Bioinformatics Approach to Predict Structure, Sequence Motif and Expression Level of Breast Cancer Biomarker Molecules:

Protein and miRNA

B.S. THESIS

ADISSERTATIONSUBMITTEDTOBRACUNIVERSITYINPARTIAL FULFILMENTOFTHEREQUIREMENTS FOR THE BACHELOR OF SCIENCE IN BIOTECHNOLOGY

Submitted by: Jannatul Ferdous

Student ID: 12136012

Biotechnology Program

Department of Mathematics and Natural Sciences

BRAC University

Bangladesh

April 2016

Declaration

I hereby solemnly declare that the research work embodying the results reported in this thesis entitled **"Bioinformatic Approach to Predict Structure, Sequence Motif and Expression Level of Breast Cancer Biomarker Molecule: Protein and miRNA"** submitted by the undersigned has been carried out under the supervision of **Ms. Romana Siddique**, Senior Lecturer, Biotechnology Program, Department of Mathematics and Natural Sciences, BRAC University, Dhaka. It is further declared that the research work presented here is original and any part of this thesis has not been submitted to any other institution for any degree or diploma.

Author:

_____

Jannatul Ferdous

Certified:

_____

Ms. Romana Siddique

Supervisor

Senior Lecturer

Biotechnology Program

Department of Mathematics and Natural Sciences

BRAC University, Dhaka

Acknowledgement

First of all, I would like to express my utmost gratitude to the Almighty and His blessings, for helping me with the strength and perseverance needed to successfully complete this project.

My sincere gratitude to **Professor A. A. Ziauddin Ahmad**, Chairperson,

Department of Mathematics and Natural Sciences, BRAC University and **Professor Naiyyum Choudhury**, former Coordinator of the Biotechnology and Microbiology Program of the Department of Mathematics and Natural Sciences, BRAC University for their exemplary guidance and support during my undergraduate life in the Biotechnology Program at BRAC University.

I would like to show my gratitude to my thesis project supervisor **Ms. Romana Siddique** for allowing me to work on this thesis project under her supervision and for his inspiration, ideas and suggestions to improve this work. She has offered me help and support in every step of my project whenever I needed them.

April, 2016 **Jannatul Ferdous**

ABSTRACT

Breast cancer is an important public health issue. It is both a heterogeneous disease and most common cancer affecting women worldwide. It has become the reason of 69% cancer death in women throughout the whole world and 15% of cancer death in Bangladesh. Three reasons can be attributed to this disease, they are: Poor diagnosis, inefficient treatment and continuous recurrence. Within the borders of our own country in 50% cases the diagnosis miss early detection, 40% of the cases face relapse of the disease and most of the patients have to go through inefficient treatment. Geneticists, molecular biologists, cell biologists, oncologists all are trying to find a solution to this burning question of reducing breast cancer mortality rate, and according to recent findings biomarker studies can provide help in this vital research. Surprisingly, even though there has been many discovered biomarker molecules but due to lack of ample information about specific biomarker molecules most of these have not seen any clinical usage, which is ultimately no help to this critical development.

The project was designed to find the basic properties -structure, sequence motif and expression level of potential biomarker molecules of breast cancer. Only protein and miRNA are selected as biomarkers even though there are others, because these two can be found in easily collectable body fluid. 11 protein and 7 miRNA were selected, as they are the one showing high specificity and sensitivity as biomarkers. The protein molecules are - ER, ER Beta, PR, TTR, Ki67, HSP60, Her2, CyclinD1, Cyclin E, P53 and CEA. The miRNAs were- miR10b, miR21, miR145, miR155, miR191, miR 382 and miR425. While doing this project bioinformatics approach was taken to find out properties. For structure SWISS MODEL Workspace (protein), mfold (miRNA), for sequence motif MEME, and for expression level GEO Profiles were used.

This study about the biomarkers can help in the betterment of diagnosis, treatment and recurrence. Because knowing about the important properties of biomarker molecules can help in constructing a biomarker panel for diagnosis, treatment and recurrence. Currently mammography is used for diagnosis which give false results at times, can be replaced by biomarker panel that will detect breast cancer even before any symptoms show up, and for treatment, biomarkers can help in forming anti miRNA targets, site directed mutagenesis, specific inhibitors and virtual screening. For the recurrence part biomarker panel can check people for their risk of local or regional relapse. To sum it up the future of breast cancer treatment is in the hand of a good, specific and sensitive panel of biomarkers.

Table of Contents

LIST OF TABLES

LIST OF FIGURE

LIST OF ABBREVIATIONS

1. BRCA1- Breast Cancer Susceptibility gene 1
2. BRCA2- Breast Cancer Susceptibility gene 2
3. BMI- Body Mass Index
4. LR- Local Relapse
5. RR- Regional Relapse
6. HRT- Hormone Replacement Therapy
7. CBE- Clinical Breast Examination
8. MRI- Magnetic Resonance Imaging
9. FNA- Fine Needle Aspiration
10. ANC- Axillary Lymph Node Clearance
11. ASR- Age Standardized Incidence Rate
12. dsRNA- double stranded Ribonucleic Acid
13. RISC- RNA Induced Silencing Complex
14. HSP- Heat Shock Protein
15. TTR- Transthyretin
16. CEA- Carcinoembryonic Antigen
17. ER- Estrogen Receptor
18. PR- Progesterone Receptor
19. HER2- Human Epidermal Growth Factor Receptor 2
20. Neu2- Neuraminidase 2
21. miRNA- Micro RNA

CHAPTER 1:

INTRODUCTION

In the age of modern invention and advancements, humankind is winning over the world with science. Science and its research has always been working with the problem that the world is facing and solving them. Talking of the problem, there is no doubt that diseases with no recovering medicine, are on the top of all. As days are passing new diseases are emerging, throwing new challenges to the face of science. Each disease needs years of studies to find a remedy and within this time these lethal diseases are wiping people off the earth like dust particle.

Cancer, is the most dangerous of them all. This is not just one disease but a combination of many. It starts with unnatural cell division. (NCI, 2016). Studies, researches- a lot has been going on regarding it, but there's always something missing that keeps mankind slow enough to not win this race over. Surprisingly, Cancer has classification among it, almost 100 types of them are there - some kills people right away and some of them make people suffer till they count their last of breaths. (NCI, 2016) .Breast cancer is one very popular disease while considering killing people, as it is 3rd biggest reason of women death every year. (Waige, 2010).

To save mankind from such a drastic extinction that cancer might cause with time, new types of scientific researches are being designed. Different sections of science works together along with medical science to work out a possible way to fight back with Cancer. Computational biology or in other name, bioinformatics is one very popular sector that is helping these days medical research like no other sections can ever do. (Bolstad, 2003).

Hence this thesis project is inspired from the concept of bioinformatics helping medical science to win over breast cancer. The basic concern of this thesis is to know about breast cancer and its biomarkers- protein and miRNA, then finding their structures, motifs and expression level that might help fighting with the breast cancer in an easier way.

1.1 Breast Cancer

Breast cancer is a type of cancer that happens in the breast tissue. Mostly it means a tumor in the lobules or milk producing ducts. If it happens in the lobular area then it is known as lobular carcinoma and if it happens in the ductal area it is known as ductal carcinoma. Besides there are 18 subtypes of breast cancers. If breast cancer occurs in one cell, then over time it passes through other normal cells and can get into lymph nodes etc. Breast cancer is caused by the physical reasons but in 5-10%cases they happen genetically because of the presence of BRCA1 and BRCA2 genes. (NHS, 2014).

Outcomes of this disease depends on when it was detected (early or late stage), extent of the disease in the body cell. This disease is more common in developed countries and surviving rates are surprisingly low. Breast cancer is consisting 25% the reason of death among all cancer patients. (Li, 2002).

