Microarray Analysis and In silico Drug Designing for Inhibition of Survivin Expression for Treatment of Colon Cancer

Glory Basumata, Tanima Shree

Abstract— Colon cancer is the most common malignancy and the leading cause of cancer related death worldwide. Recent microarray data in the public database demonstrate significant gene expression and epigenetic alterations in tumor conditions. The present study aims at identifying a potential therapeutic target for colon cancer through microarray data analysis and suggests promising leads which could be developed as drugs. To obtain a significant target, a complete list of differentially expressed genes was derived by statistical analysis of colon cancer against normal gene expression data. Using computational network biology approach, a network of all significant genes was built and analyzed on the basis of association of the nodes with cancer pathways. On analysis of nodes predominantly associated with colon cancer pathways, NCI cancer gene index annotation, Survivin was identified as a potential target. Using structure based virtual screening approach two promising leads were identified for Survivin with appreciable docking scores favorable energies.

Index Terms— colon cancer, colorectal, microarray, quercetin, regorafenib, survivin analysis, survivin inhibition.

1 INTRODUCTION

Colon, or colorectal, cancer starts in the large intestine (colon) or the rectum (end of the colon). Other types of cancer can affect the colon, such as lymphoma, carcinoid tumors, melanoma, and sarcomas. According to the American Cancer Society, colorectal cancer is one of the leading causes of cancer-related deaths in the United States. However, early diagnosis can often lead to a complete cure.

Almost all colon cancer starts in glands in the lining of the colon and rectum. There is no single cause of colon cancer. Nearly all colon cancers begin as non-cancerous (benign) polyps, which slowly develop into cancer. Colon cancer is the fourth most common cancer globally with 639,000 deaths reported annually [1]. Typical chemotherapy is provided by injection route to reduce tumor growth and metastasis.

Recent research investigates for the therapeutic drug discovery for inhibition of expression of Survivin which is found to be over expressed in colon. Survivin is a multifunctional protein responsible for controlling cell proliferation and inhibition of apoptosis. Survivin inhibits caspases-3 and caspases-7 activity and the G2/M phase of the cell cycle is also regulated [2]. Survivin is expressed during the embryonic and fetal development which gets overexpressed in variety of cancer such as colorectal, breast, non-small lung cancer and B-cell lymphomas.

Microarray is a high throughput technique to analyze many DNA molecules at molecular level simultaneously. A microarray starts with a piece of glass, or sometimes a silicon chip, of the size of a microscope slide or smaller. Onto this substrate thousands of patches of single-stranded DNA are fixed, called probes, each patch measuring just tens of micrometers across.

Microarray is a significant advance both because they may contain a very large number of genes and because of their small size. Microarrays are therefore useful when one wants to survey a large number of genes quickly or when the sample to be studied is small. Microarray may be used to assay gene expression with a single sample or to compare gene expression in two different cell types of tissue samples, such as in healthy and diseased tissue. Because a microarray can be used to examine the expression of hundreds or thousands of genes at once, it promises to revolutionize the way scientists examine gene expression.

This technology is still considered to be in its infancy; therefore, many initial studies using microarrays have represented simple surveys of gene expression profiles in a variety of cell types. Nevertheless, these studies represent an important and necessary first step in our understanding and cataloging of the human genome.

As more information accumulates, scientists will be able to use microarrays to ask increasingly complex questions and perform more intricate experiments. With new advances, researchers will be able to infer probable functions of new genes based on similarities in expression patterns with those of known genes. Ultimately, these studies promise to expand the size of existing gene families, reveal new patterns of coordinated gene expression across gene families, and uncover entirely new categories of genes. Furthermore, because the product of any one gene usually interacts with those of many others, our understanding of how these genes coordinate will become clearer through such analysis, and precise knowledge of these inter-relationships will emerge.

Corresponding Author: *Glory Basumata, MSc Applied Genetics, Bangalore University, INDIA, PH:+919614790782.*
E-mail:glorybasumata@hotmail.com

The use of microarrays may also speed up the identification of genes involved in the development of various diseases by enabling scientists to examine a much larger number of genes. This technology will also aid the examination of the integration of gene expression and function at the cellular level, revealing how multiple gene products work together to produce physical and chemical responses to both static and changing cellular needs.

A DNA Microarray Experiment

1. Prepare your DNA chip using your chosen target DNAs.
2. Generate a hybridization solution containing a mixture of fluorescently labeled cDNAs.
3. Incubate your hybridization mixture containing fluorescently labeled cDNAs with the DNA chip.
4. Detect bound cDNA using laser technology and store data in a computer.
5. Analyze data using computational methods.

