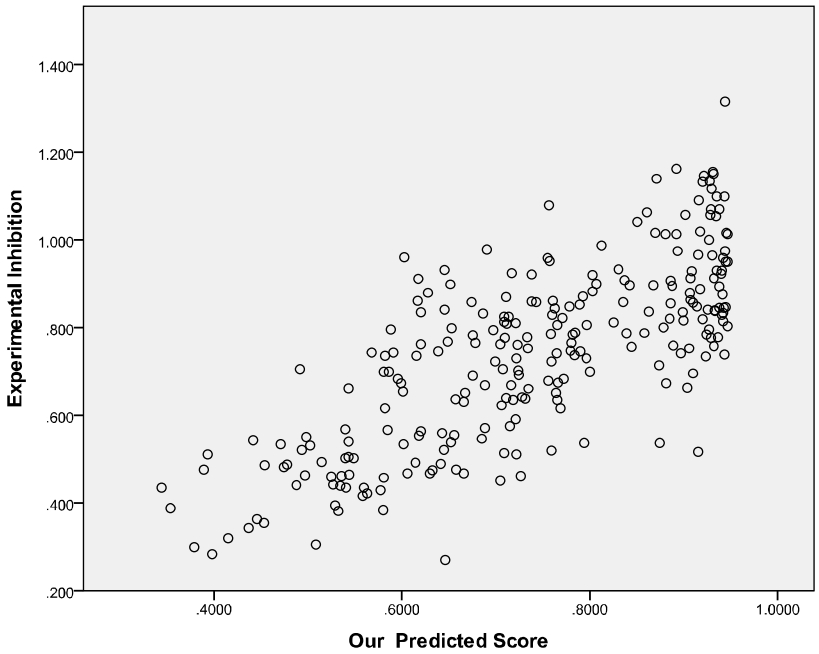


Fig. 8.6: Distribution between experimental inhibition and predicted inhibition for Data Set 1 by siRNA Designer with 6-8-8-8-1 ANN Model at ΔG >=-32.5 kcal/mol(R=0.753)

8.3.4 Comparison with siRNA Design Approaches

The inhibition capacity of siRNA for targeted mRNA is observed for each of the six scoring models, s-Biopredsi, DSIR, ThermoComposition21, i-Score, MysiRNA and OpsiD. A comparison between inhibition activities (Experimental versus Observed) for Data Set 1 by each of the five models with our model has been done. Also Pearson correlation coefficient (R) is calculated for each of the six scoring models and is shown in Fig.8.7. We achieve a Pearson correlation coefficient of R= 0.727 for Data Set 1 which is better than the other five models. The result shows the improvement of our approach in the accuracy of predicted siRNA. Also the experimental siRNAs activities of Data set 1 are plotted against the predicted siRNAs activities by each of the second generation models (s-Biopredsi, DSIR, ThermoComposition21, i-Score and MysiRNA) together with our model, which is shown in Fig. 8.8.

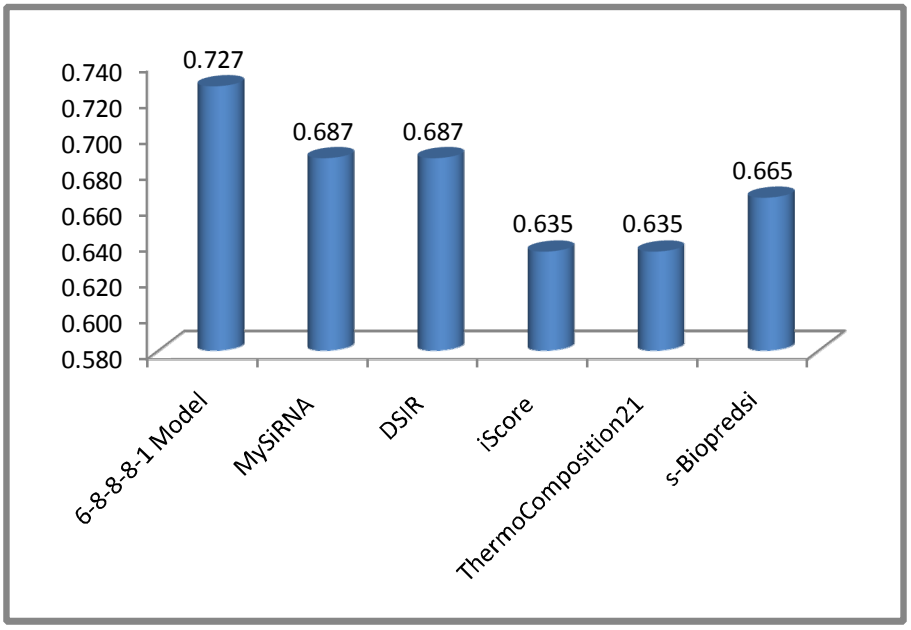


Fig. 8.7: Comparison between selected Second Generation Models and 6-8-8-8-1 ANN Model using Pearson Correlation Analysis.

In order to make a comparison of the effect of whole stacking energy on inhibition capacity of various models, Pearson Correlation coefficient is calculated for all of them at $\Delta G >= -32.5$ kcal/mol. We achieved an improved correlation coefficient of R =0.753 when whole tacking energy is greater than or equal to -32.5 kcal/mol. The value is compared with other five models and found better, which shows improvement in the performance of our model with them. The results of comparison are shown in shown in Fig.8.9.

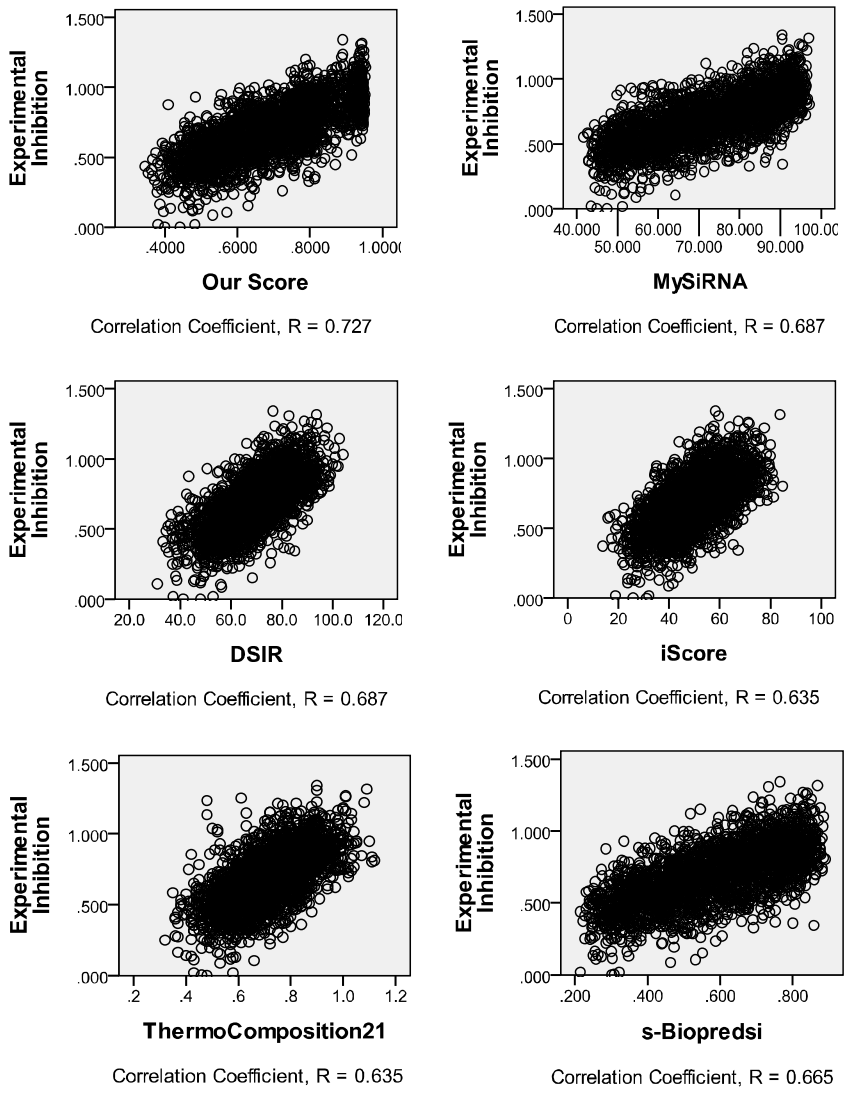


Fig.8.8: Comparative analysis of distribution between experimental inhibition and predicted inhibition of Dataset 1 for siRNA Designer with 6-8-8-8-1 ANN Model and selected second generation models.

Also the experimental siRNA activities of Dataset 1 at whole stacking energy, ΔG >=-32.5 kcal/mol are plotted against the predicted siRNA activities by each of the second generation models (s-Biopredsi, DSIR, ThermoComposition21, i-Score and MysiRNA) and is compared with our model and shown in Fig.8.10.

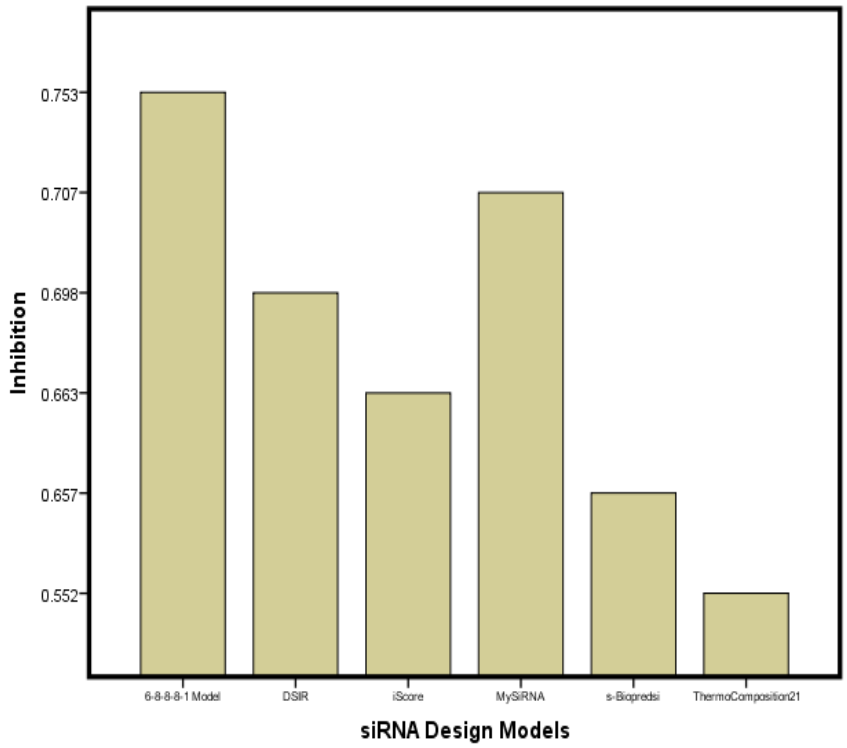


Fig. 8.9: Comparative analysis of Pearson Correlation Coefficient (R) involving second generation models and siRNA Designer with 6-8-8-8-1 Model at whole stacking energy, ΔG >= -32.5 kcal/mol for Data Set 1.

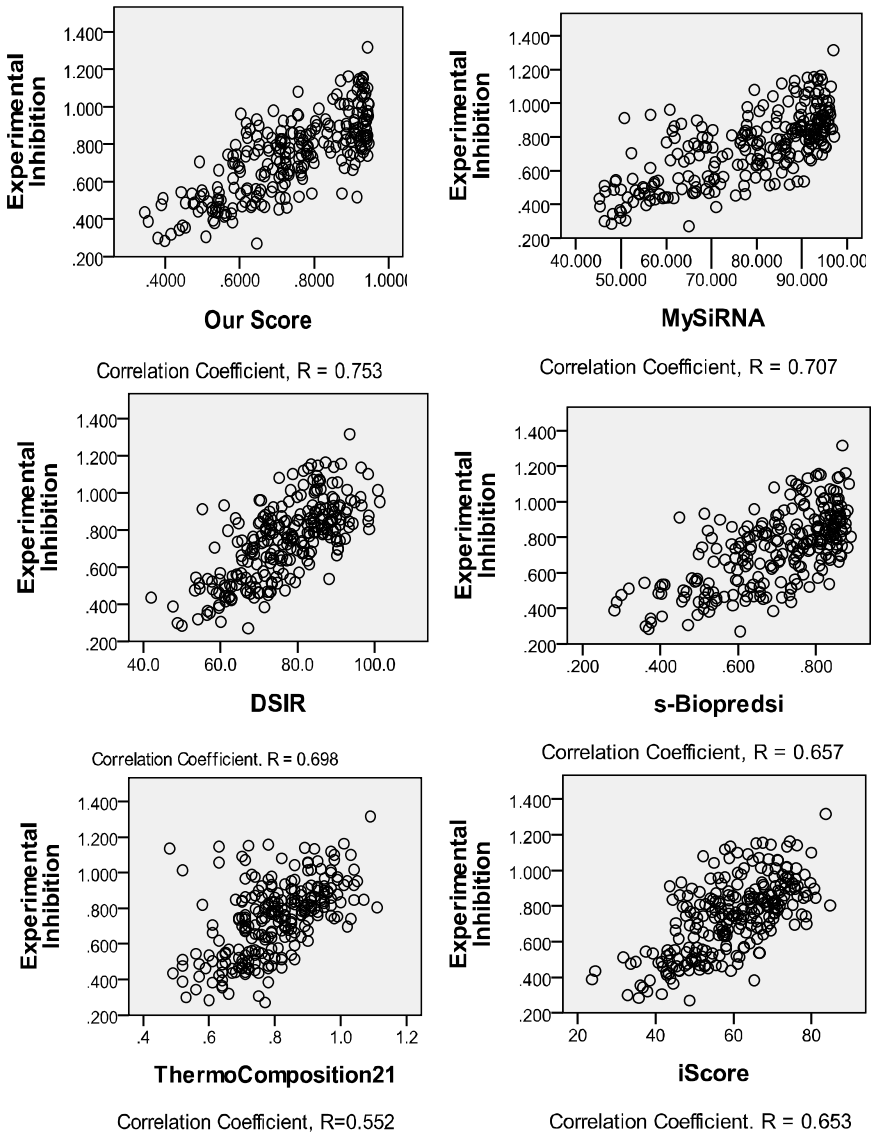


Fig. 8.10: Comparative analysis of distribution between experimental inhibition and predicted inhibition of siRNA Designer with 6-8-8-8-1 Model and selected second generation models for Dataset 1 at whole stacking energy ΔG >= -32.5 kcal/mol.

8.3.5 Discussion

Using 6-8-8-8-1 ANN model we are able to predict the percentage of inhibition efficiency of each predicted siRNA against a target mRNA or cDNA sequence. By maintaining a cut-off in inhibition efficiency (normally cut-off will be 70%-80% depending on the amount of silencing needed), one can select efficient siRNAs which are capable of inhibiting target mRNAs. The performance analysis and comparison of the approach with selected good scoring models are done. It is found that the prediction accuracy is improved in our model compared to selected existing state of the art models. The improvement in Pearson correlation coefficient shows better performance of our model. The effect of ΔG on inhibition efficiency is also understood. But when we tried to find the sensitivity and specificity of the model, it could not show better results over the existing state of the art methods. Thus using this model, the prediction efficiency is only optimized in terms of inhibition efficiency. So we moved forward to design another efficient model which can optimize the siRNA efficiency in terms of inhibition capacity, sensitivity, specificity, accuracy of prediction and off-target possibility. This is a 5-12-1 ANN model.

8.4 Optimized siRNA Designer Approach

8.4.1 Results

The results and discussion of the optimized siRNA designer, OpsiD, with 5-12-1 ANN model is described in this section. Predicted inhibition capacity of each siRNA on targeted genes and off-target possibility on non-target genes have been observed with OpsiD.

8.4.1.1 Off-Target Possibility Prediction

Even though an siRNA may have very good inhibition capacity i.e. it may have very good ability to bind to the target mRNA for gene knockdown, it may be fully unsuitable for practical therapeutic use because of its similarity to segments of other mRNAs. In such a case, the siRNA may bind to the other mRNAs instead of the intended target. In this way, it may interfere with the translation of essential genes to proteins, and cause unintended effects. This problem is known as "off target effect", and is the major barrier against the practical use of gene silencing through siRNAs in therapeutics. So an important factor to be considered while designing efficient siRNAs for therapeutical and gene silencing applications is the chance of siRNA to do off-target effects on non-target genes. Thus the inhibition capacity alone is not a reliable indicator of an

siRNA's practical utility because of the possibility of off-target effect.

Initially the OpsiD model is designed to predict the inhibition capacity of an siRNA through its neural network regression model and we were getting excellent results. But in order to improve further by avoiding off-target possibility of designed siRNAs on non-target genes, we have added BLAST search technique in our model. So, the siRNA designer model, OpsiD, mitigates the problem of "off-target effect on non-target genes" by providing the facility of running BLAST search of the generated siRNAs against standard databases of mRNAs such as the NCBI RefSeq database [163]. The BLAST score given in the OpsiD outputs for a particular siRNA's maximal match obtained for that sequence against some other mRNA subsequence in the selected gene database. A BLAST score of 19 indicates a complete match of the siRNA sequence with a subsequence of some other mRNA in the gene database. A score of 18 indicates that the siRNA sequence differs from a subsequence of some other mRNA by only a single nucleotide. The use of siRNAs with high BLAST score may lead to off-target effect, and the user must make the trade off between the "goodness" of an siRNA with respect to inhibition capacity, and its similarity to other mRNA fragments.

For example, in Table 8.1 the siRNA with best inhibition efficacy against a given target mRNA is 89% with a BLAST score 17, means even though the siRNA is best efficient to degrade the target mRNA by 89%, there is a high risk of 'off target effect' with 17 nucleotide matches to any other genes in the database. So we believe that the users will be able to eliminate those siRNA sequences with high BLAST score, even though they possess very good inhibition capacity. As per our approach, instead of selecting siRNA with the best inhibition capacity, we can consider both "inhibition efficiency and number of matches of BLAST score" to select siRNAs for gene silencing. Thus the risk of "off target effect" against unintended target sites may be avoided to a great extent. The sample screen shot showing user interface and results are shown Fig. 8.11 and Fig. 8.12 respectively.

Table 8.1: Sample siRNAs with Inhibition capacity and BLAST Score in OpsiD

siRNA Strand	Inhibition	BLAST Score
AGGGUUAUUUUUCUUUGGC	75	11
GAAAAAAACCAAAGGGUUA	67	3
AACCACUGUAGAAAAUAAC	35	0
UCUUUAUGUUUUUGGCGUC	89	17
UUCUUUAUGUUUUUGGCGU	76	9
GGGCCUUUCUUUAUGUUUU	55	7
UUAUAAAUGUCGUUCGCGG	77	12
UAAUUUUUUGGAUGAUUGG	45	4
UUAAAAUCGCAGUAUCCGG	67	8