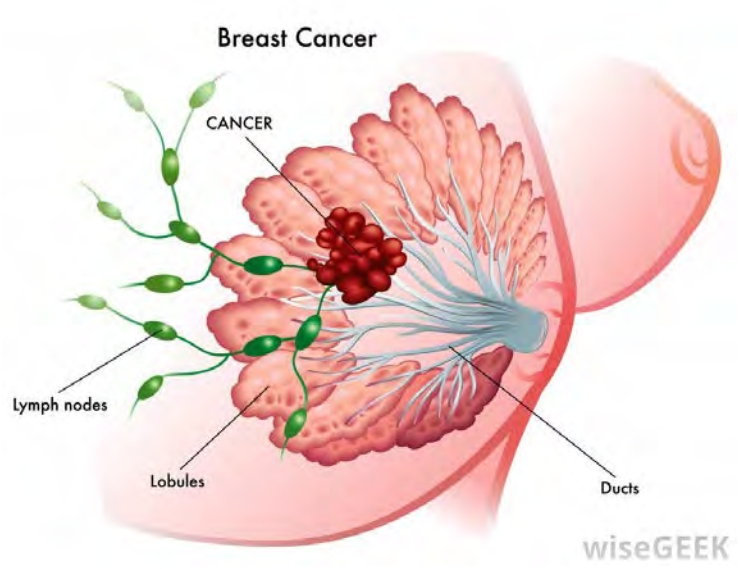


Figure 1.1: Breast Cancer anatomy. (Komen, 2016)

1.1.1 Symptoms

- lump in a breast
- change in breast shape
- fluid excreting from the nipple
- distortment of the nipular or breast skin
- discoloration of the skin from normal color
- skin itching or irritation
- breast or nipple pain
- nipple moving inward
- skin lump formed in the underarm area

1.1.2 Risk Factors

- sex with same-sex people
- overweight to obesity
- absence of physical exercise in daily routine
- early age first period
- not having children or having it late
- older age
- family history

- radiation
- hormone replacement therapy during menopause
- menopause at a late age
- having genetic situations
- having other breast diseases
- having alcohol
- having oral sexual protection

BREAST CANCER RISK FACTORS			
Parameter	**Low Risk**	**High Risk**	**RR**
Sex	male	female	150.0
Age	Young	Old	>10
Family history	No	Yes	2.6
BRCA1 mutation	No	Yes	15
DNA methylation changes in tumor	No	Yes	1.4-5.3
History of benign condition	No	Yes	4.0-5.0
Age at menarche	>14	<12	1.5
Age at first birth	<20	>30	1.9 - 3.5
Age at ovarectomy	<35	no	3.0
Age at menopause	<45	>55	2.0
BMI (postmenopausal)	<22.9	>30.7	1.6
HRT	never	current	1.2-1.4
Bone density	1st quintile	4th quintile	2.7-3.5
Breast density	10%	>75%	4.6
Serum Oestradiol	1st quintile	4th quintile	1.8-2.4
Weight gain	Low	High	1.2-2.3
Height	Low	High	1.3-1.9
Radiation	No	Yes	1.6-5.2
Alcohol	No	Yes	1.4
Smoking	No	Yes	1.13-1.50

BMI=body mass index; HRT= hormone replacement therapy

Table1.1: Predisposing factor of Breast Cancer(Evangelia, 2014)

1.1.3 Classification of Breast Cancer

Breast cancer can be classified into many subclasses based on different parameters-

First of all it can be classified according to the grade-

- well differentiated- low grade
- poorly differentiated- high grade
- moderately differentiated- intermediate grade

Secondly it can be classified based on the stage of the cancer. Tumor size, lymph node involvement and metastasis is the considering parameter here.

Thirdly protein and gene status is another parameter. All breast cancers are tested for ER, PR and neu2 proteins presence or absence and thus classified,

Finally and most importantly histological appearance based-

1. Infiltrating or Invasive Ductal Carcinoma- This is the most common type of breast cancer. It starts mostly in the milk ducts.
2. Medullary Carcinoma- This consists almost 15% of breast cancer cases. Mostly middle aged women are affected by it. And the affected tissue resemblances the color of medulla, hence the naming.
3. Lobular Carcinoma in situ- This is rather a rarer form of noninvasive tumor. It is considered more as a marker of a breast cancer.
4. Infiltrating Lobular Carcinoma- The second most common type of cancer. It starts in the lobules.
5. Tubular Carcinoma- Cancer cell appears like tubules in this case, hence the name. Mostly women above 50 are affected by it.
6. Mucinous Carcinoma-This is a rare invasive breast cancer and rarely spreads to lymph nodes.
7. Inflammatory Breast Cancer- This type of breast cancers are rare but aggressive and leads to blockage of lymph vessels. Affected area or tissues appear like a sheet rather than a lump, swollen and red.
8. Triple Negative Breast Cancer- This type of breast cancer means that it is negative for ER, PR and neu2 protein.
9. Metastatic Breast Cancer- This represents to those type of breast cancer when it spreads to other organs, like brains etc. (Mandal, 2013)

1.1.4 Breast Cancer and the world

Breast cancer is one of the major health problems facing the world today, being a significant contributor to overall morbidity and mortality (Rai, 2012). This cancer is by far the most

frequent cancer among women with an estimated 1.38 million new cancer cases (Kulasingam, 2008). It is now the most common cancer both in developed and developing regions with 690000 new cases(Kulasingam, 2008).And it is the most frequent cause of cancer death in women in developing and developed regions (Rai, 2012).

Figure1.2: Breast Cancer incidence rate worldwide. (Komen, 2016)

Figure1.3: Breast cancer incidence and mortality rate comparison worldwide. (Komen, 2016)

1.1.5 Breast cancer in Bangladesh

Breast cancer remains the most common cancer among women in Bangladesh. It has become a hidden burden which accounts for 69% of cancer death in women (Sacha, 2014).In Bangladesh, the incidence rate of breast cancer was about 22.5 per 100000 in females (Rai, 2012). Breast cancer has been reported as the highest prevalence rate (19.3 per 100,000) among Bangladeshi women between 15 and 44 years of age when compared to other types of cancer (Kulasingam, 2008).The results of the maternal mortality survey conducted by the National Institute of Cancer Research and Hospital in Bangladesh (2010) showed that 21% of total number of death among women between 15 and 49 years of age was due to breast cancer. (Rai, 2012). Apparently, breast cancer is becoming a major public health concern of the Bangladesh government, which is evidenced by the establishment of the National Institute of Cancer & Research Hospital, Bangladesh. (Sacha, 2014).But behind this situation the main possible reason is lack of public awareness for early detection of cancer, very expensive and limited treatment options which reflects the actual situation in rural areas of Bangladesh. (Forazy, 2015). Sadlythis proportions will rise in the next decades if nothing is done to make this rates changed. (Ferlay, 2008).

Features	Bangladesh	India	SouthAsian(SA) immigrantinUK/US	UK/US
Incidence(ASRper100,000)	21.4	25.8	SAversusnon-SA: 40.5versus57.4	95inUK,92.9inUS
Meanage(years)	41.8	45–49	51.8	62.8
Premenopausal	56%	50%	45%	24.5%
ERpositive	63–72%	52–60%	59–71.9%	70–79.3%
Triplenegative	9–22.4%	20–22%	19	8–12%
Histology:invasiveductal carcinoma	95%	88.5%	69.1%	65.6%
Stageatinitialdiagnosis	III–IV:90%	III–IV:60%	III–IV:16%	III–IV:11%
TumorgradeIII	63%	60%	41.9%	34.4%

ASR,age-standardizedincidencerate;ER,estrogenreceptor.

Table 1.2: Comparison of breast cancer incidence rate in different countries. (Hossain, 2014)

1.1.6 Present Screening system of Breast Cancer

Screening is a presumptive identification for disease and not a diagnostic tool. (Sacha, 2014).Screening alerts individuals for further testing the primary goal of screening is to prevent lethal, progressive disease by detecting cancer at an earlier, more treatable stage or by detecting precursor lesions that can be removed before they develop into invasive cancers. The current screening methods used to detect breast tumors either benign or malignant, include clinical breast examination (CBE), mammography and ultrasound. (Kulasingam, 2008).

Mammography is the cornerstone of breast cancer screening and early diagnosis. Calcifications, masses and distortions can be detected with mammography.But there are a number of limitations to mammography the sensitivity and specificity of mammography for women over the age of 50 is 77-84% and 90-94%, respectively but it is lower in women aged 40-49 yrs. (70% sensitivity with 90% specificity) (Kulasingam, 2008). For women under the age of 40, mammographic screening yields a poor sensitivity of only 33%.(Sacha, 2014). Mammography does not detect all breast cancers, besides it suffers from high false positive and negative rates, hazardous exposure and patient discomfort(Kulasingam, 2008).And it also does not provide information regarding the prognosis of the lesion detected.