General Steps in Microarray Data Analysis:

1. Creating raw data
2. Background correction
3. Quality control
4. Spot filtering
5. Aggregation and normalization
6. Identification of significant differential expression
7. Pattern recognition

Broadly, steps in microarray data analysis are:

Normalization

Used for standardizing the data and removing variation amongst the data. Normalization is used to calculate true intensities. It is of different types:

1. Local- For single arrays and standard value are calculated for individual spots.
2. Global- For multiple arrays and a single stranded value is calculated for the entire array.
3. Linear
4. Non-Linear

Clustering

It is done to annotate the function of unknown genes by grouping similar genes. This analysis is done on the image of microarray.

Types of Clustering:

1. **Supervised**
   Used to find genes with expression level which are significantly different from other groups of samples and the genes which accurately predict a characteristic of the sample. E.g. Decision tree, B neutral, support vector machine.
2. **Unsupervised**
   Used to find out internal relationship between the dataset. Techniques used: Self Organized Map, Hierarchical, Principal Component Analysis, K-Means, where K denotes the number of clusters.

Hierarchical is further subdivided into Divisive, in which one cluster splits into many; and Agglomerative in which many clusters combine to form on big cluster at the end. K-means comes under centroid based clustering.

Statistical Tests

These tests are done in order to find out the differentially expressed genes. Basically two types of tests are performed on microarray data.

***t*-test:** When two conditions are present for the microarray data then *t*-test is performed[3]. A *t*-test any statistical hypothesis test in which the test statistic follows a students'*t*-distribution if the null hypothesis is supported. The *t*-test assesses whether the means of two groups are statistically different form each other. This anlaysis is appropriate whenever you want to compare the means of two groups, and especially appropriate as the analysis for the post test-only two-group randomized experimental design. A one-sample location test of whether the mean of a normally distributed population has a value specified in a null hypothesis. A two sample location test of the null hypothesis that the means of two normally distributed populations are equal.All such tests are usually called Student's *t*-tests, a test of the null hypothesis that the difference between two responses measured on the same statistical unit has a mean value of zero. For example, suppose we measure the size of a cancer patient's tumor before and after a treatment. If the treatment is effective, we expect the tumor size for many of the patients to be smaller following the treatment. This is often referred to as the "paired" or "repeated measures" *t*-test. A test of whether the slope of a regression line differs significantly from 0.

Calculation: In testing the null hypothesis that the population mean is equal to a specified value μ_0, one uses the statisticical formula.

$$t = \frac{\overline{x} - \mu_0}{s / \sqrt{n}}$$

where $\overline{x}$ is the sample mean, **s** is the sample standard deviation of the sample and ***n*** is the sample size. The degrees of freedom used in this test is $n - 1$.

There are various conditions in which a *t*-test can be applied, i.e, Independent two sample *t*-test, Dependent *t*-test for paired samples etc.

A generalization of Student's statistic, called Hotelling's T-square statistic, allows for the testing of hypotheses on multiple (often correlated) measures within the same sample, i.e multivariate testing which includes One-sample T2 test and Two-sample T2 test.

For performing statistical analysis for microarray data, Multi Expression Viewer (MeV) is generally used.

ANOVA test: When more than two conditions are present for the microarray data then ANOVA test is performed. ANOVA provides a statistical test of whether or not the means of several groups are all equal, and therefore generalize *t*-test to more than two groups. Doing multiple two sample *t*-tests would result in an increased chance of committing a type I error. For this reason, ANOVAs are useful in comparing two, three or more means [4].

Drug Design

Drug discovery and development is an intense, lengthy and an interdisciplinary endeavor. Drug discovery is mostly portrayed as a liner, consecutive process that starts with target and lead discovery, followed by lead optimization and preclinical *in vitro* and *in vivo* studies to determine if such compounds satisfy a number of pre-set criteria for initiating clinical development. For the pharmaceutical industry, the number of years to bring a drug from discovery to market is approximately 12-14 years and costing up to $1.2-1.4billion.

Traditionally, drugs were discovered by synthesizing compounds in a time-consuming multi-step processes against a battery of *in vivo* biological screens and further investigating the promising candidates for their pharmacokinetic properties, metabolisms and potential toxicity. Such a development process has resulted in high attrition rates with failures attributed to poor pharmacokinetics (39%), lack of efficacy (30%), animal toxicity (11%), adverse effects in humans (10%) and various commercial and miscellaneous factors.

Today, the process of drug discovery has been revolutionized with the advent of genomics, proteomics, bioinformatics and efficient technologies like, combinatorial chemistry, high throughput screening (HTS), virtual screening, *de novo* design, *in vitro*, *in silico* ADME screening and structure-based drug design.

In silico Drug Design

In silico methods can help in identifying drug targets via bioinformatics tools. They can also be used to analyze the target structures for possible binding/active sites, generate candidate molecules, check for their drug likeness, dock these molecules with the target, rank them according to their binding affinities, further optimize the molecules to improve binding characteristics.