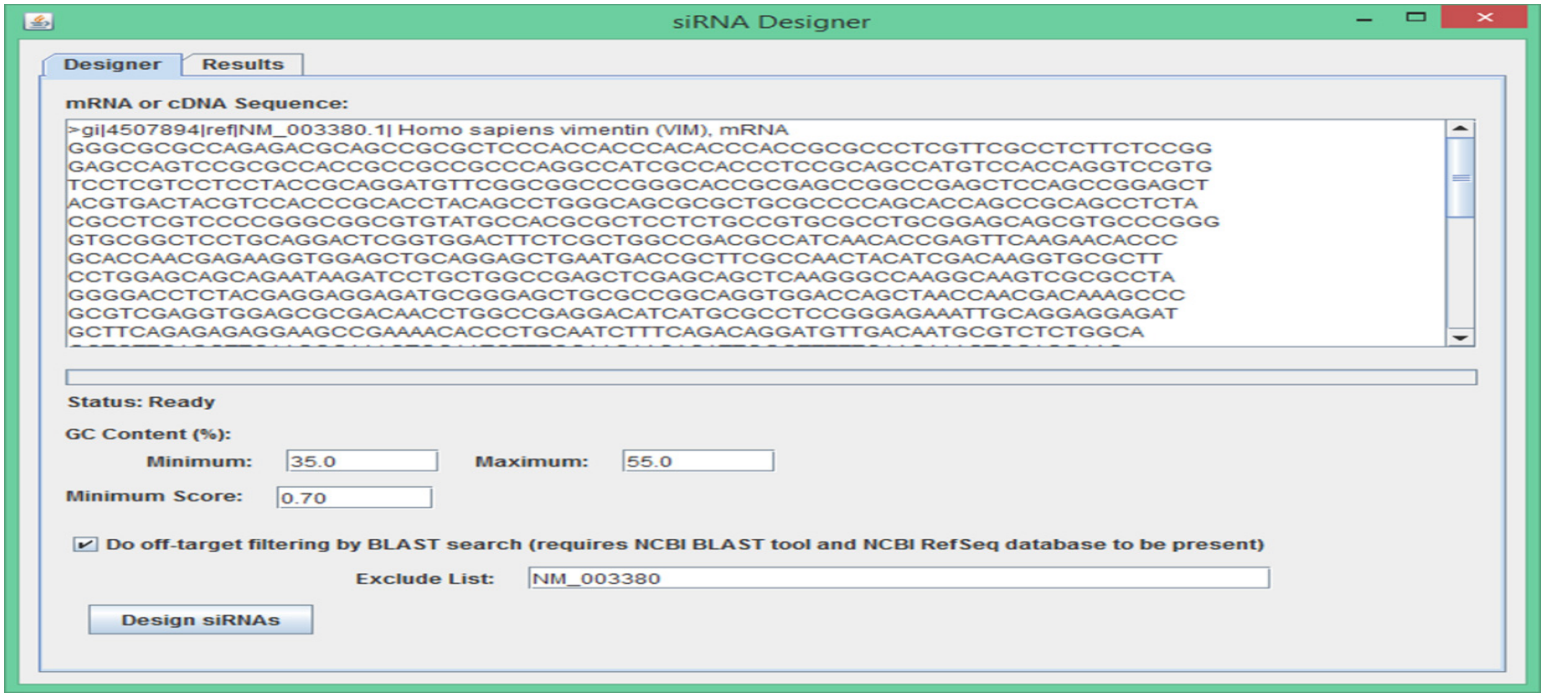
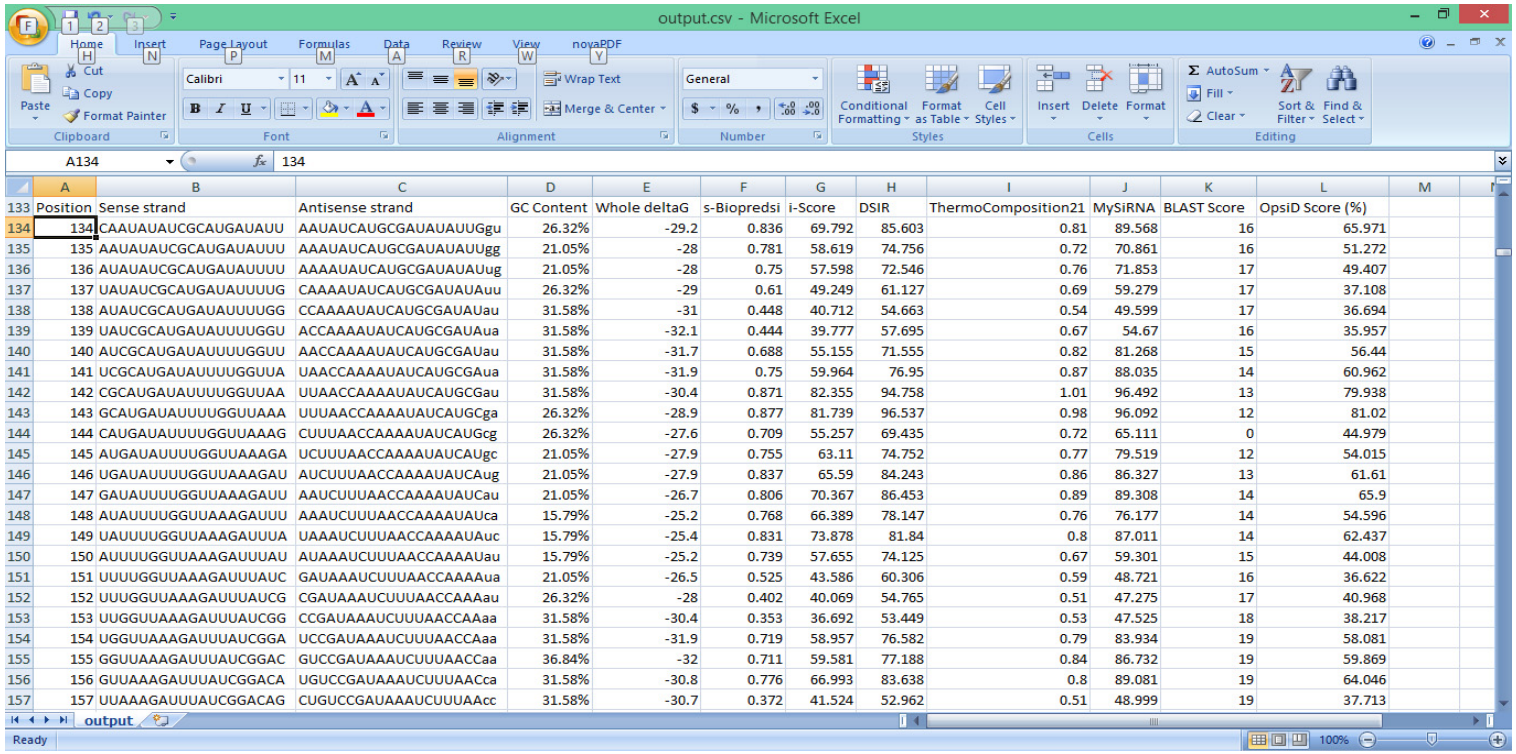


Fig. 8.11: Sample screen shot showing the user interface of OpsiD with off-target filtering



output.csv - Microsoft Excel

	Position	Sense strand	Antisense strand	GC Content	Whole deltaG	s-Biopredsi	i-Score	DSIR	ThermoComposition21	MySiRNA	BLAST Score	OpsiD Score (%)
134	134	CAAUAUAUCGCAUGAUAUU	AAUAUCAUGCGAUAUAUUGgu	26.32%	-29.2	0.836	69.792	85.603	0.81	89.568	16	65.971
135	135	AAUAUAUCGCAUGAUAUUU	AAAUAUCAUGCGAUAUAUUgg	21.05%	-28	0.781	58.619	74.756	0.72	70.861	16	51.272
136	136	AUAUAUCGCAUGAUAUUUU	AAAAUAUCAUGCGAUAUAUug	21.05%	-28	0.75	57.598	72.546	0.76	71.853	17	49.407
137	137	UAUAUCGCAUGAUAUUUUG	CAAAAUAUCAUGCGAUAUAuu	26.32%	-29	0.61	49.249	61.127	0.69	59.279	17	37.108
138	138	AUAUCGCAUGAUAUUUUGG	CCAAAAUAUCAUGCGAUAUau	31.58%	-31	0.448	40.712	54.663	0.54	49.599	17	36.694
139	139	UAUCGCAUGAUAUUUUGGU	ACCAAAAUAUCAUGCGAUAua	31.58%	-32.1	0.444	39.777	57.695	0.67	54.67	16	35.957
140	140	AUCGCAUGAUAUUUUGGUU	AACCAAAAUAUCAUGCGAUau	31.58%	-31.7	0.688	55.155	71.555	0.82	81.268	15	56.44
141	141	UCGCAUGAUAUUUUGGUUA	UAACCAAAAUAUCAUGCGAua	31.58%	-31.9	0.75	59.964	76.95	0.87	88.035	14	60.962
142	142	CGCAUGAUAUUUUGGUUAA	UUAACCAAAAUAUCAUGCGau	31.58%	-30.4	0.871	82.355	94.758	1.01	96.492	13	79.938
143	143	GCAUGAUAUUUUGGUUAAA	UUUAACCAAAAUAUCAUGCga	26.32%	-28.9	0.877	81.739	96.537	0.98	96.092	12	81.02
144	144	CAUGAUAUUUUGGUUAAAG	CUUUAACCAAAAUAUCAUGcg	26.32%	-27.6	0.709	55.257	69.435	0.72	65.111	0	44.979
145	145	AUGAUAUUUUGGUUAAAGA	UCUUUAACCAAAAUAUCAUgc	21.05%	-27.9	0.755	63.11	74.752	0.77	79.519	12	54.015
146	146	UGAUAUUUUGGUUAAAGAU	AUCUUUAACCAAAAUAUCAug	21.05%	-27.9	0.837	65.59	84.243	0.86	86.327	13	61.61
147	147	GAUAUUUUGGUUAAAGAUU	AAUCUUUAACCAAAAUAUCau	21.05%	-26.7	0.806	70.367	86.453	0.89	89.308	14	65.9
148	148	AUAUUUUGGUUAAAGAUUU	AAAUCUUUAACCAAAAUAUca	15.79%	-25.2	0.768	66.389	78.147	0.76	76.177	14	54.596
149	149	UAUUUUGGUUAAAGAUUUA	UAAAUCUUUAACCAAAAUAuc	15.79%	-25.4	0.831	73.878	81.84	0.8	87.011	14	62.437
150	150	AUUUUGGUUAAAGAUUUAU	AUAAAUCUUUAACCAAAAUau	15.79%	-25.2	0.739	57.655	74.125	0.67	59.301	15	44.008
151	151	UUUUGGUUAAAGAUUUAUC	GAUAAAUCUUUAACCAAAAua	21.05%	-26.5	0.525	43.586	60.306	0.59	48.721	16	36.622
152	152	UUUGGUUAAAGAUUUAUCG	CGAUAAAUCUUUAACCAAAau	26.32%	-28	0.402	40.069	54.765	0.51	47.275	17	40.968
153	153	UUGGUUAAAGAUUUAUCGG	CCGAUAAAUCUUUAACCAAaa	31.58%	-30.4	0.353	36.692	53.449	0.53	47.525	18	38.217
154	154	UGGUUAAAGAUUUAUCGGA	UCCGAUAAAUCUUUAACCAaa	31.58%	-31.9	0.719	58.957	76.582	0.79	83.934	19	58.081
155	155	GGUUAAAGAUUUAUCGGAC	GUCCGAUAAAUCUUUAACCaa	36.84%	-32	0.711	59.581	77.188	0.84	86.732	19	59.869
156	156	GUUAAAGAUUUAUCGGACA	UGUCCGAUAAAUCUUUAACca	31.58%	-30.8	0.776	66.993	83.638	0.8	89.081	19	64.046
157	157	UUAAAGAUUUAUCGGACAG	CUGUCCGAUAAAUCUUUAAcc	31.58%	-30.7	0.372	41.524	52.962	0.51	48.999	19	37.713

Fig. 8.12: Sample Screen Shot showing the output with BLAST Score of OpsiD.

8.4.2 Performance Evaluation

For OpsiD, predicted inhibition capacity of each siRNA for a targeted mRNA is found out and done the performance evaluation in terms of Pearson Correlation, Sensitivity-Specificity, Accuracy of Prediction, Mathews Correlation Coefficient, and Receiver Operating Characteristic analysis. These values are calculated a described in Section 4.8 of Chapter 4.

8.4.2.1 Pearson Correlation

Pearson Correlation (R) gives the correlation between the inhibition efficiency of predicted model with the original experimental inhibition efficiency. Pearson correlation coefficient (R) is calculated for each of the six scoring models s-Biopredsi, DSIR, ThermoComposition21, i-Score, MysiRNA and OpsiD. The inhibition activities (Experimental versus Observed) for Data Set1 and Data Set2 for all models are observed. We achieved a Pearson correlation coefficient of R= 0.699 for Data Set 1 and R= 0.606 for Data Set 2. The distribution between experimental inhibition and predicted inhibition for Data Set 1 and Data Set 2 is shown in Fig. 8.13 and Fig. 8.14 respectively.

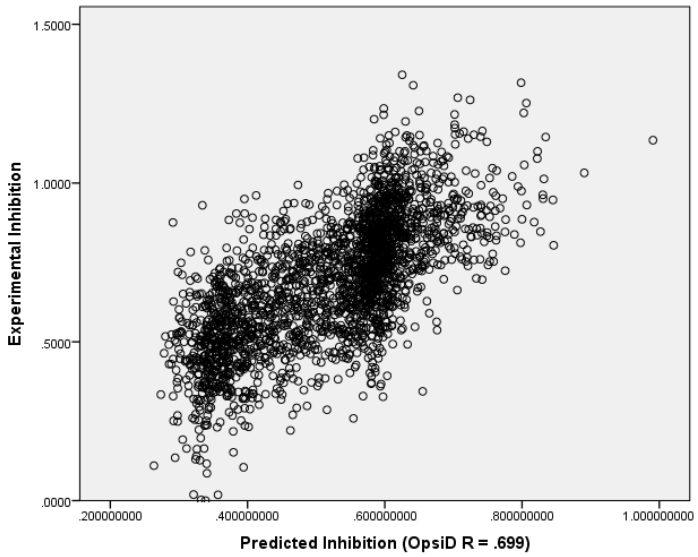


Fig. 8.13: Distribution between experimental inhibition and predicted inhibition for Data Set 1 by OpsiD (R=0.699)

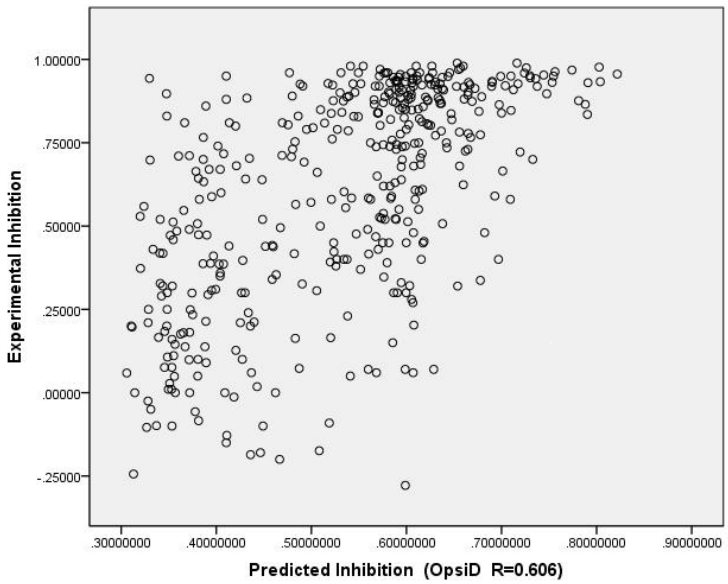


Fig. 8.14: Distribution between experimental inhibition and predicted inhibition for Data Set 2 by OpsiD (R=0.606)

8.4.2.2 Sensitivity-Specificity

Normally the siRNA design models are expected to have the ability to reject as many as false positives as possible and retain maximum true positives. OpsiD is compared against five previous siRNA design models for their ability to select efficient siRNAs and reject inefficient siRNAs. For this we used Data Set 3 which contains 476 siRNA from 9 genes. The inhibition capacity of the Data Set 3 is compared with experimental data results and the results are classified into 4 groups: True Positive (TP) and True Negative (TN) when the program could identify efficient siRNA and inefficient siRNA, and False Positive (FP) and False Negative (FN) when the program falsely identified inefficient siRNA as efficient, or efficient siRNA as inefficient, respectively. Both the sensitivity (ability to identify true positives) and specificity (ability to reject false positives) are taken into consideration. Appendix 3 shows sample TP, TN, FP, FN values calculated using Data Set3. Total count of TP, TN, FP, FN values obtained for Data Set3 in OpsiD is shown in Table 8.2.

Table 8.2: TP, TN, FP, FN Values of OpsiD

Diagnostic Test Result	Existence of disease as determined by the standard of truth		
	Positive	**Negative**	**Row Total**
Positive	164 (TP)	41 (FP)	205
Negative	74 (FN)	197 (TN)	271
Column total	238	238	576

OpsiD is found capable of designing siRNA with high level of specificity and sensitivity. It achieves a Sensitivity of 0.69 and Specificity of 0.83. These values show better prediction power of our model.

8.4.2.3 Accuracy of Prediction, Mathews Correlation Coefficient

In addition to Pearson Correlation Coefficient and sensitivity-specificity, the performance evaluation of OpsiD is also done by Accuracy of Prediction and Mathews Correlation Coefficient (MCC) as described in Chapter 4 (Section 4.8.2). For calculating MCC, a cut off value of inhibition efficiency of siRNA up to 60% is applied in Data set 3 to categorize siRNA as efficient or inefficient. The predicted siRNA is considered as efficient if predicted value is above

the threshold and inefficient if predicted value is below this threshold. OpsiD achieves the highest MCC of 0.52 with the experimentally verified data. The MCC value of 0.52 indicates a strong correlation between observed and experimental prediction OpsiD performed well with MCC = 0.52 and Accuracy of prediction = 0.76. Overall, the performance analysis indicates the improvement in performance of our model in terms of Accuracy, MCC, and Sensitivity over other models. The Table 8.3 shows the validation results for Pearson Correlation, Sensitivity, Specificity, Positive Predictive Value (PPV), Negative Predictive Value (NPV), False Positive Rate (FPR), False Negative Rate (FNR), False Discovery Rate (FDR), F-Score (F), Accuracy of prediction, and MCC of OpsiD.

Table 8.3: Validation results of OpsiD

Validation Strategy	**Results of OpsiD**
Pearson Correlation for Data Set 1	0.699
Pearson Correlation for Data Set 1	0.606
Sensitivity	0.69
Specificity	0.83
PPV	0.80
NPV	0.73
FPR	0.17

FNR	0.31
FDR	0.20
F Score	0.74
Accuracy	0.76
MCC	0.52

8.4.2.4 ROC Analysis

In addition, we used receiver operating characteristic analysis that combines both sensitivity and specificity by plotting the sensitivity (Y axis) against 1- specificity (X axis). For the ROC analysis, we considered siRNA with inhibition equal to or above 70% as efficient siRNA and below 70% as inefficient siRNA. Fig.8.15 and Fig.8.16 shows the ROC curve obtained for OpsiD model for Data Set 1 and Data Set2 respectively. It is then possible to calculate the area under the curve, known as the AUC, as a single measure of performance (for which an AUC of 1 reflects perfect classification and an AUC of 0.5 reflects random classification). We achieved an AUC of 0.862 for Data Set 1 and 0.809 for Data Set 2 which are comparatively good results and indicate better performance of our model, OpsiD.

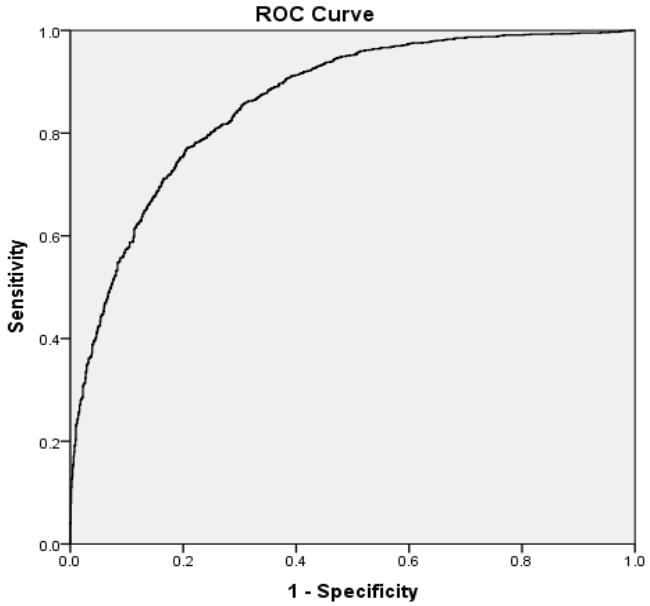


Fig. 8.15: The ROC Analysis Curve of Data Set1 by OpsiD (AUC =0.862)

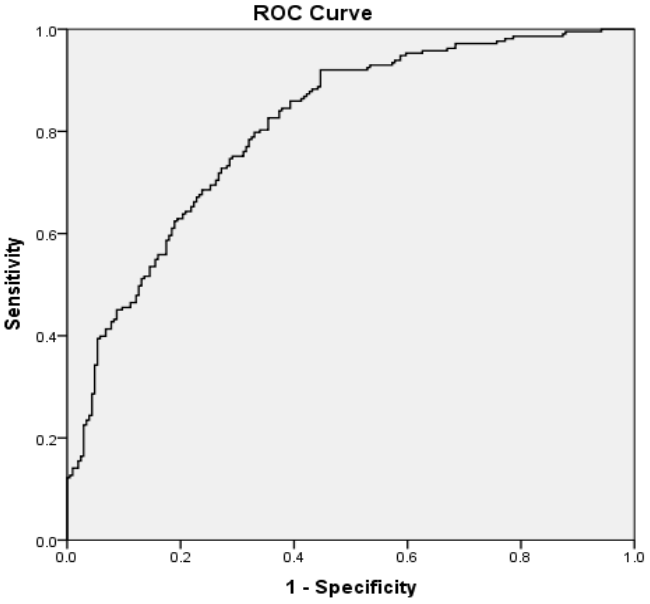


Fig. 8.16: The ROC Analysis Curve of Data Set 2 by OpsiD (AUC =0.809)

8.4.2.5 Effect of ΔG on Performance

When the four parameters (score of MysiRNA, score of DSIR, score of i-Score and ThermoComposition21 score) are combined with ΔG using a multi-layer perceptron feed-forward neural network model, considerable performance improvement in prediction accuracy is noticed. For this we divided the Data Set2 into two sets with a threshold of ΔG = -34.6. (i.e., ΔG < -34.6 kcal/mol and ΔG >= -34.6 kcal/mol). (In [21], Ichihara et al. calculated an effective threshold value of -34.6, for separating the data sets). This combination in our model results with Pearson Correlation of 0.693 for Data Set1 and 0.741 for Data Set2 between the experimental inhibition and predicted inhibition efficiencies, when the threshold of whole stacking energy, ΔG >= -34.6 kcal/mol. Table 8.4 shows the Pearson Correlation Coefficient (R) of OpsiD for Data Set 1 and Data Set 2. The R value without considering ΔG for OpsiD is 0.699 for Data Set 1 and 0.606 for Data Set 2. Even though the R value is little reduced for Data Set 1 while considering ΔG >= -34.6 kcal/mol, still it is a better result when compared with selected existing models. This improvement in Pearson correlation values show the importance and influence of whole stacking energy on inhibition efficiency of siRNA. Sample predicted inhibition values of siRNA for Data Set1

and Data Set 2 at whole tacking energy, ΔG >= -34.6 kcal/mol are shown in Appendix 4 and Appendix 5 respectively. Distribution between experimental inhibition and predicted inhibition of each siRNA for Data Set 1 and Data Set2 by OpsiD when Δ G >=-34.6 kcal/mol is shown in Fig. 8.17 and Fig 8.18.

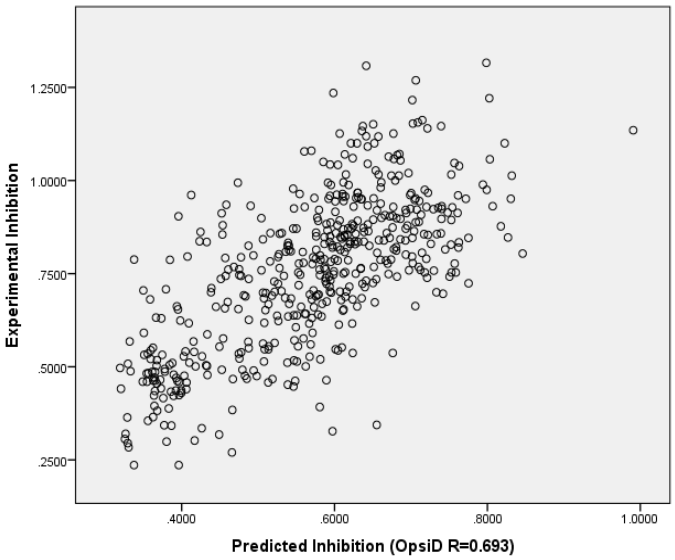


Fig. 8.17: Distribution between experimental inhibition and predicted inhibition for Data Set 1 by OpsiD when ΔG > =-34.6 kcal/mol (R=0.693).