Physical examination or CBE is also another important detection method since almost 33% of women developing breast cancer are not identified by imaging tools. (Evangelia, 2014). Screening modalities such as ultrasound and MRI are available but they are not trustworthy for use as a population screening tool due to a lack of evidence for its benefit. (Kulasingam, 2008).

1.1.7 Present Diagnosis System

Breast cancer diagnosis can be done in two ways.Fine-needle aspiration (FNA) and core needle biopsy or excisional biopsy. In FNA a needle is inserted into the mass to extract cells which are stained and observed under the microscope to investigate any abnormal cell morphology. Ithas a high diagnostic accuracy, with 10-15% false negative rate, Core needle biopsyutilizes a needle to obtain the specimen.This provides more information than FNA. (Kulasingam, 2008).

1.1.8 Present Treatment System

The current breast cancer treatment includes different therapy: surgery, radiotherapy, chemotherapy, endocrine and molecular therapy, with systemic treatment being given before (neo-adjuvant) or after (adjuvant) surgery. (Kulasingam, 2008).

Surgery has always been the primary treatment for breast cancer. Depending on the tumor size, breast conserving surgeries are decided to perform. Axillary lymph node clearance (ANC) is an important surgical procedure.

Radiotherapy is another important treatment system. Women who are at high risk of recurrence are treated with radiotherapy. Even though it is proving its impact on mortality rate but it is also increase the risk of cardiovascular events. Intraoperative radiotherapy has been suggested to be a better option among all.

Endocrine therapy is another key treatment of adjuvant therapy and for more than twenty years tamoxifen is the most commonly used drug. It is given after chemotherapy rather than at the same time, the standard therapy duration is 5 years. The main disadvantages are that it has antagonistic and agonistic functions on other organs such as the endometrium and bone and cause an increased risk of thromboembolism. After long administration breast cancer patients can become resistant. Some studies are now saying that aromatase inhibitors are also suitable for breast cancer treatment.

Chemotherapy is usually selected for women with high risk of metastatic disease given as an adjuvant treatment after surgery to increase the chance of long-term disease free survival and as neo-adjuvant treatment to reduce the size of the tumor before surgery. The most commonly used chemotherapeutic agents are the anthracyclines (doxorubicin and epirubicin).

Recentlythere is an increasing interest in antibody treatment The most well known is Herceptin, a humanized monoclonal antibody that is directed against the external domain of the *HER2* receptor . A recent study has demonstrated that yearly administration of Herceptin during or after chemotherapy can reduce recurrence risk by 50%. (Kulasingam, 2008).

1.2 The Challenge

Treatment systems are helping lessen the mortality rate but it is not lessening the number of women who develops breast cancer. To make the greatest impact on breast cancer patients the following parameters need to be taken into account:

- the identification of women predisposed to breast cancer by risk prediction markers and
- The application of a preventive or early detection strategy. This need is further magnified by the current controversies of the efficacy of breast cancer screening and the concern about over diagnosis and unnecessary treatment (Kulasingam, 2008).

If we check clearly then it can be seen that 40% of breast cancers have regional or distant spread of their disease at the time of diagnosis. (Rai, 2012). And, survival rates for people diagnosed with advanced breast cancer have changed little over the past 20 years .Without doubt, shifting all cases to early detection will have a profound impact on overall mortality and economic burden. (Rai, 2012).

Unfortunately, no diagnostic or screening test is presently suitable for the early detection of clinically relevant breast cancer. This is because sufficiently high sensitivity (the probability of the test being positive in individuals with the disease) and specificity (the probability of the test being negative in individuals without the disease) are usually both not attributes of the same test; an increase in sensitivity causes a reduction in specificity, and vice versa.so new diagnostic and prescreening methods with improved sensitivity and specificity are clearly needed to identify women with early stage breast cancer. (Kulasingam, 2008).

1.3 Objective

The criteria for effective early detection state that the disease must be common with a high mortality rate. Second, the screening test must accurately detect early-stage disease. Third, the treatment after detection through screening must demonstrate improvements in prognosis and finally, the potential benefits must outweigh the potential harms and costs of screening(Evangelia, 2014).

One of the most promising ways to achieve this is through the use of cancer biomarkers. (Kulasingam, 2008). Because biomarker can be used for detection and developing targeted therapies, besides predicting responses to treatment (Bhatt, 2010). Besides they can provide prognostic information that may facilitate treatment decisions (David, 2010). Moreover a biomarker panel can identify patients at increased risk of local and regional relapse. (David, 2010). For finding the best suited biomarker what can be better than those molecules that are easily found in the body fluids of a person. That are protein and miRNA.

But only studies of biomarkers are not enough. Between 1996 and 2009 there has been 556 publications about biomarkers. But still most of them are not in clinical use. (Baskin, 2010)

So more information's are needed about individual biomarker molecule. That is why this project is designed to be. Hence 11 protein and 7 miRNA are selected based on their biomarker value. It is hopes that a comprehensive understanding of the relevance of each biomarker will be very important not only for diagnosing the disease reliably but also help in the choice of multiple therapeutic alternatives that are currently available. (Bhat2010). It is hoped that individually the information of these biomarker molecule can help in finding a new therapeutic agent, site directed mutagenesis, virtual screening. Also together they can make a panel of biomarkers with better specificity and sensitivity that is needed the most at this moment. (Li, 2002).

1.4 Biomarker

Biomarker is a word that basically represents biological markers. Basically these are the things or molecules that can show the absence, presence, intensity of the disease in a certain condition even if that cannot be known from outside. They can be measured and studied. (Strimbu, 2010). And sometimes their upregulation and downregulation reflects to the present condition of the disease.

A joint venture on chemical safety, led by WHO with the United Nations and the International Labor Organization has defined a biomarker as "any substance, structure or process that can be measured in the body or its products and influence or predict the incidence of outcome or disease."

In fact in a broader sense it also lets us know about the after condition of treatment or chemical or environmental exposure. Examples of biomarkers can include everything from pulse or blood pressure to biomolecules.

Biomarker molecules are considerably newer addition in the medical science discovery. There was a time when specific symptoms would just work as a marker. But that does not help knowing about the stage of the disease or intensity of it. Hence modern science found biomarker molecules in human body in different disease. It is a matter of relieve that these molecule express distinctly in different diseases in different stages that's why chances of misjudgment and misprediction are very less.

An ideal biomarker has certain properties that make it a potential biomarker. Such as-

1. An ideal biomarker has to be easy and safe to measure.
2. It should be cost effective to follow up.
3. It should be easily modifiable with the treatment.
4. It must be consistent in all gender and ethnic groups. (Mandal, 2013.)

Biomarker molecules can be a protein, a nucleic acid, and a hormone or any other form of chemical substances. (Gam, 2010). In a particular disease their distinct expression which is deviant from the normal case, leads to the knowledge about specific condition of that disease in the patient. For example, in disease A, protein molecule B, works as biomarker molecule. In a normal person, B would express in X rate. But in a patient with a disease it would express in 2X amount. This upregulation of the protein molecule helps knowing about the presence of that disease in the patient even though no other symptom is expressed. This is the exact reason why biomarker molecules are being used to detect the presence of a certain disease, as it helps detecting disease where no other symptoms are expressed. This early detection of diseases ease the way to fight back with diseases like nothing else. (Li, 2005). Not only detecting diseases, this also helps knowing how the chemical or medical treatment is responding towards the patient.

Biomarkers can be off three types;

1. Diagnostic biomarker- Biomarkers that actually declares the presence or absence of the disease is known as diagnostic disease.

2. Prognostic biomarker- Biomarkers that actually projects which treatment system is being helpful for the patient.

3. Predictive biomarker- Biomarker that predicts about which patient might respond positively under which treatment system. (Gam, 2010).

Since breast cancer is a very critical disease with a high death rate, using biomarker as weapon to fight back with it can help going a long way. If a patient can be diagnosed with breast cancer at an early stage where this can be stopped or prevented, then this will make the death rate fall right along. Keeping this concern in mind now a days researches are focusing on the biomarker molecules of breast at each stage, trying to know them a little better to design drugs and treatment against them to help the patient.

