PART I: TO DEVELOP A SMALL INTERFERING RNA (SIRNA) DESIGN AND INFORMATION RESOURCE TO FACILITATE GENETIC MANIPULATION OF HUMAN CELLS.

PART II: TO PARTICIPATE IN DEVELOPING MICROARRAY BASED GENE EXPRESSION SIGNATURES FOR IN VITRO DRUG SENSITIVITY AND RESISTANCE FOR BREAST CANCER.

APPROVED BY SUPERVISORY COMMITTEE

John D .Minna, M.D.

Harold ("Skip") R. Garner, Ph.D.

Khosrow Behbehani, Ph.D.

Jerry Shay, Ph.D.

I would like to dedicate this thesis to my mother Vanita Khetsi Shah, who has been the inspiration of my life, my sister Bhavana who has been my strongest pillar of support and my friends.

PART I: TO DEVELOP A SMALL INTERFERING RNA (SIRNA) DESIGN AND INFORMATION RESOURCE TO FACILITATE GENETIC MANIPULATION OF HUMAN CELLS.

PART II: TO PARTICIPATE IN DEVELOPING MICROARRAY BASED GENE EXPRESSION SIGNATURES FOR IN VITRO DRUG SENSITIVITY AND RESISTANCE FOR BREAST CANCER.

by

JYOTI KHETSI SHAH

THESIS

Presented to the Faculty of the Graduate School of Biomedical Sciences

The University of Texas Southwestern Medical Center at Dallas

In Partial Fulfillment of the Requirements

For the Degree of

MASTER OF SCIENCE

The University of Texas Southwestern Medical Center at Dallas

Dallas, Texas

June, 2004

ACKNOWLEDGEMENTS

I would like to thank my mentor Dr. John D Minna for his constant support and guidance throughout my thesis project. I am very grateful to him for letting me have this opportunity to work with him. I have learnt a lot of things which will help me for my research work in the future. I really admire his experience and knowledge of the field.

I would also like to express my gratitude towards my committee members Dr. Harold Garner, Dr. Khosrow Behbehani and Dr. Jerry Shay for their guidance and encouragement. Dr. Garner's valuable advice regarding the computational aspects of the project has been a great help. He has also been a good mentor to me throughout my masters. Dr. Behbehani and Dr. Shay have been a very valuable source of support while guiding me in various aspects of the project. I also would like to thank Dr. Michael White, Dr. George DeMartino, Dr. Cezary Wojcik, Dr.Gazdar and Dr Shay again for allowing me to access the siRNA sequence data from their labs, which helped me to validate my thesis. I wish to acknowledge Dr. Alexander Pertsemlidis as well as David Trusty for helping me with the computational aspects of my project.

I would like to thank Dr. Luc Girard, Dr. D. Tripathy, Kimberly Tomenga and Shelly Sheridan for helping me with my chemosensitivity project. Dr. Luc Girard has also helped me with my various computational projects. I would like to thank Noriaki Sunaga as well as David Shames for helping me understand the basic principles of siRNA design. David has

also helped me understand the biochemistry better. Overall, my stay in Dr. Minna's lab was really pleasant because of the friendly atmosphere. Thanks to all the Minna lab members.

I wish to thank Dr. Peter Antich, Chairperson of the Biomedical Engineering program at The University of Texas Southwestern Medical Center at Dallas as well as Kay Emerson, Biomedical Engineering program assistant, for their help and encouragement.

A special thanks to Aalok, who has not only helped me with the technical aspects of the project but has also been a source of positive energy and support throughout my masters. I would like to thank Mike and Aalok for reviewing my thesis. I wish like to thank my friends Kiran, Amol, Dhara, Danielle, Nishat, Abhijeet, Amit, Deepa, Sowmya, Dheeru for their good wishes and encouragement.

Finally, I would like to thank my mother from the bottom of my heart for being my source of strength. I wish to thank my sister and friend for all her love and support.

PART I: TO DEVELOP A SMALL INTERFERING RNA (SIRNA) DESIGN AND INFORMATION RESOURCE TO FACILITATE GENETIC MANIPULATION OF HUMAN CELLS.

PART II: TO PARTICIPATE IN DEVELOPING MICROARRAY BASED GENE EXPRESSION SIGNATURES FOR IN VITRO DRUG SENSITIVITY AND RESISTANCE FOR BREAST CANCER.

Publication No. _____

Jyoti Khetsi Shah, M.S.

The University of Texas Southwestern Medical Center at Dallas, 2004

Supervising Professor: John D. Minna, M.D.

Part I: Small interfering RNAs (siRNAs) have revolutionized our ability to study the effects of altering the expression of single genes in mammalian (and other) cells through targeted knockdown of gene expression. In the past, there were a set of rules designed to develop siRNA which worked efficiently in most cases. There was further refinement

performed in these rules in some modern research analyses which attempted to address the question of what most closely determines siRNA functionality. I have designed and implemented a new software tool siRNA Information Resource ('sIR') that incorporates the most recent refinements in the design algorithm in order to provide fast and efficient siRNA design. sIR is a web-based computational tool which takes these existing rules for designing synthetic siRNAs and puts them in a software architecture that allows the researcher to design siRNAs for every gene. It also provides a database containing information about already developed siRNA and thus allows the researcher to access the siRNA information database consisting of siRNA information from literature and various other sources. This will ultimately help in future siRNA related discoveries. It also includes a scoring system which helps in rational selection of efficient siRNA. sIR was successfully validated using already designed and developed target siRNA sequences.

Part II: One of the major problems in using chemotherapy to treat cancer is whether patients, whose tumors do not respond to one drug, would respond to another. Thus, it would be very useful if one could rationally select the appropriate chemotherapy for each patient's tumor. We are asking is whether tumor gene "expression signatures" detected by microarray analysis could identify a set of genes correlating with sensitivity or resistance to a particular drug. A large panel of breast cancer cell lines was tested with cisplatin, paclitaxel, vinorelbine, doxorubicin and gemcitabine, *in vitro* using a colorimetric assay to determine the concentration of drug that gives 50% growth inhibition (IC_{50}). Gene expression profiles were also performed using Affymetrix chips and the two data sets were merged. It was

found that a panel of ~100 genes were significantly up regulated (4 fold or more) for each drug in resistant cells. As an alternative approach, Pearson correlations between each gene expression data and each drug IC50 across all cell lines analyzed were determined. A positive correlation for a pair of gene and drug indicates the gene may be associated with resistance to the drug whereas a negative correlation would associate that gene with sensitivity to the drug. Some of these genes might be associated with the drug mechanism of action. We conclude that gene expression signatures do exist for individual breast tumor cell chemosensitivity and these could be of clinical significance.

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

CHAPTER 1

Introduction

In the new millennium, bio-computational tools have emerged as valuable resources for experimental design and analysis of the enormous amount of data available after completion of the human genome project. The main purpose of these tools is to obtain faster results, while minimizing human errors. The effort in this thesis is to develop such tools as well as to conduct research using these tools. This thesis is composed of two parts. The first part addresses the development of a computational tool called "siRNA Information Resource", to develop a small interfering RNA (siRNA) design. It also provides a database consisting of siRNA information to facilitate genetic manipulation of human cells. The second part involves participation in developing microarray based gene expression signatures for *in vitro* drug sensitivity and resistance of breast cancer cell lines. It includes development of the relevant computational tools to facilitate experimental design and analysis.

This chapter introduces the basic concepts involved in each of these projects. The subsequent chapters describe these research and development projects in detail.

1.1 RNAi Mechanism

Within a very short time, “RNA interference” (RNAi) has become an important technique that has gained wide acceptance and use by the scientific community. RNAi is popularly used as a method to investigate gene function in a variety of organisms [1]. Dissection of signaling pathways and study of cell growth and division are also applications of RNAi in cancer biology.

RNAi can be described as a process, in which double stranded RNA (dsRNA), is known to induce posttranscriptional gene silencing. Posttranscriptional gene silencing results in a decrease in the steady state level of a specific messenger RNA (mRNA) through sequence-specific degradation of the transcribed mRNA, without changing the target gene transcription rate [2]. This can lead to a reduced expression of the target gene, also known as a “knock down”. RNAi is present in most of the eukaryotes [3]. Small interfering RNA or “siRNA” refers to synthetic small interfering RNAs constructed as dsRNA which can be transfected into cells to specifically silence the expression of mRNAs to which they are complementary.

Specific mRNA degradation can be thought of as a natural function of RNAi as it can prevent transposon and virus replication. RNAi can protect the genome against invasion

by mobile genetic elements such as transposons and viruses, which produce unnatural RNA or dsRNA in the host cell when they become active [2].

When long double stranded RNA is introduced into a cell, it enters a cellular pathway known as the RNAi pathway. This long double stranded RNA is first processed into double stranded small interfering RNA approximately 21-23 nucleotides in length by an RNase III-family enzyme called Dicer. This is known as the "Initiation" step. Then, the siRNAs assemble into endoribonuclease-containing complexes known as RNA-induced silencing complexes (RISCs), unwinding in the process. These activated RISCs are then guided by the siRNA to complementary RNA molecules. Then the "Effector" step takes place, which involves cleavage and destruction of the cognate RNA. This process is depicted in Figure1.1.1 [4].

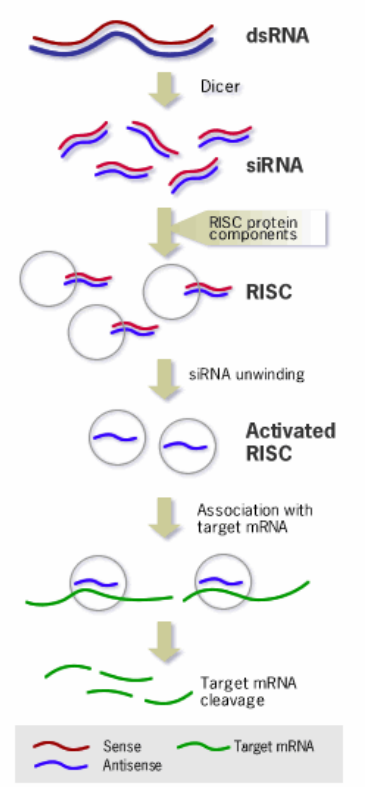


Figure 1.1.1: Induction of post-transcriptional gene silencing with the introduction of dsRNA [4]. This figure depicts the two step process of "Initiation" and "Effector" phase of RNAi mechanism. In the "Initiation" step, the small interfering RNAs (siRNAs) are formed by an enzyme called Dicer. Then, the complexes known as RNA-induced silencing complexes (RISCs) are formed. These RISCs are then guided by the siRNA to complementary RNA molecules. In the "Effector" step cleavage and destruction of the cognate RNA occurs.

1.1.1 Application in Cancer Biology

RNAi technology has proven its usefulness in systematically deciphering the functions and interactions of thousands of genes. This can be a very useful tool for cancer research [5].

Most human cancers are characterized by abnormal gene expression, some of which are important in the initiation or progression of disease [6]. It may be of therapeutic importance to silence these aberrant genes or bring down their expression to a significantly lower level. Since targeting of gene expression is very specific using siRNA; it may be possible to develop more specific and thus less toxic cancer therapies.

There have been studies where siRNA has been used to obtain dose-dependent inhibition of a gene expression in human colon cancer cells. Higher TS (thymidylate synthase) expression was found to increase drug resistance to TS-targeted compounds. siRNA developed against human TS mRNA resulted in a dose dependent inhibition of TS expression but had no effect on the expression of alpha-tubulin or topoisomerase I [7]. This study supports the target specific behavior of siRNA.

In another study, gemcitabine-induced cytotoxity, both *in vitro* and *in vivo* was increased by suppressing the expression of Focal adhesion kinase (FAK) using siRNA. It was also observed that FAK siRNA did not affect cellular proliferation or apoptosis in the

absence of gemcitabine. FAK siRNA treatment suppressed Akt gene activity, which may contribute to its chemosensitizing effect [8]. This shows that the siRNA mechanism can be used in potential therapeutic studies and as a research tool to evaluate gene function.

RNAi mechanism has been used to study the function of polo-like kinase-1(PLK1) in breast cancer, lung cancer, and cervical cancer. It has also shown that Brk (PTK6), a non-receptor protein tyrosine kinase, potentially functions as an 'adapter' by playing a role in proliferation of breast carcinoma cells. RNAi has also helped in studying responses of UCH-L1 (Neuronal ubiquitn C-terminal hydrolase) and E2F1 in lung cancer cell lines [9, 10, 11, and 12].

All these studies prove that RNAi is a very important research tool to study individual gene functions as well as has therapeutic importance in cancer research.

1.2 Introduction to siRNA Information Resource.

siRNA Information Resource ('sIR') is a "web-based" computational tool that aids in designing the target sequence for siRNA, as well as provides a database containing useful information about already developed siRNAs.

There are established rules for designing the most effective siRNAs but these methods all require "hand and eye" computation. The goal of this project was to automate

this process and ultimately pre-calculate siRNAs for all known human genes. Also it will be very useful to have all the pre-existing information on siRNA in one place. This will allow the researcher to access the siRNA information resource and will ultimately help in future siRNA related research.

sIR uses various available information from genetic databases such as Refseq [13], SOURCE [14], NCBI [15] as well as BLAST [16] alignment software to facilitate the siRNA design algorithm.

It also uses modern databases such as PostgreSQL [17] and modern as well as traditional scripting languages such as Perl [18], Bioperl [19], PHP [20], HTML, JAVA etc. to implement the siRNA target designer algorithm.

The basic architecture of siRNA Information Resource (sIR) is depicted in Figure 1.2.1.

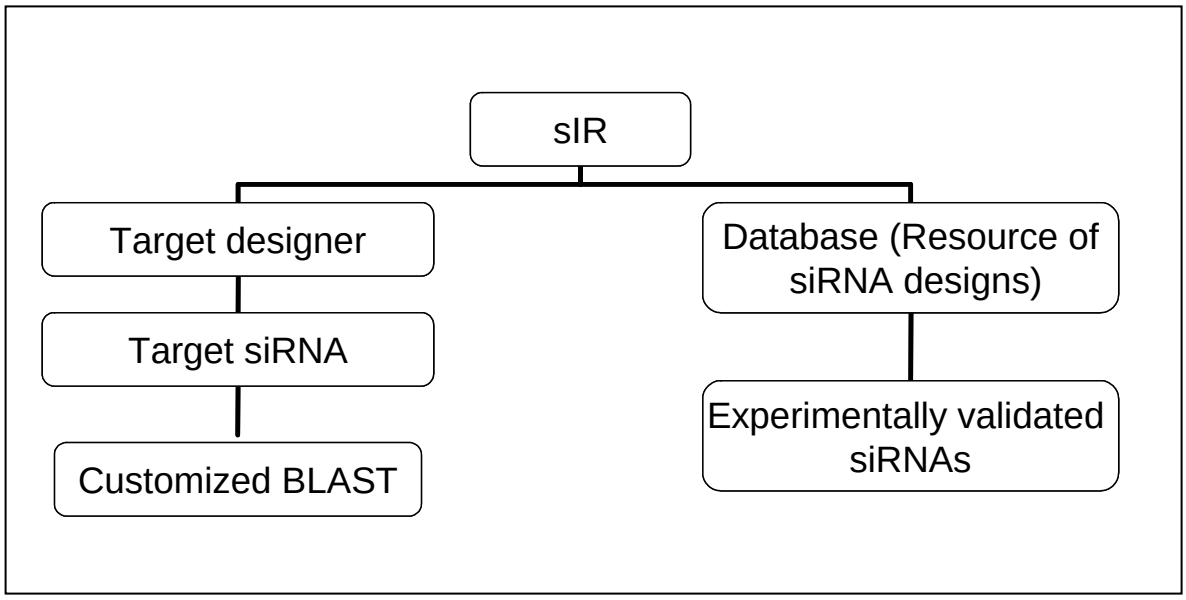


Figure 1.2.1: Overview of siRNA Information Resource ('sIR'). The block diagram above shows that sIR essentially consists of siRNA target designer, which designs siRNA and BLASTs the result automatically, and database of siRNA designs which consist of experimentally validated siRNAs.

siRNA Information Resource ('sIR') mainly consists of two modes:

a. siRNA target designer

b. siRNA resource.

The siRNA target designer determines the target siRNA depending on the user chosen parameters or default parameters. It also follows a logical choice of parameters.

Then it performs custom BLAST on the user selected target sequence against the human genomic database provided by NCBI.

The database or siRNA resource mode allows the researcher to search for existing and developed siRNAs. The subsequent chapters will include detailed information on each of these modes.

1.3 Chemosensitivity and Gene expression profiles of Breast cancer cells.

Breast cancer is the currently the second leading cause of cancer deaths in women (after lung cancer) and is the most common cancer among women [21]. Traditional as well as modern chemotherapeutic agents with novel mechanism of action such as cisplatin, paclitaxel, docetaxel, vinorelbine, gemcitabine, doxorubicin etc. are used for the treatment of breast cancer tumors.

Tumor response to these chemotherapeutic agents varies from one patient to another. One of the major problems in using chemotherapy is whether patients, whose tumors do not respond to one drug or combination of drugs, would respond to another. It would be very useful if oncologists were able to select ahead of time the chemotherapy that would be most effective for each individual patient's tumor.

This can be done by testing *in vitro* chemosensitivity of a breast tumor or by using promising modern techniques such as microarray gene expression profile. Gene expression profiling has several advantages over *in vitro* chemosensitivity testing. This will be explained in detail in subsequent chapters. One of the major goals of this project is to correlate *in vitro* sensitivity or resistance to particular gene expression profiles. This should help determine existing genetic signatures that may be predictive for future prospective studies.

A collection of computational tools was developed in order to store raw data, facilitate analysis as well as to set up experiments. These tools were mainly developed using Microsoft Excel and Visual Basic macros.

CHAPTER 2

Objectives

sIR was mainly developed for two purposes:

1. To develop and implement siRNA target design.
2. To provide a database consisting of information available on experimentally tested siRNAs.

In the past, there were a set of rules designed to develop siRNA [22]. These design rules worked efficiently in most of the cases. There was further refinement performed in these rules in some recent research papers [23, 24, 25]. The research was trying to address the question of what most closely determines siRNA functionality. sIR tries to incorporate the most recent refinements in the design algorithm in order to provide a efficient siRNA design. Moreover, since it automatically designs these target siRNAs, it saves a lot of time. sIR target designer can also be used as a research tool to find better siRNA designs as it allows the user to try user-defined patterns.

2.1 Features of siRNA Information Resource.

This section describes important features of this software.

- sIR is a "Web-based" computational tool. Hence it is easy to access and use. It can be accessed at the following URL:

  http://biotools.swmed.edu/siRNA
- It can design 23nt (nucleotide) siRNA from mRNA sequence and the pattern selection of 23nt sequence is adaptable to any future changes in the siRNA design methods.
- Input can be in the form of actual sequence data or the accession number of the Refseq database.
- The user can choose to design using standard parameters such percentage GC content, user defined pattern, avoiding nucleotide runs etc. The user can also avoid nucleotide runs or choose open reading frame (ORF). Here GC content refers to the percentage of the bases G or C in a sequence.
- There is a scoring system associated with the design of individual siRNA target sequences. This can help in choosing better siRNAs.
- Resource database: This provides information on developed and existing siRNAs. It includes both functional as well as non functional siRNA sequences. The database also stores and displays images relevant to the developed siRNA such as western blots, RTPCR etc.

- Accession finder: Allows the user to find an accession number with a gene name alias as input.
- The user can choose the target designer output and BLAST it locally.
- It provides a customized BLAST output, which helps in the quick interpretation of the BLAST output.
- It provides a variety of databases to BLAST the target siRNA sequences against. It filters the BLAST output to display sequence homology greater than 75%.
- It also gives a choice to restrict BLAST output with user defined homology.
- Updating of the resource database is possible through a password protected form.
- Images can be uploaded into the database using the update form.

CHAPTER 3

Materials and Methods

3.1 Computational Tools.

The siRNA Information Resource was built on a Linux based server. Its configuration and hardware details are mentioned in the Appendix. The databases involved were implemented locally to speed up the process. PHP and Perl (CGI) scripts were written to communicate with databases from the web pages. These scripts were able to retrieve data from the database as well as update the database using a web interface.

Database applications were performed using PostgreSQL database. PostgreSQL is one of the most advanced database servers available. Some of the important features of the PostgreSQL include unique data types, object-relational database, multiple procedural languages, extensibility, referential integrity, user defined data types etc. PostgreSQL is an open source database; hence it is available for free [17]. PostgreSQL can be used from most of the major programming languages such as C, C++, Perl, Python, Java, Tcl and PHP [20]. Thus PostgreSQL was chosen for implementing the database application. Other scripting languages such as Perl and Bioperl were used to implement the siRNA target designer algorithm. Perl is especially popular for its string operations as well as other bioinformatics applications. Bioperl is a collection of Perl modules that facilitate the development of Perl scripts for bioinformatics applications [19]. Bioperl was mainly used to parse BLAST output

in this application. PHP is a general-purpose programming language and is mostly used for building dynamic web pages. PHP and HTML scripts along with CGI module in Perl were used to create web pages.

As mentioned earlier, the siRNA Information Resource consists of two modes; target design mode and siRNA resource (database) mode. These modes are explained in detail in the subsequent sections.

3.2 Target Design mode

3.2.1 Block diagram

The block diagram below, gives the general overview of the target design mode. The various blocks in the input section of the block diagram show the various types of inputs accepted by the target design algorithm and various blocks in the output section describe the output and output post-processing. The researcher can use a sequence or accession number as input parameter. If user does not have either of the information then the accession number can be retrieved using a locally downloaded “SOURCE” database with gene aliases as input. If the user enters a sequence as input, it can be directly taken by the algorithm and processed, but if an accession number is used as input, then the program automatically queries the “RefSeq” database and retrieves the sequence for further processing. The output of this mode is in the form of target sequences. User can choose one or more of these sequences to BLAST against Unigene (Human or Mouse) or Human Genomic databases.