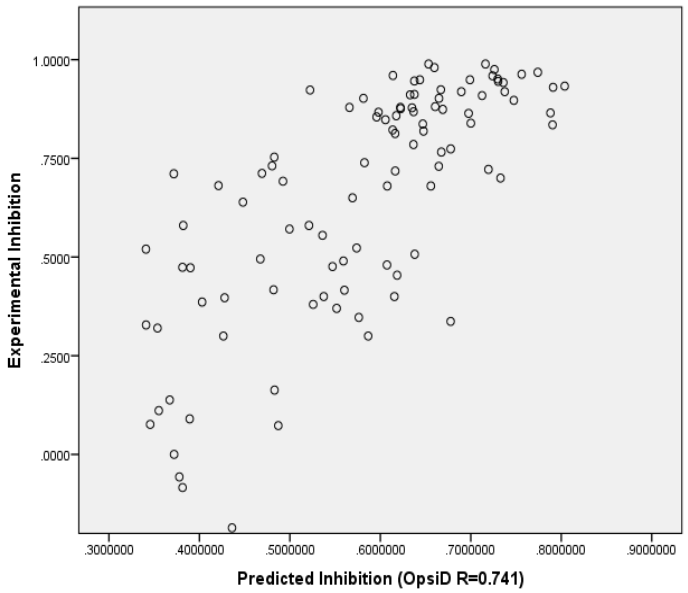


Fig. 8.18: Distribution between experimental inhibition and predicted inhibition for Data Set 2 by OpsiD when ΔG > =-34.6 kcal/mol (R=0.741)

Table 8.4: Pearson Correlation Coefficient (R) of OpsiD

Data Set	**AUC of OpsiD**
Data Set 1	0.699
Data Set 2	0.606
Data Set 1 when ΔG>= -34.6	0.693
Data Set 2 when ΔG>= -34.6	0.741

In order to find the effect of whole stacking energy on performance, the AUC values are calculated at ΔG >= -34.6 kcal/mol. With this whole stacking energy, the AUC values are also improved further and achieved 0.878 for Data Set1 and 0.906 for Data Set2. The ROC curves of Data Set1 and Data set2 at ΔG >= -34.6 kcal/mol is shown in Fig. 8.19 and Fig. 8.20. This improvement in AUC values shows the importance of the influence of whole stacking energy on the performance of siRNA. The results of performance evaluation in terms of AUC for OpsiD are shown in Table 8.5.

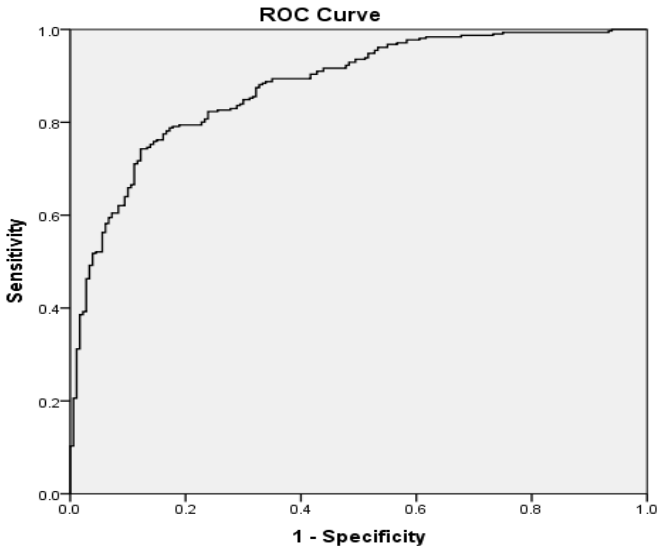


Fig. 8.19: The ROC Analysis Curve of Data Set1 by OpsiD at ΔG>=-34.6 kcal/mol (AUC = 0.878)

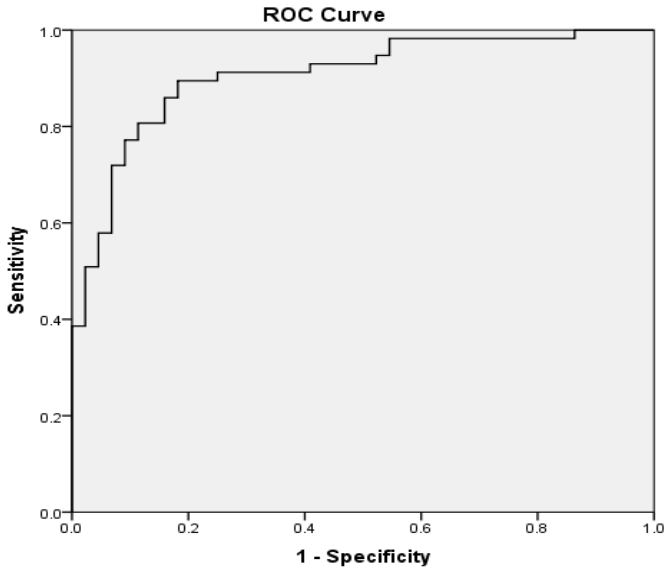


Fig. 8.20: The ROC Analysis Curve of Data Set 2 by OpsiD at ΔG>=-34.6 kcal/mol (AUC =0.906)

Table 8.5: AUC values of OpsiD

Data Set	**AUC of OpsiD**
Data Set 1	0.862
Data Set 2	0.809
Data Set 1 when ΔG>= -34.6	0.878
Data Set 2 when ΔG>= -34.6	0.906

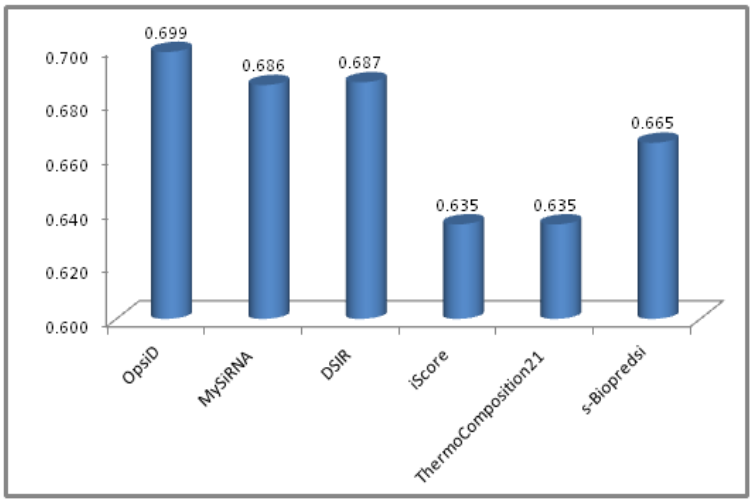


Fig. 8.21: Comparative analysis of Pearson Correlation Coefficient (R) involving OpsiD, MySiRNA, DSIR, iScore, ThermoComposition 21, s-Biopredsi for Data Set 1.

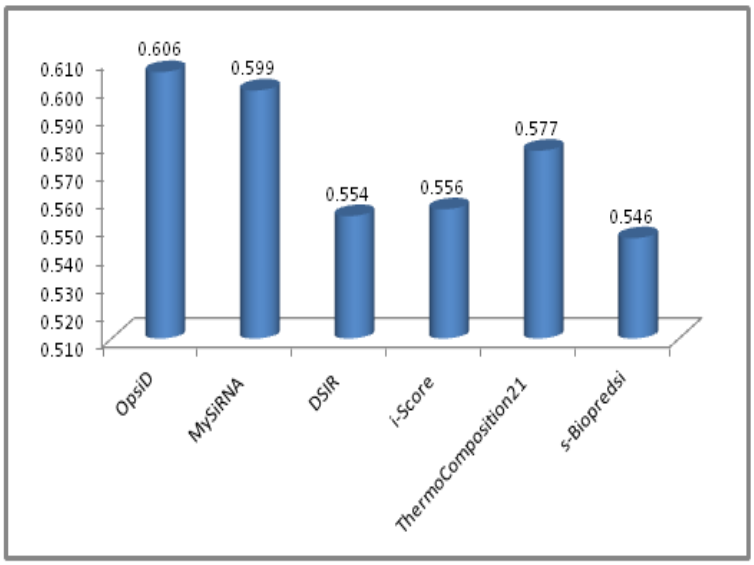
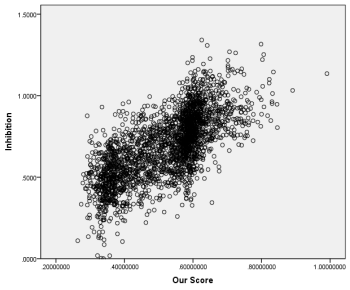


Fig. 8.22: Comparative analysis of Pearson Correlation Coefficient (R) involving OpsiD, MySiRNA, DSIR, iScore, ThermoComposition21, s-Biopredsi for Data Set 2.

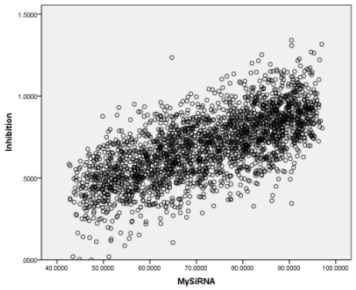
8.4.3 Comparison with siRNA Design Approaches

The results of OpsiD are compared and validated with MysiRNA, DSIR, i-Score, ThermoComposition21, and s-Biopredsi in terms of siRNA inhibition efficiency, prediction accuracy, sensitivity, specificity, MCC and AUC values. The inhibition capacity is measured in terms of Pearson Correlation, R. The Pearson Correlation Coefficient is calculated for experimental inhibition capacity versus predicted inhibition capacity for each of six models.

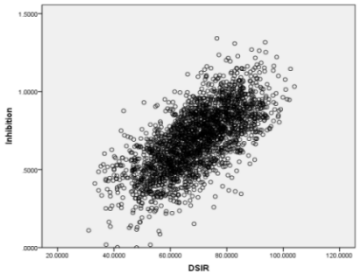
We achieved a very good correlation between the predicted and experimental siRNA inhibition efficiency for Data Set 1 and Data Set 2. In both Data Sets the correlation values are higher than MysiRNA, DSIR, i-Score, ThermoComposition21, and s-Biopredsi indicating better prediction by our model (Fig. 8.21 and Fig. 8.22). The experimental siRNA inhibition plotted against predicted inhibition with all six techniques for Data Set1 and Data Set 2 is shown in Fig. 8.23 and Fig 8.24.



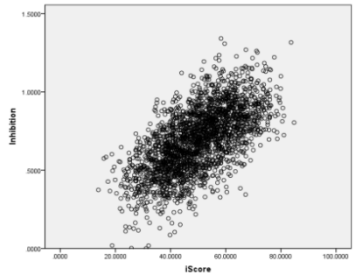
OpsiD R= 0.699



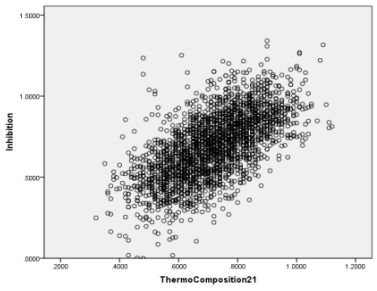
MysiRNA R= 0.686



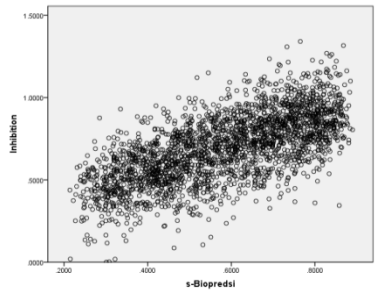
DSIR R=0.687



iScore R= 0.635



ThermoComposotion21 R=0.635



s-Biopredsi R= 0.665

Fig.8.23:Comparative analysis of distribution between experimental inhibitionandpredicted inhibition of OpsiD and second generation models for Data Set 1.

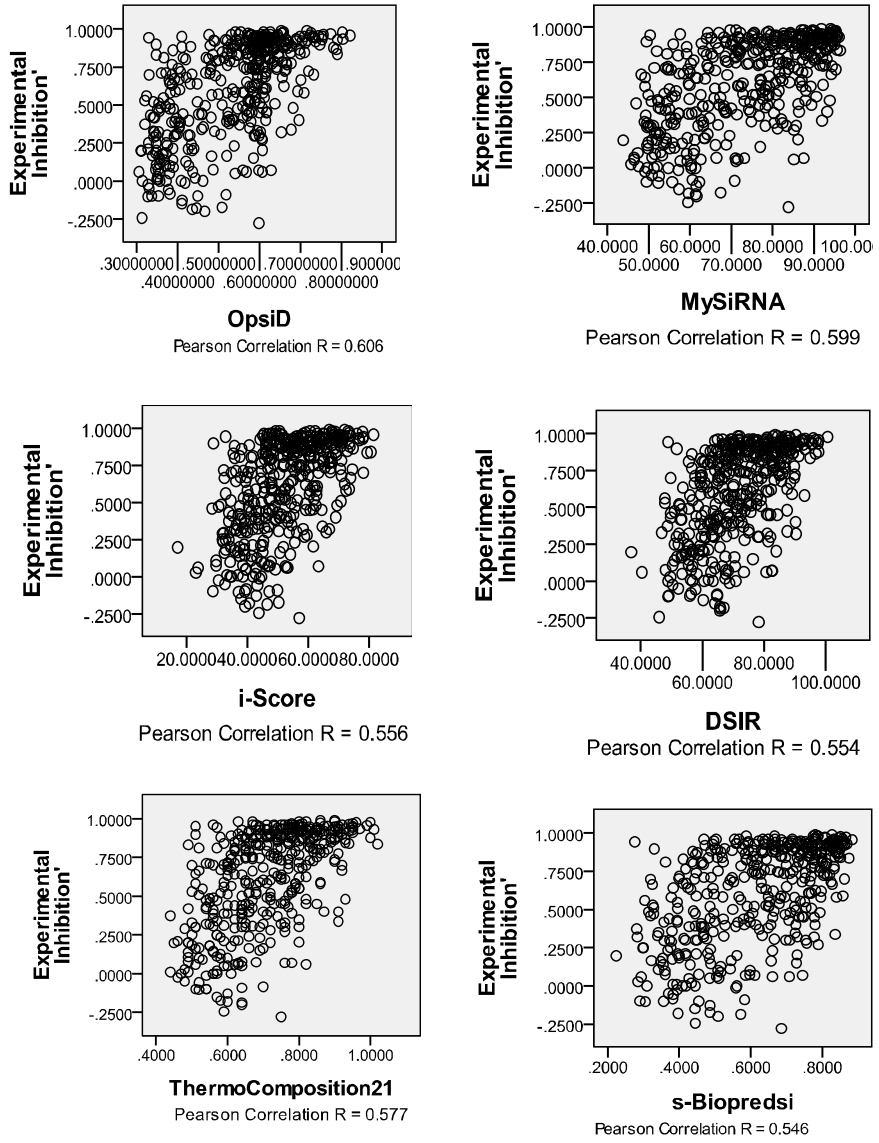


Fig. 8.24: Comparative analysis of distribution between experimental inhibition and predicted inhibition of OpsiD and second generation models for Data Set 2.

Next comparison is done in terms of sensitivity, specificity, MCC, ROC analysis and Accuracy of prediction. OpsiD is found capable of designing siRNA with good level of specificity and sensitivity. It achieves Sensitivity (Sn) of 0.69 and Specificity (Sp) of 0.83. From the results shown in Table 8.6, it is observed that the value of Sn for OpsiD is the best Sn value than all other shown models. This reflects the highest rate of predicting high efficacy siRNAs. OpsiD achieved a specificity of 0.83. Even though it is an acceptable level of specificity, it is comparatively a lower value when compared with other techniques, indicating that this aspect must be improved. But on analyzing the sensitivity-specificity values shown in Table 8.6, it is clear that even though all other models have high specificity than OpsiD, their sensitivity values are very less. So if we consider combined sensitivity-specificity effects shown by our model, we can come to the conclusion that OpsiD performs as well as or better than all other models in the list. Overall, the analysis indicates the performance improvement of our model in terms of Accuracy, MCC, and Sensitivity over other models. The Table 8.6 shows the comparison of results from 6 models in terms of TP, TN, FP and FN. Table 8.7 shows the comparison of results from 6 models in terms of accuracy, sensitivity, specificity and MCC.

Table 8.6: Comparative analysis of TP, TN, FP, FN for OpsiD, MySiRNA, DSIR, iScore, ThermoComposition21, s-Biopredsi for Data Set 3

Count	**OpsiD**	**MysiRNA**	**DSIR**	**Thermo**	**iScore**	**s-Biopredsi**
Count TP	164	153	105	96	0	43
Count TN	197	204	223	217	238	233
Count FP	41	34	15	21	0	5
Count FN	74	85	133	142	238	195

Table 8.7: Comparative analysis of Accuracy, Sensitivity, Specificity and MCC for OpsiD, MySiRNA, DSIR, iScore, ThermoComposition21, s-Biopredsi for Data Set 3

	OpsiD	**MySiRNA**	**DSIR**	**iScore**	**Thermo Composition21**	**s-Biopredsi**
Accuracy	0.76	0.75	0.69	0.66	0.5	0.58
Sensitivity	0.69	0.64	0.44	0.4	0	0.18
Specificity	0.83	0.86	0.94	0.91	1	0.98
MCC	0.52	0.51	0.43	0.37	--	0.26

For the ROC analysis, we considered siRNA with inhibition efficiency equal to or above 70% as efficient siRNA and below 70% as inefficient siRNA. Fig.8.25 and Fig.8.26 shows comparative analysis of ROC curve obtained for OpsiD, MySiRNA, DSIR, iScore, ThermoComposition21, s-Biopredsi for Data Set 1 and Data Set2 respectively. Then it is possible to calculate the area under the curve, known as the AUC, as a single measure of performance (for which an AUC of 1 reflects perfect classification and an AUC of 0.5 reflects random classification). The AUC obtained by our model has been compared with each of 5 techniques and found that we have got an AUC of 0.862 for Data Set 1 and 0.809 for Data Set 2, which are better than those obtained from MysiRNA, DSIR, i-Score, ThermoComposition21, and s-Biopredsi. The AUC values of 0.809 and 0.862 indicate better performance of our model. The results are shown in Fig. 8.27 and Fig. 8.28.

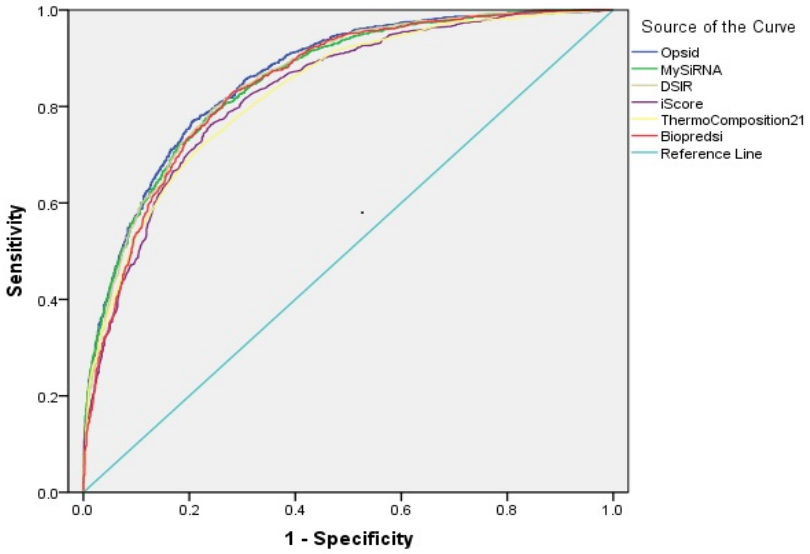


Fig. 8.25: Comparative analysis of ROC curve for OpsiD, MySiRNA, DSIR, iScore, ThermoComposition21, s-Biopredsi for Data Set 1.

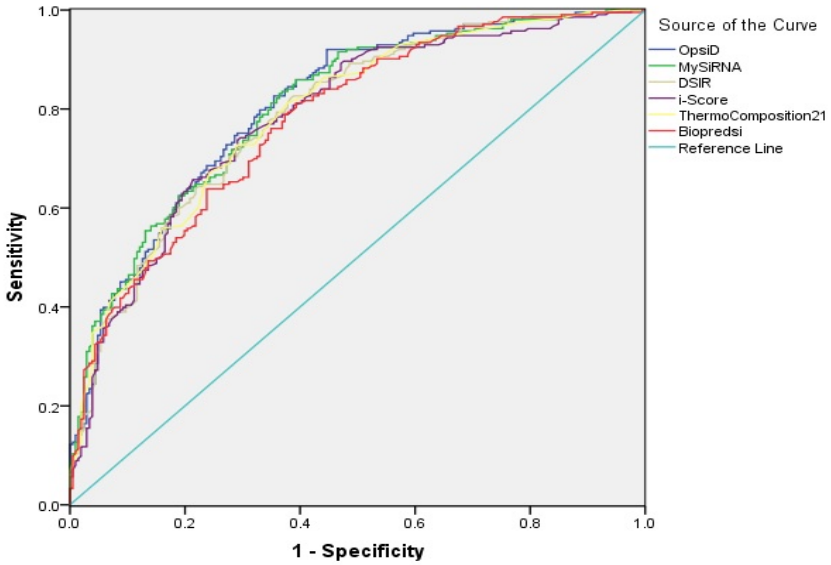


Fig. 8.26: Comparative analysis of ROC Curve for OpsiD, MySiRNA, DSIR, iScore, ThermoComposition21, s-Biopredsi for Data Set 2.