For breast cancer proteins and miRNA s are really known as great biomarkers for detecting breast cancer and also its stage. (Li, 2005). Some of these proteins are upregulated and some of these are downregulated, some miRNAs are expressed more than normal and some of that miRNAs are less expressed than a normal patient. These indications help finding more information about the condition of the patient with the disease. After knowing these situation drugs, treatment are designed for the patients. This is why this thesis project is designed to know the biomarker molecules a little better so that it can open a new gate of research in the field of drug designing against breast cancer and help in biomedical sector.

1.4.1 Protein as a Biomarker

Protein is one of the main biomolecules of life. To define a protein molecule one needs to say that it is an essential molecule of body which is also a part of diet and important for cell structure and other functions. Proteins are macromolecules which are long chains of amino acids. These amino acid chains are constructed in cells when cellular machinery of the ribosomes translates RNA transcripts from DNA in the cell's nucleus. Within a given human proteome, the number of proteins can be as large as 2 million (Rai, 2012).

Proteins can be organized in four structural levels viz. primary, secondary, tertiary and quaternary. The primary structure refers to the sequence of amino acids in the polypeptide chain. Local folding of amino acid sequence into α helices and β sheets is referred to as the secondary structure and 3D conformation of the amino acid sequence is the tertiary structure. Interaction between multiple proteins subunits folded in 3D gives rise to quaternary structure of a protein each level of protein structure is essential for the proper functioning of the protein molecule. Serving as important components of the physiological pathways in cells, proteins play critical roles in vital functions of the body such as, catalyzing various biochemical reactions as

enzymes; acting as messengers, e.g. neurotransmitters; acting as control elements that regulate cell reproduction; influencing growth and development of various tissues, e.g. growth factors; transporting oxygen in the blood, e.g. hemoglobin; and defending the body against disease, e.g. antibodies (Kulasingam, 2008).

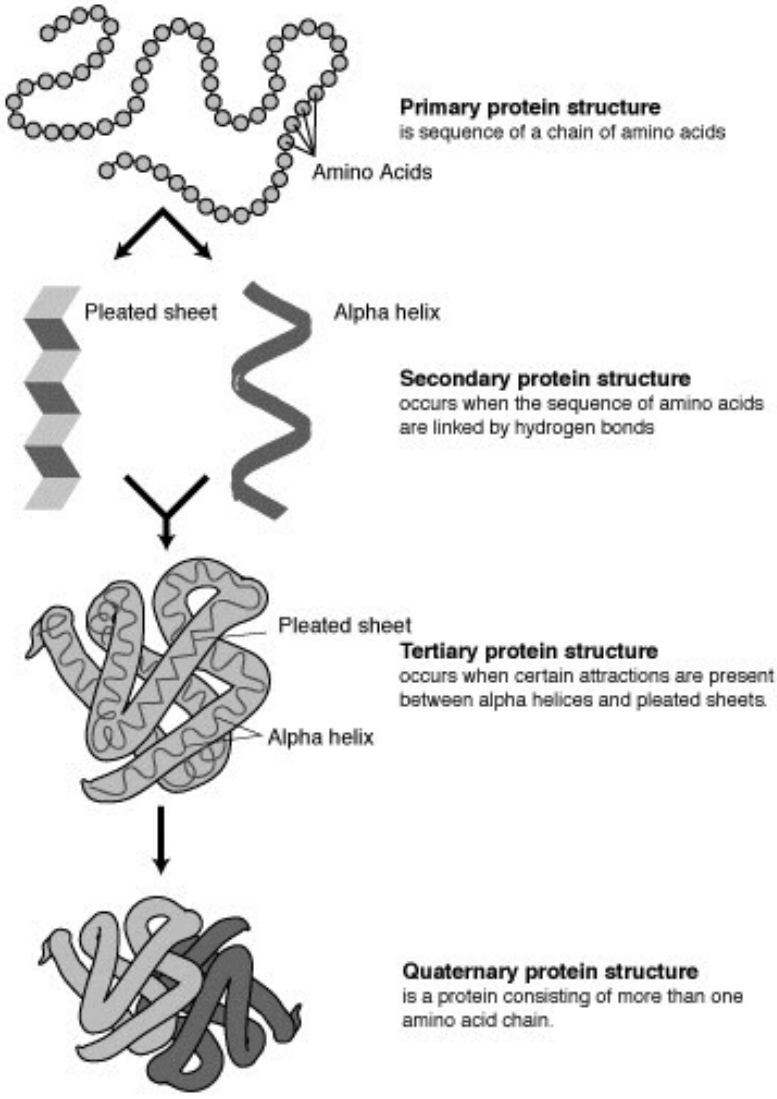


Figure1.4: Structure level of a protein molecule (Rai, 2012)

When accurately characterized, proteins represent the dynamic state of cells, reflecting pathophysiological changes arising due to certain diseases in a more timely and accurate manner than genomic and epigenetic alterations. (Rai, 2012).

Protein molecules are known as the best of biomarker molecules since they can easily be traced, studied and evaluated. (Gam, 2010). Here in this study 11 protein biomarkers are selected and been worked on. They are discussed below-

1. ki67- ki67 is a protein that is strictly associated with cell proliferation and known as marker for the cell proliferation. This is also known as MK167. This is a non-histone nuclear antigen. It can be found in the G1, S, and G2 phase of the cell cycle. And overexpressed in the mitosis and hardly expressed in the G0 phase. This is a very challenging biomarker that can perform as both predictive and prognostic biomarker molecule. (Waige, 2010).

2. CEA- CEA stands for carcinoembrogenic antigen and it is a glycoprotein, mostly found in the serum of a cancer patient. The increased level of CEA in a cancer patient makes her a candidate for breast cancer. This falls under the category of diagnostic biomarkers. (Gam, 2012).

3. TTR- TTR is a carrier protein basically which is found in the cerebrospinal fluid. This carries thyroid hormone and retinol. It is consisted of tetramer of the same subunits. This is a serum protein and their overexpression of this protein proves the women having breast cancer. (Chung, 2014).

4. ER –ER stands for Estrogen Receptor. This is an established prognostic biomarker of breast cancer. In the case of endocrine treatment the patients are always tested for this to check how the treatment is going with the patient. (Waige, 2010).

5. PR- PR stands for Progesterone Receptor. This is a prognostic biomarker. Surprisingly this is hugely associated with ER expression. (Waige, 2010).

6. p53- p53 is known as the tumor protein. Basically its work is to suppress the tumor but once the mutation is happened within a breast cancer patient it stops working and tumorigenesis occurs. In 15% cases of breast cancer this one is found and this also works as a prognostic biomarker. (Misek, 2011).

7. HSP60- Heat shock protein 60 is a chaperonin. Its function is to protein folding, refolding, transportation. When this expresses higher than it can work as a predictive, diagnostic and also as a prognostic biomarker. (Tong, 2016).

8. Her2- Her 2 stands for human epidermal growth factor receptor 2. This can effect growth of some cancer cells. High expression of her 2 is found in the breast cancer patients. This is also found in the 15% cases. (Waige, 2010).

9. Cyclin D1- Cyclin D1 is overexpressed when tumor grows or early onset of cancers as this has the ability to regulate the proliferation of important molecules. This is found in about 50% of the breast cancer cases and can work as a prognostic and predictive biomarker. (Waige, 2010).

10. Cyclin E- This protein is also a member of cyclin family. This plays an important role in the tumorigenesis as this contributes in the regulation of cell cycle transitions etc. This works as a very good prognostic biomarker as this relates with stage and grade of breast cancer while doing chemotherapy or endocrine treatment. (Waige, 2010).

11. ER beta- This molecule was first discovered in 1996. Though this protein is expressed from different chromosome and gene than the other ER molecules, but this shares 59% homology with them in the binding domain. This is downregulated in the patient with a breast cancer and can work as a diagnostic biomarker. (Waige, 2010).