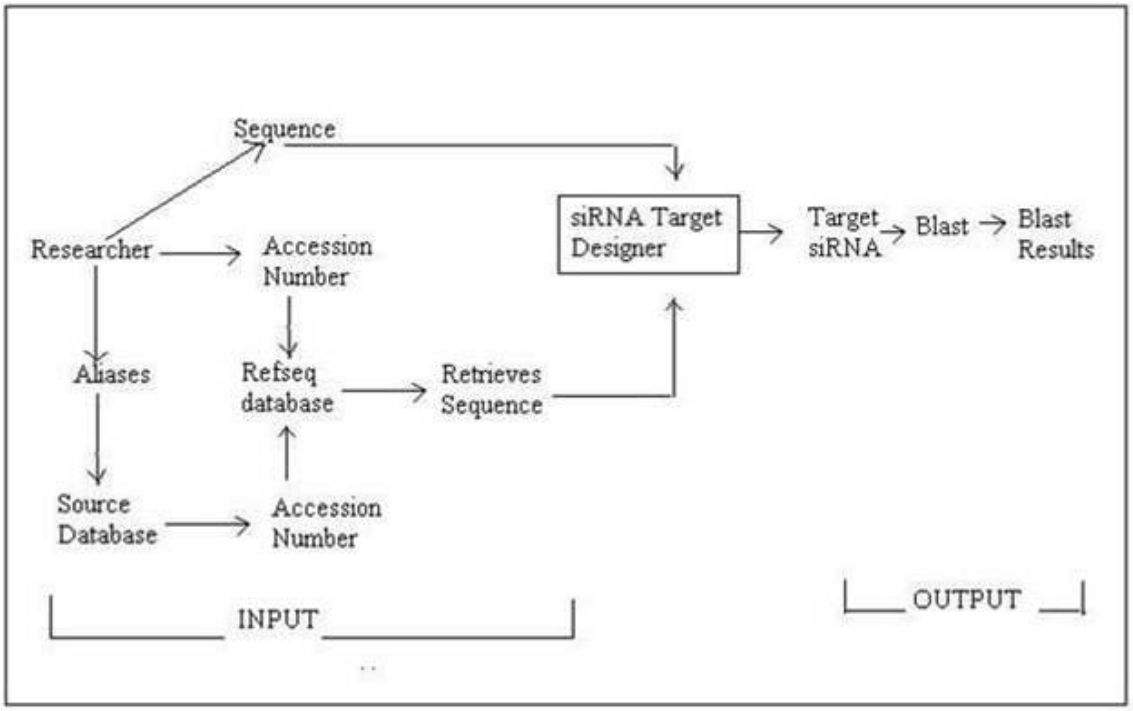


Figure 3.2.1.1 Block diagram of siRNA target designer.

3.2.2 Input to siRNA target designer

As mentioned earlier, the input to the siRNA target designer can be a plain text sequence or accession number. This input can be entered in a text box provided on the main form of siRNA information resource. The next figure is a snap shot of the siRNA Information Resource form. The Help page can activated by clicking on the links provided on the main form.

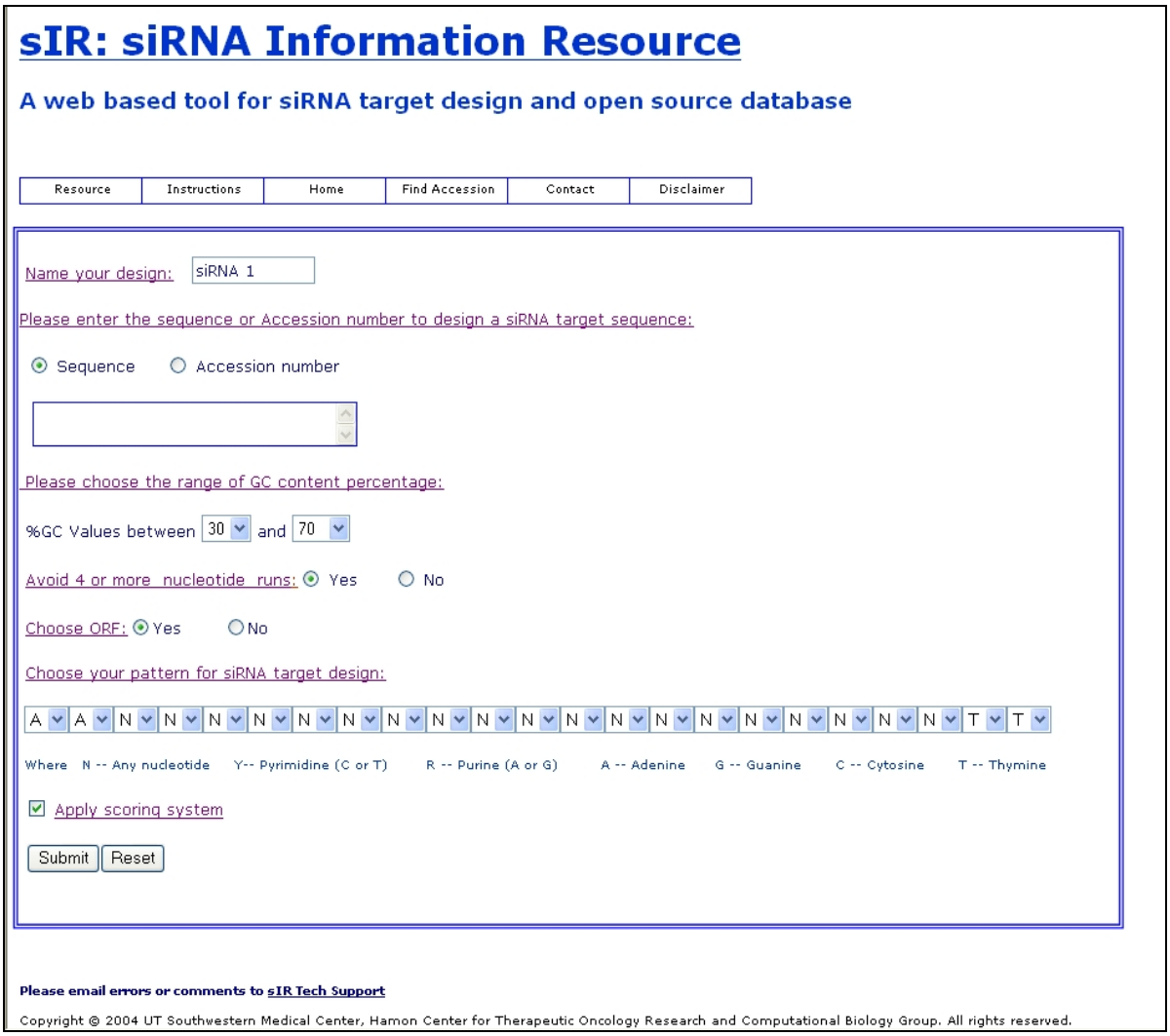


Figure 3.2.2.1: Snapshot of siRNA Information resource form.

If the user does not have either of the information, then the "Find Accession" option on the menu bar of siRNA Information Resource can be used. The next figure depicts the snapshot of the webpage used to find the accession number.

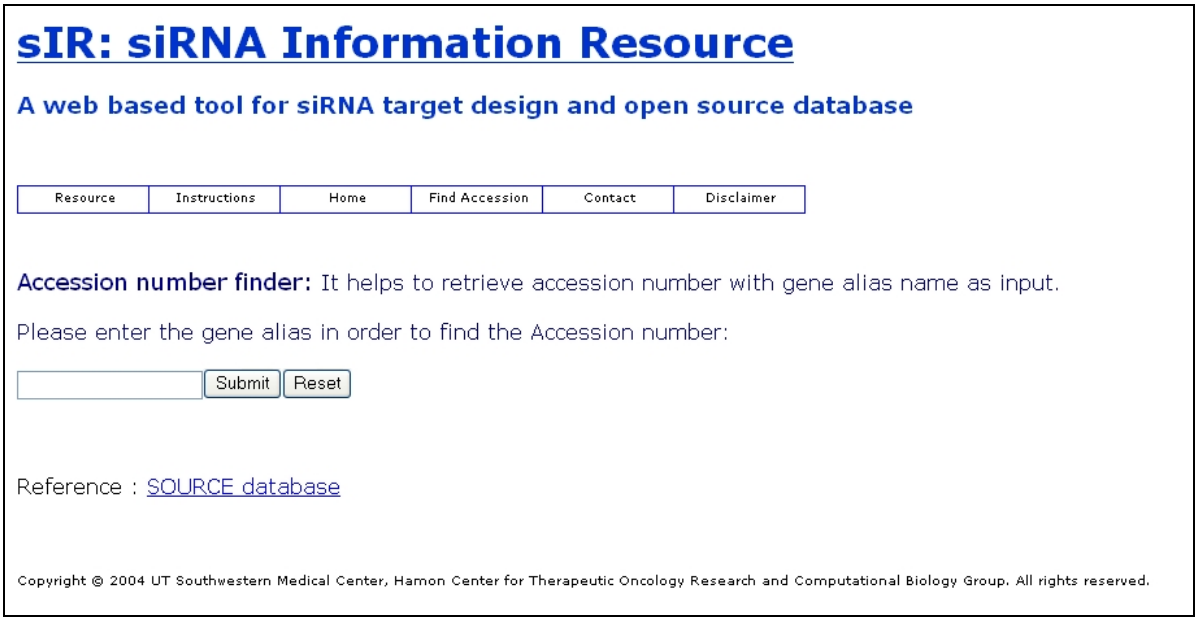


Figure 3.2.2.2 Snapshot of Accession number finder form.

The user can enter gene alias name in the text box provided on the page above and the accession number can be retrieved. The next figure shows the accession numbers retrieved with "telomerase" as the input. The user can then choose the appropriate accession number depending on the description and use it as an input.

gene_name	aliases
- NM_003219	||TERT||hTERT||TRT (LL)||TP2 (LL)||TCS1 (LL)||hEST2 (LL)||EC 2.7.7.-||Telomerase catalytic subunit||TELOMERE REVERSE TRANSCRIPTASE||Telomerase reverse transcriptase||telomerase reverse transcriptase isoform 1||telomerase reverse transcriptase isoform 4||telomerase reverse transcriptase isoform 3||telomerase reverse transcriptase isoform 2||
- NM_006601	||TEBP||Hsp90 co-chaperone||Telomerase-binding protein p23||Progesterone receptor complex p23||UNACTIVE PROGESTERONE RECEPTOR, 23-KD||unactive progesterone receptor, 23 kD||likely ortholog of mouse telomerase binding protein, p23 (LL)||
- NM_007110	||TLP1||VAULT2||TP1||p240||TEP1||telomerase-associated protein 1||telomerase protein component 1||

Figure 3.2.2.3: Output of the accession finder with 'Telomerase' as an example input.

3.2.3 Adjusting parameters

The researcher can choose from various parameters to optimize the siRNA design. The parameters are as follows:

Pattern: User can choose a user defined pattern for siRNA design. The main form has 23 scroll bars to choose a pattern. Each scroll bar has the following 7 choices.

1. A -- Adenine
2. G -- Guanine
3. C -- Cytosine
4. T – Thymine
5. N -- Any nucleotide (A, T, G or C).
6. Y-- Pyrimidine (C or T).
7. R -- Purine (A or G).

The default for pattern is "AAN (19) TT".

If the user defined pattern or default pattern is not found by the program, then it follows a logic built in the algorithm to find most commonly used patterns in a hierarchical manner.

GC content: User can choose the GC content using two scroll bars which limits the pattern selection to a GC content in that range. The default range for GC content is between 30% and 70%.

Avoid 4 or more nucleotide runs: Sometimes researchers prefer to avoid four or more nucleotide runs together in a siRNA sequence. This can be done by choosing appropriate radio button given on the form. The default for this parameter is "Yes".

Choose ORF: This radio button is used to select an open reading frame (ORF) in an mRNA sequence before looking for patterns. The default for this parameter is "Yes". This will be explained in detail in subsequent sections.

Apply scoring system: This check box is used to apply an optional scoring system while filtering target sequences for the output depending upon the input parameters. This scoring system will help the user to determine best possible siRNA. This scoring system is explained in detail in subsequent sections.

3.2.4 Target designer algorithm

The target designer algorithm takes the input as well as various parameters from the input form and screens the mRNA sequence for target sequences with input parameters. The next figure demonstrates the flowchart and basic logic of the algorithm. If the radio button for ORF is chosen, then the program pre-processes the input sequence by choosing an open reading frame 50nt below the 5' end of mRNA and 50nt above the 3' end of mRNA. Then it chooses appropriate target sequences using a user defined pattern. If a user defined pattern is not found, then it looks for the following patterns in an hierarchical order [44], where the number 'n' in parenthesis after a character means that character repeating 'n' number of times.

a. AAN (19) TT

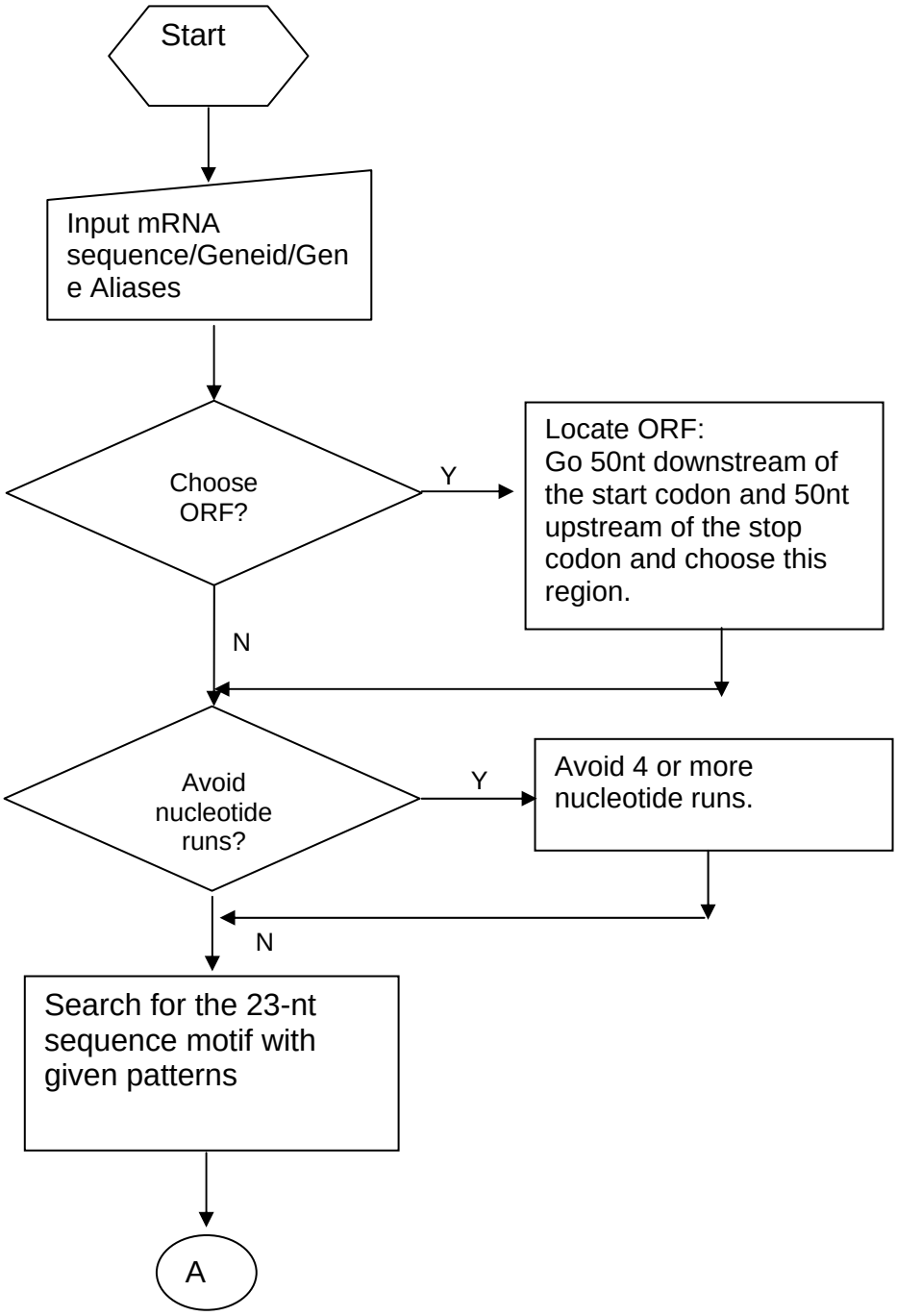
b. AAN (21)

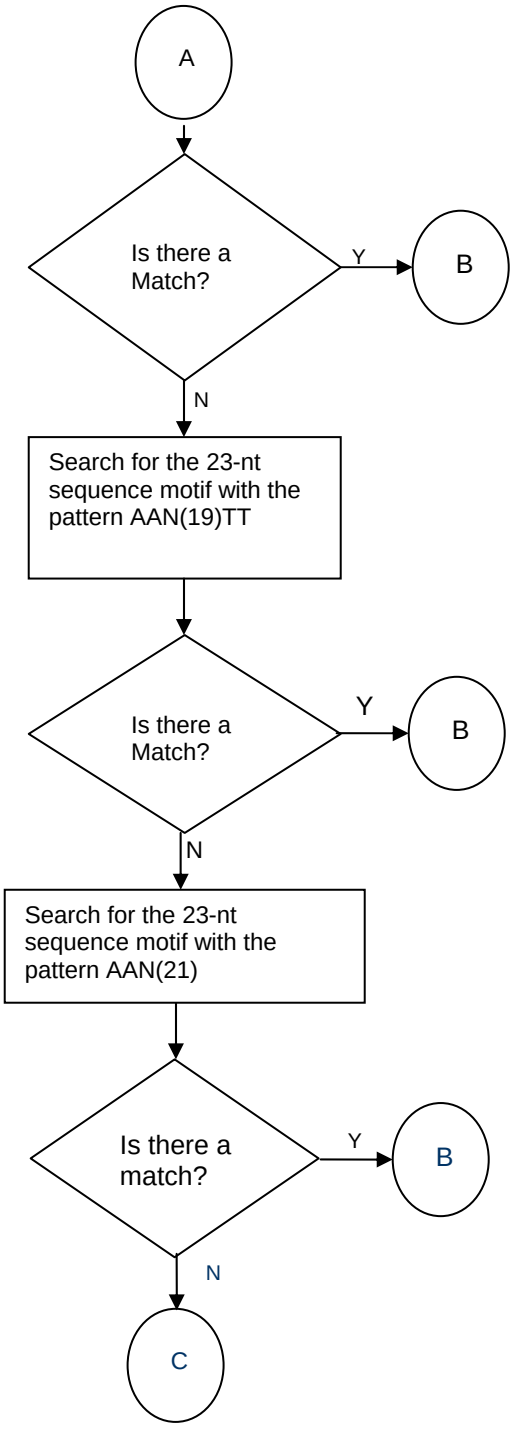
c. NAN (19) TT

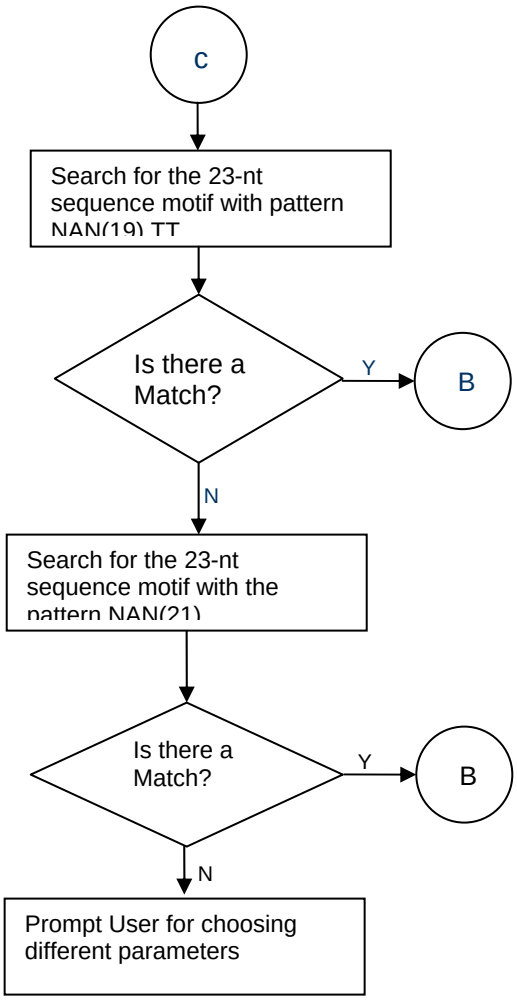
d. NAN (21)

If none of these patterns are found, then it prompts the user to change the input parameters. After a pattern is found, it filters the target sequences according to the GC content range given by user or default. It also scores the target sequences individually and sorts the target sequences in the descending order of the score. After the user has chosen target sequences to BLAST against the user chosen database, it gives a customized BLAST output displaying BLAST results with homology greater than 75%.

The figure 3.2.4.1 is a flowchart depicting the flow of logic in the algorithm used in the program.




A
Is there a Match?
Y
B
N
Search for the 23-nt sequence motif with the pattern AAN(19)TT
Is there a Match?
Y
B
N
Search for the 23-nt sequence motif with the pattern AAN(21)
Is there a match?
Y
B
N
C


C
Search for the 23-nt sequence motif with pattern NAN(19) TT
Is there a Match?
Y
B
N
Search for the 23-nt sequence motif with the pattern NAN(21)
Is there a Match?
Y
B
N
Prompt User for choosing different parameters

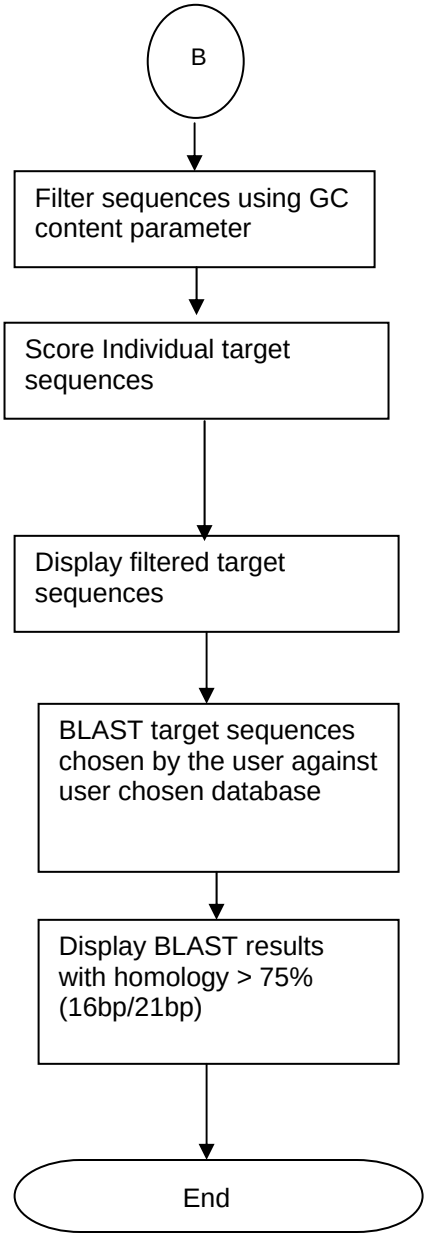


Figure 3.2.4.1: Flow chart depicting the algorithm of siRNA target designer mode.