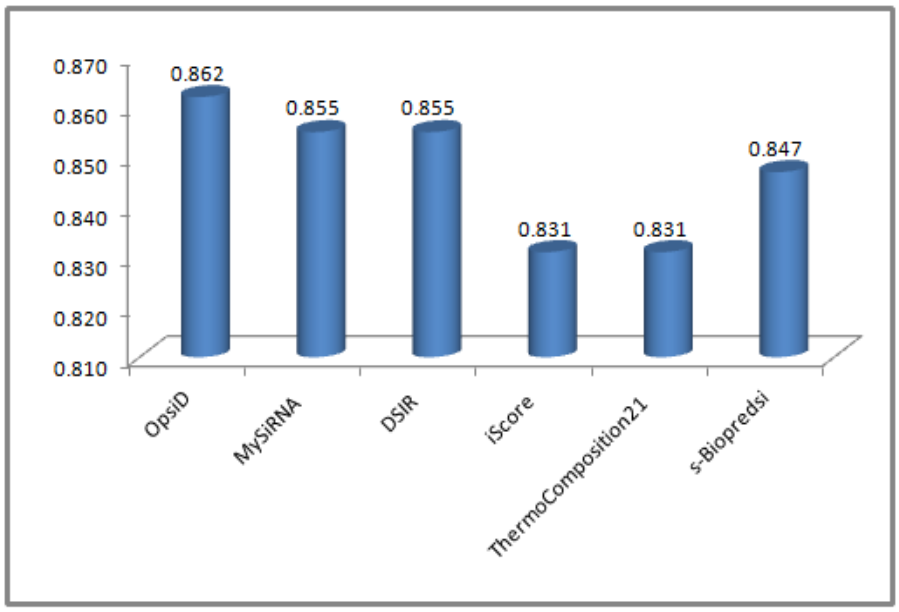


Fig. 8.27: Comparative analysis of AUC involving OpsiD, MySiRNA, DSIR, iScore, ThermoComposition21, s-Biopredsi for Data Set 1

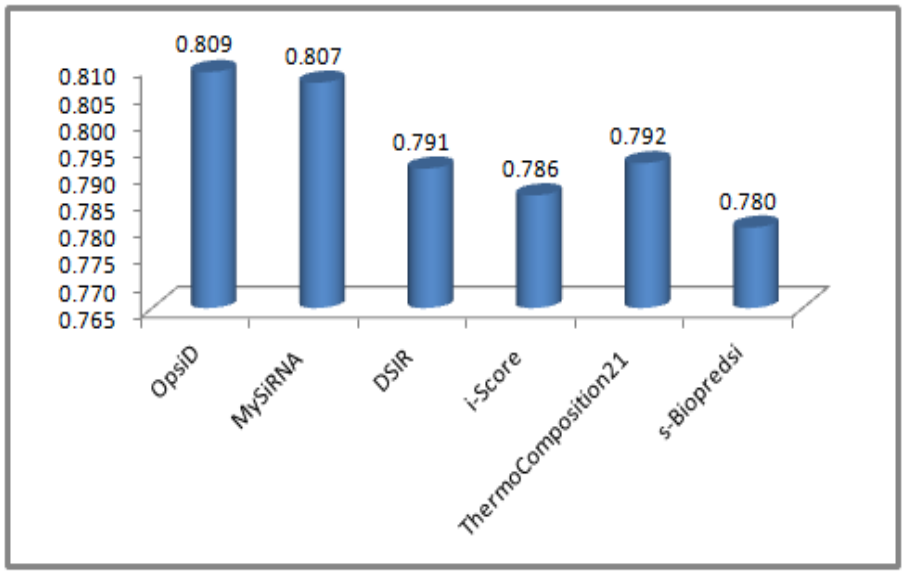


Fig. 8.28: Comparative analysis of AUC involving OpsiD, MySiRNA, DSIR, iScore, ThermoComposition21, s-Biopredsi for Data Set 2

In order to find the performance improvement with whole stacking energy, we calculated the Pearson Correlation value and AUC values of OpsiD and other five models and compared the results. When the threshold of whole stacking energy, $\Delta G >= -34.6$ kcal/mol, we achieved a Pearson Correlation of 0.693 for Data Set1 and 0.741 for Data Set2 between the experimental inhibition and predicted inhibition efficiencies. The comparative analysis of Pearson Correlation Coefficient, R, at whole stacking energy $\Delta G >= -34.6$ kcal/mol for OpsiD, MySiRNA, DSIR, iScore, Thermo Composition21, s-Biopredsi for Data Set 1 and Data Set 2 is shown in Table 8.8. The values show that our model performs better.

With this whole stacking energy, the AUC is also improved further and reached 0.878 for Data Set1 and 0.906 for Data Set2. This improvement in Pearson correlation values and AUC values show the importance of the influence of whole stacking energy on inhibition efficiency of siRNA. A comparison of improvement in Pearson Correlation, R and AUC at threshold of $\Delta G >= -34.6$ kcal/mol of OpsiD is shown in Fig 8.29.

Table 8.8: Comparative analysis of Pearson Correlation Coefficient at whole stacking energy, ΔG >= -34.6 kcal/mol for OpsiD, MySiRNA, DSIR, iScore, ThermoComposition21 and s-Biopredsi for Data Set 1 and Data Set 2

siRNA Design Approaches	**Pearson Correlation**		**Pearson Correlation when ΔG>= -34.6 kcal/mol**	
	Data Set1	**Data Set2**	**Data Set1**	**Data Set 2**
OpsiD	0.699	0.606	0.693	0.741
MysiRNA	0.686	0.599	0.668	0.737
DSIR	0.687	0.554	0.659	0.734
i-Score	0.635	0.556	0.607	0.723
ThermoComposition21	0.635	0.577	0.544	0.678
s-Biopredsi	0.665	0.546	0.622	0.724

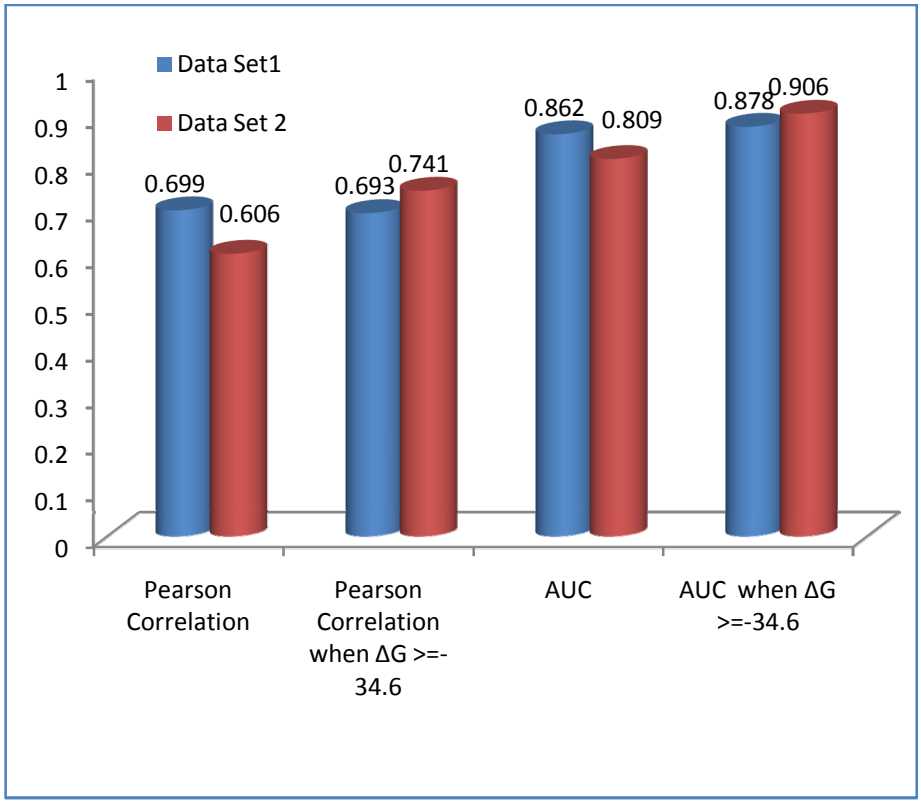


Fig 8.29: Effect of ΔG in Performance of OpsiD

8.4.4 Discussion

The complete results obtained by various validation strategies of OpsiD are summarized in Table 8.9. Using this model we are able to predict the percentage of inhibition efficiency of each predicted siRNA against a target mRNA or cDNA sequence. The performance analysis and comparison of 5-1-1 ANN (OpsiD) with selected good scoring second generation models are done. The improvement in prediction accuracy in terms of Pearson correlation coefficient shows

better performance of our model with previous good scoring siRNA design models. We tried to further optimize the inhibition efficiency in terms of sensitivity, specificity, accuracy of prediction and so on. When we compared these results with existing approaches, it is found that OpsiD achieves better performance. Thus we are able to optimize the efficacy of predicted siRNA in terms of inhibition efficiency, sensitivity, specificity and accuracy of prediction. Similarly using this model, we are able to address the problem of "off-target possibility on non-target genes" by providing the BLAST search.

Thus OpsiD provides the chance of identifying optimized siRNA with high inhibition capacity on target genes and low off-target effect on non-target genes. Also the effect of whole stacking energy (ΔG) on inhibition efficiency, by calculating the Pearson correlation coefficient at various threshold values of ΔG is noticed. The result shows an excellent improvement in Pearson correlation at ΔG >= -34.6 kcal/mol. From this, it is understood that exclusion of siRNAs with certain whole stacking energy is necessary to improve the inhibition efficiency. This reveals the importance and the influence of whole stacking energy on inhibition efficiency of siRNA.

Table 8.9: Performance of OpsiD

Validation Parameters	**Results of OpsiD**
Accuracy	0.76
Sensitivity	0.69
Specificity	0.83
MCC	0.52
Pearson Correlation for Data Set 1	0.699
Pearson Correlation for Data Set 1 when ΔG>= -34.6 kcal/mol	0.693
Pearson Correlation for Data Set 2	0.606
Pearson Correlation for Data Set 2 when ΔG>= -34.6 kcal/mol	0.741
Area Under Curve for Data Set 1	0.862
Area Under Curve for Data Set 1 when ΔG>= -34.6 kcal/mol	0.878
Area Under Curve for Data Set 2	0.809
Area Under Curve for Data Set 2 when ΔG>= -34.6 kcal/mol	0.906

8.5 Summary

The main focus of the thesis is to identify effective siRNA sequences with good inhibition efficiency and to optimize the efficiency of predicted siRNA by various efficacy parameters like sensitivity-specificity, accuracy of prediction and target specificity. We designed one SVM model and two ANN models in this study. The result and discussion of each model is presented in this chapter. From the results it is clear that one ANN model, OpsiD, performs well in terms of inhibition efficiency of siRNA against a particular target gene. Also OpsiD model is able to optimize the prediction efficacy of siRNA in terms of inhibition efficiency, sensitivity, specificity, accuracy of prediction and off-target possibility. Thus OpsiD provides the chance of identifying optimized siRNA with high inhibition capacity on target genes and low off-target effect on non-target genes. Thus the model achieves all the goals and objectives of our study.

……..✷♦✷……..

Chapter - 9
Conclusion and Future Scope

Contents

Gene silencing is an important research topic in functional genomics, biomedical research and in cancer therapeutics because of its ability to do sequence specific gene knock-down. Gene silencing is initiated by RNA interference mechanism and mediated by siRNA. siRNAs are new class of therapeutic agents which are suited for molecularly targeted gene silencing. The siRNA can be endogenous or exogenous. The use of exogenous siRNA for performing gene silencing has become an important biological milestone for mRNA target identification and drug design in various diseases, especially in cancer and AIDS. Therefore, identification of efficient siRNA capable of degrading target mRNA responsible for disease causing environment, is a key step towards the diagnosis and treatment of many serious diseases.

A significant amount of work has been undertaken over the recent past to understand the gene silencing mediated by siRNA. Many models have been proposed to predict efficient siRNAs against target mRNA. But there are many issues to be meaningfully

addressed while designing siRNA for therapeutic use. From the siRNA related studies, it is understood that among all siRNAs that can be generated against a target mRNA, only a few are found successful in causing degradation. However even those few do not perform equal knock-down effects. Also, it was earlier understood that full complementary siRNA was needed to silence a target gene. But recent studies reveal that siRNA behaves like microRNA and can suppress protein synthesis even though it is not fully complementary to the target. This shows that mismatches are allowed during target selection by siRNA. The mismatches occurring during target selection by siRNA may cause a very serious problem of "*off-target effect*" where unintended genes may be suppressed by the selected siRNA. Thus while designing exogenous siRNA therapeutically, all these issues must also be taken into consideration. Even though several algorithms and methods have been proposed to predict the efficiency of siRNA, only some of them have achieved acceptable level of efficacy.

9.1 Summary of Work

The main focus of this thesis is to develop methods to optimize the efficiency of siRNA in terms of "inhibition capacity and off-target possibility" against target mRNAs with improved sensitivity and specificity, which may be useful in the area of gene silencing and drug design for tumor development. This study aims to investigate the currently available siRNA prediction approaches and to devise a better computational approach to tackle the problem of

siRNA efficacy by inhibition capacity and off-target possibility. The strength and limitations of the available approaches are investigated and taken into account for making improved solution. Thus the approaches proposed in this study extend some of the good scoring previous state of the art techniques by incorporating machine learning and statistical approaches and thermodynamic features like whole stacking energy to improve the prediction accuracy, inhibition efficiency, sensitivity and specificity. In this thesis, we present three machine learning approaches (one SVM model and two ANN models) that enable to identify the efficiency of siRNA against target genes.

The first objective of our study is to design efficient siRNAs for any target mRNAs or cDNAs i.e. whether an siRNA is able to silence a target gene. As the first step of our study, we have selected Support Vector Machine model, to start predicting efficiency of siRNA against target mRNA or cDNA sequences. Using this model, we are able to classify a given siRNA as efficient or inefficient against a target mRNA sequence. The predicted siRNAs are analyzed and verified with existing siRNA design approaches. By carefully filtering the results, we are also able to notice the influence of thermodynamic properties like whole stacking energy (ΔG) and melting temperature of siRNA on inhibition efficiency. So we have included whole stacking energy of siRNA as one of the input parameters of our next ANN models.

The first ANN model, named siRNA Designer, is meant to achieve the second objective of predicting siRNA inhibition efficiency for a given target mRNA sequence. In this work, a 6-8-8-8-1 ANN model is designed to predict siRNA inhibition activity which is built on five previous second generation models BIOPREDsi [18], DSIR [19], ThermoComposition21 [20], i-Score [21] and MysiRNA [29] along with whole stacking energy (ΔG). It has been found that this model generates better performance than the existing state of the art techniques in terms of inhibition efficiency of predicted siRNA. Thus by 6-8-8-8-1 ANN, we are able to achieve second objective of our study, i.e., predicting the percentage of inhibition efficiency of each predicted siRNA against a target mRNA or cDNA sequence. Using this approach, one can select efficient siRNAs of user defined inhibition cut-off (normally cut-off will be 70%-80%) depending on the amount of silencing needed. But this model could not optimize the inhibition efficiency by sensitivity, specificity and accuracy of prediction.

The second ANN model is named as Optimized siRNA Designer, OpsiD, which is a 5-12-1 ANN. Using this we are able to achieve all the goals and objectives of this study, i.e., optimizing the prediction efficiency in terms of inhibition capacity, sensitivity, specificity, accuracy of prediction over the state of art techniques, with combined approach of "inhibition efficiency and off-target possibility". For finalizing the second ANN model, we have analyzed currently available best scoring models and developed a neural

network model by combining the results of selected good scoring previous models to improve the prediction accuracy. This ANN model is named Optimized siRNA Designer, OpsiD. It is built on four previous good scoring second generation models: DSIR [19], ThermoComposition21 [20], i-Score [21] and MysiRNA [29] and whole stacking energy (ΔG). The Encog machine learning frame work for Java is used to create, train and test the model and later integrated into OpsiD.

The models are trained and tested with large data sets. Pearson correlation coefficient and AUC value of ROC analysis are calculated to find the accuracy and performance of the model respectively. We achieve a Pearson Correlation Coefficient of R= 0.699 for Data Set 1 and 0.606 for Data Set 2. The AUC value for Data Set1 is 0.862 and for Data Set 2 is 0.809. Both Pearson Correlation values and AUC values are better than those of the state of the art techniques. Performance of the model is also tested with sensitivity, specificity and accuracy of prediction and found better than that of the state of the art techniques. These results show that our predicted inhibition is closer to the originally available experimental inhibition values.

The fifth input metric in our model is the whole stacking energy (ΔG) of siRNA strand, one of the important thermodynamic and stability factors of siRNA. We have analyzed the results to find its influence on performance. The inclusion of ΔG in the model

results in a performance of Pearson correlation coefficient R = 0.693 and AUC = 0.878 for Data Set 1 and R = 0.741 and AUC = 0.906 for Data Ste 2, at a specific threshold value of ΔG >= -34.6 kcal/mol. Except for R value of Data Set 1, all other values show improvement (Even though R value of 0.693 for Data Set1 is less compared to our previous result for Data Set1 (R=0.699), it is still better than the R values obtained for DSIR [19], ThermoComposition21 [20], i-Score [21] and MysiRNA [29]). These results show an excellent improvement in Pearson correlation and AUC at ΔG >= -34.6 kcal/mol. From this, it is understood that exclusion of siRNAs with certain whole stacking energy is necessary to improve the inhibition efficiency. This reveals the importance and the influence of whole stacking energy on inhibition efficiency of siRNA.

From the observations of various validations conducted in our approach, it is found that our model OpsiD, is capable of predicting the inhibition capacity of siRNA against a target mRNA with improved correlation, accuracy of prediction, MCC value, sensitivity, AUC value than those in the existing models. In addition, we have also included modules to reduce the consequence of 'off target' effect by providing facility to run BLAST search of each output siRNA sequences generated against any gene sequences in standard databases such as the NCBI RefSeq. Thus the proposed artificial neural network model 'OpsiD' can predict inhibition efficiency of a particular siRNA over a targeted mRNA sequence and can identify the similarity score of that siRNA with other genes in the database.

Using this approach, instead of selecting siRNA with the best inhibition capacity, we can consider both "inhibition efficiency and number of matches of BLAST score" to select siRNAs for gene silencing. The use of siRNAs with high BLAST score may lead to off-target effect and the user must make the trade-off between the "goodness" of siRNA with respect to inhibition capacity, and its similarity to other mRNA fragments. With this method siRNAs can be selected by carefully examining the inhibition efficiency and off target possibility. So we believe that the users will be able to eliminate those siRNA sequences with high BLAST score, even though they possess very high inhibition capacity. So we can take care of the risk of "off-target effect with unintended genes". The appraoch is available at http://opsid.in/opsid/.

This thesis introduces OpsiD, an artificial neural network model, to optimize the siRNA inhibition efficicency, built on four previous models (DSIR [19], ThermoComposition21 [20], i-Score [21] and MysiRNA [29]) and whole stacking energy (ΔG). Efforts are taken to combine different machine learning methods to improve the prediction efficiency compared to previously designed approaches. Thus using OpsiD, we are able to identify efficient siRNAs capable of performing post-transcriptional gene silencing with minimum off-target silencing and hence we achieved all the goals and objectives of our study. The proposed soft computing model may be found useful in finding exogenous siRNAs capable of effectively degrading the disease causing target mRNA and may help in drug design for cancer

treatment and other areas of bioinformatics by 'gene silencing'. In conclusion, OpsiD can design high quality siRNA that leads to gene knockdown with lower risk of "off target effect". This may be found useful in many areas of bioinformatics while designing siRNA for therapeutic and gene silencing applications.

9.2 Limitations

The computational approach for optimizing the efficiency of predicted siRNAs presented in this thesis may lead to great promise in the area of gene silencing by RNA interference. But, there are certain limitations to the method which should be acknowledged and addressed at future research directions.

Even though we have undertaken the performance evaluation of the approach in terms of certain static validation strategies, we are not able to evaluate the results biologically. This is because as far as the gene silencing by RNAi technique is concerned, the pre-clinical trials are still continuing for critical diseases like cancer. Since the research is in the pre-clinical stage, we are not able to get the actual successful gene samples for effective biological validations. In future, if the accurate and successful biological data of gene samples of particular diseases are available, the model can be tested for more accuracy and modified accordingly to assist in the development of disease treatment very effectively.

Currently, this study is aimed to identify the efficient siRNA capable of doing post-transcriptional gene silencing in mammalian cells. In this study, the machine learning method used to implement the optimized approach is artificial neural network. For accomplishing more excellent results and to identify more effective siRNAs, other concepts of machine learning may also be used. With mechanisms like Hidden Markov Models, it can be extended to identify effective exogenous siRNA which may be more accurate and able to silence various disease-associated target sites. By making more accurate predictions, the model may assist excellent disease treatment through post-transcriptional gene silencing.

9.3 Future Scope

RNAi has been successfully used to target many serious diseases like cancer on mice, with the hope of extending these approaches to treat humans. Cancer treatment will be successful if it is able to do complete removal of the tumor without making damage to any other parts of the body. This shall be achieved by doing surgery, to a certain level. But surgery is not as effective if the disease has already spread to other locations of the body. Chemotherapy is sometimes toxic to healthy tissues as it is not specific to cancer cells. Radiation also damage normal cells and tissues. By considering all these limitations of the existing cancer therapy techniques, it is very essential to develop novel target specific therapeutics for the effective treatment of cancer. The idea used in cancer treatment with RNAi is that cancer cells will be killed

through the actions of the patient's own immune system. Nowadays there are lot of insights and promises for using siRNAs as drugs targeted only into the cancer cells. Delivery of such efficient siRNAs may provide new insights into in future therapy in cancer detection and drug design.