1.4.2 MiRNA as Biomarker

miRNA are small biomolecules that stands for micro RNA. They are small, non-protein coding RNA. Their main function is the regulation of genes and their expression. Mostly they are 9-25nucleotide long in length. Many recent studies have reported that microRNA (miRNA) biogenesis and function are related to the molecular mechanisms of various clinical diseases (Kim, 2009). miRNA was discovered in C. Elegans in 1993 by the Ambros and Ruvkun laboratories.

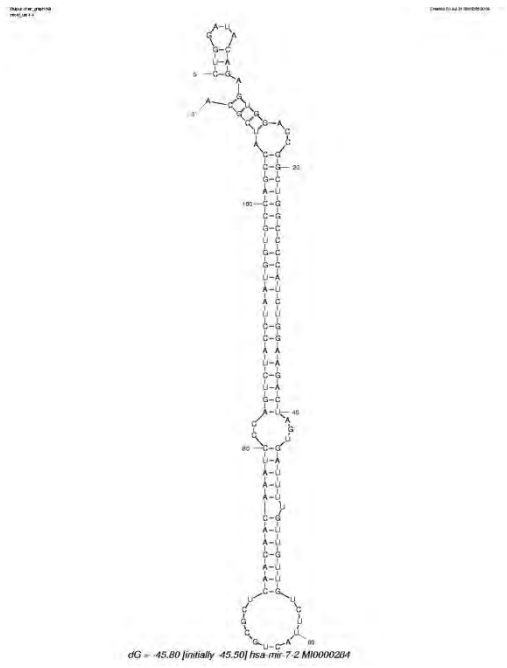


Figure1.5: Secondary structure of miRNA. (Kim, 2009).

1.4.2.1 Biogenesis if miRNA

Most of the genome sequences encoding miRNAs occur in areas of the genome that are not associated with known genes; many are found in fragile sites in human chromosomes and appear to be independently transcribed (Sacha,2014). A number of miRNAs, are

encoded in introns of primary mRNA transcripts. The excision and activation of active single-stranded miRNAs from precursor transcripts occurs through a multi-step process-

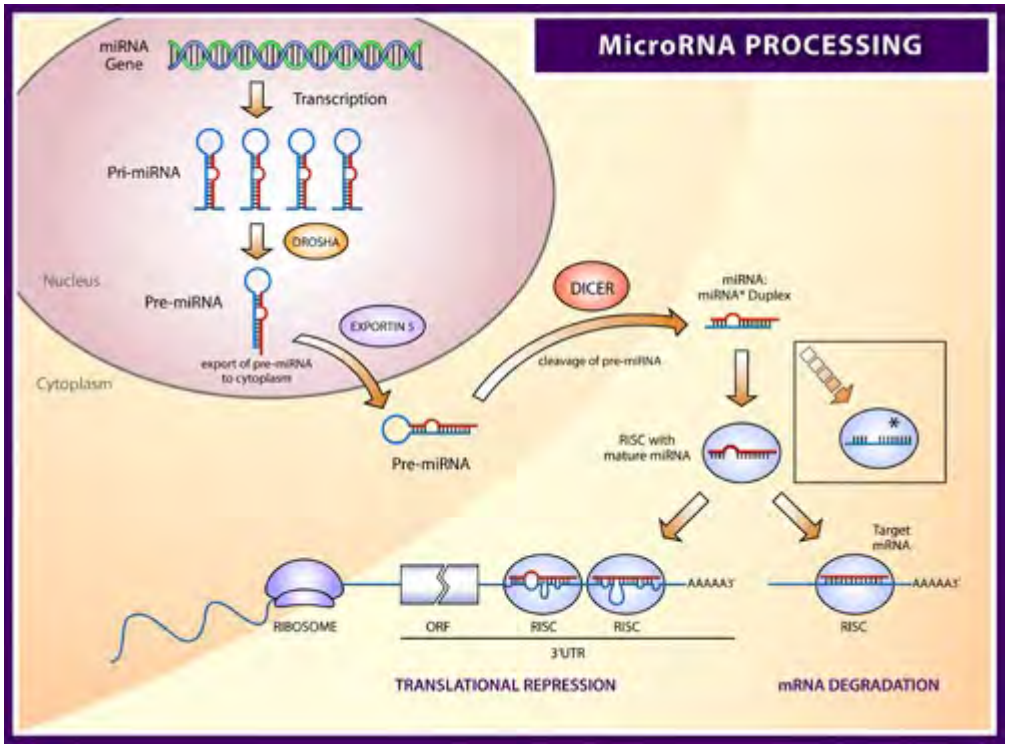


Figure 1.6: Mechanism of miRNA biogenesis. (Kim, 2009).

miRNAs are initially expressed primary miRNAs (pri-miRNAs). They are transcribed by RNA Polymerase II, and include 5' caps and 3' poly (A) tails .The miRNA portion of the pri-miRNA transcript forms a hairpin with signals for dsRNA-specific nuclease cleavage. (Bartel, 2014).

The dsRNA-specific ribonuclease Drosha digests the pri-miRNA in the nucleus to release hairpin, precursor miRNA (pre-miRNA). Pre-miRNAs appear to be approximately 70 nt RNAs with 1–4 nt 3' overhangs, 25–30 bp stems, and relatively small loops. Drosha also generates either the 5' or 3' end of the mature miRNA, depending on which strand of the pre-miRNA is selected by RISC. (Bartel, 2014).

Exportin-5 (Exp5) then exports the pre-miRNAs from the nucleus to the cytoplasm. Exp5 has been shown to bind directly and specifically to correctly processed premiRNAs. It is required for miRNA biogenesis, because of its role in coordination of nuclear and cytoplasmic processing steps. (Bartel, 2014).

Dicer is a member of the RNase III superfamily of bidentate nucleases that has been implicated in RNA interference in nematodes, insects, and plants. Once in the cytoplasm, Dicer cleaves the pre-miRNA approximately 19 bp from the Drosha cut site. The resulting double-stranded RNA has 1–4 nt 3' overhangs at either end. Only one of the two strands is the mature miRNA; some mature miRNAs derive from the leading strand of the pri-miRNA transcript, and with other miRNAs the lagging strand is the mature miRNA. (Bartel, 2014).

To control the translation of target mRNAs, the double-stranded RNA produced by Dicer must strand separate, and the single-stranded mature miRNA must associate with the RISC. Selection of the active strand from the dsRNA appears to be based primarily on the stability of the termini of the two ends of the dsRNA. The strand with lower stability base pairing of the 2–4 nt at the 5' end of the duplex preferentially associates with RISC and thus becomes the active miRNA. (Bartel, 2014).

Patterns of miRNA expression plays a very important role in oncogenesis. Because of their distinct patterns of expression associated with cancer type, remarkable stability in blood and other body fluids, miRNAs are considered to be highly promising cancer biomarkers. (Zhao, 2010).

For breast cancer biomarker 7 miRNAs are selected here that are found very promising biomarkers-

1. miRNA 10b- microRNA 10bs dysregulation is very common in breast cancer. Overexpression of this initiates invasion and metastasis in breast cancer and its expression in primary carcinoma correlates with clinical progression. They are basically found in hox gene clusters. (Harriet, 2007).

2. miRNA 21- This is most significantly upregulated in the breast cancer patient serum .It is one important prognostic biomarker molecule as its expression is related to the tumor stage, disease metastasis and patient survival chance. (Yan, 2015).

3. miRNA 145- It is encoded by mir145 gene. It is hypothesized to be a tumor suppressor. It has been seen to be downregulated in breast cancer and its downregulation contributes to the progression of breast cancer and that is why this has been marked as a potential diagnostic biomarker molecule. (Zheng, 2015).

4. miRNA 155- mir155 is an oncogenic miRNA that plays an important role in the formation of breast cancer. The expression if this miRNA is upregulated in the breast cancer patient. As the high level of miRNA 155 expression is correlated with the subtype and stage of breast cancer hence it is selected as a promising prognostic biomarker molecule. (Liu, 2015).

5. miRNA191- This is an oncogenic miRNA that plays a crucial role in forming breast cancer. Interestingly it is highly induced by the estrogen receptor. As dysregulation of this miRNA is

associated with different cancer hallmarks including metastasis and tumorigenesis, it is a great prognostic biomarker too. (Nagpal, 2013).

6. miRNA382- This is another oncogenic miRNA which is upregulated in the breast cancer patient than the serum of a normal disease free person, the fine specificity and sensitivity makes it a good candidate of breast cancer biomarker molecule. (Aguilar, 2013).