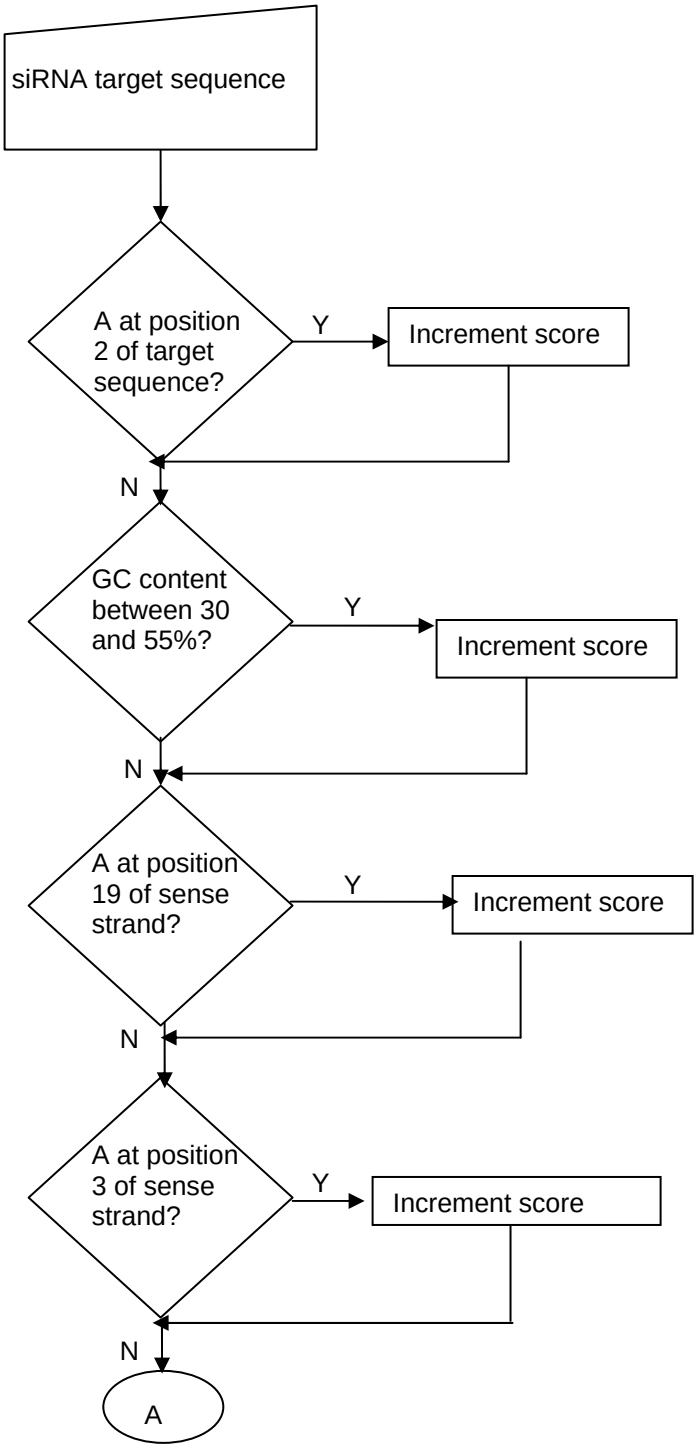
3.2.5 Scoring system

The relationship between siRNA sequence and RNAi effect was extensively analyzed in many of the recent studies. Most of the studies have shown that the RNAi effect is a function of siRNA sequence. It was also shown that presence of certain bases in a particular position, contributes more to the efficiency than other bases. Also, all the positions of a target sequence do not contribute equally to target recognition. The main goal of scoring individual target sequences was to come up with a rational design which has high probability of success. In order to achieve this goal, different rules in various research papers were compiled together to form a scoring system [22, 23, 24, 25].

It was found that the penultimate nucleotide of the antisense siRNA which is complementary to position 2 of 23nt target sequence should always be complementary to the targeted sequence. Mostly for simplification of chemical synthesis TT is used. Hence it increases the chances of having an efficient siRNA if the position 2 of the target sequence is 'A'. It was also found that moderately low amount of GC content contributes to efficiency. Hence, GC content between 30% and 55% was considered to be good for the design. Some research studies have shown that the presence of G/C content at the 5' end of the siRNA target sequence improves efficiency, whereas some studies have shown that there is no correlation between them [23, 25]. Most of the analyses have shown that, the presence of at least five A/U's at the 3' end of the sense strand increases the efficiency of the siRNA. Apart from these rules, there have been analyses based on individual positions of the siRNA

target sequence. Presence of ‘A’ at position 19 , presence of ‘A’ at position 3 and presence of ‘U’ at position 10 of the sense strand are known to positively effect the siRNA efficiency, whereas the presence of ‘G’ or ‘C’ at position 19 and presence of ‘G’ at position 13 are known to negatively effect the siRNA efficiency [23].

Based on the above, a scoring system was developed. Its logical flow chart is shown in the figure 3.2.5.1.


siRNA target sequence
A at position 2 of target sequence?
Y
Increment score
N
GC content between 30 and 55%?
Y
Increment score
N
A at position 19 of sense strand?
Y
Increment score
N
A at position 3 of sense strand?
Y
Increment score
N
A

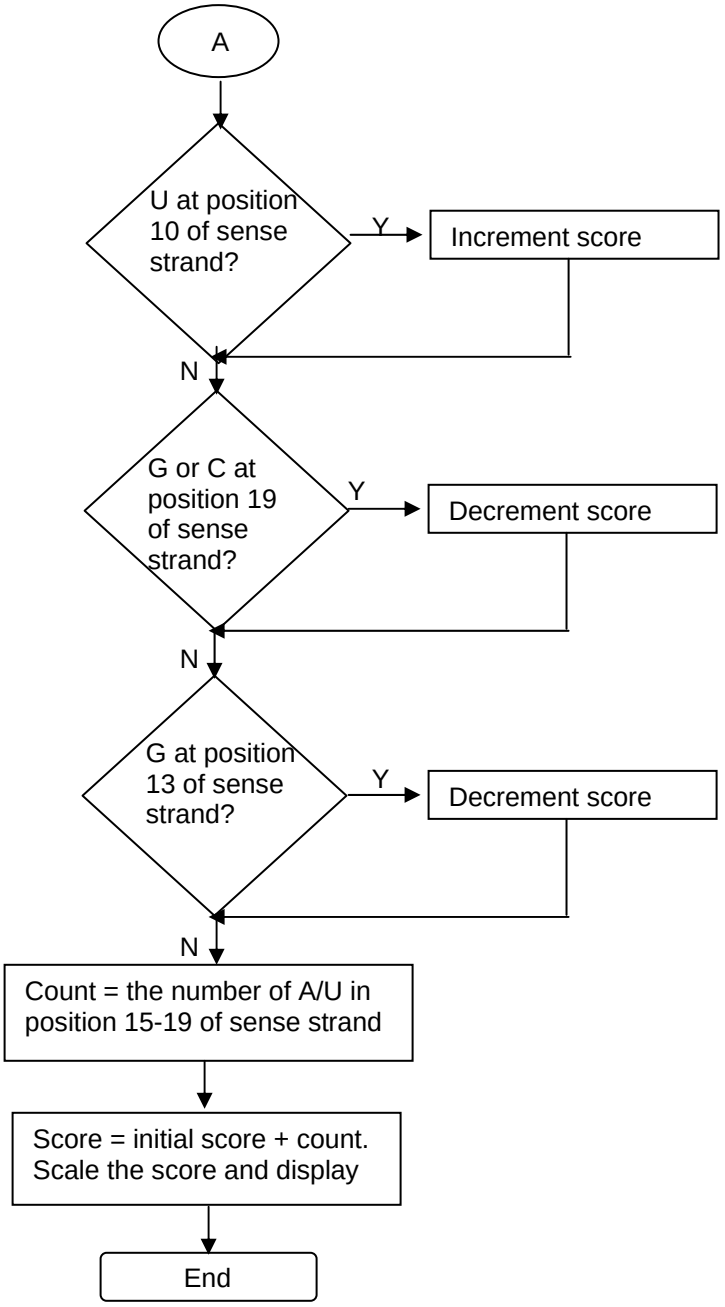


Figure 3.2.5.1 : Flowchart of scoring system for siRNA target sequences in sIR.

3.2.6 Output of siRNA target designer

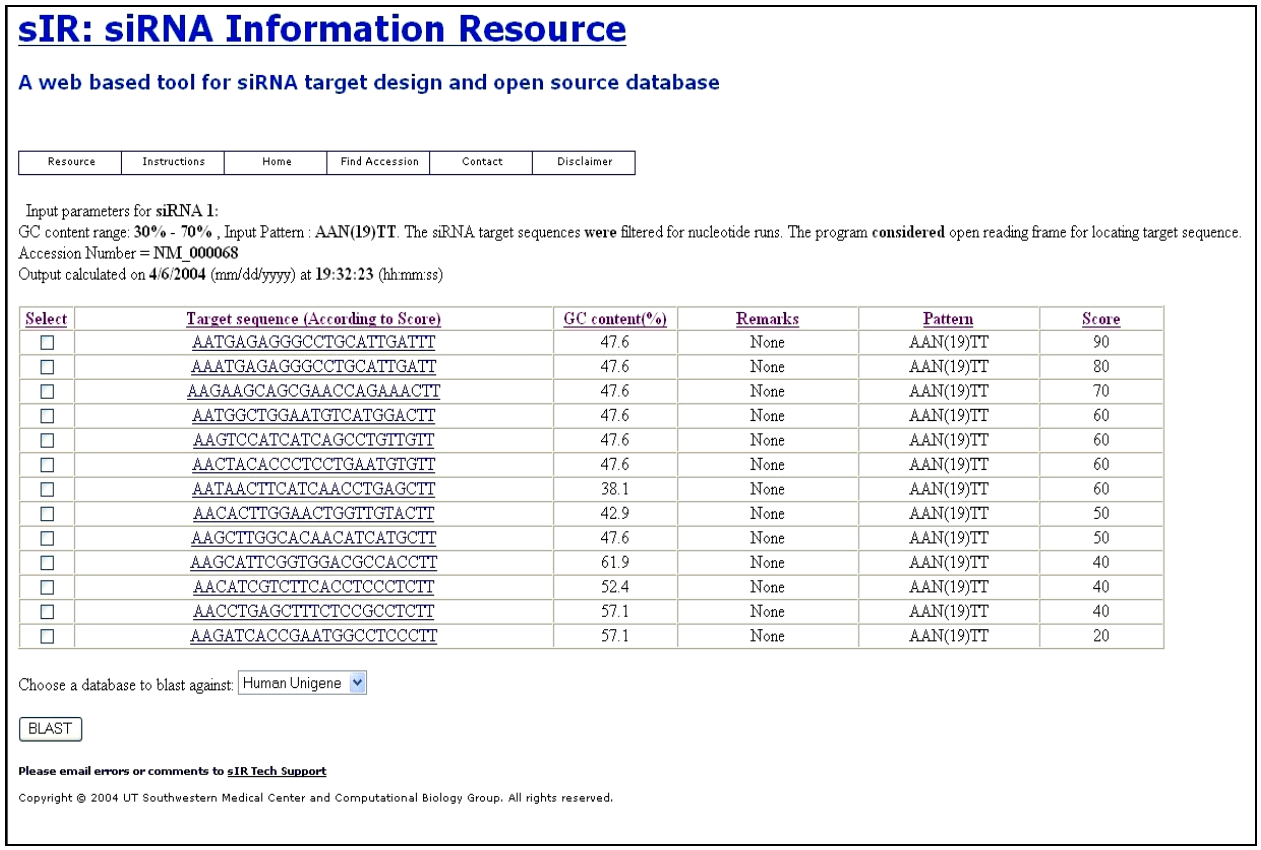
After the program has filtered target sequences and calculated scores for them individually, it displays the result in a tabular format along with GG content value, pattern, score and remarks. The 'Remarks' column in the table is used to specify if more than four nucleotide runs were found in the target sequence. The 'Pattern' column in the table is used to specify the pattern used to filter target sequences. There are also individual checkboxes provided with target sequences so that user can choose sequences to BLAST. User can sort the data using "Sort using Score" button. The program sorts the target sequences in the descending order of score, highest being the best score. The user can also choose from the following databases to BLAST the target sequences against:

a. Human Genomic.

b. Human Unigene.

c. Mouse Unigene.

The figure 3.2.6.1 is a snapshot of the output of siRNA target finder with example input as accession number 'NM_000068'.



sIR: siRNA Information Resource

A web based tool for siRNA target design and open source database

Resource | Instructions | Home | Find Accession | Contact | Disclaimer

Input parameters for **siRNA 1**:
GC content range: **30%** - **70%** , Input Pattern : **AAN(19)TT**. The siRNA target sequences **were** filtered for nucleotide runs. The program **considered** open reading frame for locating target sequence.
Accession Number = **NM_000068**
Output calculated on **4/6/2004** (mm/dd/yyyy) at **19:32:23** (hh:mm:ss)

Select	Target sequence (According to Score)	GC content(%)	Remarks	Pattern	Score
☐	AATGAGAGGGCCTGCATTGATTT	47.6	None	AAN(19)TT	90
☐	AAATGAGAGGGCCTGCATTGATT	47.6	None	AAN(19)TT	80
☐	AAGAAGCAGCGAACCAGAAACTT	47.6	None	AAN(19)TT	70
☐	AATGGCTGGAATGTCATGGACTT	47.6	None	AAN(19)TT	60
☐	AAGTCCATCATCAGCCTGTTGTT	47.6	None	AAN(19)TT	60
☐	AACTACACCCTCCTGAATGTGTT	47.6	None	AAN(19)TT	60
☐	AATAACTTCATCAACCTGAGCTT	38.1	None	AAN(19)TT	60
☐	AACACTTGGAACTGGTTGTACTT	42.9	None	AAN(19)TT	50
☐	AAGCTTGGCACAACATCATGCTT	47.6	None	AAN(19)TT	50
☐	AAGCATTCGGTGGACGCCACCTT	61.9	None	AAN(19)TT	40
☐	AACATCGTCTTCACCTCCCTCTT	52.4	None	AAN(19)TT	40
☐	AACCTGAGCTTTCTCCGCCTCTT	57.1	None	AAN(19)TT	40
☐	AAGATCACCGAATGGCCTCCCTT	57.1	None	AAN(19)TT	20

Choose a database to blast against: Human Unigene

BLAST

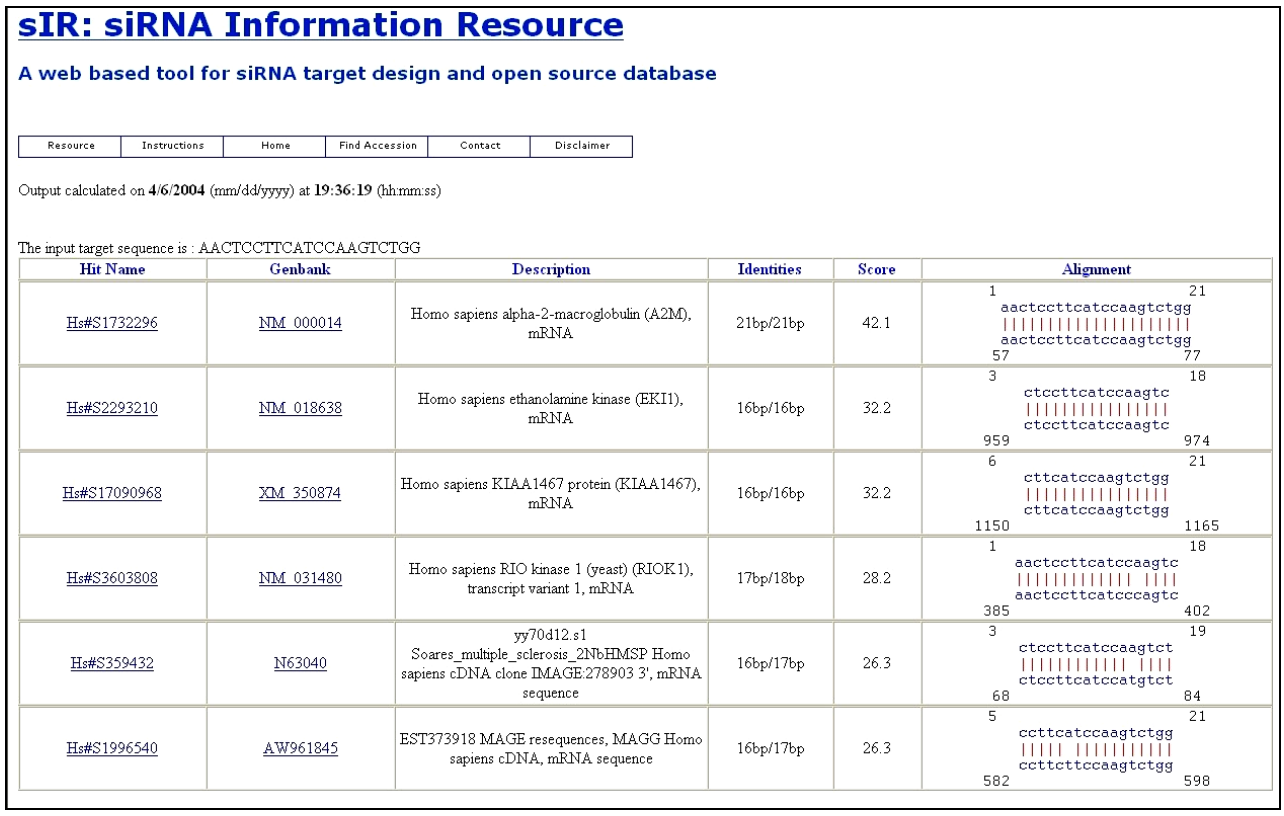
Please email errors or comments to sIR Tech Support

Figure: 3.2.6.1 Output of siRNA target finder for example accession number 'NM_000068'. The figure displays the input parameters used for this program along with the date and time information at which the output was calculated. It also provides a choice for database which can be used to BLAST the sequence against.

Each of the individual target sequences displayed in the above figure can be clicked to view information on the target sequence which includes position in mRNA, sense strand and antisense strand.

3.2.7 BLAST

The sequences chosen from the list of target siRNAs (Figure 3.2.6.1) are then BLASTed against the user selected database. This BLAST is performed locally on the server. The program then gives a customized BLAST output in easily readable tabular format. The program also filters BLAST output depending on the percent of homology found and displays the BLAST results where homology was found to be greater than 75%. Figure 3.2.7.1 depicts the BLAST output for sequence "AACTCCTTCATCCAAGTCTGGTT" and the database "Human Unigene".



sIR: siRNA Information Resource

A web based tool for siRNA target design and open source database

Resource | Instructions | Home | Find Accession | Contact | Disclaimer

Output calculated on **4/6/2004** (mm/dd/yyyy) at **19:36:19** (hh:mm:ss)

The input target sequence is : AACTCCTTCATCCAAGTCTGG

Hit Name	Genbank	Description	Identities	Score	Alignment
Hs#S1732296	NM_000014	Homo sapiens alpha-2-macroglobulin (A2M), mRNA	21bp/21bp	42.1	1 aactccttcatccaagtctgg 21 |||||||||||||||||||| aactccttcatccaagtctgg 57 77
Hs#S2293210	NM_018638	Homo sapiens ethanolamine kinase (EKI1), mRNA	16bp/16bp	32.2	3 ctccttcatccaagtc 18 |||||||||||||||| ctccttcatccaagtc 959 974
Hs#S17090968	XM_350874	Homo sapiens KIAA1467 protein (KIAA1467), mRNA	16bp/16bp	32.2	6 cttcatccaagtctgg 21 |||||||||||||||| cttcatccaagtctgg 1150 1165
Hs#S3603808	NM_031480	Homo sapiens RIO kinase 1 (yeast) (RIOK1), transcript variant 1, mRNA	17bp/18bp	28.2	1 aactccttcatccaagtc 18 ||||||||||||| |||| aactccttcatcccagtc 385 402
Hs#S359432	N63040	yy70d12.s1 Soares_multiple_sclerosis_2NbHMSP Homo sapiens cDNA clone IMAGE:278903 3', mRNA sequence	16bp/17bp	26.3	3 ctccttcatccaagtct 19 |||||||||||| |||| ctccttcatccatgtct 68 84
Hs#S1996540	AW961845	EST373918 MAGE resequences, MAGG Homo sapiens cDNA, mRNA sequence	16bp/17bp	26.3	5 ccttcatccaagtctgg 21 ||||| ||||||||| ccttcttccaagtctgg 582 598

Figure 3.2.7.1: Customized BLAST output. The ouput displays the input target information along with the date and time of the calculation details. The output also provides clickable links to Genbank data as well as UniGene (http://www.ncbi.nlm.nih.gov/) data to retrieve the individual sequence information.

3.3 Open source database

Another mode of siRNA Information Resource is “siRNA Resource” mode where a database is provided with already developed siRNA information. This database helps the researcher to find a list of already developed and tested siRNAs. This list includes the siRNA sequences which have worked as well as the ones which failed to work. The database

also has clickable images of the siRNA tests for some of the siRNAs. This is very helpful as it can prevent the user from designing and testing failed siRNAs as well as provide information on working siRNAs. The block diagram shown in the next figure shows the basic architecture of 'Resource' database.