As structures of more and more protein targets become available through crystallography, NMR and bioinformatics methods, there is an increasing demand for computational tools that can analyze active sites and suggest potential drug molecules that can bind to these sites specifically. Also to combat life-threatening diseases such as AIDS, Tuberculosis, Malaria etc., a global push is essential. Time and cost required for designing a new drug are immense and at an unacceptable level. According to some estimates it costs about $880 million and 14years to develop a new drug before it is introduced on the market intervention of computers at some plausible steps is imperative to bring down the cost and time required in the drug discovery process[5].

The use of computers and computational methods permeates all aspects of drug discovery today and forms the core

Fig1. Timeline for drug discovery and availability of drug on the market.

of structure-based drug design. High-performance computing, data management software and internet are facilitating the access of huge amount of data generated and transforming the massive complex biological data into workable knowledge in modern day drug discovery process. The use of complementary experimental and informatics techniques increases the chance of success in many stages of the discovery process, from the identification of novel targets and elucidation of their functions to the discovery and development of lead compounds with desired properties.

Computational tools offer the advantage of delivering new drug candidates more quickly and at a lower cost.
Major roles of computation in drug discovery are:

- Virtual screening & *de novo* design
- In silico ADME/T prediction
- Advanced methods for determining protein–ligand binding

Structure Based Drug Design

The crystal structure of a ligand bound to a protein provides a detailed insight into the interactions made between the protein and the ligand. Structure designed can be used to identify where the ligand can be changed to modulate the physicochemical and ADME properties of the compound, by showing which parts of the compound are important to affinity and which parts can be altered without affecting the binding. The equilibrium between target and ligand is governed by the free energy of the complex compared to the free energy of the individual target and ligand. This includes not only the interaction between target and ligand but also the salvation and entropy of the three different species and the energy of the conformation of the free species.

Virtual Screening

Virtual screening (VS) is a computational technique used in drug discovery research. By using computers, it deals with the quick search of large libraries of chemical structures in order to identify those structures which are most likely to bind to a drug target, typically a protein receptor or enzyme.

Walters, et al. define virtual screening as "automatically evaluating very large libraries of compounds" using computer programs [6]. The aim of virtual screening is to identify molecules of novel chemical structure that bind to the macromolecular target of interest. Thus, success of a virtual screen is defined in terms of finding interesting new scaffolds rather than many of these hits. Interpretations of virtual screening accuracy should therefore be considered with caution. Low hit rates of interesting scaffolds are clearly preferable over high hit rates of already known scaffolds.

There are two broad categories of screening techniques:

Ligand-based
Given a set of structurally diverse ligands that binds to a receptor, a model of the receptor can be built by exploiting the collective information contained in such set of ligands. These are known as pharmacophore models.
A candidate ligand can then be compared to the pharmacophore model to determine whether it is compatible with it and therefore likely to bind.
A popular approach to ligand-based virtual screening is based on searching molecules with shape similar to that of known actives, as such molecules will fit the target's binding site and hence will be likely to bind the target. Ligand-based methods typically require a fraction of a second for a single structure comparison operation. A single CPU is enough to perform a large screening within hours.

Structure-based
Structure-based virtual screening involves docking of candidate ligand into a protein target followed by applying a scoring function to estimate the likelihood that the ligand will bind to the protein with high affinity. A means of handling the input from large compound libraries is needed. This requires a form of compound database that can be queried by the parallel cluster, delivering compounds in parallel to the various compute nodes.

2 MATERIAL AND METHODS

2.1 PyMol
PyMol is used for visualization of molecules [7]. Fig.2 shows PyMol used for visualization of Survivin.

2.2 Marvin Sketch
Marvin Sketch is the molecule drawing tool of Marvin, a chemical structure drawing and visualizing package, including an integrated chemical file format converter [8]. Fig.3 shows Marvin Sketch used for cleaning the drugs in 3D.

2.3 Cytoscape
Cytoscape is used for visualizing biological molecular interaction networks and integrate global datasets and functional annotations [9]. Fig4. shows Cytoscape networking of gene symbols.