7. miRNA425- Another oncogenic miRNA. This is marked as an important biomarker because it is two times dysregulated in the serum of a breast cancer patient than the normal one. (Zhao, 2010).

CHAPTER 2:
METHODS

This section is divided according to the two biomarker molecule that were worked with. Since miRNA was worked with first to be checked their structure, motifs and expression level followed by protein molecules, this section is designed exactly by that order of working.

2.1 miRNA

2.1.1. Structure Prediction

While predicting structures of miRNAs, few steps had been followed. First the sequences of the miRNAs had to be taken from a database. In this project miRBase was used for that. Secondly the sequences were given as input in the secondary structure forming web tool that is Mfold. Details about these two database and website are given below:

2.1.1.1 miRBase

miRBase stands for microRNA base. The miRBase database is a searchable database of published miRNA sequences and annotation. Each entry in the miRBase Sequence database comes with a predicted hairpin portion of a miRNA transcript (termed miR in the database), with information on the location and sequence of the mature miRNA sequence (termed miR). Both hairpin and mature sequences are available for searching and browsing, and all the entries can also be retrieved by name, keyword, references and annotation. All sequence and annotation data are also available for download.miRBase is managed by the Griffiths-Jones lab at the Faculty of Life Sciences, University of Manchester with funding from theBBSRC. miRBase was previously hosted and supported by the Wellcome Trust Sanger Institute. (Kozomara, 2014). URL Link: http://www.mirbase.org/

Figure 2.1:miRBase homepage

2.1.1.2 mfolds

The mfold web server is one of the oldest web servers. It has been in full operation since the fall of 1995 and was first introduced at Washington University's School of Medicine. Mfold is a software that actually helps forming secondary structure of a nucleic acid just by the sequence of it. (Zucker, 2003). URL Link: http://unafold.rna.albany.edu/?q=mfold.

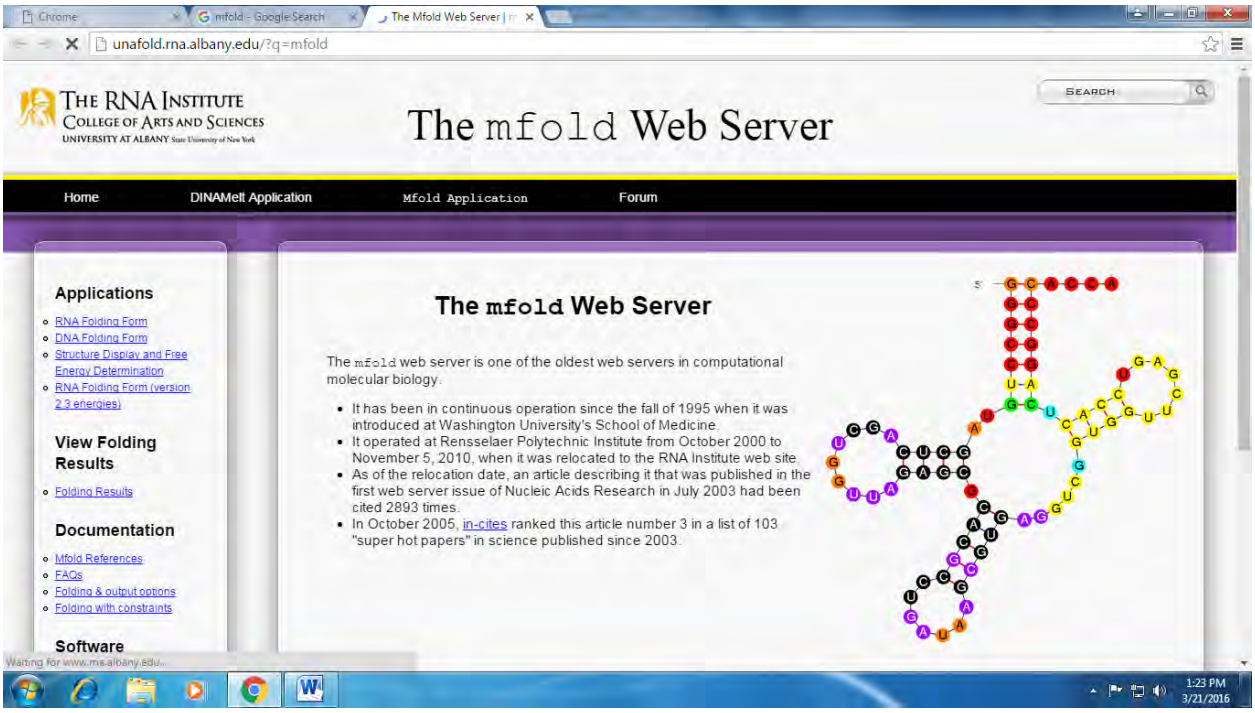


Figure 2.2 : The mfold Web Server homepage.

2.1.2 Discovering Sequence Motif

For discovering sequence motif two steps were followed – first getting the fasta format of a sequence from a database and then using the sequences as input into a software that can work best for finding motifs. For the first step miRBase and for the second MEME suite software was used.

2.1.2.1 miRBase

Mirbase stands for microRNA base. The miRBase database is a searchable database of published miRNA sequences and annotation. Each entry in the miRBase Sequence database comes with a predicted hairpin portion of a miRNA transcript (termed miR in the database), with information on the location and sequence of the mature miRNA sequence (termed miR). Both hairpin and mature sequences are available for searching and browsing, and all the entries can also be retrieved by name, keyword, references and annotation. All sequence and annotation data are also available for download.miRBase is managed by the Griffiths-Jones lab at the Faculty of Life Sciences, University of Manchester with funding from theBBSRC. miRBase was previously hosted and supported by the Wellcome Trust Sanger Institute. (Kazamara, 2014). URL Link: http://www.mirbase.org/

Figure 2.3: miRBase homepage.

2.1.2.2 MEME

MEME stands for Multiple Expectation maximization for Motif Elicitation. MEME discovers novel, ungapped motifs (recurring, fixed-length patterns) in nucleic acid or protein sequences (sample output from sequences).

MEME splits variable-length patterns into two or more separate motifs. MEME represents motifs as position-dependent letter-probability matrices which describes the probability of each possible letter at each position in the pattern. Individual MEME motifs do not contain gaps. Patterns with variable-length gaps are split by MEME into two or more separate motifs.

MEME takes as input a sequence or a group of sequences and outputs as many motifs as requested. This tool can choose the best width, number of occurrences, and description for each motif by the help of statistical modelling. MEME on the web can take a second (control) set of input sequences and then discovers motifs that are enriched in the primary set relative to the control set. This discovery is called discriminative motif discovery.The MEME Suite was developed byTimothy Bailey at the Institute for Molecular Bioscience at the University of Queensland and William Stafford Noble in the Department of Genome Sciences at the University of Washington. . This Suite have previously been supported byColumbia University, the Computational Biology Research Center at the National Institute of Advanced Industrial Science and Technology, the National Biomedical Computation Resource, and the San Diego Supercomputer Center. Maintenance and development of the MEME Suite is funded by

the National Institutes of Health.it also receives support from Amazon and Google. (Timothy, 2009). URL Link: http://meme-suite.org/tools/meme