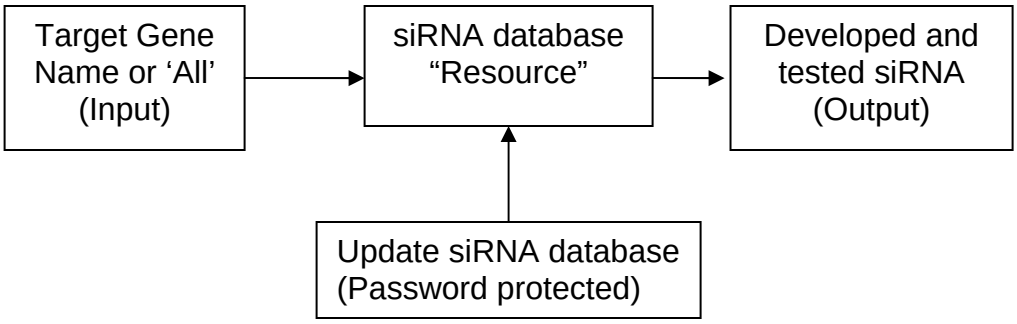


Figure 3.3.1: Basic architecture of siRNA "Resource" database.

As shown above, the input to the form can be a target gene name or the user can choose "All" option to display the whole database. The siRNA database is a PostgreSQL database. This database has a collection of developed and tested siRNAs along with their information. It also consists of information on working efficiency of the siRNA, relevant pictures (if any) along with publishing details. This database is then queried with the target gene name and the resultant output is displayed in a tabular format. User can click on the links with figures to get more details on the siRNA. The next figure shows the input form for accessing this database. This web page can be reached using the menu bar on the siRNA Information Resource web page by clicking on the "Resource" tab.

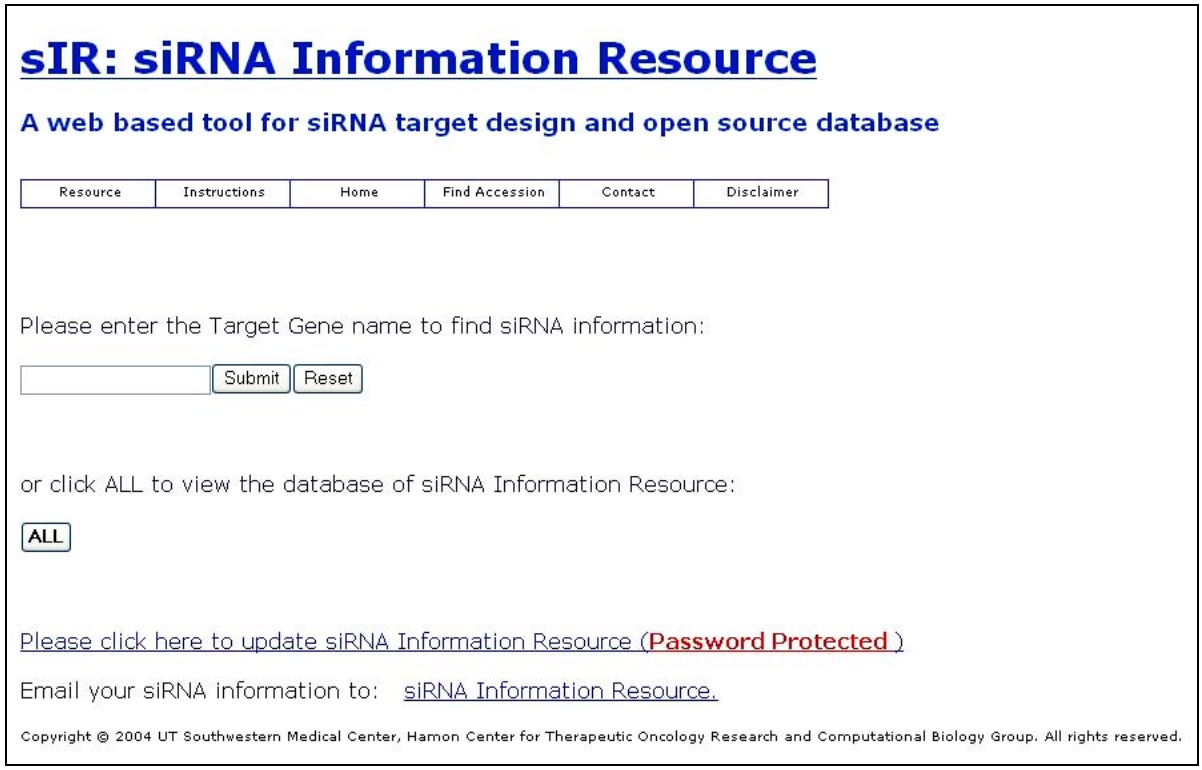


Figure 3.3.2: Snap shot of siRNA resource database input form.

The output for target gene name "Cav" for caveolin is depicted in figure 3.3.3 (a).

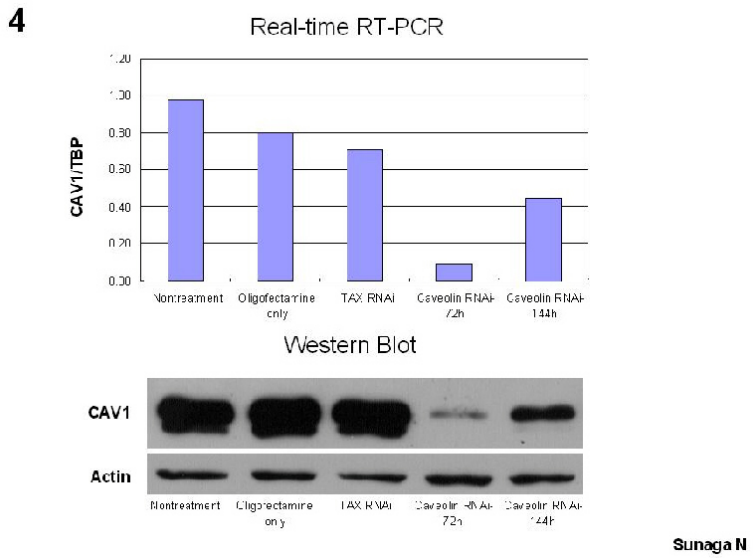
sIR: siRNA Information Resource

A web based tool for siRNA target design and open source database

Resource | Instructions | Home | Find Accession | Contact | Disclaimer

name	target_gene	target_sequence	sense_sirna	antisense_sirna	percent_gc	designer	efficiency	knock_down_efficacy_mrna	knock_down_efficacy_protein	publication	reference
CAV1	CAV1	GC AGA CGA GCT GAG CGA GAA GCA	AGA CGA GCU GAG CGA GAA GCA	CUU CUC GCU CAG CUC GUC UGC	57	White MA	Works (protein and mRNA)	CAV1	CAV1	Published	"Biochemistry 42:7967,2003"
CAV1-347	CAV1	AA CAT CTA CAA GCC CAA CAA CAA	CAU CUA CAA GCC CAA CAA CTT	GUU GUU GGG CUU GUA GAU GTT	57	Sunaga N	Works (protein)	CAV1-347	CAV1-347	Submitted	

(a)



(b)

Figure 3.3.3: (a) Output of the siRNA database for query "CAV". The links can be clicked to view siRNA figures. **(b) Figure for Caveolin knock down efficiency (mRNA).** This figure opens in a new browser window.

3.3.1 Update siRNA database.

It is very important to update the database with new siRNA information as well as figures. Hence, a webpage for updating siRNA database is provided and is password protected to maintain the authenticity of the data. The input form for updating the database is shown in the next figure.

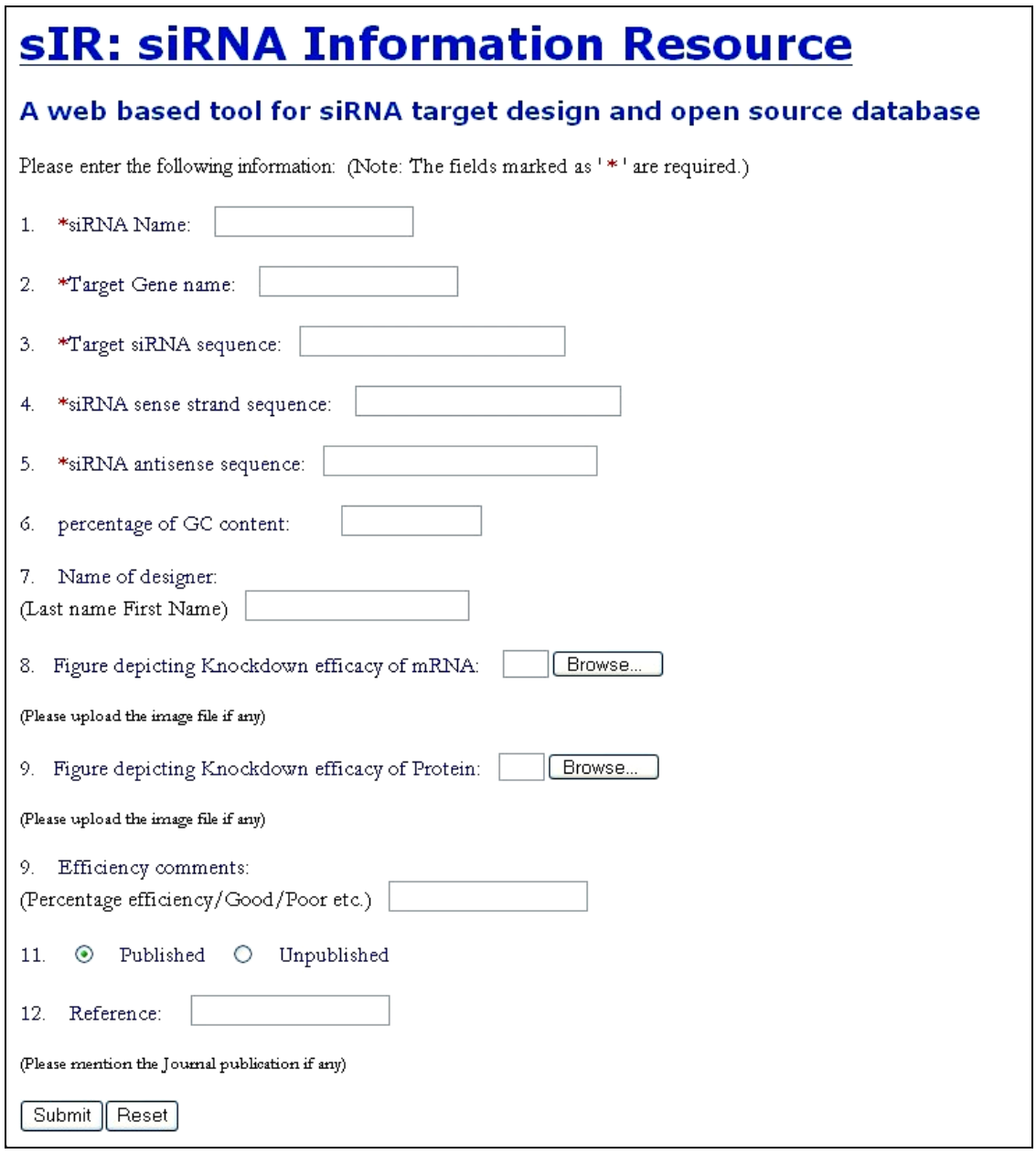


Figure 3.3.1.1: Snapshot of siRNA database update form. The first five fields are mandatory. Relevant figures can be uploaded in the database using the upload tool provided with the form.

The first five fields, namely siRNA name, Target gene name, Target siRNA sequence, sense strand sequence and anti-sense strand sequence are the required fields. Other fields such as percent GC content, Designer name, Figures, efficiency comments, reference are optional. Efficiency comments can contain details about the working efficiency of the siRNA sequence. It can be either Good, Average, Poor or in the form of percentage values. If the sequence is published then respective journals can be referenced in the ‘Reference’ section.

CHAPTER 4

Implementation

This chapter covers the details on the software implementation of various modules and objects explained in chapter 3. Subsequent sections will explain the various databases and applications implemented in this software.

4.1 Databases

4.1.1 RefSeq

The Reference Sequence (RefSeq) database was downloaded and made available for this application locally on the server. RefSeq provides a biologically non-redundant collection of DNA, RNA, and Protein sequences. Each RefSeq entry represents a naturally occurring molecule of an organism. RefSeqs are processed information and not a primary piece of research data itself as are Genbank records [13]. RefSeq was mainly used for retrieving mRNA sequences given 'Accession numbers' as input. Since designing siRNA deals with mRNA sequence, 'refmRNA' database consisting of only mRNA (Accession prefix 'NM_') sequences was used.

This database was implemented using PostgreSQL database and the sequences were retrieved from the database using SQL queries. A perl DBI module was used in CGI

script in order to communicate with the database using a CGI script. DBI module provides a consistent interface for database applications [18].

4.1.2 SOURCE

SOURCE is a web based database and it pools information together from various sources. Its report includes information on aliases, functional description, annotations etc. SOURCE collects information from various sources such as Online Mendelian Inheritance in Man (OMIM), SwissProt, LocusLink, UniGene, GenBank, PubMed as well as many others [14]. Source database was made available locally using the 'Batch SOURCE' mode, where gene aliases of a batch of accession numbers were downloaded. The SOURCE data was used to obtain various gene alias names with accession numbers as input. The program retrieves gene accession number with gene alias name as input. This is very helpful, if the user does not have both the sequence information and accession number information.

This database was implemented using PostgreSQL database. A PHP script was used to connect to the database as well as to produce dynamic web pages with the results.

4.13 UniGene

The main goal of UniGene database is to produce an organized view of the transcriptome. UniGene attempts to partition Genbank sequences into a non-redundant set of

gene-oriented clusters. The sequences in each of these cluster contains sequences that represent a unique gene [26]. One can avoid redundancies using UniGene database as it consist of only gene and expressed sequence tag (EST) sequences. When a researcher designs siRNA target sequence, it is quite useful to perform BLAST against UniGene database as a preliminary step. This will give an initial idea about homology of target sequence with other known genes in the UniGene database.

The sIR software provides both human as well as mouse UniGene databases as options to BLAST the target siRNA sequence against. It uses a CGI script with Bioperl to access the database and perform the BLAST operation.

4.1.4 Human Genomic database

Human Genomic database was downloaded from the NCBI database resource and made available locally on the server. It consists of data obtained from the sequencing of the human genome from 23 pairs of chromosomes. This database is much bigger than Unigene database and consists of additional information as compared to Unigene database. Some researchers may be interested in testing the siRNA sequences for homology against human genomic database. Hence, this database was provided as one of the options to BLAST siRNA target sequences against. sIR uses a CGI script with Bioperl to access the database and perform the BLAST operation.

4.2 Implementation of target design mode.

Implementation of target design mode can be divided into two stages. The first stage is to design siRNA target sequences using input parameters and the second stage is to produce customized BLAST output on user chosen target sequences.

Perl with CGI module was used to implement the first stage. Perl is a very popular tool for string operations. Various modules or separate programs were written individually to calculate GC content, avoid nucleotide runs, check the sequence, and sort the output according to score. DBI module was used to connect to PostgreSQL database with the CGI script.

BLAST is a short form for Basic Local Alignment Search Tool [16]. BLAST was performed locally using NCBI BLAST program. BLAST tool consists of various programs such as BLASTN, BLASTP, BLASTX, TBLASTN, and TBLASTX. BLASTN is used to compare nucleotide sequences to one another. All other programs are used for protein sequences. In this application, Bioperl was used to call 'BLASTALL' function, which performs BLAST operation. Since only nucleotides were involved, the 'BLASTN' program for nucleotide BLAST was used. The BLAST program implemented in this software takes the user selected target siRNA sequence and performs BLAST against the user selected database. Again, Bioperl and Perl were used to parse the BLAST output to

provide a customized BLAST output. This BLAST output consists of details on BLAST hit name, Genbank ID (if any), description, identities, score and alignment. Only the BLAST results with more than 75% homology (> 16bp/21bp) are displayed.

4.3 Implementation of siRNA Information database: Open source database.

siRNA database consists of information on target gene, siRNA target sequence, sense and anti-sense strands, GC content, designer, efficiency of siRNA, relevant figures, journal publication etc. Implementation of this database can be divided into two parts. The first part was to create the database and retrieve the information and the second part was to update the database with new information using a user friendly input form.

PostgreSQL was used to implement the database. It is easier to manage images and data if they are stored in the database as objects. BLOBs (Binary Large Object) method was implemented to store image files as objects in the database. A BLOB is stored as an object in the database and not in the table itself. Every BLOB has a unique identifier or object id which is stored in the table of the database [20].

The database items were retrieved using a PHP script which is very useful for creating dynamic web pages. A special PHP script with image headers was used to view

images in the database in a new web browser window. Database rows could be selectively retrieved using target gene name or completely retrieved using ‘ALL’ option.

The database could be updated using an update form. This form is password protected and has restricted access. When the researcher enters the information in the form and submits it, a PHP script with SQL query runs in the background, which updates the database with new entry. Images can be uploaded using the image upload tool. The uploaded images are first uploaded to the server and then imported into the database creating a unique object ID for them each time they are uploaded. This object’s ID is then stored in the database.

CHAPTER 5

Discussion, Results and Validation

5.1 Discussion

siRNA (short interfering RNA) is known to induce post-transcriptional gene silencing by a process called RNAi (RNA interference). It is known to cause sequence specific degradation of mRNA. RNAi technology has proven its usefulness in many fields such as cancer, gene therapeutics, functional genomics etc.

There have been many recent studies to determine the relationship between specific siRNA sequences and the RNAi effect [22, 23, 24, 25]. It was found that the efficiency of siRNA is a function of target sequences and its content. It was also found that all the positions of siRNA do not contribute equally to the target recognition [24]. Hence, having a scoring system that can predict siRNA efficiency to some extent can be very useful. Although, the scoring system used in this application is based on the analysis of the siRNA sequence and its potency [22, 23], it may not accurately predict the siRNA activity. It can only help the process of siRNA design.

5.2 Reproducibility of existing siRNA target sequences.

In order to test the validity of the software, a database of 107 tested siRNA target sequences was used. These siRNA sequences were collected from various laboratories in University of Texas at Southwestern Medical Center (UTSW) as well as from published papers. Out of these 107 siRNA sequences, 83 were functional and 24 were not functional as documented by the principal investigators.

The validation was performed in a two step process:

1. First, all the mRNA sequences or accession numbers of the siRNAs in the database were tested in the software to see if the software was able to design these siRNA target sequences. The software was able to design these target sequences with appropriate input parameters.
2. Next, scores were calculated for each of these siRNA sequences using the scoring system described in chapter 3. It was found that after sorting the results of siRNA target designer using the score, approximately 70% of the functional siRNAs in the database had a score of more than 60 and approximately 65 % of the non functional siRNAs in the database had a score of less than 40. This is demonstrated in the Figure 5.2.1.

Table 5.2.1 shows the scores obtained by siRNA target sequence ('Functional' as well as 'Non functional'). It should be noted that the total number of 'functional' siRNAs (83) was much greater than the total number of 'non functional' siRNAs (24).

Score	Functional siRNAs	Non functional siRNAs
10	0	2
20	1	2
30	2	5
40	11	6
50	14	4
60	20	4
70	14	1
80	14	0
90	6	0
100	1	0
Total	83	24

Table 5.2.1:Scores obtained by siRNA target sequence ('Functional' as well as 'Non functional'). This table clearly shows that approximately 70% of the working siRNA have scores greater than 60 where as approximately 65% of the not working siRNA have scores less than 40.

This is also demonstrated in the graphical format in Figure 5.2.1.

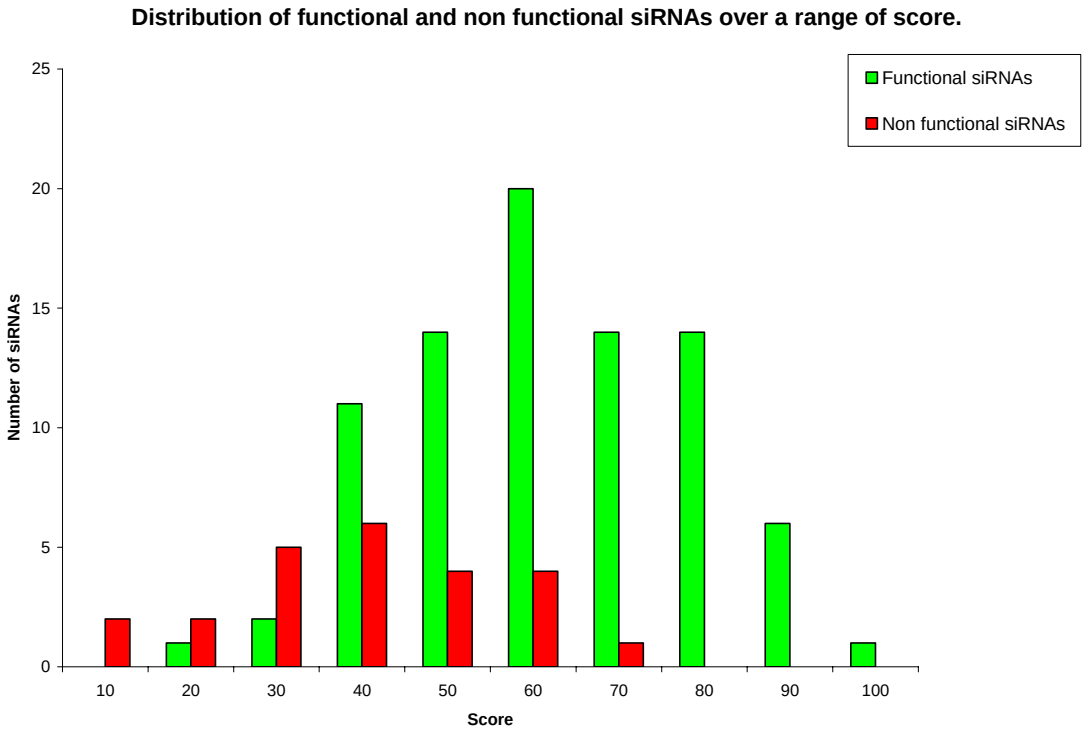


Figure 5.2.1: Distribution of tested and 'Functional' and 'Non functional' siRNAs. This figure is a plot of number of siRNAs against scores obtained by those siRNAs. It can be observed in the figure that the distribution of 'functional' siRNAs is prominent for scores greater than 60 and the distribution of 'non functional' siRNAs is prominent for scores less than 40.