Gene specific silencing has allowed systematic approach of designing new drugs, and for enhancing the effect of already existing drugs. RNAi could enable gene silencing with high specificity and improved efficiency than with any other techniques. In principle any gene may be knocked-down by a synthetic siRNA with exact complementary sequence. There by, any disease caused by abnormally enhanced activity of one or more genes may, in theory be regulated by RNAi-based therapies. Many of the siRNA therapies are at preclinical stage. The methods for delivering siRNA drugs should concentrate on maximizing the specificity of siRNA and minimizing the toxicity and degradation effects that compromise drug efficacy. Thus RNAi has great potential in future gene therapy applications since it has the potential to regulate disease related genes. Hence in the post-genomic era, siRNAs is considered as an important tool for validating gene function and drug targeting.

……..✷♦✷……..

References

[1]. Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E., & Mello, C. C. (1998). Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. nature, 391(6669), 806-811.

[2]. Elbashir, S. M., Harborth, J., Lendeckel, W., Yalcin, A., Weber, K., & Tuschl, T. (2001). Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. Nature, 411(6836), 494-498.

[3]. Martínez, M. A., Gutiérrez, A., Armand-Ugón, M., Blanco, J., Parera, M., Gómez, J., Clotet, B., & Esté, J. A. (2002). Suppression of chemokine receptor expression by RNA interference allows for inhibition of HIV-1 replication. Aids, 16(18), 2385-2390.

[4]. Xia, H., Mao, Q., Eliason, S. L., Harper, S. Q., Martins, I. H., Orr, H. T., Paulson, H. L., Yang, L., Kotin, R. M., & Davidson, B. L. (2004). RNAi suppresses polyglutamine-induced neurodegeneration in a model of spinocerebellar ataxia. Nature medicine, 10(8), 816-820.

[5]. Soutschek, J., Akinc, A., Bramlage, B., Charisse, K., Constien, R., Donoghue, M., El-bashir, S., Geick, A., Hadwiger, P., Harborth, J., John, M., Kesavan, V., Lavine, G., Pandey, R.K., Racie, T., Rajeev, K.G., Rohl, I., Toudjarska, I., Wang, G., Wuschko, S., Bumcrot, D., Koteliansky, V., Limmer, S., Manoharan, M., & Vornlocher, H. P. (2004). Therapeutic silencing of an endogenous gene by systemic administration of modified siRNAs. Nature, 432(7014), 173-178.

[6]. Borkhardt, A. (2002). Blocking oncogenes in malignant cells by RNA interference - New hope for a highly specific cancer treatment?. Cancer cell, 2(3), 167-168.

[7]. Dykxhoorn, D. M., & Lieberman, J. (2006). Running interference: prospects and obstacles to using small interfering RNAs as small molecule drugs. Annu. Rev. Biomed. Eng., 8, 377-402.

[8]. Dykxhoorn, D. M., Palliser, D., & Lieberman, J. (2006). The silent treatment: siRNAs as small molecule drugs. Gene therapy, 13(6), 541-552.

[9]. McManus, M. T., & Sharp, P. A. (2002). Gene silencing in mammals by small interfering RNAs. Nature reviews genetics, 3(10), 737-747.

[10]. Meister, G., & Tuschl, T. (2004). Mechanisms of gene silencing by double-stranded RNA. Nature, 431(7006), 343-349.

[11]. Hannon, G. J., & Rossi, J. J. (2004). Unlocking the potential of the human genome with RNA interference. Nature, 431(7006), 371-378.

[12]. Holen, T., Amarzguioui, M., Wiiger, M. T., Babaie, E., & Prydz, H. (2002). Positional effects of short interfering RNAs targeting the human coagulation trigger Tissue Factor. Nucleic acids research, 30(8), 1757-1766.

[13]. Ameres, S. L., Martinez, J., & Schroeder, R. (2007). Molecular basis for target RNA recognition and cleavage by human RISC. Cell, 130(1), 101-112.

[14]. Doench, J. G., Petersen, C. P., & Sharp, P. A. (2003). siRNAs can function as miRNAs. Genes & development, 17(4), 438-442.

[15]. Burchard, J., Jackson, A. L., Malkov, V., Needham, R. H., Tan, Y., Bartz, S. R., & Linsley, P. S. (2009). MicroRNA-like off-target transcript regulation by siRNAs is species specific. Rna, 15(2), 308-315.

[16]. Jackson, A. L., Bartz, S. R., Schelter, J., Kobayashi, S. V., Burchard, J., Mao, M., Li, B., Cavet, G., & Linsley, P. S. (2003). Expression profiling reveals off-target gene regulation by RNAi. Nature biotechnology, 21(6), 635-637.

[17]. Jackson, A. L., Burchard, J., Schelter, J., Chau, B. N., Cleary, M., Lim, L., & Linsley, P. S. (2006). Widespread siRNA "off-target" transcript silencing mediated by seed region sequence complementarity. Rna, 12(7), 1179-1187.

[18]. Huesken, D., Lange, J., Mickanin, C., Weiler, J., Asselbergs, F., Warner, J., Hall, J. (2005). Design of a genome-wide siRNA library using an artificial neural network. Nature biotechnology, 23(8), 995-1001.

[19]. Vert, J. P., Foveau, N., Lajaunie, C., & Vandenbrouck, Y. (2006). An accurate and interpretable model for siRNA efficacy prediction. BMC bioinformatics, 7(1), 520.

[20]. Shabalina, S. A., Spiridonov, A. N., & Ogurtsov, A. Y. (2006). Computational models with thermodynamic and composition features improve siRNA design. BMC bioinformatics, 7(1), 65.

[21]. Ichihara, M., Murakumo, Y., Masuda, A., Matsuura, T., Asai, N., Jijiwa, M., Ishida, M., Ishida, M.M., Shinmi, J., Yatsuya, H., Qiao, S., Takahashi, M., & Ohno, K. (2007). Thermodynamic instability of siRNA duplex is a prerequisite for dependable prediction of siRNA activities. Nucleic acids research, 35(18), e123.

[22]. Matveeva, O., Nechipurenko, Y., Rossi, L., Moore, B., Sætrom, P., Ogurtsov, A. Y., Atkins, J.F., & Shabalina, S. A.

(2007). Comparison of approaches for rational siRNA design leading to a new efficient and transparent method. Nucleic acids research, 35(8), e63.

[23]. Ladunga, I. (2007). More complete gene silencing by fewer siRNAs: transparent optimized design and biophysical signature. Nucleic acids research, 35(2), 433-440.

[24]. Gong, W., Ren, Y., Zhou, H., Wang, Y., Kang, S., & Li, T. (2008). siDRM: an effective and generally applicable online siRNA design tool. Bioinformatics, 24(20), 2405-2406.

[25]. Tafer, H., Ameres, S. L., Obernosterer, G., Gebeshuber, C. A., Schroeder, R., Martinez, J., & Hofacker, I. L. (2008). The impact of target site accessibility on the design of effective siRNAs. Nature biotechnology, 26(5), 578-583.

[26]. Ren, Y., Gong, W., Zhou, H., Wang, Y., Xiao, F., & Li, T. (2009). siRecords: a database of mammalian RNAi experiments and efficacies. Nucleic acids research, 37(suppl 1), D146-D149.

[27]. Horn, T., & Boutros, M. (2010). E-RNAi: a web application for the multi-species design of RNAi reagents—2010 update. Nucleic acids research. 2010 Jul;38(Web Server issue):W332-9.

[28]. Mysara, M., Garibaldi, J. M., & ElHefnawi, M. (2011). MysiRNA-designer: a workflow for efficient siRNA design. PLoS One, 6(10), e25642.

[29]. Mysara, M., Elhefnawi, M., & Garibaldi, J. M. (2012). MysiRNA: Improving siRNA efficacy prediction using a machine-learning model combining multi-tools and whole stacking energy (ΔG). Journal of biomedical informatics, 45(3), 528-534.

[30]. Filhol, O., Ciais, D., Lajaunie, C., Charbonnier, P., Foveau, N., Vert, J. P., & Vandenbrouck, Y. (2012). DSIR: assessing the design of highly potent siRNA by testing a set of cancer-relevant target genes. PLoS ONE 7(10): e48057.

[31]. Mazur, S., Csucs, G., & Kozak, K. (2012). RNAiAtlas: a database for RNAi (siRNA) libraries and their specificity. Database, 2012, bas027.

[32]. Boudreau, R. L., Spengler, R. M., Hylock, R. H., Kusenda, B. J., Davis, H. A., Eichmann, D. A., & Davidson, B. L. (2013). siSPOTR: a tool for designing highly specific and potent siRNAs for human and mouse. Nucleic acids research, 41(1), e9-e9.

[33]. Weiss, S.M., & Kapouleas, I. (1989) An empirical comparison of pattern recognition, neural nets, and machine learning classification methods. In: In Proceedings of the Eleventh International Joint Conference on Artificial Intelligence, Morgan Kaufmann, pp 781–787

[34]. Xu, L., Krzyzak, A., & Oja, E. (1991). Neural nets for dual subspace pattern recognition method. International Journal of Neural Systems, 2(03), 169-184.

[35]. Oja, E. (1989). Neural networks, principal components, and subspaces. International journal of neural systems, 1(01), 61-68.

[36]. Alberts, B., Johnson, A., Lewis, J., Raff, M., & Roberts, K. (2002). Molecular biology of the cell, 4 ed. New York, NY: Garland Science.

[37]. Ridley, M. (2006). Genome. New York, NY: Harper Perennial.

[38]. Watson, J. D., & Crick, F. H. C. (1953). A structure for deoxyribose nucleic acid. Nature, 421(6921), 397-3988.

[39]. Saenger, W. (1984). Principles of Nucleic Acid Structure; Springer: Tokyo.

[40]. Alberts, B., Bray, D., Hopkin, K., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2009). Essential cell biology. 3rd ed., Garland Science, New York.

[41]. Butler JM. Forensic DNA typing: biology, technology, and genetics of STR markers. 2nd ed. Elsevier: New York, 2005.

[42]. Ghosh, A., & Bansal, M. (2003). A glossary of DNA structures from A to Z.Biological Crystallography, 59(4), 620-626.

[43]. Blevins, T., Rajeswaran, R., Shivaprasad, P. V., Beknazariants, D., Si-Ammour, A., Park, H. S., Vazquez, F., Robertson, D., & Pooggin, M. M. (2006). Four plant Dicers mediate viral small RNA biogenesis and DNA virus induced silencing. Nucleic acids research, 34(21), 6233-6246.

[44]. Jana, S., Chakraborty, C., Nandi, S., & Deb, J. K. (2004). RNA interference: potential therapeutic targets. Applied microbiology and biotechnology, 65(6), 649-657.

[45]. Schultz, U., Kaspers, B., & Staeheli, P. (2004). The interferon system of non-mammalian vertebrates. Developmental & Comparative Immunology, 28(5), 499-508.

[46]. Whitehead, K. A., Dahlman, J. E., Langer, R. S., & Anderson, D. G. (2011). Silencing or stimulation? siRNA delivery and the immune system. Annual review of chemical and biomolecular engineering, 2, 77-96.

[47]. Mattick, J. S., & Gagen, M. J. (2001). The evolution of controlled multitasked gene networks: the role of introns and other noncoding RNAs in the development of complex organisms. Molecular Biology and Evolution, 18(9), 1611-1630.

[48]. Mattick, J. S. (2001). Non-coding RNAs: the architects of eukaryotic complexity. EMBO reports, 2(11), 986-991.

[49]. Mattick, J. S. (2003). Challenging the dogma: the hidden layer of non protein coding RNAs in complex organisms. Bioessays, 25(10), 930-939.

[50]. Mattick, J. S. (2004). The hidden genetic program of complex organisms. Scientific American, 291(4), 60-67.

[51]. Crick, F. H. (1958). On protein synthesis. In Symposia of the Society for Experimental Biology (Vol. 12, p. 138).

[52]. Crick, F. (1970). Central dogma of molecular biology. Nature, 227(5258), 561-563.

[53]. Leavitt, S. A. (2010). Deciphering the Genetic Code: Marshall Nirenberg. Office of NIH History. Available from: http://history. nih.gov/exhibits/nirenberg/, Retrieved 10 June 2014.

[54]. Redberry, G.W. (2006). Gene silencing: new research. Science Publishers: New York.

[55]. NCBI. (2013). Gene Silencing. National Center for Biotechnology Information. Available from: http://www.ncbi.nlm.nih.gov/genome/probe/doc/ApplSilencing.shtml, Retrieved11 November 2013.

[56]. Shrivastava, N., & Srivastava, A. (2008). RNA interference: an emerging generation of biologicals. Biotechnology journal, 3(3), 339-353.

[57]. Rácz, Z., & Hamar, P. (2006). Can siRNA technology provide the tools for gene therapy of the future?. Current medicinal chemistry, 13(19), 2299-2307.

[58]. Gewirtz, A. M. (2007). On future's doorstep: RNA interference and the pharmacopeia of tomorrow. The Journal of clinical investigation, 117(12), 3612.

[59]. Smith, C.J.S., Watson, C.F., Ray, J., Bird, C.R., Morris, P.C., Schuch, W., and Grierson, D. (1988). Antisense RNA inhibition of polygalacturonase gene expression in transgenic tomatoes. Nature 334, 724-726.

[60]. Tuschl, T., Zamore, P. D., Lehmann, R., Bartel, D. P., & Sharp, P. A. (1999). Targeted mRNA degradation by double-stranded RNA in vitro. Genes & development, 13(24), 3191-3197.

[61]. Zamore, P. D., Tuschl, T., Sharp, P. A., & Bartel, D. P. (2000). RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals. Cell, 101(1), 25-33.

[62]. Elbashir, S. M., Lendeckel, W., & Tuschl, T. (2001). RNA interference is mediated by 21-and 22-nucleotide RNAs. Genes & development, 15(2), 188-200.

[63]. Taylor, P. R., Gordon, S., & Martinez-Pomares, L. (2005). The mannose receptor: linking homeostasis and immunity through sugar recognition. Trends in immunology, 26(2), 104-110.

[64]. Mao, C. P., Lin, Y. Y., Hung, C. F., & Wu, T. C. (2007). Immunological research using RNA interference technology. Immunology, 121(3), 295-307.

[65]. Tiemann, K., & Rossi, J. J. (2009). RNAi based therapeutics–current status, challenges and prospects. EMBO molecular medicine, 1(3), 142-151.

[66]. Ritprajak, P., Hashiguchi, M., & Azuma, M. (2008). Topical application of cream-emulsified CD86 siRNA ameliorates

allergic skin disease by targeting cutaneous dendritic cells. Molecular Therapy, 16(7), 1323-1330.

[67]. Hickerson, R. P., Smith, F. J., Reeves, R. E., Contag, C. H., Leake, D., Leachman, S. A., Milstone, L.M., McLean, W.H.I., & Kaspar, R. L. (2008). Single-nucleotide-specific siRNA targeting in a dominant-negative skin model. Journal of Investigative Dermatology, 128(3), 594-605.

[68]. Kehren, J., Desvignes, C., Krasteva, M., Ducluzeau, M. T., Assossou, O., Horand, F., Hahne, M., Kagi, D., Kaiserlian, D., & Nicolas, J. F. (1999). Cytotoxicity Is Mandatory for CD8+ T Cell–mediated Contact Hypersensitivity. The Journal of experimental medicine,189(5), 779-786.

[69]. Gonzalez-Gonzalez, E., Ra, H., Hickerson, R. P., Wang, Q., Piyawattanametha, W., Mandella, M. J., Kino, G.S., Leake, D., Avilion, A.A., Solgaard, O., Doyle, T.C., Contag, C.H., & Kaspar, R. L. (2009). siRNA silencing of keratinocyte-specific GFP expression in a transgenic mouse skin model. Gene therapy, 16(8), 963-972.

[70]. Harding, C. R. (2004). The stratum corneum: structure and function in health and disease. Dermatologic therapy, 17(s1), 6-15.

[71]. Hengge, U. R., Walker, P. S., & Vogel, J. C. (1996). Expression of naked DNA in human, pig, and mouse skin. Journal of Clinical Investigation, 97(12), 2911.

[72]. Wyatt, L. S., Belyakov, I. M., Earl, P. L., Berzofsky, J. A., & Moss, B. (2008). Enhanced cell surface expression, immunogenicity and genetic stability resulting from a spontaneous truncation of HIV Env expressed by a recombinant MVA. Virology, 372(2), 260-272.

[73]. Rissmann, R., Oudshoorn, M. H., Hennink, W. E., Ponec, M., & Bouwstra, J. A. (2009). Skin barrier disruption by acetone:

observations in a hairless mouse skin model. Archives of dermatological research, 301(8), 609-613.

[74]. Fluhr, J. W., Darlenski, R., & Surber, C. (2008). Glycerol and the skin: holistic approach to its origin and functions. British Journal of Dermatology, 159(1), 23-34.

[75]. Uribe, S., & Sampedro, J. G. (2003). Measuring solution viscosity and its effect on enzyme activity. Biological procedures online, 5(1), 108-115.

[76]. Jorgensen, R. A., Cluster, P. D., English, J., Que, Q., & Napoli, C. A. (1996). Chalcone synthase cosuppression phenotypes in petunia flowers: comparison of sense vs. antisense constructs and single-copy vs. complex T-DNA sequences. Plant molecular biology, 31(5), 957-973.

[77]. Romano, N., & Macino, G. (1992). Quelling: transient inactivation of gene expression in Neurospora crassa by transformation with homologous sequences. Molecular microbiology, 6(22), 3343-3353.

[78]. Cogoni, C., & Macino, G. (2000). Post-transcriptional gene silencing across kingdoms. Current opinion in genetics & development, 10(6), 638-643.

[79]. Ratcliff, F., Harrison, B. D., & Baulcombe, D. C. (1997). A similarity between viral defense and gene silencing in plants. Science, 276(5318), 1558-1560.

[80]. Burnett, J. C., Rossi, J. J., & Tiemann, K. (2011). Current progress of siRNA/shRNA therapeutics in clinical trials. Biotechnology journal, 6(9), 1130-1146.

[81]. Amarzguioui, M., & Prydz, H. (2004). An algorithm for selection of functional siRNA sequences. Biochemical and biophysical research communications, 316(4), 1050-1058.

[82]. Tuschl, T. (2001). RNA interference and small interfering RNAs. Chembiochem,2(4), 239-245.

[83]. Martinez, J., Patkaniowska, A., Urlaub, H., Lührmann, R., & Tuschl, T. (2002). Single-stranded antisense siRNAs guide target RNA cleavage in RNAi. Cell, 110(5), 563-574.

[84]. Hamilton, A. J., & Baulcombe, D. C. (1999). A species of small antisense RNA in posttranscriptional gene silencing in plants. Science, 286(5441), 950-952.

[85]. Lee, R. C., Feinbaum, R. L., & Ambros, V. (1993). The *C. elegans* heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. cell, 75(5), 843-854.

[86]. Carthew, R. W. (2006). Gene regulation by microRNAs. Current opinion in genetics & development, 16(2), 203-208.

[87]. Lewis, B. P., Burge, C. B., & Bartel, D. P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. cell, 120(1), 15-20.

[88]. Friedman, R. C., Farh, K. K. H., Burge, C. B., & Bartel, D. P. (2009). Most mammalian mRNAs are conserved targets of micro RNAs. Genome research,19(1), 92-105.

[89]. Schickel, R., Boyerinas, B., Park, S. M., & Peter, M. E. (2008). MicroRNAs: key players in the immune system, differentiation, tumorigenesis and cell death. Oncogene, 27(45), 5959-5974.

[90]. Baek, D., Villén, J., Shin, C., Camargo, F. D., Gygi, S. P., & Bartel, D. P. (2008). The impact of microRNAs on protein output. Nature, 455(7209), 64-71.

[91]. Selbach, M., Schwanhäusser, B., Thierfelder, N., Fang, Z., Khanin, R., & Rajewsky, N. (2008). Widespread changes in protein synthesis induced by microRNAs. Nature, 455(7209), 58-63.

[92]. Zhang, B., Pan, X., Cobb, G. P., & Anderson, T. A. (2007). microRNAs as oncogenes and tumor suppressors. Developmental biology, 302(1), 1-12.

[93]. Paddison, P. J., Caudy, A. A., & Hannon, G. J. (2002). Stable suppression of gene expression by RNAi in mammalian cells. Proceedings of the National Academy of Sciences, 99(3), 1443-1448.

[94]. Couzin, J. (2002). Breakthrough of the year: small RNAs make big splash. Science, 298, 2296-2297.

[95]. Shi H., Djikeng A., Mark T., Wirtz E., Tschudi C., & Ullu E. (2000) Genetic interference in Trypanosoma brucei by heritable and inducible double-stranded RNA. RNA 6:1069–1076.

[96]. Pai, S. I., Lin, Y. Y., Macaes, B., Meneshian, A., Hung, C. F., & Wu, T. C. (2006). Prospects of RNA interference therapy for cancer. Gene therapy, 13(6), 464-477.

[97]. Chen, Y., Zhu, X., Zhang, X., Liu, B., & Huang, L. (2010). Nanoparticles modified with tumor-targeting scFv deliver siRNA and miRNA for cancer therapy. Molecular Therapy, 18(9), 1650-1656.

[98]. Wu, Y., Wang, W., Chen, Y., Huang, K., Shuai, X., Chen, Q., & Lian, G. (2010). The investigation of polymer-siRNA nanoparticle for gene therapy of gastric cancer in vitro. International journal of nanomedicine, 5, 129.

[99]. Zenke, K., Nam, Y. K., & Kim, K. H. (2010). Development of siRNA expression vector utilizing rock bream β-actin

promoter: a potential therapeutic tool against viral infection in fish. Applied microbiology and biotechnology, 85(3), 679-690.

[100]. Tang, Y., Li, Y. B., Wang, B., Lin, R. Y., van Dongen, M., Zurcher, D. M., & Qi, R. (2012). Efficient in vitro siRNA delivery and intramuscular gene silencing using PEG-modified PAMAM dendrimers. Molecular pharmaceutics, 9(6), 1812-1821.

[101]. Hood, E. (2004). RNAi: What's all the noise about gene silencing? Environmental Health Perspectives, 112(4), A224.