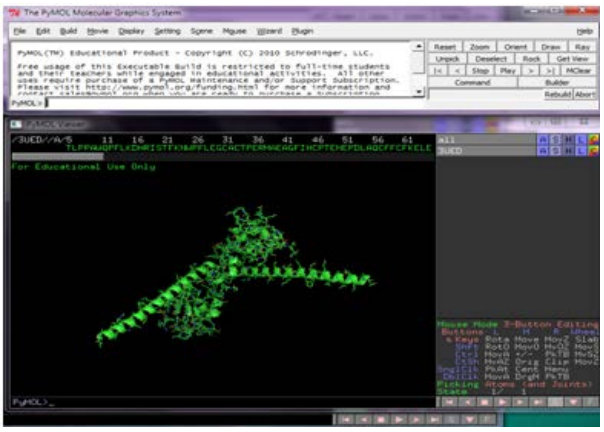


Fig2. PyMol: Visualization of Survivin (PDB ID: 3UED)

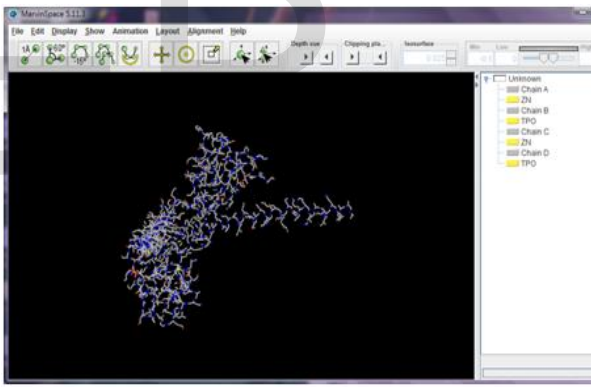


Fig.3 Anti-cancerous drug in 3D cleaned using Marvin Sketch

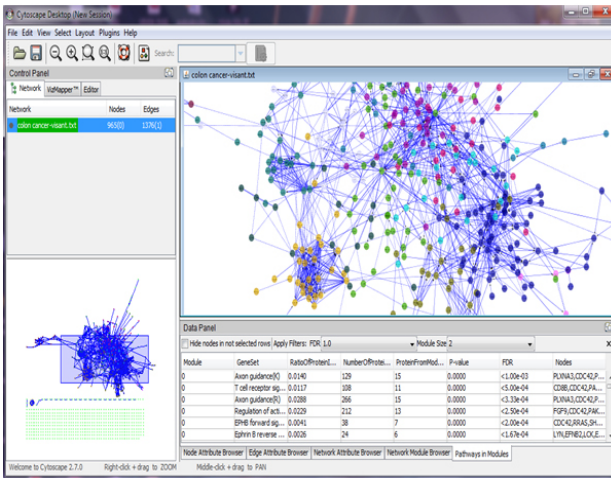


Fig4. Cytocscape: Neworking of gene symbols

2.4 Autodock Tools 4.0

In the field of molecular docking modeling, docking is a method which predicts the preferred orientation of one molecule to a second when bound to each other to form a stable complex. Knowledge of the preferred orientation in turn may be used to predict the strength of association or binding affinity between two molecules using for example scoring functions [10].

Receptor or host or lock- The "receiving" molecule, most commonly a protein or other biopolymer

Ligand or guest or key- The complementary partner molecule which binds to the receptor. Ligands are most often small molecules but could also be another biopolymer.

Docking- Computational simulation of a candidate ligand binding to a receptor

Binding mode- The orientation of the ligand relative to the receptor as well as the conformation of the ligand and receptor when bound to each other.

Pose- A candidate binding mode

Scoring- The process of evaluating a particular pose by counting the number of favorable intermolecular interactions such as hydrogen bonds and hydrophobic contacts.

Ranking- The process of classifying which ligands are most likely to interact favorably to a particular receptor based on the predicted free-energy of binding.

Docking is frequently used to predict the binding orientation of small molecule drug candidates to their protein targets in order to in turn predict the affinity and activity of the small molecules. Hence, docking plays an important role in the rational design of drugs.

To perform a docking screen, the first requirement is a structure of the protein of interest. Usually, the structure has been determined using a biophysical technique such as x-ray crystallography, or less often, NMR spectroscopy. This protein structure and a database of potential ligands serve as inputs to a docking program.

2.4.1 Search Algorithm

The search space in theory consists of all possible orientations and conformations of a protein paired with the ligand. Most docking programs in use account for a flexible ligand, and several attempt to model a flexible protein receptor.
Each "snapshot" of the pair is referred to as pose.

A variety of conformational search strategies have been applied to the ligand and to the receptor. These include:

- Systematic or stochastic torsional searches about rotatable bonds
- Molecular dynamics simulations
- Genetic algorithms to "evolve" new low energy conformation.

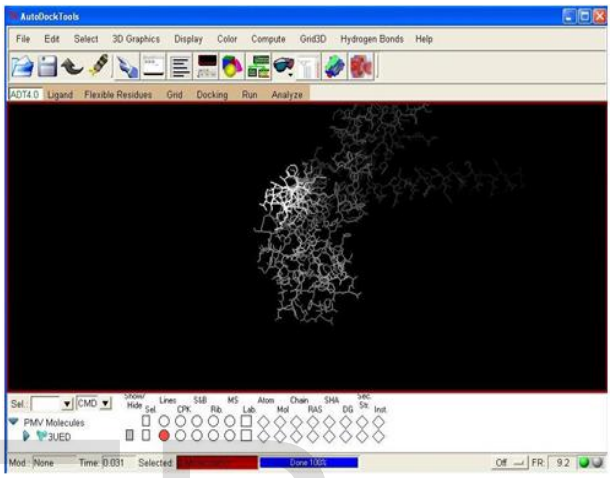


Fig.5 Autodock view of Survivin

2.4.2 Scoring Functions

The scoring function takes a pose as input and returns a number indicating the likelihood that the pose represents a favorable binding interaction.