There was approximate efficiency data available for 61 siRNAs (Both functional as well as non functional). Average score was calculated for siRNA sequences with similar efficiency.

Table 5.2.2 shows these values.

Percentage efficiency	Average score at that efficiency
<40%	40.43
<60%	57.14
<80%	65.38
<100%	67.06

Table 5.2.2: Percentage efficiency and average score at that percentage efficiency.

This can be demonstrated using a graphical plot given in the Figure 5.2.2.

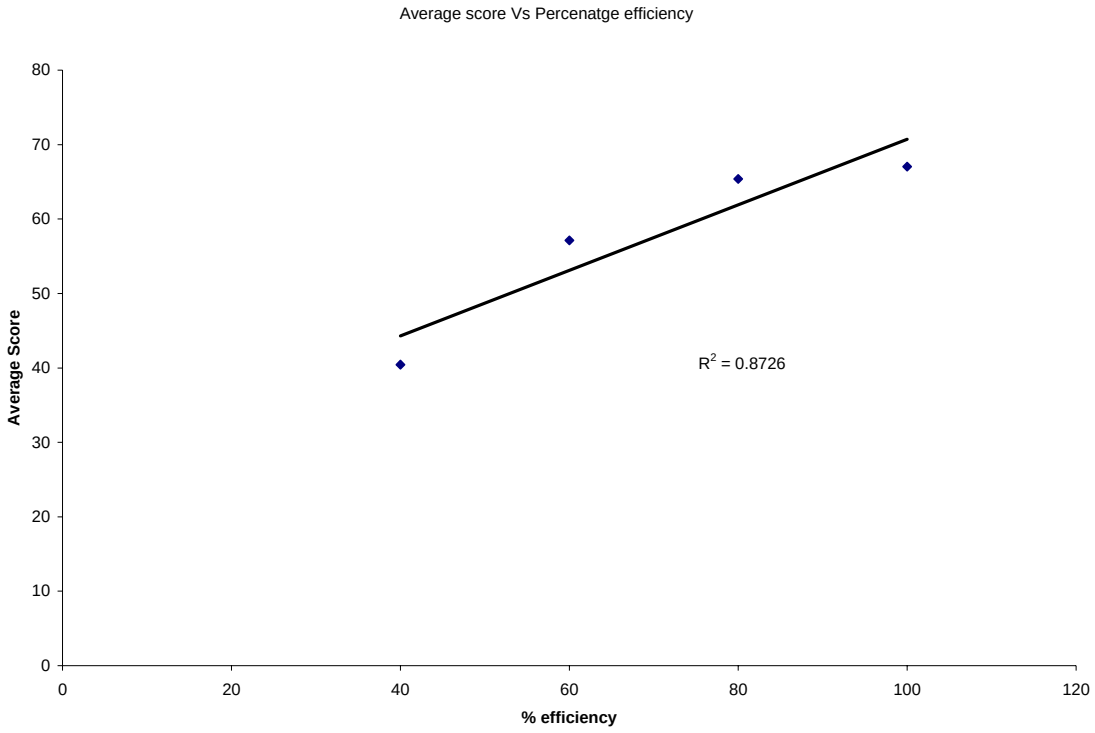


Figure 5.2.2: correlation between average score and percent efficiency of siRNA sequences. This figure shows that there is some amount of correlation (correlation = 0.87) between average score and percent efficiency of siRNA sequences.

5.3 Conclusion

RNAi technology is a popular research methodology to decode various functions of genes. It is very important to have computational tools to perform the task of designing the best possible siRNAs as well as to store them, in an effort to facilitate research. It is very important to have all the available siRNAs in one place for future reference.

- sIR was able to show its capability to design siRNA as well to provide a database consisting of already developed and tested siRNAs.
- This software tool was able to provide a web based interface that is user friendly and easily accessible.
- It was able to provide a customized BLAST output in an easily readable and interpretable format.
- BLAST operation could be performed against a variety of databases provided by the software.
- Scoring system of this software may predict the siRNA activity completely, but it can be helpful in designing better siRNA sequences.
- The sIR software was validated successfully using working and tested siRNA target sequences.

CHAPTER 6

Maintenance and Future work.

6.1 Maintenance of sIR (siRNA Information Resource).

The databases used in this application are locally downloaded. Hence, these databases have to be updated at least once in month to ensure that the most recent changes are included in the database. This job will be performed by system administrator of the server on which these programs are available.

6.2 Future work.

Batch processing of siRNA target sequences:

siRNA design tool is able to receive one accession number or one sequence information at one time. This is quite adequate most of the times. However, because of its growing popularity; the researchers may want to find siRNA target sequences for more than one gene at a time. This can be accomplished by adding a batch processing feature to this software, where a batch of accession numbers can be given as input and output can be emailed to the researcher after completion of the process.

Part II: CHAPTER 7

Gene expression profiles and Chemosensitivity data of breast cancer cell lines.

7.1 Objectives and application:

As mentioned earlier, one of the major problems in using chemotherapy to treat cancer is whether patients, whose tumors do not respond to one drug or a combination of drugs, would respond to another. Hence, one of the long term objectives of this research is to be able to rationally select the chemotherapy for each patient's tumor that would be the most effective.

In the past, there have been extensive studies to determine if the *in vitro* sensitivity or resistance of tumor cells from a patient is able to predict their response *in vivo* [27]. Although contentious, it appears there is some correlation [28]. Based on this, some *in vitro* drug responses have been tried, but tumor cells should be exposed to chemotherapeutic agent in conditions similar to those in human body determined by pharmacokinetic parameters [29]. It will be beneficial to find out if there is any other method to accurately predict the drug responses in a patient.

Single gene mutation or mRNA and protein expression in tumor cells were also used to predict the response to chemotherapy. It was found that there was an association in some studies [30, 31] whereas no association was found in others [29, 32-34]. This can be argued based on the possibility that the failure to predict by these methods may be because drug sensitivity of tumor cells could be determined by many genes instead of one gene that influence overall sensitivity.

One potential approach to resolve the apparent discordant findings is to examine if tumor cell line "gene expression signatures" detected by microarray analysis prior to treatment could identify a set of genes correlating with sensitivity or resistance to a particular drug.

Gene expression profiles have several potential advantages:

- It is possible to perform profiles using small samples from clinically available specimens.
- Multiple genes are examined at the same time
- Expression profiles can be done in a short time.

Hence, it would be beneficial if one is able to sample a patient's tumor before treatment, perform gene expression profiling and from this profile predict the sensitivity and resistance of that individual tumor to various chemotherapy agents.

Therefore, the objectives of this study were as follows:

1. To perform chemosensitivity tests on a large number of breast cancer cell lines with agents commonly used in the treatment of breast cancer.
2. Perform microarray analysis on the same tumor cell lines.
3. Determine if a correlation between *in vitro* sensitivity or resistance and gene expression profiles exist.

This information would be useful to start clinical trials.

Also in this process, large amounts of data were produced. Hence, it was practical to develop a collection of computational tools to facilitate this research. This also became one of the objectives of this project. This set of tools will be used for similar studies in the future.

CHAPTER 8

Materials and Methods

8.1 Cell lines.

Chemosensitivity tests as well as gene expression profiles were performed on 17 breast cancer cell lines. The cell lines used were: HCC1143, HCC1395, HCC1419, HCC1428, HCC 1569, HCC1806, HCC1937, HCC1954, HCC2688, HCC3153, HTB122, HTB126, HTB131, HTB22 (MCF7), HTB24, HTB25, HTB26 (MDA-MB 231).

These cell lines were mainly developed at the University of Texas at Southwestern Medical Center, from primary tumors. The cell lines used for this study were of an adherent nature. All cell lines were maintained in RPMI-1640 (Invitrogen, Carlsbad, CA) supplemented with 10% fetal bovine serum and incubated in 5% CO_2 at 37˚C in a humidified atmosphere.

8.2 Chemotherapeutic drugs.

Five anti-cancer drugs commonly used in the treatment of breast cancer tumors were used for this study. They include the DNA-damaging agents cisplatin and gemcitabine, the anthracycline antibiotic doxorubicin and the anti-microtubule agents paclitaxel and vinorelbine.

Cisplatin: Cisplatin is an inorganic complex formed by an atom of platinum surrounded by chlorine and ammonia atoms in the cis position of a horizontal plane. Intracellularly, water displaces the chloride to form highly reactive charged platinum complexes. These complexes inhibit DNA replication through covalent binding, leading to intrastrand, interstrand, and protein cross-linking of DNA. Experimental and clinical data suggest that cisplatin enhances radiation therapy effects. Figure 8.2.1(a) shows the chemical structure of cisplatin.

Paclitaxel: Paclitaxel is known to prevent cell division process by promoting disassembly of microtubules - cytoskeletal structures that assemble and divide throughout the life of a cell. At the start of cell division, a large number of microtubules are formed, and as cell division comes to an end, these microtubules are normally broken down. However, paclitaxel prevents microtubules from breaking down. In the presence of this drug, cancer cells, which attempt to divide frequently, cease to grow and divide as they become clogged with microtubules [35]. The chemical structure of paclitaxel is shown in figure 8.2.1(b).

Gemcitabine: Gemcitabine exhibits cell cycle specificity, primarily killing cells undergoing DNA synthesis (S-phase) and also blocking the progression of cells through the G1/S-phase boundary. The chemical structure of gemcitabine is shown in figure 8.2.1(c).

Vinorelbine (Vinorelbine Tartrate): Vinorelbine is a Vinca alkaloid that interferes with microtubule assembly. The anti-tumor activity of vinorelbine is thought to be due to primarily inhibition of mitosis at metaphase through its interaction with tubulin. It may also interfere with:

a. amino acid, cyclic AMP, and glutathione metabolism.
b. calmodulin-dependent calcium transport ATPase activity .
c. cellular respiration .
d. nucleic acid and lipid biosynthesis.

The chemical structure of vinorelbine is shown in figure 8.2.1(d).

Doxorubicin: Doxorubicin is a cytotoxic anthracycline antibiotic isolated from cultures of *Streptomyces peucetius var. caesius*. It binds to nucleic acids, presumably by specific intercalation of the planar anthracycline nucleus with the DNA double helix. The chemical structure of doxorubicin is shown in figure 8.2.1(e).

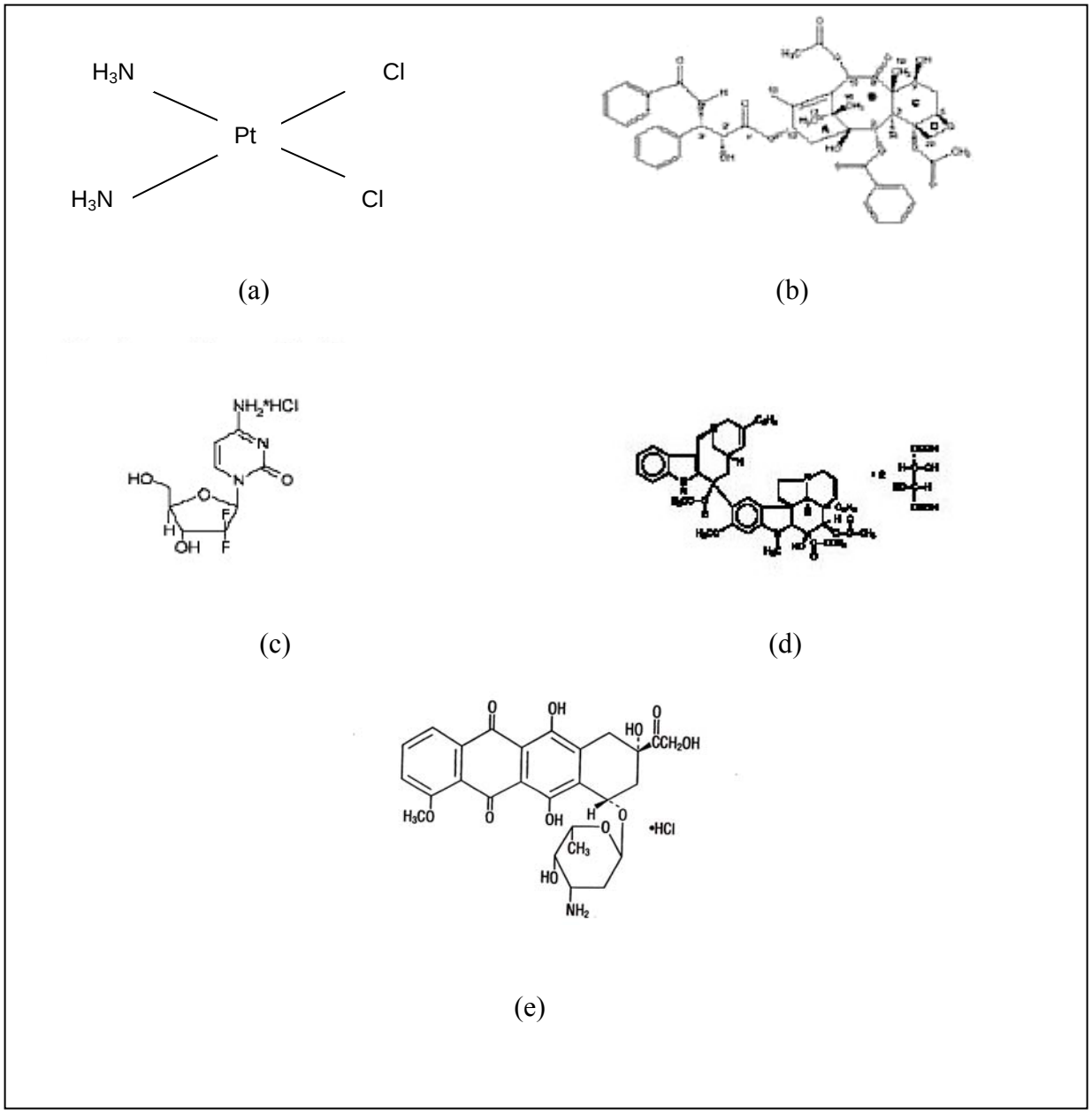


Figure 8.2.1: Chemical structures of the drugs involved in the chemosensitivity tests. This figure shows chemical structures of the drugs involved in the chemosensitivity tests. (a)Cisplatin (b) Paclitaxel (c) Gemcitabine (d) Vinorelbine (e) Doxorubicin.

Table 8.2.1 below gives details about the drug concentration, molecular weight, storage conditions, stock solutions etc of the drug used. It also gives details about the 4-fold dilutions of the drug with culture medium and the final concentration of the drugs.

Drug	M.W	Storage	Stock solution concentrations	Working concentrations	Final concentrations
Cisplatin	300.05	Room temperature	N/A (Freshly prepared)	0.12 μM to 2,000 μM	0.06 μM to 1,000 μM
Paclitaxel	853.93	Room temperature	8000nM (Diluted in PBS)	0.12 nM to 2,000 nM	0.06 nM to 1,000 nM
Gemcitabine	299.50	Room temperature	N/A (Freshly prepared)	0.24 nM to 4,000 nM	0.12 nM to 2,000 nM
Vinorelbine	1079.12	3 - 4° F	8000nM (Diluted in PBS)	0.12 nM to 2,000 nM	0.06 nM to 1,000 nM
Doxorubicin	579.99	3 - 4° F	N/A (Freshly prepared)	0.74 nM to 12,000 nM	0.37 nM to 6,000 nM

Table 8.2.1: Information on drugs. This table describes molecular weight (M.W), storage conditions of the respective drugs. The 'stock solution concentrations' column shows the concentration of drug diluted in PBS to create a stock solution. N/A (not applicable) means that the drug was freshly prepared. The 'working concentrations' column signifies the concentrations used in 4 – fold dilution with culture media during the experiments. The 'final concentration' column gives the actual value of concentrations of the drug after dilution in each well of the 96-well plates used for chemosensitivity experiments.

8.3 Chemosensitivity data.

Chemosensitivity of cell lines were determined using the MTT (Colorimetric) assay kit (Chemicon International, Temecula, CA).

MTT, chemically known as 3-[4, 5-dimethylthiazol-2-yl]-2, 5-diphenyl tetrazolium bromide, is a water-soluble yellow dye which can be readily taken up by viable cells and reduced by the action of mitochondrial dehydrogenases. The product of this reduction process is a water-insoluble blue formazon that can be dissolved in an isopropanol (0.04 N HCl) solution. Formazan production is directly proportional to the number of viable cells over the range of 200 to 50,000 cells per well [36]. Hence MTT can be used to determine optimal cell density and chemosensitivity is determined from the reduction of cell number; due to the inhibition of growth, with increasing drug concentration.

Since the growth rate of every cell line varies, it was important to find the optimal number of cells per well required to plate. Hence, as a pre-processing technique, “Plating Assays” were performed before performing the actual “Chemosensitivity Assays”.

Plating Assays: In order to perform plating assays, different number of cells per well were plated and incubated for 5 days. At the end of the incubation period, colorimetric assay was performed in order to determine the cell growth. The initial cell density with

optimal optical density reading and linear growth at the end of the 5^{th} day was chosen as the initial optimal cell number of cells/well. The protocol below describes this process in detail:

Protocol for Plating Assay:

Day 0: Preparing cell suspension and plating the cells.

Day 5: Harvest of cells and MTT assay.

a. Cells were grown in a 100 mm dish or T-75 flasks and kept in a sub-confluent state.

b. Day0: Cells were trypnisinized with 1~2ml of trypsin/EDTA solution, and incubated at 37° C until cells float in the dish. Approximately, 8~9 ml of media was added and the cell suspension was collected.

c. The number of cells in 1 mL of medium was counted using a cell counter device (Beckman Coulter Z1 particle counter).

d. Five serial dilutions of cells in media were prepared (e.g.: 4×10^4, 2×10^4 , 1×10^4, 0.5×10^4 , 0.25×10^4).

e. 100µl of a cell suspension (with appropriate dilutions) was plated in a 96 well plate.

f. The plate was incubated in a humidified incubator in 5% carbon dioxide for 5 days at 37° C.

g. Day5: 10µl of MTT solution was added to each well.

h. The plate was then incubated for 4~5 hours.

i. The cells were viewed periodically under a microscope for presence of formazon formation. When the purple precipitate was clearly visible, media was removed with

a multi-channel aspirator, and then 100 µl of Isopropanol (0.04 N, HCl) solution was added to each well.

j. The absorbance was then measured on a micro-plate reader (EL312, Bio-Tek Instruments, Inc.) with a test wavelength of 570 nm and a reference wavelength of 750 nm.

k. The average values were determined from 6 replicate wells after the 2 extreme values were excluded. Absorbance against number of cells per well was then plotted.

l. The number of cells to use in further experiments was determined from the linear portion of the plot that yielded an absorbance of 0.2-0.9.

The analysis part, after getting the raw data from plate reader can be performed using a Microsoft Excel program called "Optimal cell density calculator", specially designed to analyze and store the data obtained from plating assays.

Protocol for Chemosensitivity assay:

1. Cells were grown in a 100 mm dish or T-75 flasks and kept in a sub-confluent state.
2. Day 0: Cells were trypnisinized with 1~2ml of trypsin/EDTA solution, and incubated at 37° C until cells float in the dish. Approximately, 8~9 ml of media was added and the cell suspension was collected.
3. The number of cells in 1 mL of medium was counted using a cell counter device (Beckman Coulter Z1 particle counter)..

4. The cell suspension was dispensed with appropriate cell density calculated using the data from plating assays.
5. After repeated pipetting, 50 µl of cell suspension was plated in a 96 well plate.
6. The cells were allowed to adhere approximately 24 hours in the incubator.
7. Day 1: 4 fold range dilutions of drugs were prepared. For stock solutions, drugs were diluted with PBS and filtered using a 0.45 µm sterile filter. For working drug solutions, drugs were diluted with complete medium just before use.
8. Fifty µl of appropriate concentration drug solution was added to 96 well plates except for the control. Fifty µl of medium was added to the control.
9. The plate was incubated in a humidified incubator in 5% carbon dioxide for 4 days at 37° C.
10. Day5: Ten µl of MTT solution was added to each well.
11. The plate was then incubated for 4~5 hours.
12. The cells were viewed periodically under a microscope for presence of formazon formation. When the purple precipitate was clearly visible, media was removed with a multi-channel aspirator, and then 100 µl of Isopropanol (0.04 N, HCl) solution was added to each well.
13. The absorbance was then measured on a micro-plate reader (EL312, Bio-Tek Instruments, Inc.) with a test wavelength of 570 nm and a reference wavelength of 750 nm.
14. The average values were determined from 6 replicate wells after the 2 extreme values were excluded. Absorbance was plotted against number of cells per well.

15. Percentage of each value against control value was calculated.
16. A graph of absorbance percentage (Y-axis) against drug concentration (X-axis) was plotted.
17. The IC50 concentration was determined as the drug concentration required to reduce the absorbance percentage (cell number) to 50%.

The analysis part, after getting the raw data from plate reader was performed using a Microsoft Excel program called "MTT database", specially designed to analyze and store the data obtained from chemosensitivity assays.

8.4 Gene expression profiles.

Gene expression profiles were evaluated using Affymetrix cDNA microarray (U133A and B chips). For this process, RNA was extracted from the cell lines on which chemosensitivity tests were to be performed. RNA was prepared using the RNeasy Midi kit (Qiagen, Valencia, CA). Extracted RNA was analyzed for quality using agarose-formaldehyde gels or using the RNA 6000 Nano kit (Agilent Technologies, Palo Alto, CA) with Agilent Bioanalyzer software. The HG-U133B chip from Affymetrix has 22,283 genes (13,794 Unigene clusters) and the HG-U133A chip has 22,645 genes (17179 Unigene clusters). Five micrograms of total RNA was used in a single round of amplification. The Affymetrix protocol starts with cDNA synthesis, using a poly (T) primer with a T7 promoter. The double stranded cDNA generated is then used to prime the synthesis of cRNA

using biotinylated ribonucleotides (UTP and CTP). After the labeled cRNA is synthesized it is fragmented and hybridized to the GeneChip at 45 C for 16 hours in a rotary incubator. After hybridization the analyte solution is removed and the array is washed and stained with Streptavidin Phycoerythrin in the Affymetrix GeneChip fluidics station. After washing, the array is scanned (Agilent GeneArray Scanner) and the data extracted with the MicroArray Suite 5.0 software (Affymetrix) which also provides scaling and other data normalization prior to analysis.