[102]. Dave, R. S., & Pomerantz, R. J. (2004). Antiviral effects of human immunodeficiency virus type 1-specific small interfering RNAs against targets conserved in select neurotropic viral strains. Journal of virology, 78(24), 13687-13696.

[103]. Wilson, J. A., Jayasena, S., Khvorova, A., Sabatinos, S., Rodrigue-Gervais, I. G., Arya, S., Sarangi, F., Harris-Brandts, M., Beaulieu, S., & Richardson, C. D. (2003). RNA interference blocks gene expression and RNA synthesis from hepatitis C replicons propagated in human liver cells. Proceedings of the National Academy of Sciences, 100(5), 2783-2788.

[104]. Chen, J., Wall, N. R., Kocher, K., Duclos, N., Fabbro, D., Neuberg, D., Griffin, J.D., Shi, Y., & Gilliland, D. G. (2004). Stable expression of small interfering RNA sensitizes TEL-PDGFβR to inhibition with imatinib or rapamycin. Journal of Clinical Investigation, 113(12), 1784.

[105]. Martinez, L. A., Naguibneva, I., Lehrmann, H., Vervisch, A., Tchénio, T., Lozano, G., & Harel-Bellan, A. (2002). Synthetic small inhibiting RNAs: efficient tools to inactivate oncogenic mutations and restore p53 pathways. Proceedings of the National Academy of Sciences, 99(23), 14849-14854.

[106]. Lapteva, N., Yang, A. G., Sanders, D. E., Strube, R. W., & Chen, S. Y. (2005). CXCR4 knockdown by small interfering RNA abrogates breast tumor growth in vivo. Cancer gene therapy, 12(1), 84-89.

[107]. July, L. V., Beraldi, E., So, A., Fazli, L., Evans, K., English, J. C., & Gleave, M. E. (2004). Nucleotide-based therapies targeting clusterin chemosensitize human lung adenocarcinoma cells both in vitro and in vivo. Molecular Cancer Therapeutics, 3(3), 223-232.

[108]. Ning, S., Fuessel, S., Kotzsch, M., Kraemer, K., Kappler, M., Schmidt, U., Taubert, H., Wirth, M.P., & Meye, A. (2004). siRNA-mediated down-regulation of survivin inhibits bladder cancer cell growth. International journal of oncology, 25(4), 1065-1136.

[109]. Qin, X. F., An, D. S., Chen, I. S., & Baltimore, D. (2003). Inhibiting HIV-1 infection in human T cells by lentiviral-mediated delivery of small interfering RNA against CCR5. Proceedings of the National Academy of Sciences, 100(1), 183-188.

[110]. Li, M. J., Bauer, G., Michienzi, A., Yee, J. K., Lee, N. S., Kim, J., Shirley, L., Castanotto, D., Zaia, J., & Rossi, J. J. (2003). Inhibition of HIV-1 infection by lentiviral vectors expressing Pol III-promoted anti-HIV RNAs. Molecular Therapy, 8(2), 196-206.

[111]. Giladi, H., Ketzinel-Gilad, M., Rivkin, L., Felig, Y., Nussbaum, O., & Galun, E. (2003). Small interfering RNA inhibits hepatitis B virus replication in mice.Molecular Therapy, 8(5), 769-776.

[112]. Randall, G., Grakoui, A., & Rice, C. M. (2003). Clearance of replicating hepatitis C virus replicon RNAs in cell culture by small interfering RNAs. Proceedings of the National Academy of Sciences, 100(1), 235-240.

[113]. Randall, G., & Rice, C. M. (2004). Interfering with hepatitis C virus RNA replication. Virus research, 102(1), 19-25.

[114]. Leung, R. K., & Whittaker, P. A. (2005). RNA interference: from gene silencing to gene-specific therapeutics. Pharmacology & therapeutics, 107(2), 222-239.

[115]. Popescu, F. D., & Popescu, F. (2007). A review of antisense therapeutic interventions for molecular biological targets in asthma. Biologics: targets & therapy, 1(3), 271.

[116]. HOPES. (2013). Gene Silencing. HOPES - Huntington's Outreach Project for Education, at Stanford, Stanford University. Available from: http://web.stanford.edu/ group/hopes/cgi-bin/hopes_test/,Retrieved 13December 2013.

[117]. Kim, D. H., & Rossi, J. J. (2008). RNAi mechanisms and applications. Biotechniques, 44(5), 613.

[118]. Davidson, B. L., & McCray, P. B. (2011). Current prospects for RNA interference-based therapies. Nature Reviews Genetics, 12(5), 329-340.

[119]. Rettig, G. R., & Behlke, M. A. (2012). Progress toward in vivo use of siRNAs-II. Molecular Therapy, 20(3), 483-512.

[120]. Behlke, M. A. (2006). Progress towards in vivo use of siRNAs. Molecular Therapy, 13(4), 644-670.

[121]. Guo, P., Coban, O., Snead, N. M., Trebley, J., Hoeprich, S., Guo, S., & Shu, Y. (2010). Engineering RNA for targeted siRNA delivery and medical application. Advanced drug delivery reviews, 62(6), 650-666.

[122]. Mantha, N., Das, S. K., & Das, N. G. (2012). RNAi-based therapies for Huntington's disease: delivery challenges and oppo rtunities. Therapeutic delivery, 3(9), 1061-1076.

[123]. Birmingham, A., Anderson, E. M., Reynolds, A., Ilsley-Tyree, D., Leake, D., Fedorov, Y., Baskervilla, S., Maksimova, E., Robinson, K., Karpilow, J., Marshall, W.S., & Khvorova, A. (2006). 3′ UTR seed matches, but not overall identity, are associated with RNAi off-targets. Nature methods, 3(3), 199-204.

[124]. Aagaard, L., & Rossi, J. J. (2007). RNAi therapeutics: principles, prospects and challenges. Advanced drug delivery reviews, 59(2), 75-86.

[125]. Wianny, F., & Zernicka-Goetz, M. (2000). Specific interference with gene function by double-stranded RNA in early mouse development. Nature cell biology, 2(2), 70-75.

[126]. Gil, J., & Esteban, M. (2000). Induction of apoptosis by the dsRNA-dependent protein kinase (PKR): mechanism of action. Apoptosis, 5(2), 107-114.

[127]. Brummelkamp, T. R., Bernards, R., & Agami, R. (2002). A system for stable expression of short interfering RNAs in mammalian cells. science, 296(5567), 550-553.

[128]. Lee, N. S., Dohjima, T., Bauer, G., Li, H., Li, M. J., Ehsani, A., & Rossi, J. (2002). Expression of small interfering RNAs targeted against HIV-1 rev transcripts in human cells. Nature biotechnology, 20(5), 500-505.

[129]. Scherr, M., Morgan, M. A., & Eder, M. (2003). Gene silencing mediated by small interfering RNAs in mammalian cells. Current medicinal chemistry, 10(3), 245-256.

[130]. Ryther, R. C. C., Flynt, A. S., Phillips, J. A., & Patton, J. G. (2005). siRNA therapeutics: big potential from small RNAs. Gene Therapy, 12(1), 5-11.

[131]. Heale, B. S., Soifer, H. S., Bowers, C., & Rossi, J. J. (2005). siRNA target site secondary structure predictions using local stable substructures. Nucleic acids research, 33(3), e30-e30.

[132]. Luo, K. Q., & Chang, D. C. (2004). The gene-silencing efficiency of siRNA is strongly dependent on the local structure of mRNA at the targeted region. Biochemical and biophysical research communications, 318(1), 303-310.

[133]. Reynolds, A., Leake, D., Boese, Q., Scaringe, S., Marshall, W. S., & Khvorova, A. (2004). Rational siRNA design for RNA interference. Nature biotechnology, 22(3), 326-330.

[134]. Ui-Tei, K., Naito, Y., Takahashi, F., Haraguchi, T., Ohki-Hamazaki, H., Juni, A., Ueda, R., & Saigo, K. (2004). Guidelines for the selection of highly effective siRNA sequences for mammalian and chick RNA interference. Nucleic Acids Research, 32(3), 936–948.

[135]. Chalk, A.M., Wahlestedt, C., & Sonnhammer, E.L.L. (2004). Improved and automated prediction of effective siRNA. Biochemical and Biophysical Research Communications, 319(1), 264–274.

[136]. Khvorova A., Reynolds A., and Jayasena S.D. (2003). Functional siRNAs and miRNAs Exhibit Strand Bias. Cell, 115, 209–216.

[137]. Takasaki, S., Kotani S., and Konagaya A. (2004). An Effective Method for Selecting siRNA Target Sequences in Mammalian Cells. Cell Cycle, 3, 790–795.

[138]. Hohjoh, H. (2004). Enhancement of RNAi activity by improved siRNA duplexes. FEBS Letters, 557, 193–198.

[139]. Hsieh A.C., Ronghai B., Manola J., Vazquez F., Bare O., Khvorova A., Scaringe S., and Sellers W.R. (2004). A library of siRNA duplexes targeting the phosphoinositide 3-kinase

pathway: determinants of gene silencing for use in cell-based screens. Nucleic Acids Research, 32:893–901.

[140]. Scholkopf, B., Guyon, I., & Weston, J. (2003). Statistical learning and kernel methods in bioinformatics. Nato Science Series Sub Series III Computer and Systems Sciences, 183, 1-21.

[141]. Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features (pp. 137-142). Springer Berlin Heidelberg.

[142]. Kumar, V. P., & Poggio, T. (2000). Learning-based approach to real time tracking and analysis of faces. In Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on (pp. 96-101). IEEE.

[143]. Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992 July). A training algorithm for optimal margin classifiers. In Proceedings of the fifth annual workshop on Computational learning theory (pp. 144-152). ACM.

[144]. Ding, Y., Chan, C. Y., & Lawrence, C. E. (2004). Sfold web server for statistical folding and rational design of nucleic acids. Nucleic acids research, 32(suppl 2), W135-W141.

[145]. Henschel, A., Buchholz, F., & Habermann, B. (2004). DEQOR: a web-based tool for the design and quality control of siRNAs. Nucleic acids research,32(suppl 2), W113-W120.

[146]. Naito, Y., Ui-Tei, K., Nishikawa, T., Takebe, Y., & Saigo, K. (2006). siVirus: web-based antiviral siRNA design software for highly divergent viral sequences.Nucleic acids research, 34(suppl 2), W448-W450.

[147]. Holen, T. (2006). Efficient prediction of siRNAs with siRNArules 1.0: an open-source JAVA approach to siRNA algorithms. Rna, 12(9), 1620-1625.

[148]. Lu, Z. J., & Mathews, D. H. (2008). OligoWalk: an online siRNA design tool utilizing hybridization thermodynamics. Nucleic acids research, 36(suppl 2), W104-W108.

[149]. Park, Y. K., Park, S. M., Choi, Y. C., Lee, D., Won, M., & Kim, Y. J. (2008). AsiDesigner: exon-based siRNA design server considering alternative splicing. Nucleic acids research, 36(suppl 2), W97-W103.

[150]. Thang, B. N., Ho, T. B., & Kanda, T. (2015). A semi–supervised tensor regression model for siRNA efficacy prediction. BMC bioinformatics, 16(1), 80.

[151]. Hamada, M., Ohtsuka, T., Kawaida, R., Koizumi, M., Morita, K., Furukawa, H., Imanishi, T., Miyagishi, M., & Taira, K. (2002). Effects on RNA interference in gene expression (RNAi) in cultured mammalian cells of mismatches and the introduction of chemical modifications at the 3'-ends of siRNAs. Antisense and Nucleic Acid Drug Development, 12(5), 301-309.

[152]. Jain, C. K., & Prasad, Y. (2009). Feature selection for siRNA efficacy prediction using natural computation. In Nature & Biologically Inspired Computing, 2009. NaBIC 2009. World Congress on (pp. 1759-1764). IEEE Press.

[153]. Wang, X., Wang, X., Varma, R. K., Beauchamp, L., Magdaleno, S., & Sendera, T. J. (2009). Selection of hyperfunctional siRNAs with improved potency and specificity. Nucleic acids research, 37(22), e152.

[154]. Patzel, V. (2007). In silico selection of active siRNA. Drug discovery today,12(3), 139-148.

[155]. Panjkovich, A., & Melo, F. (2005). Comparison of different melting temperature calculation methods for short DNA sequences. Bioinformatics, 21(6), 711-722.

[156]. Sugimoto, N., Nakano, S. I., Katoh, M., Matsumura, A., Nakamuta, H., Ohmichi, T., Yoneyama, M., & Sasaki, M. (1995). Thermodynamic parameters to predict stability of RNA/DNA hybrid duplexes. Biochemistry, 34(35), 11211-11216.

[157]. Vickers, T. A., Koo, S., Bennett, C. F., Crooke, S. T., Dean, N. M., & Baker, B. F. (2003). Efficient reduction of target RNAs by small interfering RNA and RNase H-dependent antisense agents A comparative analysis. Journal of Biological Chemistry, 278(9), 7108-7118.

[158]. Chang, C. C., & Lin, C. J. (2011). LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST),2(3), 27.

[159]. Neuroph 2.92 [software]. (2013). Retrieved from http://neuroph.sourceforge.net

[160]. Encog Machine Learning Framework [software]. (2013). Retrieved from http://www.heatonresearch.com/encog

[161]. Harborth, J., Elbashir, S. M., Vandenburgh, K., Manninga, H., Scaringe, S. A., Weber, K., & Tuschl, T. (2003). Sequence, chemical, and structural variation of small interfering RNAs and short hairpin RNAs and the effect on mammalian gene silencing. Antisense and Nucleic Acid Drug Development, 13(2), 83-105.

[162]. Fellmann, C., Zuber, J., McJunkin, K., Chang, K., Malone, C. D., Dickins, R. A., Xu, Q., Hengartner, M.O., Elledge, S.J., Hannon, G.J., & Lowe, S. W. (2011). Functional identification of optimized RNAi triggers using a massively parallel sensor assay. Molecular cell, 41(6), 733-746.

[163]. NCBI Reference Sequence Database. (2015). Assed at http://www.ncbi.nlm.nih.gov/refseq/.

[164]. Hinton, G. E. (1989). Connectionist learning procedures. Artificial intelligence, 40(1), 185-234.

[165]. Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1985). Learning internal representations by error propagation (No. ICS-8506). CALIFORNIA UNIV SAN DIEGO LA JOLLA INST FOR COGNITIVE SCIENCE.

[166]. Fletcher, R. (1975). Practical Methods of Optimization, John Wiley & Sons.

[167]. Gill, P.E., Murray, W., & Wright, M.H. (1980). Practical Optimization, Academic Press inc.

[168]. Hestenes, M. (1980). Conjugate Direction Methods in Optimization, Springer Verlag, New York.

[169]. Powell, M. (1977). Restart Procedures for the Conjugate Gradient Method, in: Mathematical Programming, pp 241-254.

[170]. Johansson, E.M., Dowla, F.U., Goodman D.M. (1990). Backpropagation Learning for Multi-Layer Feed-Forward Neural Networks Using the Conjugate Gradient Method, Lawrence Livermore National Laboratory, Preprint UCRL-JC-104850.

[171]. Battiti, R., Masulli, F. (1990). BFGS Optimization for Faster and Automated Supervised Learning, INCC 90 Paris, International Neural Network Conference, pp 757-760.

[172]. Moller, M.F. (1990), Learning by Conjugate Gradients, The 6th International Meeting of Young Computer Scientists, Czechoslovakia.

[173]. Moller, M. F. (1993). A scaled conjugate gradient algorithm for fast supervised learning. Neural networks, 6(4), 525-533.

[174]. Vihinen, M. (2012). How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis. BMC genomics, 13(Suppl 4), S2.

[175]. Lund, O., Nielsen, M., Lundegaard, C., Kesmir, C., Brunak, S. (2005). Immunological Bioinformatics. Cambridge, MA, USA: MIT Press.

[176]. Apache Ver2.0 [software]. (2013). Retrieved from http://www.apache.org/licenses/LICENSE-2.0.

[177]. Han, J., & Kamber M. (2006). Data mining: concepts and techniques, 2 ed. Morgan Kaufmann, Elsevier, New York, 2006.

[178]. Blastall [software]. (2013). Retrieved from http://nebc.nox.ac.uk/bioinformatics/docs/blastall.html.

……..✷♦✷……..

Published Papers Related to the Research

International Journal

1. Reena Murali, Philips George John, David Peter S, "Soft computing model for optimized siRNA design by identifying off target possibilities using Artificial Neural Network Model", GENE, ELSEVIER, Volume 562(2), pp 152-158, 2015, doi: 10.1016/j.gene.2015.02.067.

2. Reena Murali, David Peter S, "Computational Model for Predicting Effective siRNA Sequences using Whole Stacking Energy (ΔG) for Gene Silencing", International Journal of Biological, Veterinary, Agricultural and Food Engineering, Volume 9(1), pp 6-12, 2015.

3. Reena Murali, David Peter S, "Mechanism of RNAi in Genomics and Therapeutics: A Review", International Journal of Computer Applications, Volume 48(4), pp 21-24, 2012.

4. Reena Murali, David Peter S, "siRNA Efficiency Prediction Using Support Vector Machine", International Journal of Emerging Technology and Advanced Engineering, Volume 2(4), pp 605-611, 2012.

International Conference

5. Reena Murali, David Peter S, "Off -target Possibility Prediction of Efficient siRNA: A Workflow", International Conference on Systems Biology and Bioengineering (ICSBB'15), Proceedings of The World Congress on Engineering 2015, 1-3 July, 2015, London, U.K., pp586-591.

6. Reena Murali, David Peter S, "Computational Model for Predicting Effective siRNA Sequences using Whole Stacking Energy (ΔG) for Gene Silencing", 13th International Conference on Computational Models for Life Sciences (ICCMLS2015), Singapore, Jan 8-9- 2015, pp 126-132.

7. Reena Murali, David Peter S, "Computational Model for Predicting Effective siRNA Sequences for Gene Silencing", Eight International Conference on Data Mining and Warehousing (ICDMW-2014), Bangalore, July 25-27, 2014, pp 138-142.

8. Reena Murali, David Peter S, "Predicting efficient siRNAs using Thermodynamic Properties", International Conference on Recent Innovations in Technology (ICRIT 2012), January12-14, 2012, pp 86-89.

……..✷♦✷……..