Most scoring functions are physics-based molecular mechanics force fields that estimate the energy of the pose; a low (negative) energy indicates a stable system and thus a likely binding interaction.

2.4.3 Application of Docking

Docking is most commonly used in the field of drug designmost drugs are small organic molecules, and docking may be applied to:

1. Hit Identification: Docking combined with a scoring function can be used to quickly screen large databases of potential drugs in silico to identify molecules that are likely to bind to protein target of interest (virtual screening).
2. Lead optimization: Docking can be used to predict in where and in which relative orientation a ligand binds to a protein (also referred to as the binding mode or pose). This information may in turn be used to design more potent and selective analogs.
3. Bioremediation- Protein ligand docking can also be used to predict pollutants that can be degraded by enzymes.

3 METHODOLOGY

3.1 MICROARRAY DATA ANALYSIS

3.1.1 Data Collection

Data Collection from Gene expression omnibus (GEO-NCBI)
For microarray, GEO DataSets Record GDS756 was taken.

Title: Colon cancer progression

Summary: Comparison of gene expression in SW480, a primary tumor colon cancer cell line, to that in SW620, an isogenic metastatic colon cancer cell line. Cell lines derived from one individual. Results provide insight the progression of cancer from primary tumor growth to metastasis [11].

Platform: GPL96: [HG-U133A] Affymetrix Human Genome U133A Array

Sample Count: 6

Steps for Data Collection

1. URL for GEO: www. ncbi.nlm.nih.gov/geo
2. Select Datasets: type "colon cancer" and then click on "Go".
3. Record No. GDS "Progression in colon cancer" was selected and .cel files for each record under it was extracted under two heads i.e normal and diseased conditions separately.

3.1.2 Normalization

Normalization is done using **R** (Bioconductor Package) [12]. R – Bioconductor software is shown in Fig.6

For R language, the following commands were used to normalize and cluster the microarray data as extracted from GEO.

1. Load the package "affy" into R session: >library(affy)
2. Read data and store in object raw data: >rawdata<-ReadAffy()
3. For normalization of data: >normal<-rma(rawdata)
4. Write expression values from object to a text file: >write.exprs(normal,file="Normalized.txt")
5. Making variables accessible by name, hence each variable is now accessible by name of CEL files. >attach(normal_table)
6. This would give the list of variables: >names(normal_table)
7. For seeing the variation in raw data and normal data:.boxplot(normal_table)

3.1.3 Statistical Tests (*t*-test)

Statistical test carried out using Multi Expression Viewer (MeV) [13]. To carry out this test, t-test was selected after loading the cell files onto the MeV software. The steps are mentioned below:

1. Open MeV-> Load data -> upload the "Normalized.txt" file as obtained after normalization.
2. From the analysis option on the menu bar-> statistics-> *t*-test
3. Group the data(.cel files) into two groups according to diseased and normal data-> ok
4. Select *t*-tests->table views->significant
5. Save the list of "gene symbols" as a text file.