Affymetrix data was analyzed using in-house MATRIX (MicroArray TRansformation In eXcel) 1.24 [37]. First, the samples to be analyzed were divided into two groups, one consisting of samples resistant to the drug, and the other consisting of samples sensitive to the drug. These groups were formed on the basis of the chemosensitivity tests. Log ratios were calculated using the log ratio mode of the software. All signals for gene duplicates were pooled by averaging and the resulting data was normalized so that all samples had the same median. In the calculation of these log ratios, a threshold of 100 for the expression signals was used to avoid spurious differential expression due to background. As an alternative comparison between affymetrix expression data and IC50 values, Pearson correlations between log-transformed affymetrix signals for each gene and log transformed IC50 values for each drug was calculated across all cell lines. Thus correlation between gene expression data and chemosensitivity data was determined. These results are explained in chapter 10. Scatter plots were obtained where each spot on the plot represented a gene and their coordinates represent the expression level being compared in each group. The cut-off

for expression change was chosen as 4-fold or more. Clustering analysis was performed for both samples as well as genes using this software.

8.5 Computational tools.

In order to analyze drug sensitivity data as well as the gene expression data, several computational tools developed in-house were used. The chemosensitivity data was analyzed and stored using “MTT database 1.10” program [38]. This program stores MTT assay data in a database format, displays charts, calculates IC50 values for each assay. It also automatically summarizes and updates the data in different table formats. The gene expression data was analyzed using “MATRIX 1.2.4”. This software imports microarray gene expression data and performs several analyses including clustering, scatter plots, log ratios, color displays, sample averaging, correlations etc. The plating assay data was stored and analyzed using “Optimal cell density calculator” program developed as a part of this project. The experimental setup was made easier using “Drug Information and Calculator” program developed as a part of this project. The last two programs are explained in detail in the next chapter.

CHAPTER 9

Development of drug sensitivity tools

9.1 Block Diagram

A set of computational tools were developed in order to analyze chemosensitivity data as well as to perform chemosensitivity experiments. The block diagram below gives an overview of the tools developed.

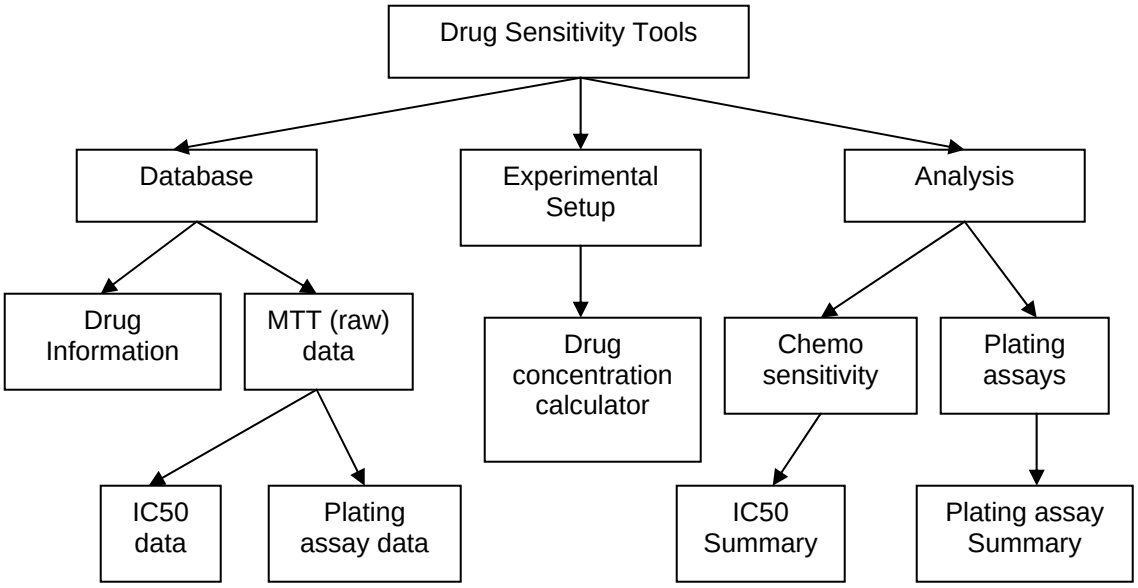


Figure: 9.1.1 Overview of the drug sensitivity computational tools. The figure above shows the computational tools used for storing data, setup experiments as well as to analyze the data.

Database: The databases are Microsoft Excel database. There are mainly two types of databases, one for storing drug information, and other for storing raw MTT data.

Drug Information database: It is an Excel database and it consists of information on the various drugs used for drug sensitivity tests. It consists of information such as molecular weight, chemical structure, mechanism of action, solubility and concentrations.

MTT (raw) data: It is very important to store the raw MTT data from the plate reader for future reference. Two separate Excel databases were created in order to store this information for chemosensitivity tests as well as for plating assays. Along with the raw data, they consist of information on cell line, drug, respective plots etc.

Experimental setup: An Excel program called "Drug Information and calculator" was developed to pre-calculate drug and solvent information before every chemosensitivity experiment. This program is explained in detail in the next section.

Analysis: Microsoft Excel programs were developed to analyze the raw MTT data and summarize their results. "MTT Database 1.10" program was developed to calculate the IC50 values from the MTT data. "Optimal Cell Density Calculator 4.0" was developed to calculate the optimal number of cells/well required to plate for chemosensitivity experiments. This program is explained in detail in the next section.

9.2 Optimal cell density and drug concentration calculator and database.

9.2.1 Optimal cell density calculator.

This program was especially developed to automatically process and store the data obtained from MTT assay. This program was developed using Microsoft Excel (Visual Basic). Its main objective was to calculate the optimal number of cells to be plated for chemosensitivity experiments.

This program consists of three sheets:

1. Plating Assay sheet.
2. List of optimal cell density.
3. Summary sheet.

Plating assay sheet: This is the main database sheet where the raw data is stored along with absorbance plots and result. It consists of a form which takes inputs from the user and delivers the output result. The user can import the raw file using “Import” button. The user can choose the cell line from the list of cell lines provided with the scroll bar option. Finally, the user needs to enter the information about the highest concentration of cells/well plated. The “Plot and Calculate” button gives the desired result. Figure 9.2.1.1 shows a screenshot of this sheet.

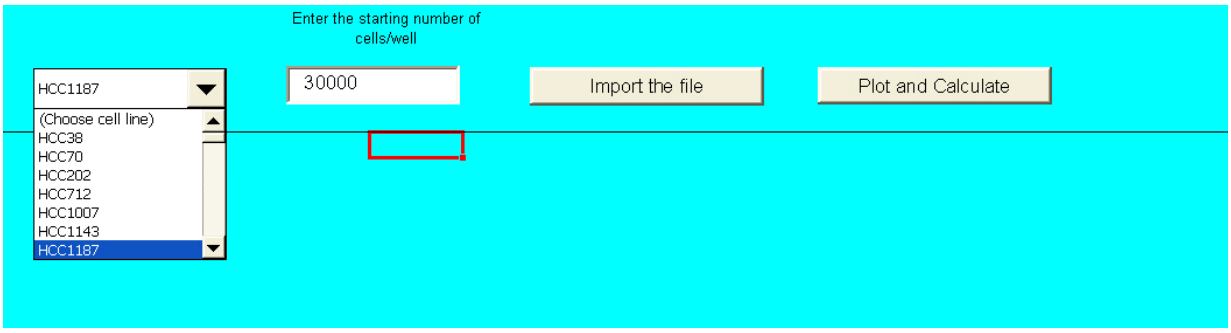


Figure 9.2.1.1: Optimal Cell Density Calculator Form.

The program automatically calculates the 5 serial dilutions, given the highest concentration of cells per well plated. For example, if the user enters 4 x 10^4 as the highest concentration, the program automatically calculates the concentrations in the other columns as 2 x10^4, 1 x10^4, 0.5x10^4 and 0.25x10^4 respectively.

The average value of the absorbance in each column is determined from 6 replicate wells after excluding two extreme wells. The program then plots this absorbance against number of cells per well.

The program then looks for a linear region in the plot. This is determined by calculating slope of the curve between various points. The linear region is the region where difference between two consecutive slopes is minimum. The slope between any two points A1 and A2 is calculated using the following formula:

$$\text{Slope } (A1, A2) = \frac{(Y_{a1} - Y_{a2})}{(X_{a1} - X_{a2}).}$$

Where, X and Y are the coordinates of points A1 and A2. This is explained using the curve shown in figure 9.2.1.2.

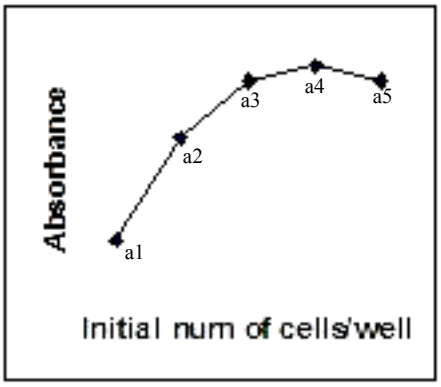


Figure 9.2.1.2: Linear range logic. The program calculates slope between all the consecutive points namely a1, a2, a3, a4 and a5. Then it calculates the difference between slopes. The points which show minimum difference between slopes define the linear range. Slope (a2, a3) – slope (a3, a4) = minimum. Therefore, the curve is considered linear between a2 and a4.

Finally, the result box displays the approximate linear range where the optimal cell density per well can be chosen. The sheet also stores raw MTT data along with the plot and results. When the user wants to import another data file, the database sheet automatically appends the new data and graph below the latest data. This way the user can store the results of all the plating assays in one sheet for future reference. The program also automatically updates the summary page. Figure 9.2.1.3 shows the database and plot layout of the optimal cell density calculator program.

List of optimal cell density: This is a reference sheet and it consists of information on the optimal cell densities of both Lung as well as Breast cancer cell lines, used in the past.

Summary sheet: This sheet is automatically updated with experiment ID as well as the cell line; every time the user uses this program to calculate optimal cell density. The user can decide on the final number of cells per well chosen from the resultant range and can manually enter this information in the last column.

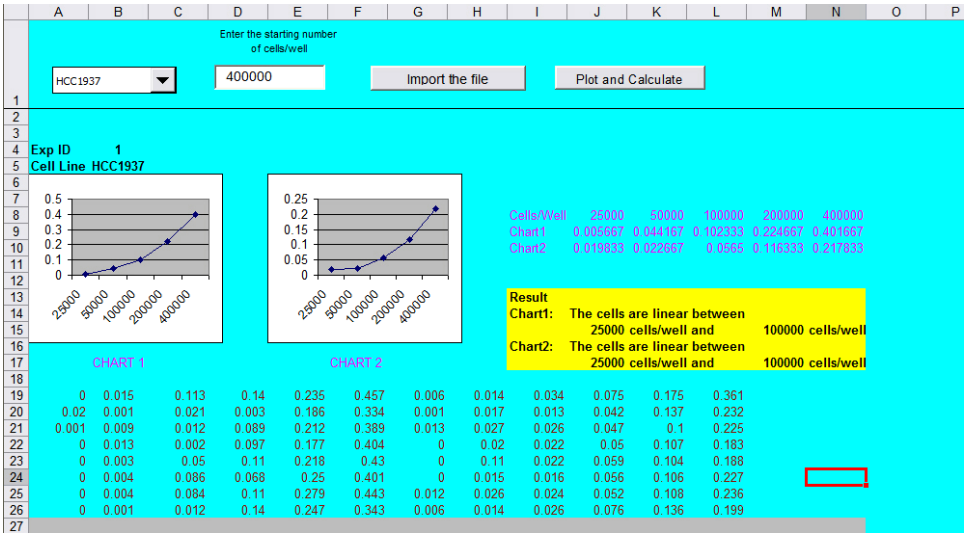


Figure 9.2.1.3: The screen shot of optical cell density calculator database. This figure shows the typical layout of plots as well as the raw MTT data.

Requirements to use this program:

1. This program was designed and tested on a PC with Microsoft Excel on Windows

XP machine. It will not work on a Macintosh.

2. Macros must be enabled to use this program.
3. The input file can be either a text file or Microsoft excel file.
4. The arrangement of cells per well in the 96 plate should as shown in figure 9.2.1.4. The first and seventh column should be blank. The 2nd, 3rd, 4th, 5th and 6th columns should have cell densities in the increasing order of number of cells per well in 2-fold serial dilutions. This should be repeated for columns 8th, 9th, 10th, 11th and 12th respectively.

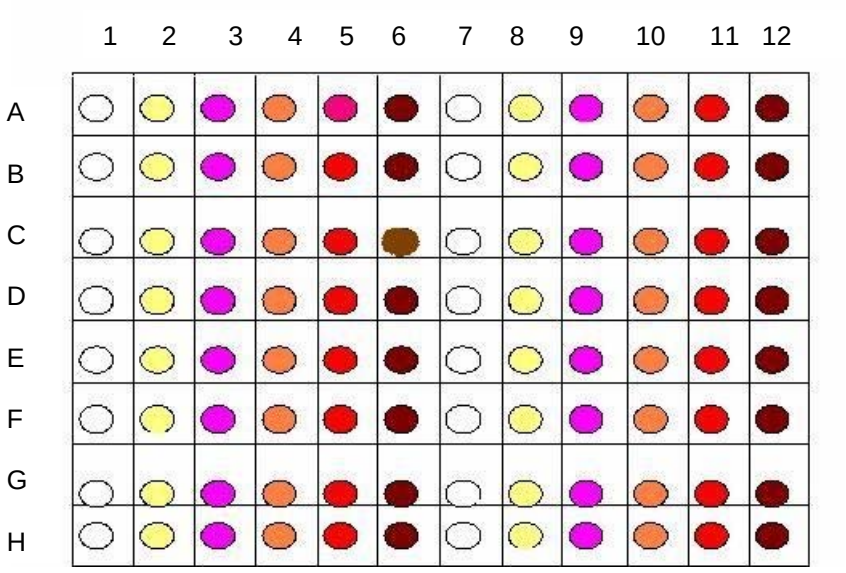


Figure 9.2.1.4: Plating Assay layout. This figure shows the experimental setup compatible with the program, for plating assay in a 96 well plate. Different colors indicate the different number of cells/well plated in that particular column.

9.2.2 Drug Information and Concentration Calculator:

This program was designed to automatically calculate the solvent and drug quantity needed to achieve required concentration to make the experimental setup easier. It was developed using Microsoft Excel (Visual Basic). Information on all the drugs involved in this study is also included as a feature. This program is also able to add a new drug and drug information for future studies.

It consists of 3 modes:

1. Calculator mode.
2. Drug Information database sheet.
3. "Add new drug" mode.

Calculator mode: It optimally calculates the quantity of drug and solvent to be added for all the chemosensitivity tests, given the number of plates as input. The user can choose the drug as well as enter the number of plates required for the chemosensitivity tests. When the user activates the "Calculate" button, the output is displayed in the 2 boxes named solvent and drug giving solvent quantity as well as drug quantity respectively. The program uses the information previously stored in the "Drug Information Sheet" to calculate the stock solution information. This information can be modified by the user at any point of time. Stock solution and dilution concentration for the particular drug (entered previously into the database) are also displayed on the worksheet for reference.

The formula used for calculation is as follows:

C1 x V1 = C2 x V2

Where C1=Initial concentration of drug, V1=Initial volume of drug, C2=Final concentration of drug, V2=Final volume of drug solution (solvent +drug), C1= number of moles= (mass in liters /Molecular weight)

Drug solution: V1 and Solvent Solution: V2-V1.

A screenshot of "Drug Information and Calculator" is shown in the figure 9.2.2.1

.

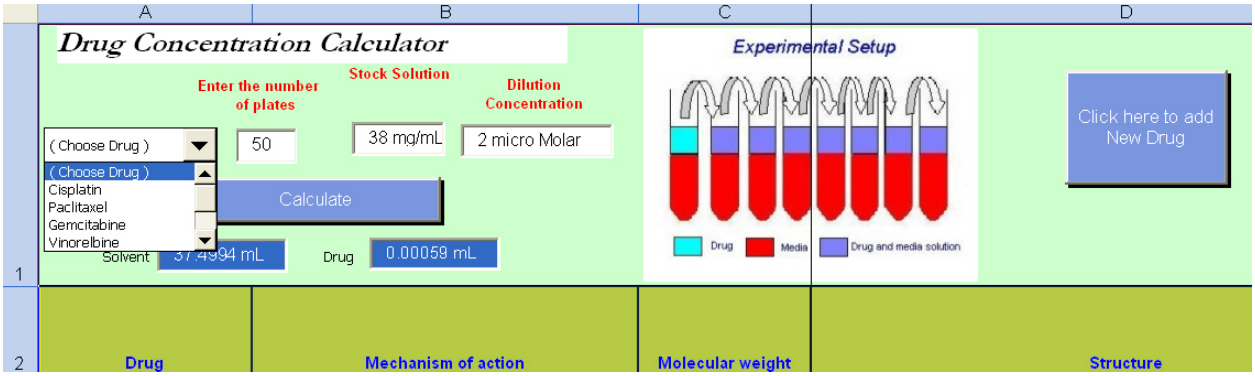


Figure 9.2.2.1: Drug Information and Concentration Calculator

Drug Information database sheet: This sheet contains information on the drug, its mechanism of action, structure, molecular weight, storage and stability information as well as concentration details. This sheet can be updated by the user anytime. Also new drug information can be added in this sheet using the "Add new drug" option.

Add New Drug mode: This mode provides the facility to add a new drug, to be used along with the calculator. The program allows the user to automatically append the new drug information to the existing database. When the "Add New Drug" button is activated, the user sees the screen shown in figure 9.2.2.2. The program prompts the user for new drug name.

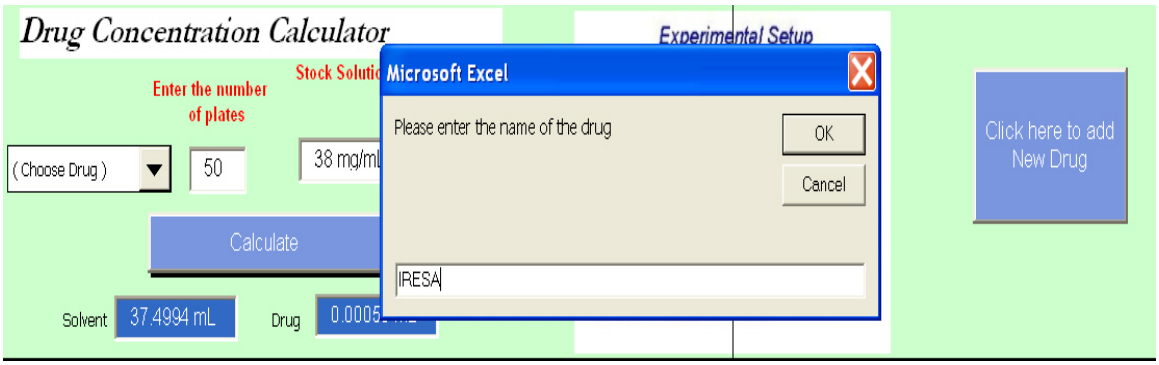


Figure 9.2.2.2: "Add New Drug" mode of Drug Information and Concentration Calculator. "Add New Drug" mode, prompts the user to enter the name of a new drug. Then the program guides the user to a space allocated for the new drug information to be stored. After this process, the "Calculator" mode is ready to calculate stock solution for this new drug.

Requirements to use this program:

1. This program was designed and tested on a PC with Microsoft Excel on Windows XP machine. It will not work on a Macintosh.
2. Macros must be enabled to use this program.

CHAPTER 10

Results, Discussion and Conclusion.

10.1 Results and Discussion.

Chemosensitivity was analyzed using MTT assays and Gene expression data was obtained using Affymetrix. Several observations were made during this study. The IC50 values of each drug varied greatly from one cell line to another (100 to 1000 fold), confirming that there were different phenotypes, sensitive and resistant cell lines. This is demonstrated in Figure 10.1.1.

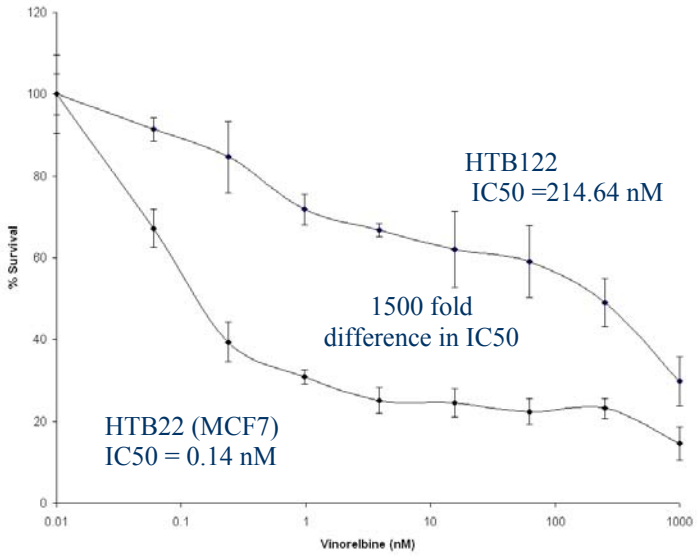


Figure 10.1.1: Sensitivity of Breast Cancer Lines to Vinorelbine. This figure shows a wide range of different IC50 values in breast cancer cell lines (HTB122 and HTB22) for drug Vinorelbine, depicting different phenotypes, sensitive and resistant cell lines.

In order to classify cell lines into different phenotypes for further analysis, the IC50 values of all the cell lines were plotted on a logarithmic scale for the respective drugs. This is shown in the next figure.