Appendix

Appendix 1

Sample cDNA Sequences for siRNA design

NCBI gene symbol	cDNA insert sequence used for siRNA design	Length (nt)
Hs TC10 **(BD135193)**	CTTCCTTGTTTGGTCTGCTGTGGATCTGCCTTATTGCATATGCCATGCATCAGATAATGGATGCAT CAGATAATGGTGTTAGACAAAGCTTCATTGTGAACAACCTAATGCATTTTAGAGAAACAATCTC ATCACATTTTTTCTAGCCTTTCCTACATTTAAACTTGCTGTTGCCCAAATTATAATTTTTTAAATGT CTTTGGTGGGCTTCTGTTAATTCACATGACTTGAGCTTATAGCTATGTCTACTGCACAGATTGGG TAATGGAACACTAAACTTTTATACTTGAAAATGACAGCCTTAAATGCTCATATCAGTCACAAATC TAGGATGTACTGTCTTGTTGTATGTGAGCTTTGTAGAGATTTTTAAAAATATAAGCATCACCTTC CCATTGAAGAGTGGAGAGAGTCTACTGGATGACTGGCCAGGAACTTTCTCTCTGAATCGGACAT TTGGATGTCTTCTTTCTTCCAAGAAATGGTGGTTCACATTAAAGTATCATGGCCTTATGTATGCTC AAATGGAATCTTATGTAACTTTCTTATTTAATTTTGGTCTGCTTATTTTTAGATAAAATTGAAAG GAATTGTATAAATCAATTAACATATTAGCTGAGTTG	**623**
Hs CDC34 **(NM_004359)**	CAAGGGGCTGCAGGAAGAGCCGGTCGAGGGATTCCGCGTGACACTGGTGGACGAGGGCGATCT ATACAACTGGGAGGTGGCCATCTTCGGGCCCCCCAACACCTACTACGAGGGCGGCTACTTCAAGG CGCGCCTCAAGTTCCCCATCGACTACCCATACTCTCCACCAGCCTTTCGGTTCCTGACCAAGATGT GGCACCCTAACATCTACGAGACGGGGGACGTGTGTATCTCCATCCTCCACCCGCCGGTGGACGAC CCCCAGAGCGGGGAGCTGCCCTCAGAGAGGTGGAACCCCACGCAGAACGTCAGGACCATTCTCC TGAGTGTGATCTCCCTCCTGAACGAGCCCAACACCTTCTCGCCCGCAAACGTGGACGCCTCCGTGA TGTACAGGAAGTGGAAAGAGAGCAAGGGGAAGGATCGGGAGTACACAGACATCATCCGGAAG CAGGTCCTGGGGACCAAGGTGGACGCGGAGCGTGACGGCGTGAAGGTGCCCACCACGCTGGCCG AGTACTGCGTGAAGACCAAGGCGCCGGCGCCCGACGAGGGCTCAGACCTCTTCTACGACGACTA CTACGGGACGGCGAGGTGGAGGAGGAGG	**607**
Hs UBE2D3 **(NM_003340)**	TCCAGCACAATGTTCTGCAGGTCCAGTTGGGGATGATATGTTTCATTGGCAAGCCACAATTATGG GACCTAATGACAGCCCATATCAAGGCGGTGTATTCTTTTTGACAATTCATTTTCCTACAGACTACC CCTTCAAACCACCTAAGGTTGCATTTACAACAAGAATTTATCATCCAAATATTAACAGTAATGGC AGCATTTGTCTCGATATTCTAAGATCACAGTGGTCGCCTGCTTTAACAATTTCTAAAGTTCTTTTA TCCATTTGTTCACTGCTATGTGATCCAAACCCAGATGACCCCCTAGTGCCAGAGATTGCACGGAT CTATAAAACAGACAGAG	**344**
Hs UBE2B **(NM_003337)**	CTCATGCGGGATTTCAAGCGGTTACAAGAGGACCCACCTGTGGGTGTCAGTGGCGCACCATCTG AAAACAACATCATGCAGTGGAATGCAGTTATATTTGGACCAGAAGGGACACCTTTTGAAGATG GTACTTTTAAACTAGTAATAGAATTTTCTGAAGAATATCCAAATAAACCACCAACTGTTAGGTTT TTATCCAAAATGTTTCATCCAAATGTGTATGCTGATGGTAGCATATGTTTAGATATCCTTCAGAA TCGATGGAGTCCAACATATGATGTATCTTCTATCTTAACATCAATTCAGTCTCTGCTGGATGAAC CGAATCCTAACAGTCCAGCCAATAGCCAGGCAGCACAGCTTTATCAGGAAAACAAACGAGAATA TGAGAAAAGAGTTTCGGCCATTGTTGAACAAAGC	**420**
Hs UBE2M **(NM_003969)**	CAGCAGAAGAAGGAGGAGGAGTCGGCGGGCGGCACCAAGGGCAGCAGCAAGAAGGCGTCGGC GGCGCAGCTGCGGATCCAGAAGGACATAAACGAGCTGAACCTGCCCAAGACGTGTGATATCAGC TTCTCAGATCCAGACGACCTCCTCAACTTCAAGCTGGTCATCTGTCCTGATGAGGGCTTCTACAAG AGTGGGAAGTTTGTGTTCAGTTTTAAGGTGGGCCAGGGTTACCCGCATGATCCCCCCAAGGTGA AGTGTGAGACAATGGTCTATCACCCCAACATTGACCTCGAGGGCAACGTCTGCCTCAACATCCTC AGAGAGGACTGGAAGCCAGTCCTTACGATAAACTCCATAATTTATGGCCTGCAGTATCTCTTCTT GGAGCCCAACCCCGAGGACCCACTGAACAAGGAGGCCGCAGAGGTCCTGCAGAACAACCGGCG GCTGTTTGAGCAGAACGTGCAGCGCTCCATGCGGGGTGGCTACATCGGCTCCACCTACTTTG	**511**
Hs C6orf110 **(XM_371822)**	AGGAATATGTGGCTTCAGCTATGCACGGGGACAGCCATGACCGGTATGAGCGTCTCACCTCTGTC TCCAGCTCCGTTGACTTTGACCAAAGGGACAATGGTTTCTTCCTGGCTGACAGCCATCTTCAGGA TAAAGGATGATGAGATCCGGGACAAATGTGGGGGCGATGCCGTGCACTACCTGTCCTTTCAGCG GCACATCATCGGGCTGCTGGTGGTTGTGGGCGTCCTCTCCGTAGGCATCGTGCTGCCTGTCAACTT CTCAGGGGACCTGCTGGAGAACAATGCCTACAGCTTTGGGAGAACCACCATTGCCAACTTGAAA TCAGGGAACAACCTGCTATGGCTGCACACCTCCTTCGCCTTCCTGTATCTGCTGCTCACCGTCTAC AGCATGCGTAGACACACCTCCAAGATGCGCTACAAGGAGGATGATCTGGTGAAGCGGACCCTCT TCATCAATGGAATCTCCAAATATGCAGAGTCAGAAAAGATCAAGAAGCATTTTGAGGAAGCCT ACCCCAACTGCACAGTTCTCGAAGCCCGCCCGTGTTACAACGTGGCTCGCCTAATGTTCCTCGATG CAGAGAGGAAGAAGGCCGAGCGGGGAAAGCTGTACTTCACAAACCTCCAGAGCAAGGAGAAC GTGCCTACCATGATCAACCCCAAGCCCTGTGGCCACCTCTGCTGCTGTGTGGTGCGAGGCTGTGA GCAGGTGGAGGCCATTGAGTACTACACAAAGCTGGAGCAGAAGCTGAAGGAAGACTACAAGCG GGAGAAGGAGAAGGTGAATGAGAAGCCTCTTGGCATGGCCTTTGTCACCTTCCACAATGAGACT ATCACCGCCATCATCCTGAAGGACTTCAACGTGTGTAAATGCCAGGGCTGCACCTGCCGTGGGGA GCCACGCCCCTCATCCTGCAGCGAGTCCCTGCACATCTCCAACTGGACCGTGTCCTATGCCCCTGA CCCTCAGAACATCTACTGGGAGCACCTCTCCATCCGAGGCTTCATCTGGTGGCTGCGCTGCCTGGT CATCAATGTCGTCCTCTTCATCCTCCTCTTCTTCCTCACCACTCCAGCCATCATCATCACCACCATGG ACAAGTTCAACGTCACCAAGCCTGTGGAGTACCTCAACAACCCCATCATCACCCAGTTCTTCCCCA CCCTGCTGCTGTGGTGCTTCTCGGCCCTCCTTCCCACCATCGTCTACTACTCAGCCTTCTTTGAAGC CCACTGGACACGCTCTGGGGAGAACAGGACAACCATGCACAAGTGCTACACTTTCCTCATCTTCA TGGTGCTGCTCCTACCCTCGCTGGGACTGAGCAGCCTGGACCTCTTCTTCCGCTGGCTCTTTGATA AGAAATTCTTGGCTGAGGCAGCTATTCGGTTTGAGTGTGTGTTCCTGCCCGACAACGGCGCCTTC TTCGTGAACTACGTCATTGCCTCAGCCTTTATCGGCAACGCCATGGACCTGCTGCGCATCCCAGGC CTGCTCATGTACATGATCCGGCTCTGCCTGGCGCGCTCGGCCGCCGAGAGGCGCAACGTGAAGCG GCATCAGGCCTACGAGTTCCAGTTTGGCGCAGCCTACGCCTGGATGATGTGCGTCTTCACGGTGG TCATGACCTACAGTATCACCTGCCCCATCATCGTGCCCTTCGGGCTCATGTACATGCTGCTGAAGC ACCTGGTAGACAGGTACAATCTCTACTACGCCTACCTGCCGGCCAAGCTGGACAAGAAGATCCAC TCGGGGGCTGTGAACCAGGTGGTGGCCGCGCCCATCCTCTGCCTCTTCTGGCTGCTCTTCTTTTCC ACCATGCGCACGGGGTTCCTAGCTCCCACGTCTATGTTCACATTTGTGGTCCTGGTCATCACCATC GTCATCTGTCTCTGCCACGTCTGCTTTGGACACTTCAAATACCTCAGTGCCCACAACTACAAGATT GAGCACACGGAGACAGATACTGTGGACCCCAGAAGCAATGGACGGCCCCCCACTGCTGCTGCTG TCCCCAAATCTGCGAAATACATCGCTCAGGTGCTGCAGGACTCAGAGGTGGACGGGGATGGGGA	**2186**

	TGGGGCTCCTGGGAGCTCAGGGGATGAGCCCCCATCATCCTCATCCCAAGATGAGGAGTTGCTG ATGCCACCCGACGCCCTCACGGACACAGACTTCCAGTC	
Hs CD81P3 **(AY044845)**	ATGAGGGGCAGGAGCTGGCACTGGGCTGCCTGGCGAGGACAAGCACACAGAAGCACACACACC TGGCAGTGTCCTTTGGGCGATCTGTGCCCGAGGCACCAGTTGGGCGGTCAACTCTGCAGGAAGT GGTGGGAATCCGGTCAGACTTGGCCGTGGAGGCTGGAGCTCCCTATGCTGAGCGATTGGCTGCA GGGGAGCTTCGTCTGGGCAAGGAAGGGACCGATCGGTACCGCATGGTAGTAGGGGGTGCCCAG GCAGGGGACGCAGGCACCTACCACTGCACTGCCGCTGAGTGGATTCAGGATCCTGATGGCAGCT GGGCCCAGATTGCAGAGAAAAGGGCCGTCCTGGCCCACGTGGATGTGCAGACGCTGTCCAGCCA GCTGGCAGTGACAGTGGGGCCTGGTGAACGTCGGATCGGCCCAGGGGAGCCCTTGGAACTGCTG TGCAATGTGTCAGGGGCACTTCCCCCAGCAGGCCGTCATGCTGCATACTCTGTAGGTTGGGAGAT GGCACCTGCGGGGGCACCTGGGCCCGGCCGCCTGGTAGCCCAGCTGGACACAGAGGGTGTGGGC AGCCTGGGCCCTGGCTATGAGGGCCGACACATTGCCATGGAGAAGGTGGCATCCAGAACATACC GGCTACGGCTAGAGGCTGCCAGGCCTGGTGATGCGGGCACCTACCGCTGCCTCGCCAAAGCCTAT GTTCGAGGGTCTGGGACCCGGCTTCGTGAAGCAGCCAGTGCCCGTTCCCGGCCTCTCCCTGTACAT GTGCGGGAGGAAGGTGAGAGGGGGCTGGGCCCTGGACGGGGTTCGAGGATTGTCAACCCCTTT TCCTTCTGTTTCCATGACCCCTCCCTCTCCGTCAGCTGT	**872**
Hs RAB6IP1 **(NM_015213)**	TCGCCGACTACTTTGTCATCTGCGGACTGGACACGGAGACCGGGCTGGAGCCGGACGAGCTGTC GGCATTATGCCAGTACATACAGGCTTCTAAAGCCAGGGATGGTGCCAGCCCTTTCATTTCAAGTA CGACTGAAGGAGAAAATTTTGAGCAGACACCATTGAGAAGAACATTCAAATCTAAGGTCCTTGC ACGATATCCTGAGAACGTAGAATGGAATCCCTTTGACCAAGATGCAGTAGGAATGCTATGTATG CCGAAAGGGCTGGCATTCAAGACCCAGGCTGATCCCAGGGAGCCCCAATTCCATGCCTTTATTAT CACAAGGGAGGATGGCTCTCGGACATTTGGGTTTGCCCTCACATTTTATGAAGAGGTGACTAGC AAGCAGATCTGCAGTGCAATGCAGACCCTCTACCACATGCACAATGCTGAGTATGATGTCCTACA TGCTCCCCCTGCTGATGACAGAGACCAGAGCAGCATGGAGGATGGTGAAGACACTCCTGTGACC AAACTGCAGCGCTTCAACTCCTATGACATTAGCCGGGACACTCTCTACGTCTCTAAGTGCATCTGC CTCATCACACCCATGTCTTTCATGAAGGCATGTCGGAGCGTGCTGGAGCAACTCCACCAGGCAGT CACTTCACCTCAGCCCCCTCCACTGCCCCTTGAGAGCTACATATACAACGTACTCTACGAGGTGCC GCTCCCACCTCCTGGCCGGTCCTTGAAGTTTTCTGGGGTCTATGGGCCAATAATCTGCCAGAGACC AAGTACCAATGAGCTTCCCCTATTTGACTTTCCTGTCAAAGAGGTTTTTGAACTGCTCGGGGTGG AGAATGTGTTTCAGCTTTTTACTTGTGCCCTTCTGGAGTTTCAAATCCTGCTCTACTCACAGCATT ACCAGAGACTGATGACTGTGGCGGAGACGATTACAGCTCTCATGTTTCCTTTCCAGTGGCAGCAT GTCTATGTCCCTATTCTCCCAGCTTCTCTCCTGCATTTCTTAGATGCTCCTGTTCCATACCTGATGG GTTTGCATTCCAATGGCCTGGATGACCGGTCAAAGCTGGAGCTGCCTCAAGAGGCTAACCTCTGC TTTGTGGACATTGACAACCACTTCATTGAGTTGCCAGAGGACTTGCCACAGTTCCCCAACAAATT GGAGTTTGTCCAGGAAGTCTCTGAGATTCTCATGGCATTTGGAATTCCCCCTGAAGGGAATCTTC ATTGCAGTGAGAGTGCCTCCAAGCTGAAGAGGCTGCGGGCCTCTGAGCTTGTCTCGGACAAGAG GAATGGGAACATTGCTGGCTCCCCTTTGCATTCCTACGAGCTTCTTAAGGAGAATGAAACTATTG CCCGGCTGCAAGCCTTGGTCAAGAGAACTGGGGTGAGCCTGGAAAAGTTGGAAGTGCGTGAAG ACCCCAGCAGCAATAAGGATCTCAAAGTTCAGTGTGATGAAGAAGAACTCAGGATTTACCAGCT AAACATTCAGATCCGGGAAGTTTTTGCAAATCGTTTCACTCAGATGTTTGCAGATTATGAGGCGT TTGTCATCCAACCCAGCCAGGATAAGGAATCCTGGTTTACCAACAGGGAGCAAATGCAAAACTT TGATAAAGCATCTTTTCTGTCAGATCAGCCTGAGCCCTACCTGCCCTTCCTCTCAAGATTCCTGGA GACCCAGATGTTTGCATCTTTCATTGACAACAAAATAATGTGTCATGATGATGATGATAAAGAC CCTGTACTCCGGGTATTTGATTCCCGAGTTGACAAGATCAGGCTGTTGAATGTTCGGACACCTAC TCTCCGTACATCCATGTACCAGAAGTGTACCACTGTGGATGAAGCAGAGAAAGCAATTGAGCTG CGTCTGGCAAAAATTGACCATACTGCAATTCACCCACATTTACTTGACATGAAGATTGGACAAG GGAAATATGAGCCGGGCTTCTTCCCTAAGCTGCAGTCTGATGTACTTTCCACTGGGCCAGCCAGC AACAAGTGGACGAAAAGGAATGCCCCTGCCCAGTGGAGGCGGAAAGATCGGCAGAAGCAGCAC ACAGAACACCTGCGTTTAGATAATGACCAGAGGGAGAAGTACATCCAGGAAGCCAGGACTATG GGCAGCACTATCCGCCAGCCCAAACTGTCCAACCTCTCTCCATCAGTGATTGCCCAGACCAATTGG AAGTTTGTAGAGGGCCTGCTGAAGGAATGCCGCAATAAGACCAAGAGGATGCTGGTGGAAAAG ATGGGCCGAGAAGCTGTGGAGCTAGGGCATGGGGAGGTGAACATCACAGGGGTGGAAGAGAA CACCCTGATTGCCAGCCTTTGTGATCTCCTGGAAAGGATCTGGAGTCATGGACTACAAGTGAAAC AGGGGAAATCAGCCTTATGGTCCCACCTGTTACATTATCAGGACAACCGGCAGAGAAAACTCAC ATCAGGAAGCCTCAGTACCTCAGGAATACTTCTTGATTCAGAACGTAGGAAGTCTGATGCCAGC TCACTCATGCCTCCCCTGAGGATCTCCCTGATTCAGGATATGAGGCACATCCAGAACATCGGGGA AATCAAGACTGATGTGGGAAAGGCCAGAGCATGGGTGCGACTGTCCATGGAAAAAAAGTTACT TTCCAGACACCTGAAGCAGCTCCTCTCAGACCATGAGCTCACCAAAAAGTTATATAAGCGCTATG CCTTCCTGCGCTGTGATGACGAGAAGGAGCAGTTCCTCTATCACCTCCTGTCTTTCAATGCCGTCG ATTACTTTTGCTTCACCAATGTCTTCACAACTATCCTGATCCCGTACCACATTCTGATCGTACCAA GCAAGAAGCTGGGGGGCTCCATGTTCACTGCCAACCCATGGATCTGTATATCAGGAGAATTGGG TGAGACACAGATCATGCAGATTCCCAGGAATGTGCTAGAGATGACCTTCGAGTGCCAGAACTTG GGGAAGCTTACTACTGTCCAGATTGGCCATGATAACTCTGGGCTGTATGCCAAATGGCTGGTGG AGTATGTGATGGTCAGGAATGAGATCACAGGACATACCTACAAGTTCCCGTGTGGCCGGTGGTT AGGGAAGGGCATGGATGATGGAAGCCTGGAGCGGATCCTAGTTGGGGAGCTGCTCACATCCCA GCCTGAGGTGGATGAGAGGCCATGCCGGACCCCGCCGCTGCAGCAGTCCCCCAGTGTCATCCGGA GGCTTGTTACCATCTCACCCAACAACAAGCCCAAGCTGAACACTGGGCAGATCCAGGAGTCCATC GGGGAGGCAGTCAATGGCATTGTGAAGCACTTCCATAAGCCTGAGAAAGAGCGAGGCAGTCTG ACGCTGTTGCTCTGTGGAGAGTGTGGCCTTGTCTCGGCCTTGGAACAGGCTTTCCAGCATGGATT TAAATCGCCCCGGCTCTTCAAAAATGTCTTCATTTGGGATTTCCTGGAAAAAGCACAAACCTATT ATGAGACATTAGAGAAGAATGAAGTAGTCCCTGAGGAAAACTGGCATACAAGAGCCCGGAACT TCTGCCGATTTGTCACTGCAATCAACAATACTCCCCGGAACATCGGCAAGGATGGCAAGTTTCAG ATGCTGGTGTGCTTGGGAGCCAGAGATCACCTCCTACACCACTGGATTGCCCTGCTGGCTGACTG CCCCATCACTGCACACATGTATGAGGATGTGGCACTGATCAAAGACCATACACTTGTCAATTCCT TGATTCGTGTGCTGCAGACATTGCAGGAGTTCAACATCAC	**3784**

Appendix 2

Sample siRNAs of Data Set1 showing inhibition efficiency at ΔG >= -32.5 kcal/mol using 6-8-8-8-1 ANN Model

Antisense Stamd	ΔG	iScore	s- Biopre	Thermo	DSIR	MySiRNA	Original Experimental	Our Score (6-8-8-8-1)
CUAAUAUGUUAAUUGAUUUau	-24.6	57	0.71	0.7	69.8	58.49	0.46	0.54
AAUAUGUUAAUUGAUUUAUac	-23.6	65.3	0.69	0.78	71.4	71.03	0.38	0.58
GAUUUAUACAAUUCCUUUCaa	-28.2	61.5	0.74	0.86	74.8	82.84	0.51	0.71
CAAUUCCUUUCAAUUUUAUcu	-26.2	44.5	0.50	0.64	56.7	49.72	0.36	0.45
CAGACCAAAAUUAAAUAAGaa	-27.4	56.4	0.66	0.82	68	71.96	0.52	0.65
AGACCAAAAUUAAAUAAGAaa	-27.7	51.3	0.54	0.71	62.4	59.71	0.44	0.53
ACCAAAAUUAAAUAAGAAAgu	-25	48.6	0.62	0.68	61.1	51.45	0.44	0.49
CAAAAUUAAAUAAGAAAGUua	-23.8	59.3	0.74	0.75	68.7	61.63	0.44	0.56
UAAGAAAGUUACAUAAGAUuc	-28.2	64.2	0.77	0.86	75.1	85.58	0.59	0.72
AAGUUACAUAAGAUUCCAUuu	-30.4	54.2	0.74	0.82	73.4	77.19	0.51	0.72
ACAUAAGAUUCCAUUUGAGca	-31.4	52	0.60	0.72	71.5	70.68	0.56	0.66
CCAUUUGAGCAUACAUAAGgc	-32.5	46	0.46	0.64	62.4	60.34	0.44	0.53
CAUUUGAGCAUACAUAAGGcc	-32.5	63.2	0.66	0.76	73	86.61	0.65	0.76
UUAAUGUGAACCACCAUUUcu	-32	76.4	0.85	0.91	91.2	94.92	0.86	0.91
AUUUCUUGGAAGAAAGAAGac	-30.8	66.7	0.75	0.85	78.1	90.44	0.54	0.79
GAAAGAAGACAUCCAAAUGuc	-32.3	50.6	0.54	0.71	64.4	70.29	0.74	0.62
UUUAAGGCUGUCAUUUUCAag	-32	71.2	0.85	0.89	84.6	93.59	0.79	0.86
UUAAGGCUGUCAUUUUCAAgu	-32	55.5	0.71	0.74	71.1	77.77	0.87	0.71
AAGUAUAAAAGUUUAGUGUuc	-27.6	65.3	0.77	0.88	77.3	86.30	0.76	0.72
UAAAAGUUUAGUGUUCCAUua	-29.8	70.5	0.85	0.9	86.2	92.69	0.82	0.89
CUCAAGUCAUGUGAAUUAAca	-31.3	37.5	0.36	0.56	53.6	48.84	0.54	0.44
GCAACAGCAAGUUUAAAUGua	-31.6	47.8	0.60	0.71	68.7	64.84	0.75	0.64
ACAGCAAGUUUAAAUGUAGga	-30.8	55.4	0.67	0.78	75.2	77.39	0.83	0.71
AUUGUUUCUCUAAAAUGCAuu	-29.8	65.2	0.80	0.85	80.4	88.50	0.85	0.78
GUGAACAAAUGGAUAAAAGaa	-29.8	48.7	0.61	0.77	67.2	65.00	0.27	0.65
UGGAUAAAAGAACUUUAGAaa	-30.2	57.8	0.68	0.85	73.7	82.59	0.64	0.73
GAUAAAAGAACUUUAGAAAuu	-26.6	41.6	0.47	0.75	60.1	50.98	0.31	0.51
AAAAGAACUUUAGAAAUUGuu	-25.9	68.5	0.84	0.95	86.3	88.56	0.81	0.77