TABLE 1
AFFYMETRIX HUMAN GENOME U133A ARRAY

Sl. No.	SW 480	Sl. No.	SW 620
1	SW-480-1	4	SW-620-1
2	SW-480-2	5	SW-620-2
3	SW-480-3	6	SW-620-3

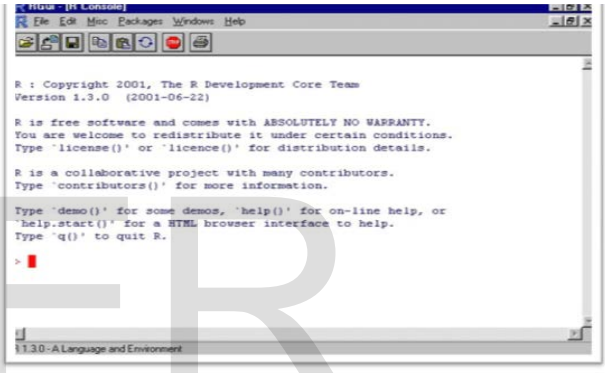


Fig.6 R- Bioconductor

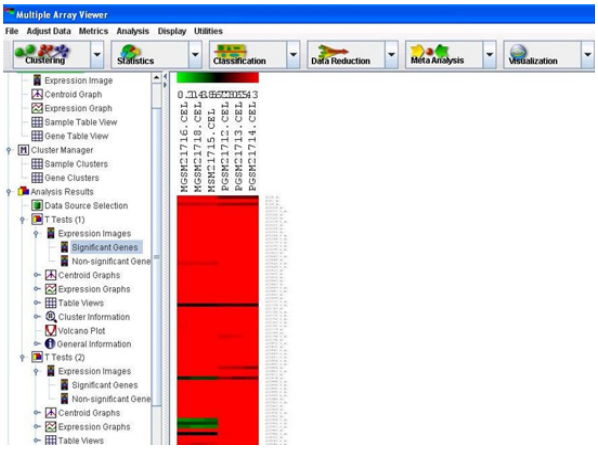


Fig7. MeV *t*-test result for the 6 cell files.

3.2 TARGET IDENTIFICATION

Builing a network in Cytocape

1. Open cytoscape
2. Select plugins-> Reactome-> Gene set mutation. Up-

load the statistical text file as obtained from MeV.
3. Once the network appears-> right click-> Load cancer gene index
4. Right click-> cluster FI network.
5. Right click-> Analyze Network Annotation-> pathway enrichment
6. Choose "pathways in cancer" from the list.
7. Select the nodes highlighted one by one.
   Right click-> Reactome FI-> Load cancer Gene Index.
8. Study by filtering the results.

3.3 FINDING THE LEAD

Finding lead using offline softwares like Autodock 4.0, Hex 6.3, Marvin Sketch and online tools i.e, ADME-Tox.

1. A compound library was made which had 101 ancancerous compounds and there 3D structures were downloaded from Zinc database or Pubchem compound as SDF files.
2. The target selected for colon cancer was Survivin with 3UED as PDB-ID based on annotations from cytoscape network
3. Minimum energy conformation of the compounds were derived using Marvin Sketch
4. Using AutoDock Vina each compound was docked one by one with the target.
5. Open AutoDock and select target molecule and add a grid box with particular dimensions by preparing a configuration.txt file.
6. Select the ligand and adjust torsions and save the file as.pdbqt files.
7. Docking scores were noted.
8. According to the existing drugs dock scores for the disease, the best dock scores were selected.
9. The best dock scoring ligands were redocked with Hex 6.3 software and there energy values were noted and then they were ranked.
10. The complexes obtained after this docking were analyzed and the interactions were studies using PyMol.
11. The result of dock scores, toxicity values, binding site residues has been attached in the result section
12. Finally, the complexes were viewed in PyMol as to see the interaction residues (target and ligand).

Anti-cancerous compounds that are important to inhibit the expression of Survivin are shown in Table2.

Lipinski rule of 5 helps in distinguishing between drug like and non drug like molecules. It predicts high probability of success or failure due to drug likeness for molecules complying with two or more of the following rules:

1. Molecular weight must not exceed 500 dalton
2. Log P value should not exceed 5
3. H-Bond donor is limited to 5
4. H-bond acceptor is limited to 10

TABLE 2
ANTI-CANCEROUS COMPOUND

Sl No.	Anti-cancerous Compound	Sl No.	Anti-cancerous Compound
1.	ABT-751	50.	Herbascutellariabar-bata
2.	Abarelix	51.	Ibritumomabtiuxetan
3.	Abatacept	52.	Ibrutinib
4.	Abexinostat	53.	Imatinibmesylate
5.	Acadesine	54.	Ipilimumab
6.	Acitretin	55.	Laetrile
7.	Alemtuzumab	56.	Lapatinibditosylate
8.	Alitretinoin	57.	Letrozole
9.	Anastrozole	58.	Limonene
10.	APR-246	59.	Liposomal paclitaxel
11.	Baclofen	60.	Lithocholic Acid
12.	Bafetinib	61.	Macitentan
13.	Bardoxolone	62.	Marizomib
14.	Basiliximab	63.	Masoprocol
15.	Benzamide	64.	Nilotinib
16.	Bevacizumab	65.	O6-benzylguanine
17.	Bexarotene	66.	Obatoclaxmesylate
18.	Bortezomib	67.	Obinutuzumab
19.	Bosutinib	68.	Octreotidepamoate
20.	Brentuximabvedotin	69.	Ofatumumab
21.	C-11 choline	70.	Panitumumab
22.	Cabazitaxel	71.	Pazopanib hydrochloride
23.	Cabergoline	72.	Pertuzumab
24.	Cabozantinib	73.	Pralatrexate
25.	Caffeine	74.	Quercetin
26.	Calcitriol	75.	Quizartinib
27.	Calcium citrate	76.	Regorafenib
28.	Calcium gluconate	77.	Retinyl acetate
29.	Cantuzumabravtan ine	78.	Rexinoid NRX194204
		79.	Ridaforolimus
30.	Captopril	80.	Rindopepimut
31.	Carbon C 14 ombrabulin	81.	Rituximab
		82.	Romidepsin
32.	Carboplatin	83.	Sabarubicin
33.	Carfilzomib	84.	Saracatinib
34.	Chloroquine	85.	Sodium dichloroacetate
35.	c-Met inhibitor INCB028060	86.	Sorafenibtosylate
		87.	Sunitinib malate
36.	Crizotinib	88.	Talmapimod
37.	Dasatinib	89.	Tamoxifen
38.	Denileukindiftitox	90.	Temsirolimus
39.	Everolimus	91.	Toremifene
40.	Exemestane	92.	Trastuzumab
41.	Fulvestrant	93.	Tretinoin
42.	Ganitumab	94.	Vandetanib
43.	Gefitinib	95.	Vandetanib
44.	Gimatecan	96.	Varlitinib
45.	Gimeracil	97.	Vemurafenib
46.	Ginger extract	98.	Volasertib
47.	Gossypol	99.	Vorinostat
48.	Halofuginonehydro-bromide	100.	Zalcitabine
		101.	Zanolimumab
49.	Henatinib maleate		