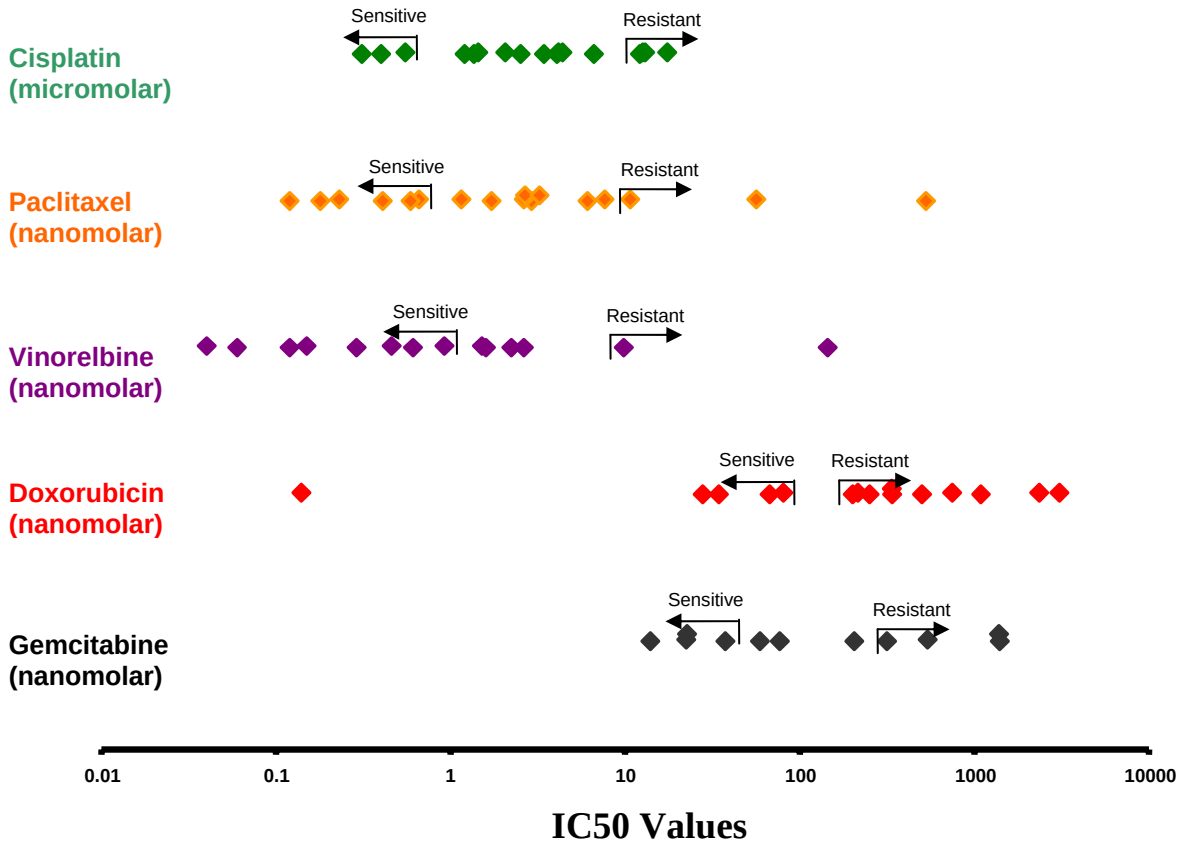


Figure: 10.1.2 This figure represents the different logarithmic scale variations of IC50 values for various cell lines. IC50's of Cisplatin differed by 10-50 fold, IC50's of Gemcitabine differed by 10-100 fold, while those of doxorubicin, paclitaxel and vinorelbine varied by over 1,000 fold for breast cancer cell lines. For breast cancer cell lines IC50's of cisplatin and paclitaxel IC50s varied by up to 1,000 fold.

The next table shows the numerical readings and their classification into resistant and sensitive phenotypes. “Red” represents resistant cell lines and “Green” represents sensitive cell lines.

	IC50s Cisplatin		IC50s Doxorubicin		IC50s Gemcitabine		IC50s Paclitaxel		IC50s Vinorelbine	
Cell Line	Average	SD	Average	SD	Average	SD	Average	SD	Average	SD
HCC1143	6.61	4.84	1092.44	421.85			6.08	6.23	0.29	0.15
HCC1395	0.55	0.01	0.14	0.04	22.71	9.05	7.61	0.57	0.92	0.4
HCC1419	12.15	6.55	338.55	146.82	315.45	126.69	526.72	414.61	9.83	5.87
HCC1428	17.46	8.77	215.52	155.39	540.3	502.99	0.66	0.23	0.46	0.14
HCC1569	3.43	2.88	34.41	20.07	59.25		0.41		2.62	0.82
HCC1806	2.05	1.99	80.46	12.6	22.47		1.15	0.07	1.51	0.57
HCC1937	0.31	0.25	335.3	515.27	205.38	132.95	2.9	1.62	1.59	0.93
HCC1954	4.14	3.23					56.52	79.3		
HCC2688	0.4	0.61	27.92	17.71	1405.65	807.17	0.59	0.14	0.61	0.27
HCC3153	1.44	1	745.85	185.16			0.23	0.03	0.04	0.01
HTB122	1.2	0.68	250.53	31.07	76.84	9.92	1.72	0.16	144.87	124.78
HTB126	12.65	9.94	2356.49	1159.62			10.71	6.21		
HTB131	5.43		200.91	36.12	13.97		0.18		0.06	0.01
HTB22 (M	13.05	10.12	3083.81	1938.32	1390.24	892.08	2.62	1.16	0.15	0.01
HTB24	1.36	1.28	500.8	858.25			0.12		0.12	0.12
HTB25	4.37	1.33					2.66	1.95		
HTB26 (MI	4.06	2.83	67.36	49.07	37.61	17.08	3.22	1.17	2.22	0.84

Resistant
Sensitive

Table 10.1.1: *In Vitro* drug sensitivity and resistance phenotypes for the breast cancer line panel across different drugs. The figure shows the average IC50 values obtained by treating different breast cancer cell lines with various chemotherapy agents along with standard deviation (SD) values.

The expression intensity data was compared between the resistant and sensitive cell lines for every gene. About 100 to 200 genes were upregulated by 4-fold or higher, and 50 to 200 genes were downregulated by 1/4-fold or lower in resistant cell lines when compared

with gene expression in sensitive cell lines. This can be observed in Figure 10.1.3 (a, b, c, d and e). The genes that are upregulated in resistant groups are therefore potential drug resistance genes and the genes that show downregulation are potential drug sensitivity genes.

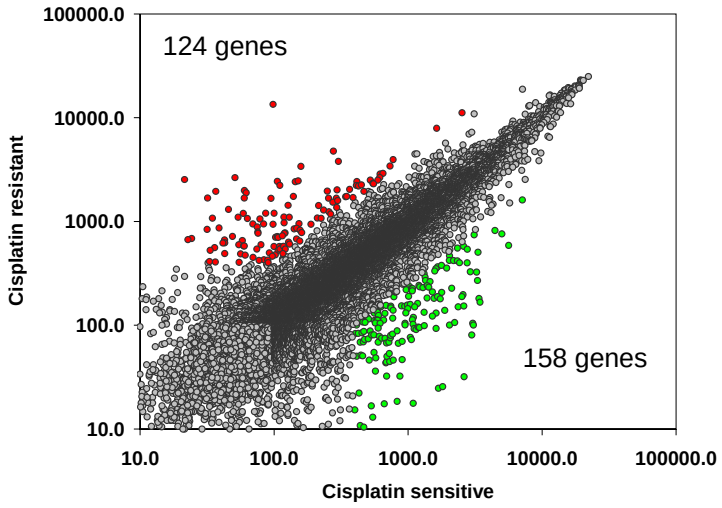


Figure 10.1.3 (a)

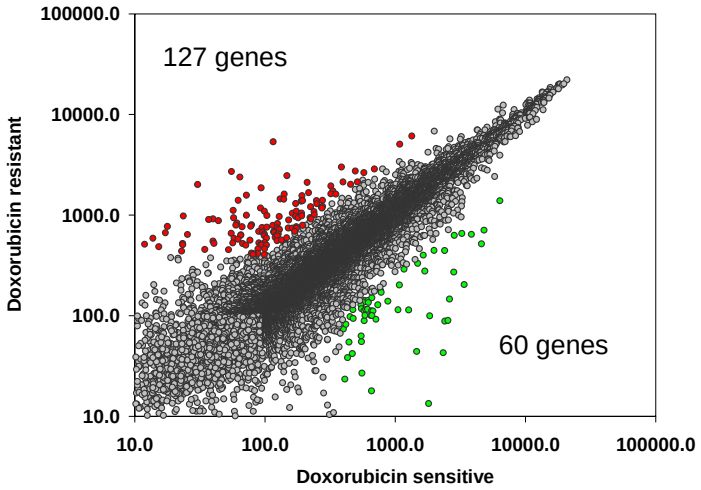


Figure 10.1.3(b)

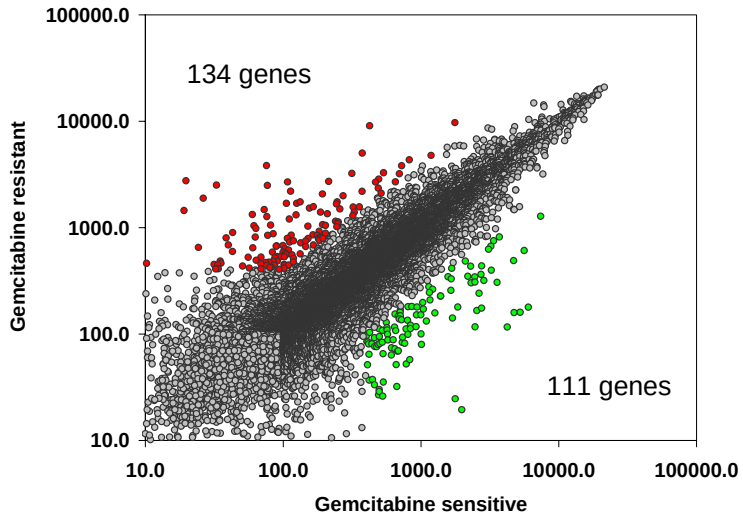


Figure 10.1.3 (c)

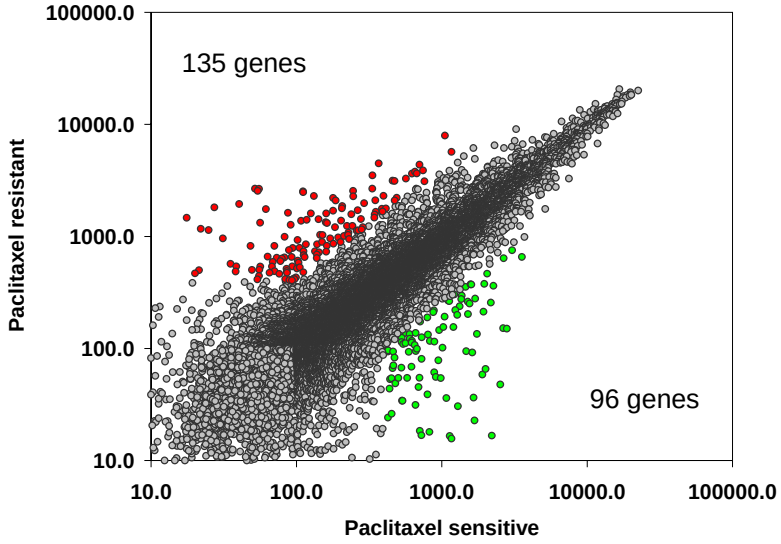
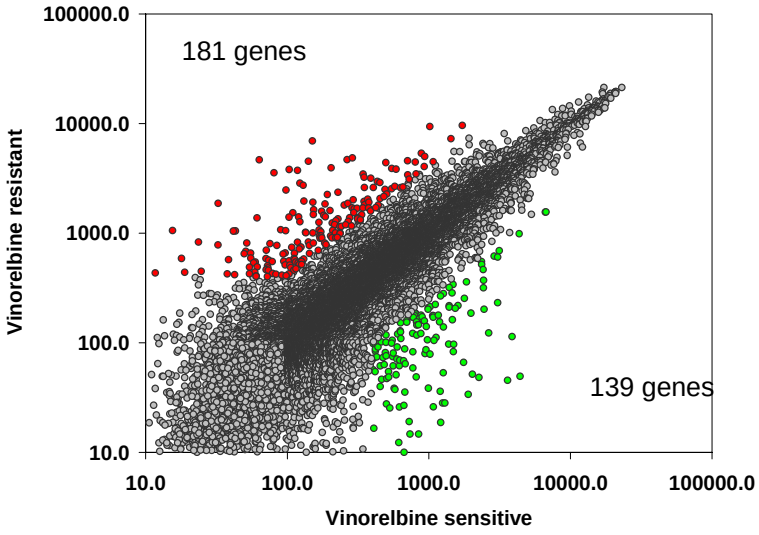


Figure 10.1.3 (d)



96 genes

Figure 10.1.3 (e)

Figure 10.1.3 (a, b, c, d, e): Scatter plots of gene expressions in resistant and sensitive cell lines. Here each spot is representative of a gene on the array and its coordinates represent the expression levels of the two groups being compared.(a) 124 genes are significantly (4 fold or more) up-regulated for and 158 genes are significantly down-regulated in cisplatin resistant cell lines when compared with gene expression in sensitive cell lines. (b) 127 genes are significantly up-regulated for and 60 genes are significantly down-regulated in doxorubicin resistant cell lines when compared with gene expression in sensitive cell lines. (c) 134 genes are significantly up-regulated for and 111 genes are significantly down-regulated in gemcitabine resistant cell lines when compared with gene expression in sensitive cell lines. (d) 135 genes are significantly up-regulated for and 96 genes are significantly down-regulated in paclitaxel resistant cell lines when compared with gene expression in sensitive cell lines. (e) 181 genes are significantly up-regulated for and 139 genes are significantly down-regulated in vinorelbine resistant cell lines when compared with gene expression in sensitive cell lines.

As an alternative approach, and to further narrow down these lists of genes, a correlation table was constructed that displayed Pearson correlations between each gene expression data and each drug IC50 across all cell lines analyzed, as previously described.

Cisplatin log IC50s	0.82	-0.26	1.08	1.24	0.54	0.31	-0.51	0.62	-0.40	0.16	1.12	0.13	0.40	1.10	0.40
LOC257152 expression	25.55	9.28	33.22	27.33	15.51	23.76	16.48	17.76	8.09	11.81	19.01	18.90	21.04	23.25	15.44

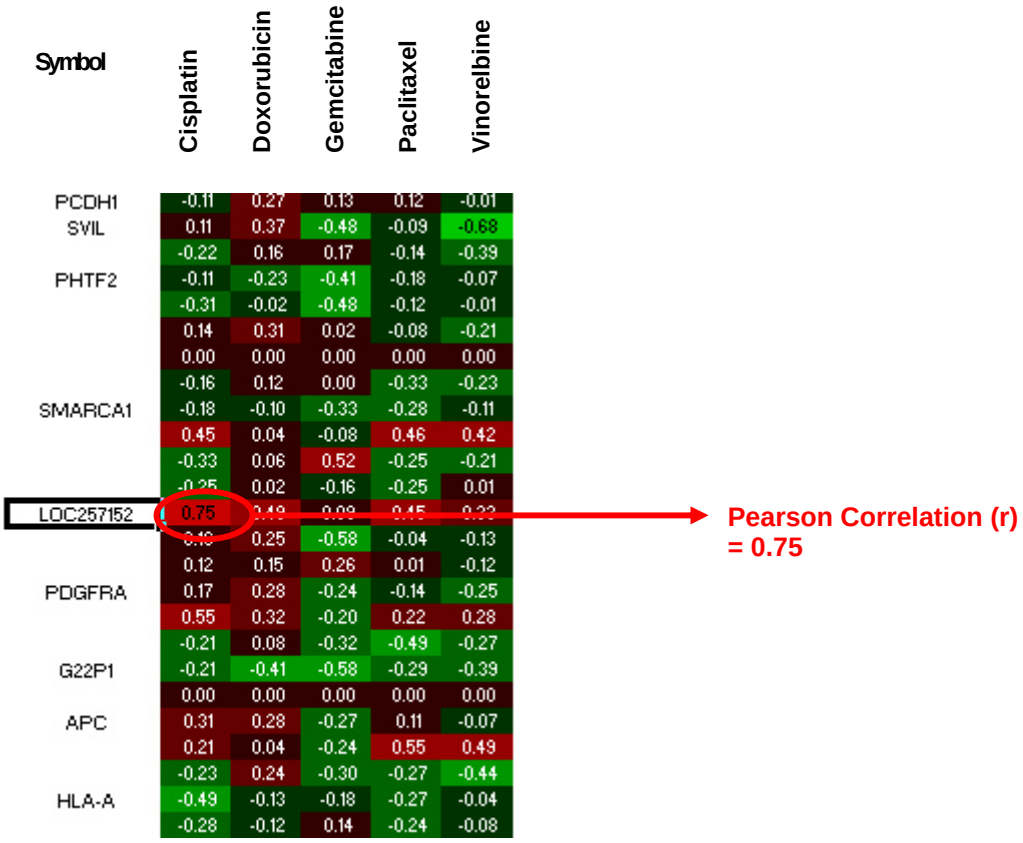


Figure 10.1.4: Correlations between Microarray Data and Drug Assays.

This figure shows that there is a range of correlations between microarray data and chemosensitivity data. A positive correlation for a pair of gene and drug indicates the gene may be associated with resistance to the drug whereas a negative correlation would associate that gene with sensitivity to the drug. A range of correlations (-0.8 to 0.78) was found for all such pairs.

Clustering analysis was performed on this correlation data. The drugs with similar mechanism of action clustered together. DNA damaging agents cisplatin, doxorubicin and gemcitabine clustered together whereas anti-microtubule agents paclitaxel and vinorelbine clustered together. This is shown in figure 10.1.5.

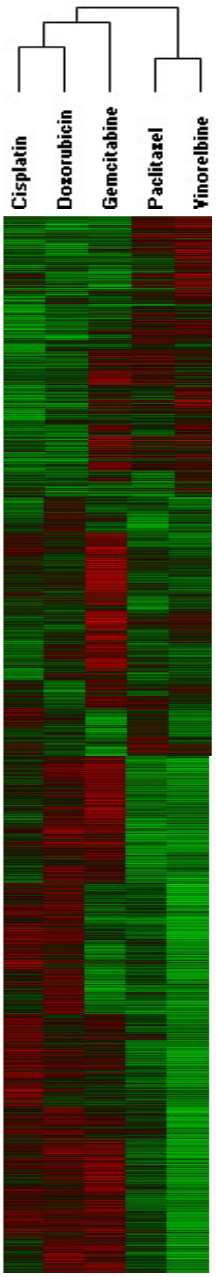


Figure 10.1.5 Clustering of correlation data suggests that sensitivities to the different drugs are associated with unique gene expression profiles. The drugs with similar mechanism of action e.g.: paclitaxel, vinorelbine and cisplatin, doxorubicin, gemcitabine have clustered together.

In order to select genes more precisely, significant correlations (|r| >= 0.5) were compared with genes that were found differentially regulated by the first method (Figure: 10.1.3). Genes were selected and classified into two groups, namely, resistant and sensitive based on the correlation (|r| >= 0.5) between gene expression data and IC50 data and differential regulation (4 fold or more) (Fig. 10.1.6 a, b, c, d and e and Table 10.1.2).

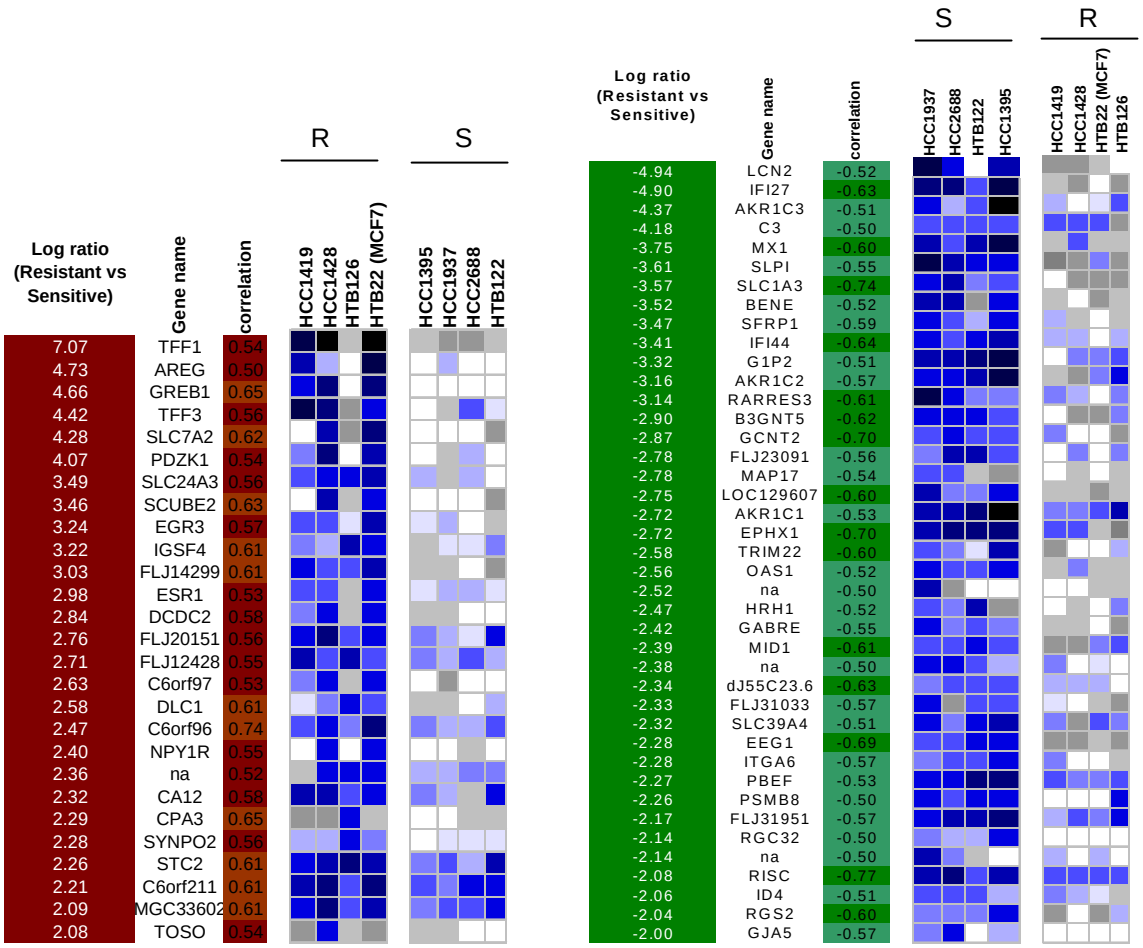


Figure 10.1.6 (a) Groups of genes for cisplatin.