Antisense Stand	ΔG	iScore	s-Biopre	Thermo	DSIR	MySiRNA	Original Experimental	Our Score (6-8-8-8-1)
GUCUGUAGGAAAAUGAAUUgu	-31.3	49.8	0.69	0.77	73	70.56	0.62	0.71
CUGUAGGAAAAUGAAUUGUca	-31	49.9	0.58	0.71	68.5	66.18	0.76	0.62
UGUAGGAAAAUGAAUUGUCaa	-31.3	72.2	0.82	0.95	90.8	94.63	0.92	0.94
GUAGGAAAAUGAAUUGUCAaa	-31.3	44.9	0.49	0.75	64.4	62.68	0.55	0.62
CGAAACUCUUUUCUCAUAUuc	-30.6	45.4	0.46	0.59	59.6	53.92	0.46	0.50
AAACUCUUUUCUCAUAUUCuc	-29.1	70.5	0.79	0.88	84.2	91.75	0.79	0.84
ACUCUUUUCUCAUAUUCUCgu	-31.8	63.6	0.78	0.86	83.6	89.85	0.84	0.90
CUCUUUUCUCAUAUUCUCGuu	-32	64.4	0.83	0.9	84.1	91.50	0.74	0.92
UCUUUUCUCAUAUUCUCGUuu	-32.1	67.2	0.84	0.87	88.6	91.96	0.84	0.93
AUUCUCGUUUGUUUUCCUGau	-32.2	64.2	0.81	0.91	84.3	91.77	0.84	0.93
UUUGUUUUCCUGAUAAAGCug	-30.8	84.8	0.89	1.11	98.6	97.12	0.80	0.95
CUGAAUUGAUGUUAAGAUAga	-29.4	45.7	0.52	0.7	64.2	56.80	0.51	0.54
UUAAGAUAGAAGAUACAUCau	-30.1	81.1	0.88	1.07	98.3	96.73	0.85	0.94
AUAGAAGAUACAUCAUAUGuu	-30.2	67.4	0.82	0.95	87.9	92.72	0.84	0.93
CUGAAGGAUAUCUAAACAUau	-31.8	51.3	0.64	0.7	73	69.41	0.47	0.67
UGAAGGAUAUCUAAACAUAug	-31	58.3	0.76	0.78	82.5	81.22	0.64	0.77
CACAUUUGGAUGAAACAUUuu	-30.8	53.7	0.69	0.79	75.8	75.79	0.64	0.72
UUUGGAUGAAACAUUUUGGau	-30.5	78	0.86	1.04	93.4	96.30	0.85	0.94
GGAUGAAACAUUUUGGAUAaa	-31.4	40.7	0.40	0.64	57.9	53.20	0.48	0.47
GUUUAUUUGGAUAUUCUUCag	-28.2	62.6	0.81	0.79	78.4	80.67	0.78	0.71
UUAUUUGGAUAUUCUUCAGaa	-29.3	71.9	0.84	0.89	87.7	92.72	0.80	0.88
UUUGGAUAUUCUUCAGAAAau	-30.2	56.3	0.67	0.81	73.2	78.79	0.83	0.71
UUGGAUAUUCUUCAGAAAAuu	-30.2	64.1	0.80	0.86	80	88.55	0.85	0.79
UGGAUAUUCUUCAGAAAAUuc	-30.4	56.2	0.62	0.77	65.7	76.88	0.64	0.66
UCAGAAAAUUCUAUUACUAgu	-28.4	58.8	0.75	0.81	76.3	77.50	0.72	0.70
AAAAUUCUAUUACUAGUUUaa	-25.5	61.1	0.73	0.78	76.8	70.72	0.48	0.63
AAUUCUAUUACUAGUUUAAaa	-25.9	55.5	0.66	0.82	72.9	66.22	0.47	0.63
AUUACUAGUUUAAAAGUACca	-26.8	69.8	0.81	0.91	83.3	89.63	0.86	0.76

Appendix 3

Sample siRNAs of Data Set 3 showing calculation of TP, TN, FP, FN values using 5-12-1 ANN (OpsiD) Model

siRNA Strand	OpsiD	Actual Activity	Predicted Activity	Classified as
CUGAUAUUAACAAAGCUUA	0.69	1	1	TP
AAUGCAAAGCACAUCCAAU	0.60	1	0	FN
GGGAGAACAGGGUAUGAUA	0.78	1	1	TP
GCGUGUCUCUCUGCCCUGG	0.34	1	0	FN
GGUUGGGAUUCCUACGGAU	0.70	1	1	TP
AGUUGUAAAGAAGCUUAUA	0.80	1	1	TP
CGAUGGGUGUGGUUGCCUUA	0.82	1	1	TP
UUAGAAAAUACAGUAAGUU	0.37	1	0	FN
CCUCUCAACGACAGCAGCU	0.57	1	0	FN
GCGACUCUGAGGAGGAACA	0.61	1	1	TP
AAAUCGAUGUUGUUUCUGU	0.57	1	0	FN
AUGUUGUUUCUGUGGAAAA	0.67	1	1	TP
GUCUGGAUCACCUUCUGCU	0.54	1	0	FN
GGGCUUUAUCUAACUCGCU	0.62	1	1	TP
ACCAGAUCCCGGAGUUGGA	0.57	1	0	FN
AACGACGAGAACAGUUGAA	0.61	1	1	TP
GACGAGAACAGUUGAAACA	0.69	1	1	TP
ACGAGAACAGUUGAAACAC	0.48	1	0	FN
ACGGAACUCUUGUGCGUAA	0.67	1	1	TP
CCAGCGAGGAUAUCUGGAA	0.62	1	1	TP
ACCGGAAACAGGUGGUCAU	0.58	1	0	FN
GCAGGUCAAGAAGAGUAUA	0.83	1	1	TP
CCACUAUAGAGGACUCCUA	0.73	1	1	TP
AGACUGUUCACAAAGGCUU	0.61	1	1	TP
UUAGCUAAAUUACUGUAAG	0.37	1	0	FN
UAAGGCAGACCCAGUAUGA	0.55	1	0	FN
UUGAAUUUCUGUUGUUGAA	0.61	1	1	TP

siRNA Strand	OpsiD	Actual Activity	Predicted Activity	Classified as
UGCACUUAAAGCAACCCUU	0.60	1	1	TP
ACCCUUCACAUCAUUUGAA	0.62	1	1	TP
ACAUUAACCUCCCAGCCGU	0.54	1	0	FN
AUGUUGACACACCUAGUUA	0.66	1	1	TP
UUGACACACCUAGUUAUCA	0.43	1	0	FN
ACAAUCUGUAACUGUUUUA	0.62	1	1	TP
GCAAAUUGUGCAAGAGGUA	0.69	1	1	TP
UGACGAUACAGCUAAUUCA	0.69	1	1	TP
ACUUCAAACUUUGUUUCCA	0.59	1	0	FN
CGAUACAGCUAAUUCAGAA	0.77	1	1	TP
CUAAGGUAGGUACAUCAAA	0.74	1	1	TP
ACACCAUAGUGGGAUUUAA	0.67	1	1	TP
AAACACUUCGAAGUUUCAA	0.59	1	0	FN
GGUGAAUUUUAAGUCUUAA	0.79	1	1	TP
GUGAAUUUUAAGUCUUAAA	0.64	1	1	TP
GUAAUACUUCUCAUUUCUA	0.70	1	1	TP
CAAAUACAAUUUGAGUAAA	0.65	1	1	TP
AUUGCACAAUUUUGUCCUA	0.67	1	1	TP
GCAGGUCAAGAGGAGUACA	0.67	1	1	TP
CAAGUUCUGUGAGGCAUGU	0.59	1	0	FN
AGGUCAAGAGGAGUACAGU	0.58	1	0	FN
AAACAUACAUACAUACAUA	0.69	1	1	TP
GAAGUACAUUAGUUUUCAA	0.84	1	1	TP
GUAGGCAUUUUAGAUGAAU	0.60	1	0	FN
UAGGCAUUUUAGAUGAAUU	0.56	1	0	FN
GAUGAAUUUGGGAAAAAUA	0.77	1	1	TP
AAAAUAAAGUUCUGCAGAA	0.63	1	1	TP
AAUCCCGUUUCAUGUCUUC	0.46	1	0	FN

Appendix 4

Sample siRNAs of Data Set 1 showing inhibition efficiency at ΔG>= -34.6 kcal/mol using 5-12-1 ANN (OpsiD) Model

siRNA Strand	Whole ΔG	s-Biopredsi	iScore	Thermo	DSIR	My SiRNA	OpsiD	Original Inhibition
CUAAUAUGUUAAUUGAUUUau	-24.60	0.71	57.00	0.70	69.80	58.49	0.39	0.46
AAUAUGUUAAUUGAUUUAUac	-23.60	0.69	65.30	0.78	71.40	71.03	0.47	0.38
GAUUUAUACAAUUCCUUUCaa	-28.20	0.74	61.50	0.86	74.80	82.84	0.55	0.51
CAAUUCCUUUCAAUUUUAUcu	-26.20	0.50	44.50	0.64	56.70	49.72	0.33	0.36
CAGACCAAAAUUAAAUAAGaa	-27.40	0.66	56.40	0.82	68.00	71.96	0.48	0.52
AGACCAAAAUUAAAUAAGAaa	-27.70	0.54	51.30	0.71	62.40	59.71	0.37	0.44
ACCAAAAUUAAAUAAGAAAgu	-25.00	0.62	48.60	0.68	61.10	51.45	0.32	0.44
CAAAAUUAAAUAAGAAAGUua	-23.80	0.74	59.30	0.75	68.70	61.63	0.40	0.44
UAAGAAAGUUACAUAAGAUuc	-28.20	0.77	64.20	0.86	75.10	85.58	0.57	0.59
AAGUUACAUAAGAUUCCAUuu	-30.40	0.74	54.20	0.82	73.40	77.19	0.54	0.51
ACAUAAGAUUCCAUUUGAGca	-31.40	0.60	52.00	0.72	71.50	70.68	0.51	0.56
AAGAUUCCAUUUGAGCAUAca	-32.60	0.54	50.80	0.69	68.50	69.92	0.51	0.55
CCAUUUGAGCAUACAUAAGgc	-32.50	0.46	46.00	0.64	62.40	60.34	0.41	0.44
CAUUUGAGCAUACAUAAGGcc	-32.50	0.66	63.20	0.76	73.00	86.61	0.61	0.65
AUAAGGCCAUGAUACUUUAau	-32.90	0.61	51.70	0.68	64.90	71.06	0.50	0.76
GCCAUGAUACUUUAAUGUGaa	-32.60	0.49	45.20	0.67	60.40	61.36	0.40	0.62
UUAAUGUGAACCACCAUUUcu	-32.00	0.85	76.40	0.91	91.20	94.92	0.74	0.86
AUUUCUUGGAAGAAAGAAGac	-30.80	0.75	66.70	0.85	78.10	90.44	0.62	0.54
GAAAGAAGACAUCCAAAUGuc	-32.30	0.54	50.60	0.71	64.40	70.29	0.49	0.74
GGAAGGUGAUGCUUAUAUUuu	-34.00	0.43	38.00	0.59	57.70	52.34	0.37	0.52
ACAGUACAUCCUAGAUUUGug	-33.90	0.64	49.50	0.78	76.00	76.74	0.57	0.79
GCAUUUAAGGCUGUCAUUUuc	-33.20	0.46	44.30	0.56	56.20	55.30	0.36	0.42
UUUAAGGCUGUCAUUUUCAag	-32.00	0.85	71.20	0.89	84.60	93.59	0.67	0.79
UUAAGGCUGUCAUUUUCAAgu	-32.00	0.71	55.50	0.74	71.10	77.77	0.55	0.87
AAGUAUAAAAGUUUAGUGUuc	-27.60	0.77	65.30	0.88	77.30	86.30	0.58	0.76
UAAAAGUUUAGUGUUCCAUua	-29.80	0.85	70.50	0.90	86.20	92.69	0.68	0.82
CUCAAGUCAUGUGAAUUAAca	-31.30	0.36	37.50	0.56	53.60	48.84	0.36	0.54
UAUAAUUUGGGCAACAGCAag	-34.50	0.85	72.20	0.94	90.40	94.55	0.72	0.86

siRNA Strand	Whole ΔG	s-Biopredsi	iScore	Thermo	DSIR	My SiRNA	OpsiD	Original Inhibition
AUUGUUUCUCUAAAAUGCAuu	-29.80	0.80	65.20	0.85	80.40	88.50	0.61	0.85
AUGCAUUAGGUUGUUCACAau	-34.50	0.73	59.70	0.82	79.40	88.05	0.61	0.75
UGCAUUAGGUUGUUCACAAug	-34.30	0.73	54.50	0.79	76.70	83.01	0.59	0.83
UAACACCAUUAUCUGAUGCau	-34.10	0.84	71.60	0.95	94.70	94.67	0.75	0.93
CAUUAUCUGAUGCAUCCAUua	-34.20	0.56	44.90	0.68	68.00	64.92	0.49	0.83
UAUCUGAUGCAUCCAUUAUcu	-33.40	0.75	61.30	0.79	80.60	87.24	0.61	0.81
AUCCAUUAUCUGAUGCAUGgc	-34.20	0.78	66.50	0.77	81.30	89.28	0.64	0.72
GUCUGUUUUAUAGAUCCGUgc	-33.30	0.60	51.40	0.75	67.40	76.02	0.54	0.68
UUUAUAGAUCCGUGCAAUCuc	-33.40	0.81	71.10	1.03	86.90	95.61	0.70	0.92
CAUAGCAGUGAACAAAUGGau	-34.40	0.74	57.50	0.84	79.20	87.51	0.61	0.95
GCAGUGAACAAAUGGAUAAaa	-33.50	0.35	35.50	0.64	52.80	51.93	0.34	0.24
GUGAACAAAUGGAUAAAAGaa	-29.80	0.61	48.70	0.77	67.20	65.00	0.47	0.27
UGGAUAAAAGAACUUUAGAaa	-30.20	0.68	57.80	0.85	73.70	82.59	0.56	0.64
GAUAAAAGAACUUUAGAAAuu	-26.60	0.47	41.60	0.75	60.10	50.98	0.33	0.31
AAAAGAACUUUAGAAAUUGuu	-25.90	0.84	68.50	0.95	86.30	88.56	0.66	0.81
UUUAGAAAUUGUUAAAGCAgg	-27.30	0.84	67.60	0.93	85.90	89.38	0.66	0.92
UUAGAAAUUGUUAAAGCAGgc	-28.50	0.84	75.80	0.93	84.50	94.29	0.69	0.90
UGAUCUUAGAAUAUCGAGAca	-32.90	0.77	66.10	0.85	85.70	91.34	0.66	0.98
GAAUAUCGAGACAAAUGCUgc	-33.30	0.48	46.60	0.69	63.60	65.91	0.46	0.79
AAUAUCGAGACAAAUGCUGcc	-33.00	0.74	67.70	0.86	82.70	92.17	0.65	0.95
UGCUGCCAUUACUGUUAAUau	-34.60	0.48	41.10	0.73	61.10	64.17	0.45	0.58
CUGUUAAUAUUUGGAUGAUaa	-29.30	0.38	37.90	0.66	54.20	49.78	0.33	0.32
UAAUAUUUGGAUGAUAAAUuc	-26.20	0.81	70.40	0.83	78.80	86.46	0.59	0.76
UUUGGAUGAUAAAUUCUUGuu	-28.90	0.86	73.20	1.01	91.00	94.84	0.74	0.93
GAUAAAUUCUUGUUGUAAAug	-26.60	0.52	41.60	0.70	61.30	49.94	0.33	0.49
AUAAAUUCUUGUUGUAAAUgc	-25.30	0.68	59.70	0.76	72.20	66.96	0.46	0.74

Appendix 5

Sample siRNAs of Data Set 2 showing inhibition efficiency at ΔG >= -34.6 kcal/mol using 5-12-1 ANN (OpsiD) Model

SiRNA Strand	Whole ΔG	s-Biopre	i-Score	Thermo	DSIR	My SiRNA	OpsiD	Original Inhibition
GCCAAAGAAAAAUAACCCU	-32.90	0.78	63.00	0.80	76.30	88.32	0.61	0.96
UAACCCUUUGGUUUUUUUC	-30.30	0.49	40.30	0.58	52.50	49.89	0.34	0.52
GUUAUUUUCUACAGUGGUU	-31.10	0.79	62.30	0.90	87.00	89.67	0.66	0.98
GACGCCAAAAACAUAAAGA	-32.40	0.78	62.90	0.83	79.10	88.96	0.62	0.86
ACGCCAAAAACAUAAAGAA	-30.90	0.83	70.30	0.81	81.70	91.16	0.64	0.95
AAAACAUAAAGAAAGGCCC	-32.90	0.54	41.30	0.57	55.50	52.80	0.36	0.11
CCGCGAACGACAUUUAUAA	-33.60	0.83	73.90	0.90	87.70	94.31	0.70	0.95
CCAAUCAUCCAAAAAAUUA	-28.60	0.87	77.90	0.96	88.40	95.22	0.73	0.98
CCGGAUACUGCGAUUUUAA	-34.40	0.86	76.10	0.94	93.40	94.99	0.76	0.96
GGUUUUGGAAUGUUUACUA	-31.00	0.86	76.70	0.97	97.60	95.60	0.80	0.93
GAUUUCGAGUCGUCUUAAU	-32.50	0.78	66.90	0.86	83.70	91.76	0.65	0.99
CCUUCAGGAUUACAAGAUU	-33.60	0.83	72.00	0.90	90.40	94.01	0.72	0.99
GACAAAUACGAUUUAUCUA	-29.00	0.85	74.80	0.93	96.80	94.23	0.79	0.87
CAAAUACGAUUUAUCUAAU	-26.40	0.80	70.70	0.89	83.50	89.10	0.63	0.88
GGUAAAGUUGUUCCAUUUU	-30.60	0.83	72.30	0.90	91.50	93.68	0.73	0.95
GAUUAUGUCCGGUUAUGUA	-33.60	0.85	73.50	0.95	96.00	94.97	0.77	0.97
CGACGCAAGAAAAAUCAGA	-34.00	0.83	68.90	0.92	92.40	93.76	0.72	0.96
GCAAGAAAAAUCAGAGAGA	-33.60	0.83	70.30	0.91	92.50	93.87	0.73	0.95
AAAACUCGACGCAAGAAAA	-32.40	0.75	69.10	0.79	80.20	90.71	0.64	0.95
AAAAAGAGAUCGUGGAUUA	-31.80	0.77	70.20	0.78	79.90	90.69	0.64	0.91
AACUCGACGCAAGAAAAAU	-32.60	0.76	57.10	0.74	79.90	80.59	0.58	0.90
ACGCAAGAAAAAUCAGAGA	-33.70	0.78	61.50	0.78	83.00	87.14	0.62	0.88
GAAAAAUCAGAGAGAUCCU	-34.00	0.80	63.60	0.81	83.70	89.43	0.64	0.87
GUAACAACCGCGAAAAAGU	-33.60	0.74	57.80	0.80	77.10	85.43	0.60	0.86
AAGUAACAACCGCGAAAAA	-32.30	0.83	67.00	0.80	86.90	90.07	0.67	0.77
AAAGACGAUGACGGAAAAA	-32.80	0.62	57.10	0.67	73.60	76.77	0.58	0.74
AAAGGUCUUACCGGAAAAC	-34.60	0.47	38.70	0.54	57.70	51.37	0.37	0.71

SiRNA Strand	Whole ΔG	s-Biopre	i-Score	Thermo	DSIR	My SiRNA	OpsiD	Original Inhibition
UUACUGACUUUCCUUGAGU	-34.60	0.38	39.70	0.70	56.30	60.56	0.38	-0.08
AAGAUGCCAUGAAAGCUUA	-34.60	0.77	70.80	0.77	79.10	90.54	0.64	0.51
UGAAAGCUUACAUCAACAA	-32.10	0.84	64.70	0.91	87.60	91.92	0.68	0.34
AAAGUAGAAGAGCUAAAGA	-32.50	0.77	58.90	0.80	85.10	85.11	0.62	0.72
AAGAGCUAAAGAAAAAAUA	-27.50	0.83	70.90	0.81	85.70	88.24	0.66	0.68
AGAAUGAAUACAGUAUUCC	-31.90	0.38	40.00	0.51	56.00	49.07	0.38	0.58
UAUUUCCGCAAUUACAUGA	-31.90	0.68	51.20	0.68	69.50	68.21	0.50	0.57
UUUGAGACUUCUUGCCUAA	-34.50	0.78	60.50	0.80	79.70	87.71	0.61	0.82
CAUGAGAAGUAUGACAACA	-33.80	0.81	65.50	0.80	87.70	89.87	0.67	0.92
CUUGACCUAUAUUUAUCCA	-31.90	0.71	62.50	0.82	73.90	87.94	0.61	0.48
AUACAGGAACAAUAUUGAU	-30.60	0.67	59.20	0.71	73.00	77.58	0.56	0.49
UGACACCGCCAAAUUUAAU	-33.50	0.70	54.90	0.73	70.20	79.09	0.57	0.65
CCUUUUGUGAAGAUCUUGA	-33.30	0.82	67.50	0.83	86.00	91.41	0.66	0.73
AAAGGUGAAGAUAUAUUCC	-31.80	0.52	45.60	0.53	64.10	53.19	0.39	0.09
GGGUAAAUACAUUCUUCAU	-31.60	0.86	78.20	0.92	89.90	95.30	0.73	0.70
AUAAAGACAAAGCCAACCG	-34.50	0.41	34.80	0.56	56.80	49.61	0.37	0.00
UAAAACACCAUGAAAAUAA	-27.50	0.72	62.70	0.79	68.20	79.10	0.52	0.58
UUACCAGUUAUAGGAACAA	-32.60	0.61	52.90	0.72	70.00	74.59	0.54	0.40
UUCCCGUUUUAUUCCAGUU	-33.50	0.65	49.40	0.71	69.50	71.33	0.53	0.38
UUCCGAAAGGUUUUGCUAC	-34.60	0.28	35.90	0.54	52.40	49.62	0.35	0.32
CUCAGAAAGGAAAUAAUUU	-28.70	0.75	66.00	0.80	78.00	85.67	0.59	0.30
GCUACAGUUAUUAAUCUGG	-32.80	0.62	53.30	0.74	73.30	76.72	0.55	0.37
AAGAAACACAGCAACAAUG	-32.70	0.45	46.50	0.65	63.80	61.96	0.43	0.30
AAGAAUGGGCUUGAAACAU	-34.20	0.73	54.20	0.87	71.70	86.35	0.62	0.40
GCACUCUGAUUGACAAAUA	-33.80	0.87	78.00	1.01	95.90	96.00	0.79	0.93
CUCGACGCAAGAAAAAUCA	-34.00	0.77	61.60	0.86	82.10	90.14	0.63	0.91

……..✷♦✷……..