5. Molar refractivity should be between 40-130

Compound passing Lipinski Rule of Five is shown in Table 3.

TABLE 3
ANTI-CANCEROUS COMPOUND THAT PASSED LIPINSKI'S RULE OF FIVE

Sl No.	Anti-cancerous Compound	Sl No.	Anti-cancerous Compound
1.	ABT-751	21.	Lithocholic Acid
2.	Abexinostat	22.	Macitentan
3.	Acadesine	23.	Marizomib
4.	Acitretin	24.	Masoprocol
5.	APR-246	25.	Nilotinib
6.	Baclofen	26.	O6-benzylguanine
7.	Benzamide	27.	Ofatumumab
8.	Bosutinib	28.	Pralatrexate
9.	Caffeine	29.	Quercetin
10.	Captopril	30.	Quizartinib
11.	Crizotinib	31.	Regorafenib
12.	Dasatinib	32.	Rexinoid NRX194204
13.	Exemestane	33.	Romidepsin
14.	Gefitinib	34.	Sorafenibtosylate
15.	Gimatecan	35.	Sunitinib malate
16.	Gimeracil	36.	Talmapimod
17.	Imatinibmesylate	37.	Tretinoin
18.	Laetrile	38.	Vandetanib
19.	Lapatinibditosylate	39.	Vemurafenib
20.	Letrozole	40.	Volasertib

4 RESULT

4.1 Selection of Target

The target selected for colon cancer was Survivin with 3UED as PDB-ID based on annotations from cytoscape network

4.2 Normalization

Normalization of the cell files were done in R Bioconductor. Data is shown in Fig.8 and Fig.9

4.3. Indentification of Lead

The leads identified after docking and energy minimization were

- Quercetin with Compound ID 5280343 and dock score -7.7 is shown in Fig.11
- Regarofenib with Compound ID 11167602 and dock score -7.6 is shown in Fig.12