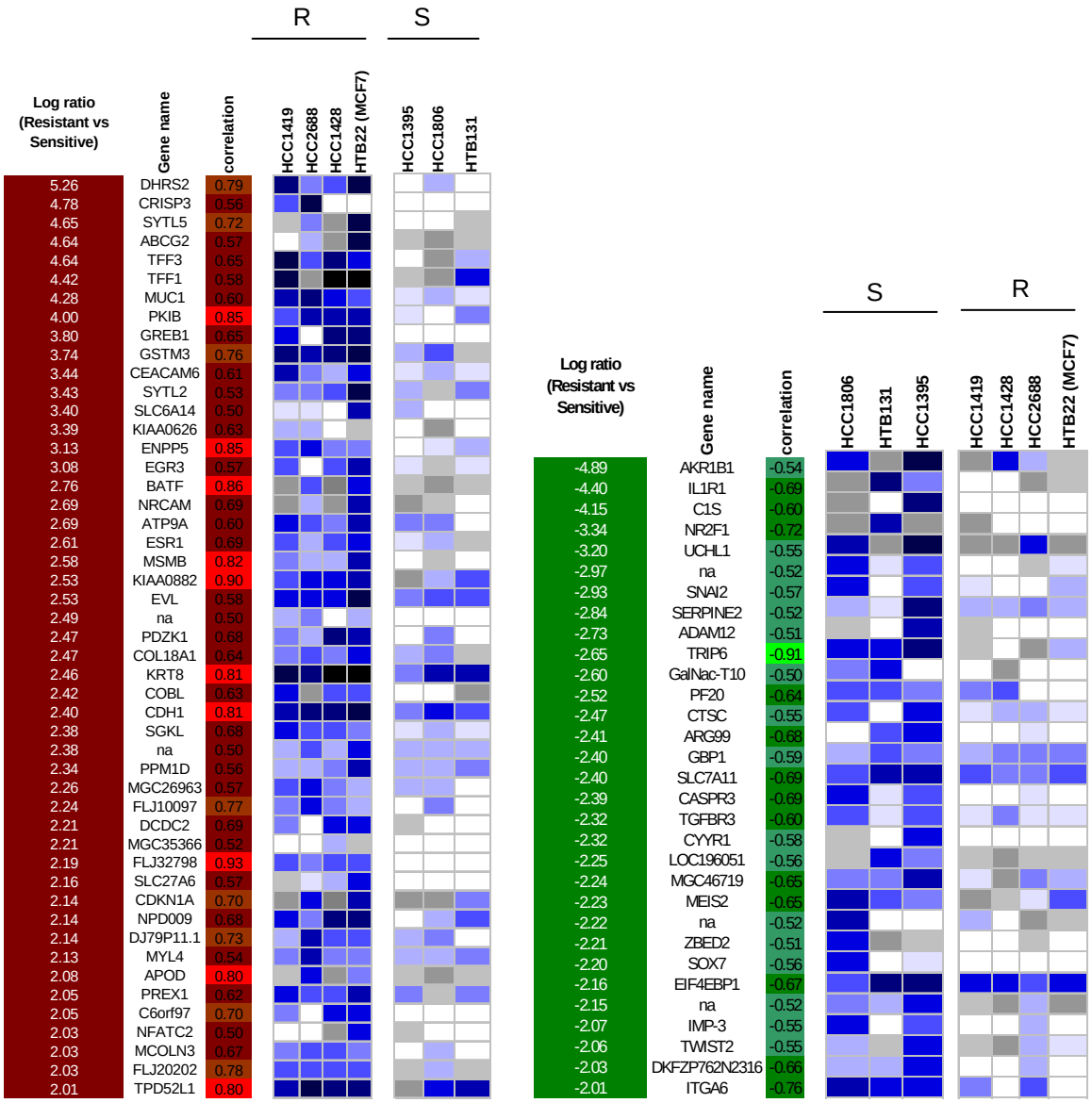


Figure 10.1.6 (b) Groups of genes for gemcitabine.

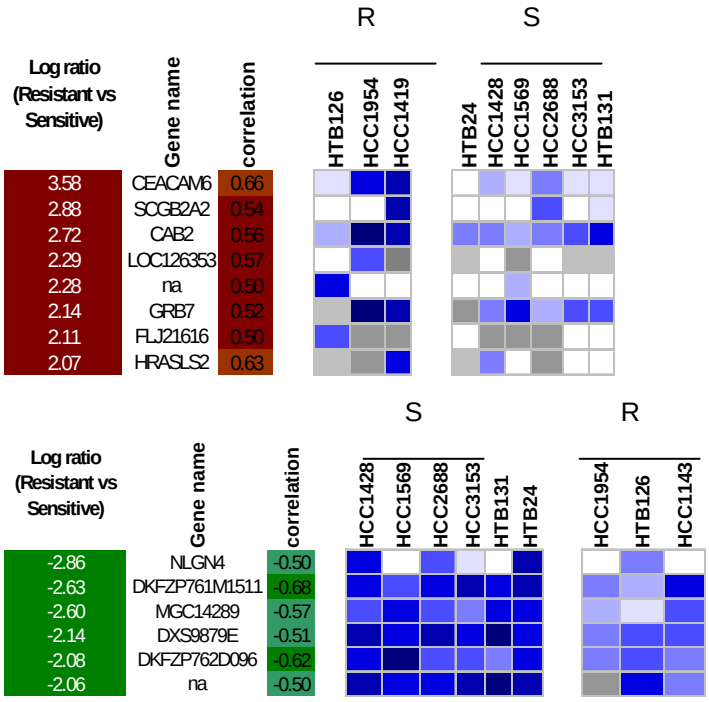


Figure 10.1.6 (c) Groups of genes for paclitaxel

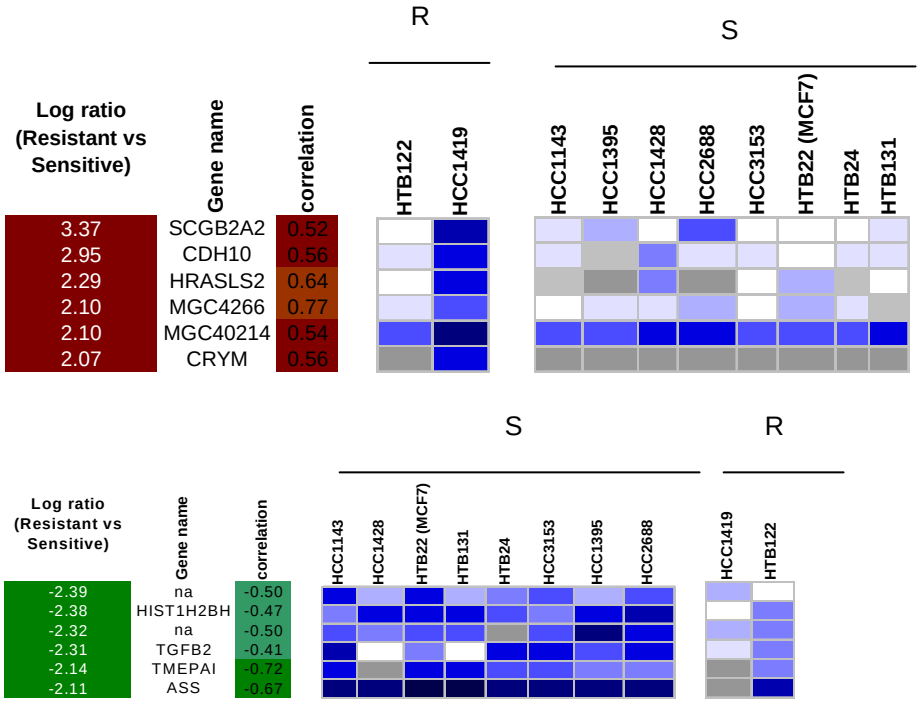


Figure 10.1.6 (d) Groups of genes for vinorelbine.

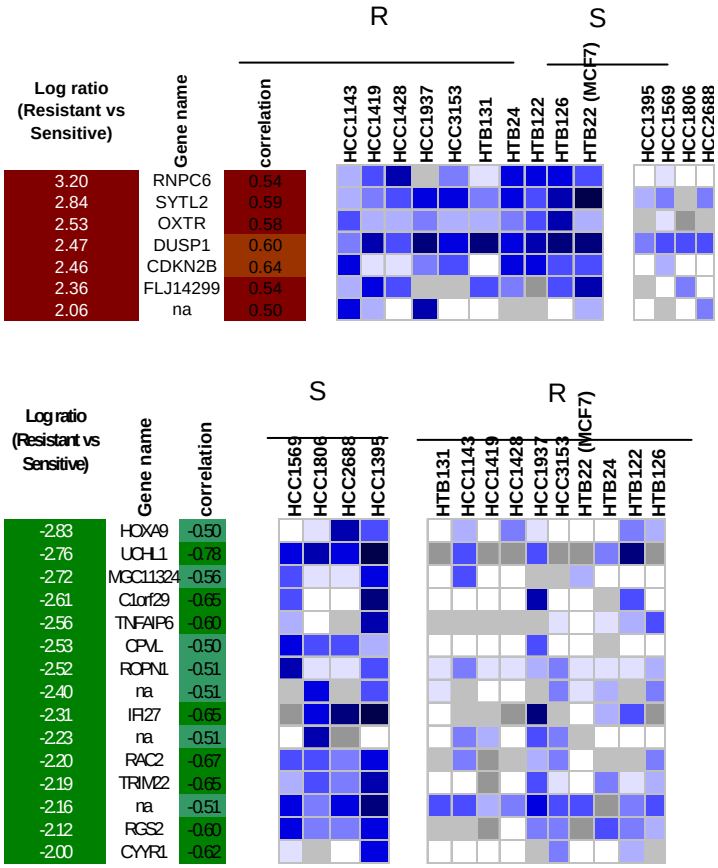


Figure 10.1.6(e): Groups of genes for doxorubicin.

Figure 10.1.6 (a, b, c, d and e): Gene Signatures Associated with Sensitivity or Resistance to Breast Cancer. This figure represents genes with significant correlation of |r| > 0.5 and |Log2 (R/S)| > 2.0 (a) 27 genes correlating with resistance and 42 genes correlating with sensitivity for cisplatin. (b) 49 genes correlating with resistance and 31 genes correlating with sensitivity for gemcitabine(c) 8 genes correlating with resistance and 6 genes correlating with sensitivity for paclitaxel (d) 6 genes correlating with resistance and 6 genes correlating with sensitivity for vinorelbine (e) 7 genes correlating with resistance and 15 genes correlating with sensitivity for Doxorubicin. ('S' stands for sensitive group and 'R' stands for resistant group). The color display in the above figure shows higher expressions with dark colors and lower expression with lighter colors.

The table below lists the number of genes found, after considering significant correlation between drug sensitivity and gene expression data and differential expression data of 4 – fold or more. It also lists the number of common genes found in 2 or more drugs whose mechanism of action was common.

Drugs	Number of genes for Sensitivity	Number of genes for Resistance
Cisplatin	42	27
Doxorubicin	15	7
Gemcitabine	31	49
Paclitaxel	6	8
Vinorelbine	6	6
Cisplatin + Doxorubicin	3	1
Gemcitabine + Doxorubicin	2	1
Cisplatin + Gemcitabine	1	8
Cisplatin + Gemcitabine + Doxorubicin	0	0
Vinorelbine + paclitaxel	0	2

Table 10.1.2: Number of genes associated with breast cancer sensitivity or resistance to one or more drugs(drugs with common mechanism of action).

Many of the genes classified as resistant and sensitive may have common pathways. This information is very useful for future studies. Although, drug resistant mechanisms and

all of the involved genes still remain unrevealed, it was found that the expression of as many as 100 genes were significantly increased or decreased in resistant compared to sensitive breast tumor cell lines. It will be important to define which genes play a direct role in determining chemosensitivity or resistance. One approach would be to use RNAi to silence the gene expression of significantly involved gene and studying the chemosensitivity behavior of the cell line after silencing of this expression. Another approach would be to identify genes upregulated in common to two or more sets of cell lines resistant to different drugs. This may be interesting, because a set of gene expressions were commonly elevated in a set of cell lines resistant to drugs whose mechanisms of action is in common. They are thought to be good candidate genes for future studies. This information can be very useful before starting clinical trials.

10.2 Conclusions

- This study demonstrated that there were distinct chemosensitivity phenotypes of breast cancer cell lines, sensitive and resistant.
- The breast cancer cell lines varied by 100-1,000 folds in their sensitivity to the various drugs.
- A breast tumor sensitive or resistant to one drug often had a different profile to another drug
- Expressions of a set of genes were commonly elevated in a set of cell lines resistant to drugs whose mechanism of action is common.
- Some of these genes might be associated with the drug mechanism of action, and they are good candidate genes for the future mechanistic studies.
- Thus we conclude that gene expression signatures do exist for individual Breast Cancer cell chemosensitivity and these be tested in clinical trials.

Appendix

Hardware configuration

The software program runs on a UNIX based server. The processor type is a 1.2 Ghz Pentium with 500MB of RAM.

Program files and organization

Target designer mode:

File name (Language)	**Purpose**	**Location (Computer)**
index.htm	Input web page	/biotools.swmed.edu/web/siRNA
alias.html	Accession Finder. Retrieves accession number given gene alias.	/biotools.swmed.edu/web/siRNA
index_instructions.htm	Instructions to use the tool.	/biotools.swmed.edu/web/siRNA
FAQ.htm	Help File. (Frequently asked questions about the tool).	/biotools.swmed.edu/web/siRNA
siRNA_Information_Resource_Disclaimer.html	Disclaimer	/biotools.swmed.edu/web/siRNA

sirna.cgi	To design target sequences	/biotools.swmed.edu/cgi-bin/siRNA
blast1.cgi	Perform BLAST operations	/biotools.swmed.edu/cgi-bin/siRNA
aa.cgi	Displays Sense, Anti-sense and position information	/biotools.swmed.edu/cgi-bin/siRNA
refseq	Consists of Refseq database	PostgreSQL database in refmrna database.
Gene_aliases	Consists of Source database	PostgreSQL database in sirna database.

Database ('Resource') Mode:

/index_sirna_info_resource.html	Input web page for resource mode.	/biotools.swmed.edu/web/siRNA
sirnaresource.php	To retrieve all the records of the database	/biotools.swmed.edu/web/siRNA
sirnaresource_genename.php	To retrieve the records of the database with target gene name as input.	/biotools.swmed.edu/web/siRNA

new_sirna_info.php	Input web page for new information. (Password protected).	/biotools.swmed.edu/web /siRNA
detail.php	File to view mRNA images in database.	/biotools.swmed.edu/web /siRNA
detail1.php	File to view protein images in database.	/biotools.swmed.edu/web /siRNA
sirna_source	Consists of siRNA database	PostgreSQL database in sirna database.

REFERENCES

1. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans Nature 391, 806 - 811 (1998); doi: 10.1038/35888 Andrew Fire, Siqun Xu, Mary K. Montgomery, Steven A. Kostas, Samuel E. Driver & Craig C. Mello.

2. RNA Interference and small Interfering RNAs .Thomas Tuschl CHEMBIOCHEM 2001, 2,239-245

3. Characterization of RNA interference in an *Anopheles gambiae* cell line by N. T. Hoa, K. M. Keene, K. E. Olson and L. Zheng

4. Literature review on RNAi mechanism on www.ambion.com.

5. RNA interference and its possible use in cancer therapy. Ait-Si-Ali S, Guasconi V, Harel-Bellan A. Bull Cancer. 2004 Jan;91(1):15-8.

6. Progress in the development of Nucleic Acid Therapeutics for Cancer. Kalota A, Shetzline SE, Gewirtz AM. Cancer Biol Ther. 2004 Jan; 3(1).

7. Small interfering double-stranded RNAs as therapeutic molecules to restore chemosensitivity to thymidylate synthase inhibitor compounds. Schmitz JC, Chen TM, Chu E. Cancer Res. 2004 Feb 15; 64(4): 1431-5.

8. RNA interference targeting focal adhesion kinase enhances pancreatic adenocarcinoma gemcitabine chemosensitivity. Duxbury MS, Ito H, Benoit E,

Zinner MJ, Ashley SW, Whang EE. Biochem Biophys Res Commun. 2003 Nov 21;311(3):786-92.

9. Effect of RNA silencing of polo-like kinase-1 (PLK1) on apoptosis and spindle formation in human cancer cells. Spankuch-Schmitt B, Bereiter-Hahn J, Kaufmann M, Strebhardt K. J Natl Cancer Inst. 2002 Dec 18;94(24):1863-77.

10. Use of RNA interference to validate Brk as a novel therapeutic target in breast cancer: Brk promotes breast carcinoma cell proliferation. Harvey AJ, Crompton MR. Oncogene. 2003 Aug 7;22(32):5006-10.

11. Discovery of inhibitors that elucidate the role of UCH-L1 activity in the H1299 lung cancer cell line. Liu Y, Lashuel HA, Choi S, Xing X, Case A, Ni J, Yeh LA, Cuny GD, Stein RL, Lansbury PT Jr. Chem Biol. 2003 Sep;10(9):837-46

12. Flavopiridol-induced apoptosis is mediated through up-regulation of E2F1 and repression of Mcl-1. Ma Y, Cress WD, Haura EB. Mol Cancer Ther. 2003 Jan;2(1):73-81

13. Refseq : NCBI handbook

14. SOURCE : http://source.stanford.edu/cgi-bin/source/sourceSearch

15. NCBI : http://www.ncbi.nlm.nih.gov

16. BLAST By Joseph Bedell, Ian Korf, Mark Yandell

17. PostgreSQL by Korry Douglas, Susan Douglas

18. Learning Perl Objects, References & Modules By Randal L. Schwartz

19. Bioperl : http://www.bioperl.org

20. PHP and PostgreSQL: Advanced Web Programming By Ewald Geschwinde, Hans-Jürgen Schönig

21. American Cancer Society Breast Cancer Facts & Figures.

22. The siRNA user guide (revised Feb,2004) by Tom Tuschl

23. Rational siRNA design for RNA interference Angela Reynolds,Devin Leake, Queta Boese, Stephen Scaringe, William S Marshall & Anastasia Khovorova. Nature biotechnology Volume 22 Number 3 March 2004.

24. Functional anatomy of siRNAs for mediating efficient RNAi in Drosophila melanogaster embryo lysate. Sayda M.Elbashir,Javier Martinez,Agnieszka Patkaniowska, Winfred Lendeckel and Thomas Tuschl.

25. Guidelines for the selection of highly effective siRNA sequences for mammalian and chick RNA interference. Kumiko Ui-Tei, Yuki Naito et al.

26. Unigene: A Unified View of the Transcriptome (The NCBI Handbook) by Joan Pontius,Lukas Wagner, and Gregory D. Schuler.

27. Fruehauf JP. *In vitro* assay-assisted treatment selection for women with breast or ovarian cancer. Endocr Relat Cancer.2002;9(3):171-82.

28. Cortazar P, Johnson BE. Review of the efficacy of individualized chemotherapy selected by *in vitro* drug sensitivity testing for patients with cancer. J Clin Oncol. 1999; 17(5):1625-31

29. Perez-Soler R, Kemp B, Wu QP, Mao L, Gomez J, Zeleniuch-Jacquotte A, Yee H, Lee JS, Jagirdar J, Ling YH. Response and determinants of sensitivity to paclitaxel in human non-small cell lung cancer tumors heterotransplanted in

nude mice. Clin Cancer Res 2000 Dec;6(12):4932-8

30. Kandioler-Eckersberger D, Kappel S, Mittlbock M, Dekan G, Ludwig C, Janschek E, Pirker R, Wolner E, Eckersberger F. The TP53 genotype but not immunohistochemical result is predictive of response to cisplatin-based neoadjuvant therapy in stage III non-small cell lung cancer. J Thorac Cardiovasc Surg 1999;117(4):744-50

31. Kawasaki M, Nakanishi Y, Kuwano K, Takayama K, Kiyohara C, Hara N. Immunohistochemically detected p53 and P-glycoprotein predict the response to chemotherapy in lung cancer. Eur J Cancer 1998 Aug;34(9):1352-7

32. Johnson EA, Klimstra DS, Herndon JE 2nd, Catalano E, Canellos GP, Graziano SL, Kern JA, Green MR. Aberrant p53 staining does not predict cisplatin resistance in locally advanced non-small cell lung cancer. Cancer Invest 2002;20(5-6):686-92

33. King TC, Akerley W, Fan AC, Moore T, Mangray S, Hsiu Chen M, Safran H. p53 mutations do not predict response to paclitaxel in metastatic nonsmall cell lung carcinoma. Cancer 2000 Aug 15;89(4):769-73

34. Graziano SL, Tatum A, Herndon JE 2nd, Box J, Memoli V, Green MR, Kern JA. Use of neuroendocrine markers, p53, and HER2 to predict response to chemotherapy in patients with stage III non-small cell lung cancer: a Cancer and Leukemia Group B study. Lung Cancer 2001 Aug-Sep;33(2-3):115-23

35. Paclitaxel Story : http://www.21cecpharm.com/px/story.htm

36. Tetrazolim (MTT) Assay for cellular Viability and activity. Methods in Molecular Biology, Vol.79:Polyamine Protocols Edited by: D.Morgan Humana Press Inc.,Totowa,NJ.

37. MATRIX (MicroArray TRansformation In eXcel) 1.24 by Dr. Luc Girard. (Manuscript under preparation).

38. MTT Database 1.10 by Dr. Luc Girard (Manuscript under preparation).

39. Beginning databases with PostgreSQL by Richard Stones and Neil Matthew.

VITAE

Ms. Jyoti Shah was born in Mumbai, India, on July 12, 1979, the daughter of Vanita Shah and Khetsi Shah. She graduated with a Bachelor's Degree in Biomedical Engineering from the University of Mumbai, India in May 2001. She obtained a Masters Degree in Biomedical Engineering from the Joint Biomedical Engineering Program at The University of Texas Southwestern Medical Center at Dallas and The University of Texas at Arlington in June 2004.

During the course of her Masters degree, she served as the President of the Biomedical Engineering Student Society on campus. She received the Alfred and Janet Potvin Award for the Outstanding Biomedical Engineering Graduate Student in 2003.