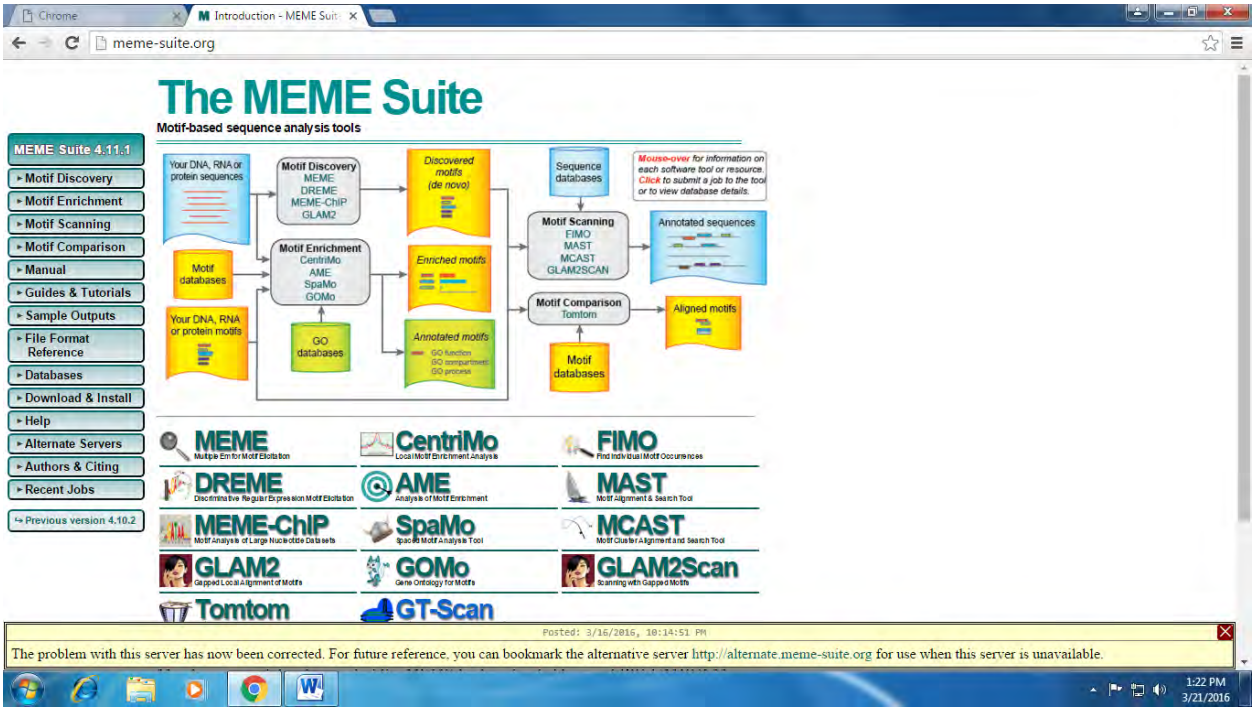


Figure 2.4: MEME Suite homepage

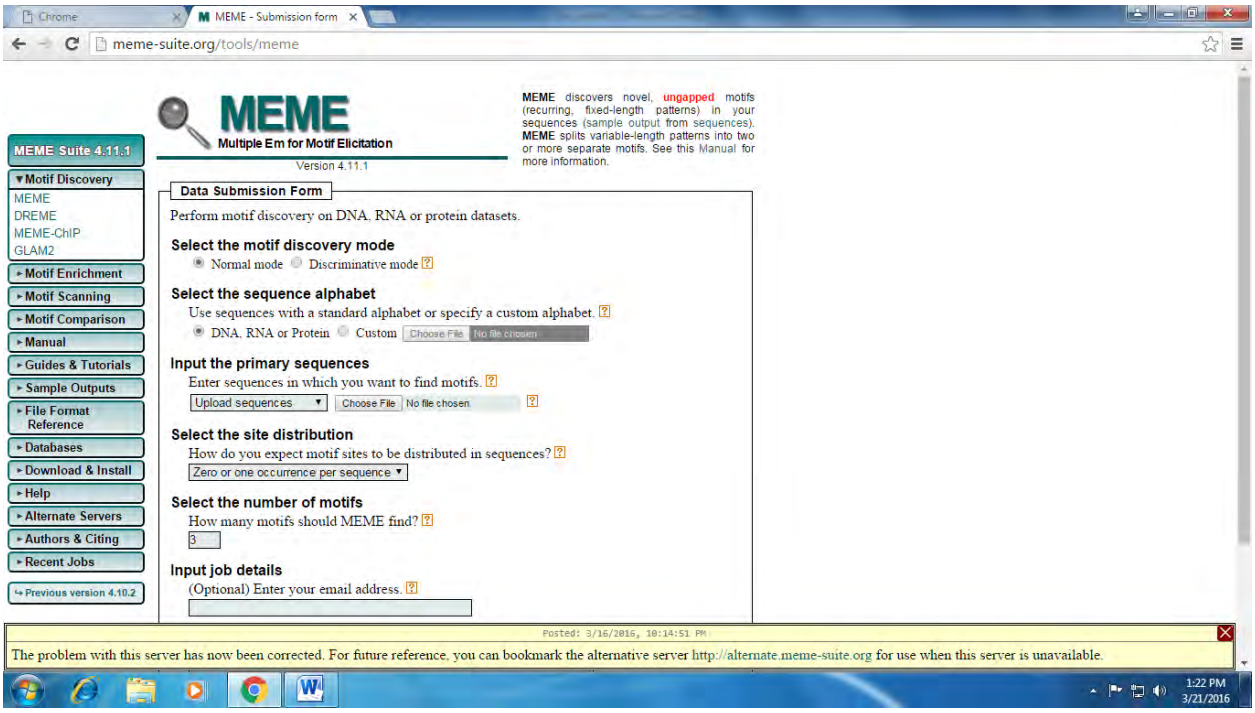


Figure 2.5: MEME Suite data submission page

2.1.3 Finding Expression Level

To find out the expression level of these miRNA in different conditions of breast cancer GEO tool of NCBI was used. Just the name of the miRNA name and little clue “breast cancer” was given as input and possible options appear.

2.1.3.1 GEO Profile

GEO represents to Gene Expression Omnibus. The GEO Profiles database stores gene expression profiles derived from curated GEO Datasets. Each Profile is presented as a chart that displays the expression level of one gene across all Samples within the Dataset. Experimental parameter is provided in the bars along the bottom of the charts to see whether a gene is differentially expressed across different experimental conditions. Every Profile have various types of links including internal links that connect genes that exhibit similar behavior, and external links to relevant records in other NCBI databases.GEO Profiles can be searched using many different attributes including keywords, gene symbols, gene names, GenBank accession numbers, or Profiles flagged as being differentially expressed.(Barrett, 2013).
URL Link: https://www.ncbi.nlm.nih.gov/geoprofiles/

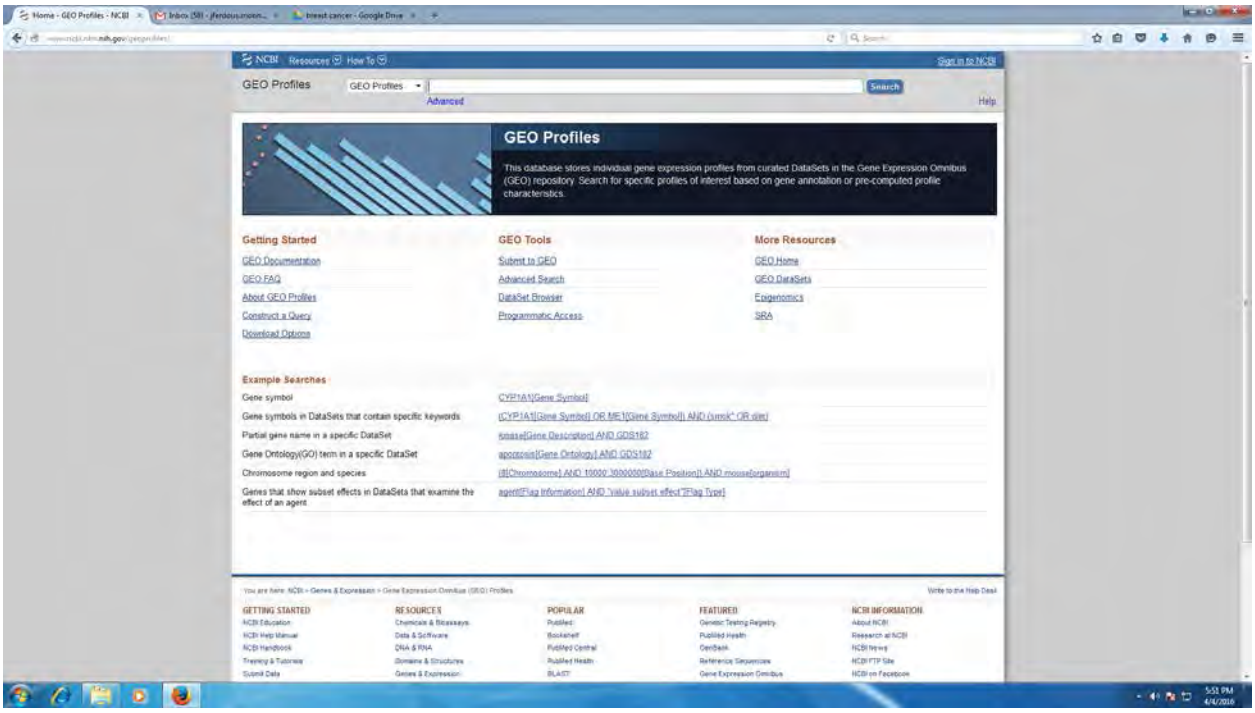


Figure 2.6: GEO Profile homepage.