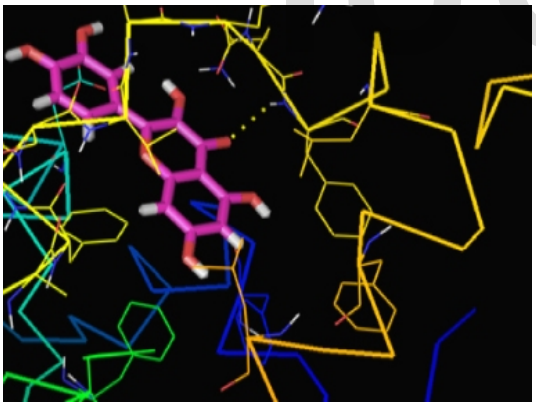


Fig10. PyMol: Quercetin-Survivin interaction

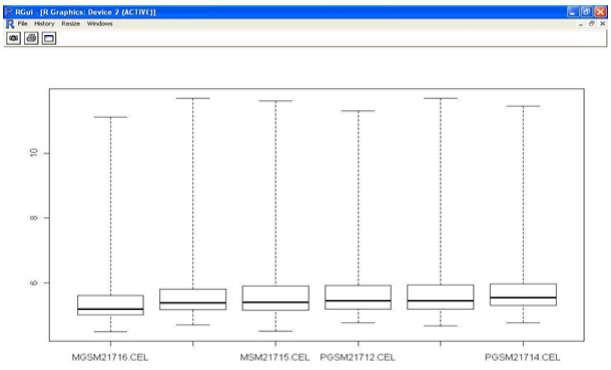
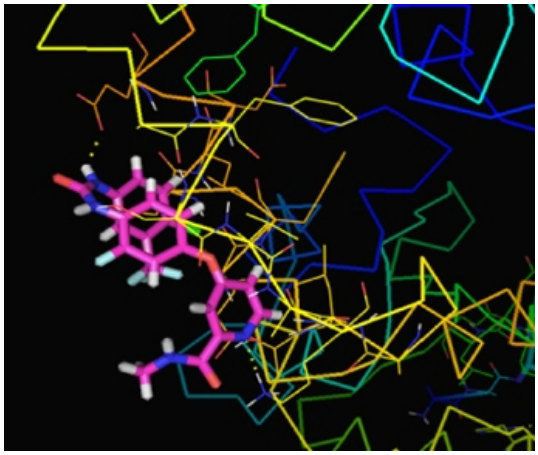
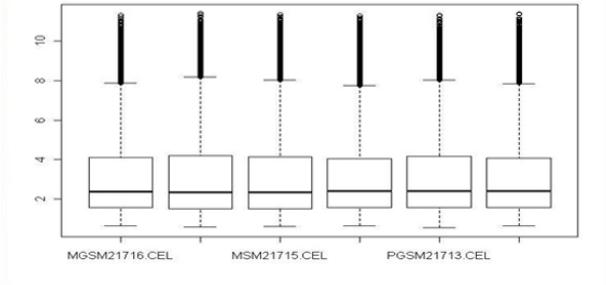


Fig8. Raw table for cell files





[11] GEO-NCBI (URL: www.ncbi.nlm.nih.gov) 2012.
[12] Bolstad, B.M., Irizarry R. A., Astrand, M., and Speed, T.P. (2003), A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and Variance. Bioinformatics 19(2):185-193
[13] MeV. (URL: www.tm4.org/mev) 2012.

4.4 Affinity Value Result

The result for higest affinity value of ligand is shown in Table 4.

4.5. Hex 6.3 Docking Result

The result for docking of ligand with highest E value is shown in Table 5.

5 CONCLUSION

Through microarray data analysis and computational network biology approach we could identity Survivin as a potential target for colon cancer treatment. The leads identified were Quercetin and Regorafenib. These lead are anti-cancerous compound and suggests inhibition of Survivin expression.

The leads that have been identified through this project would be further optimized so that they may be used in the field of drug discovery for colon cancer.

TABLE 4
TOP 10 LIGAND WITH HIGH AFFINITY VALUE

Sl No.	Ligand	Aff. (kcal/mol)
1.	Volasertib	-7.8
2.	Quercetin	-7.7
3.	Regorafenib	-7.6
4.	Lapatinibditosylate	-7.5
5.	Talmapimod	-7.5
6.	ABT-751	-7.5
7.	Imatinibmesylate	-7.4
8.	Quizartinib	-7.3
9.	Pralatrexate	-7.3
10.	Sorafenibtosylate	-7.3

TABLE 5
TOP 10 LIGAND WITH HIGH E VALUE

Sl No.	Ligand	E Value
1.	Volasertib	-299.46
2.	Lapatinibditosylate	-279.64
3.	Imatinibmesylate	-266.09
4.	Quizartinib	-261.84
5.	Pralatrexate	-252.95
6.	Talmapimod	-252.10
7.	ABT-751	-231.13
8.	Regorafenib	-224.86
9.	Sorafenibtosylate	-220.51
10.	Quercetin	-193.6

ACKNOWLEDGMENT

We are extremely grateful to Mr. Ravi Khandelwal, Business Relationship Manager, Institute of Computational Biology, Bangalore, for providing us excellent research facility and guidance during research work.

REFERENCES

[1] National Cancer Institute. (URL:www.cancer.gov) 2012.
[2] Altieri DC. Survivin, cancer networks and pathway-directed drug discovery. Nat Rev Cancer 2008; 8: 61-70
[3] Dudoit, S., Y.H. Yang, M.J. Callow, and T. Speed (2000).Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Technical report 2000 Statistics Department, University of California, Berkeley
[4] ANOVA - One-way Analysis of Variance Zar, J.H. 1999. Biostatistical Analysis. 4th ed. Prentice Hall, NJ
[5] Supercomputing Facility for Bioinformatics & Computational Biology, IIT Delhi; (URL:http://www.scfbioiitd.res.in/tutorial/drugdiscovery.htm) 2013.
[6] Walters WP, Stahl MT, Murcko MA (1998). "Virtual screening – an overview". Drug Discov. Today 3 (4): 160-178. doi:10.1016/S1359-6446(97)01163-X
[7] PyMol (URL: www.pymol.org) 2012.
[8] Marvin Sketch. (URL:www.chemaxon.com/products/marvin/marvinsketch) 2012.
[9] Cytoscape. (URL: www.cytoscape.org) 2012.
[10] Autodock Tools. (UR: www.autodock.scripps.edu) 2012.