2.2 Protein

2.2.1 Structure Prediction

Since protein is one complicated molecule while considering its structure, here one type of structure prediction that is homology modelling was focused on. For doing this few steps were followed; first the sequences were taken from a database. Here Uniprot was used for sequences. Then blast was done with these sequences to find their best suited templates. After that alignment was checked with this sequence and their templates. Clustal omega was the software that was used here. And finally this alignment result was given as an input in the homology modelling website, which here was the Swiss model workspace. Description of the databases and softwares are below:

2.2.1.1 UniProtKB

The UniProtKnowledge Base (UniProtKB) is a central hub for collecting any information about proteins with specific, accurate, rich and reviewed annotation. Here in this website along with the main information about a protein molecule (sequence, protein name, description, citation data or taxonomic information) other possible annotation information is also provided. This is a

collaboration between PIR, SIB and EMBL-EBI. The objective of this website was to provide scientific world with high quality, authentic and easily accessible platform of protein sequences and functional information. (Uniprot Consortium, 2015). URL Link: http://www.uniprot.org/

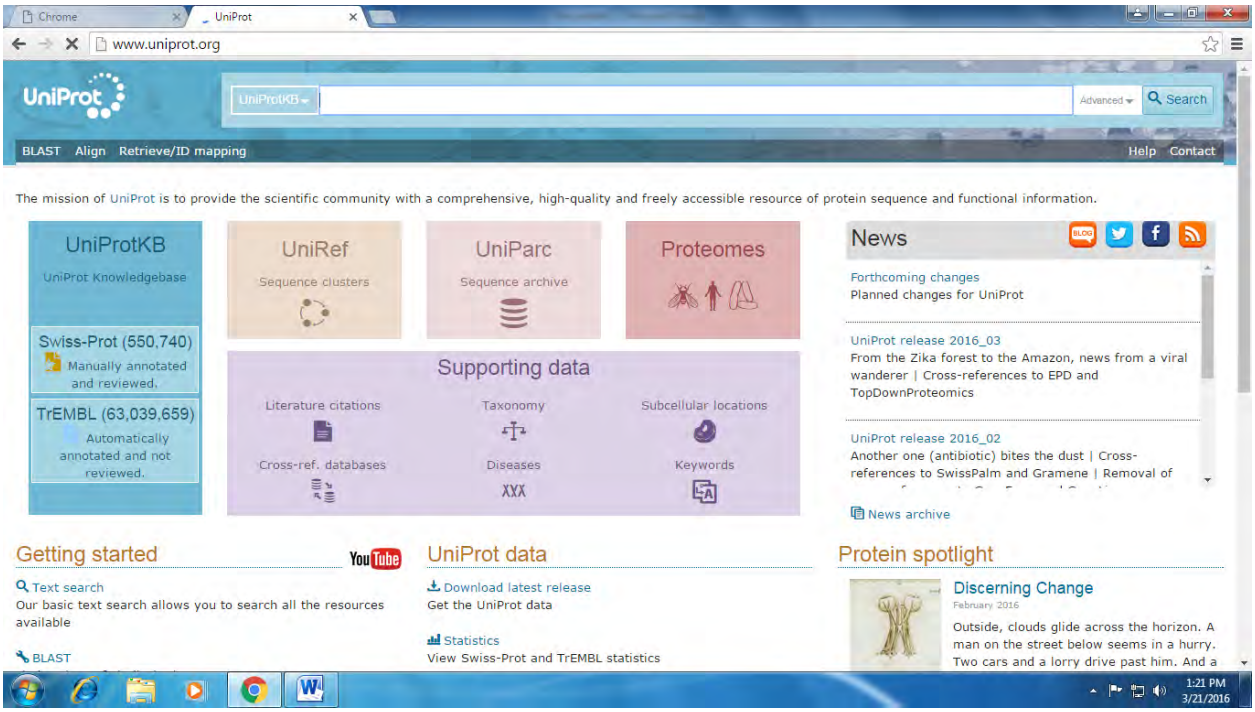


Figure 2.7: UniProtKB homepage.

2.2.1.2 BLAST

Blast stands for Basic Local Alignment Search Tool. Basic function of this tool is to find significant local similarity between sequences showing the result in e value and percentages. The program does its function by comparing the protein or nucleotide sequences with the sequences of the database and calculates the statistical significance of matches. Blast has subsections like BlastP (works with protein sequence), BlastN (works with nucleotide sequence). BLAST is such a tool that is used to study the functional and evolutionary relationships between the given sequences. Also this tool is a great help in identifying members of gene families. (Altschul, 1990). URL Link: http://blast.ncbi.nlm.nih.gov/

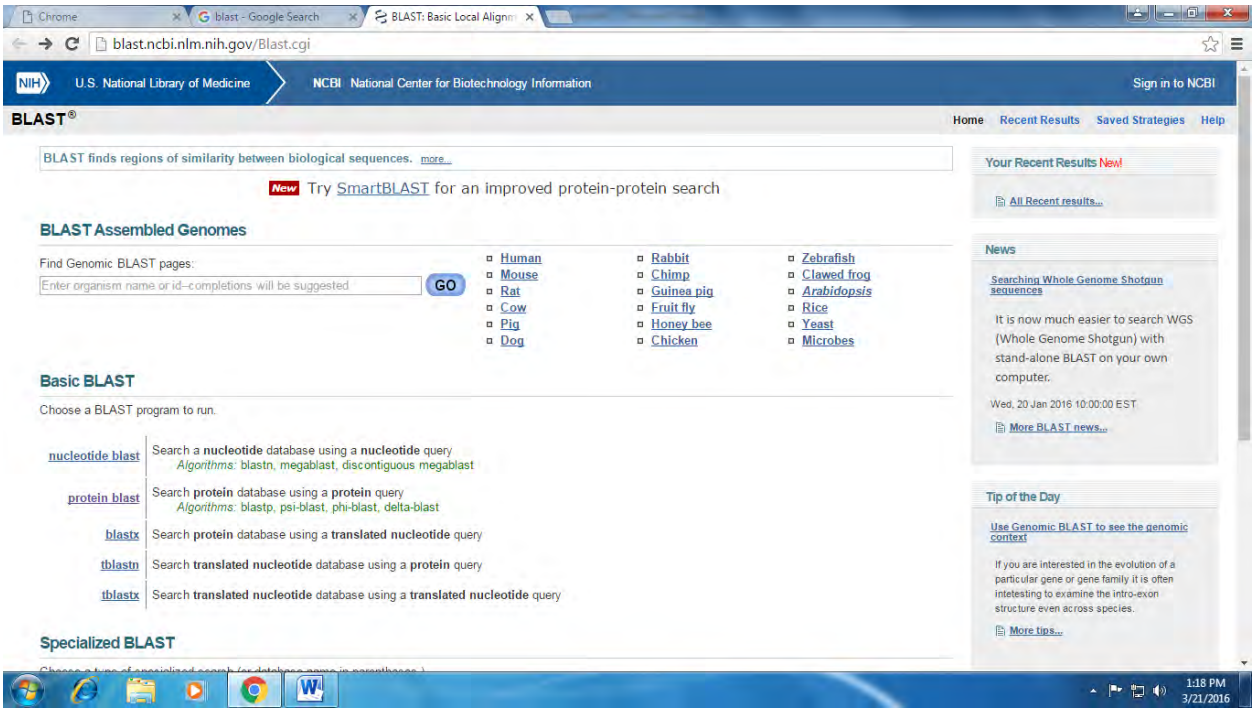


Figure 2.8 : BLAST homepage.

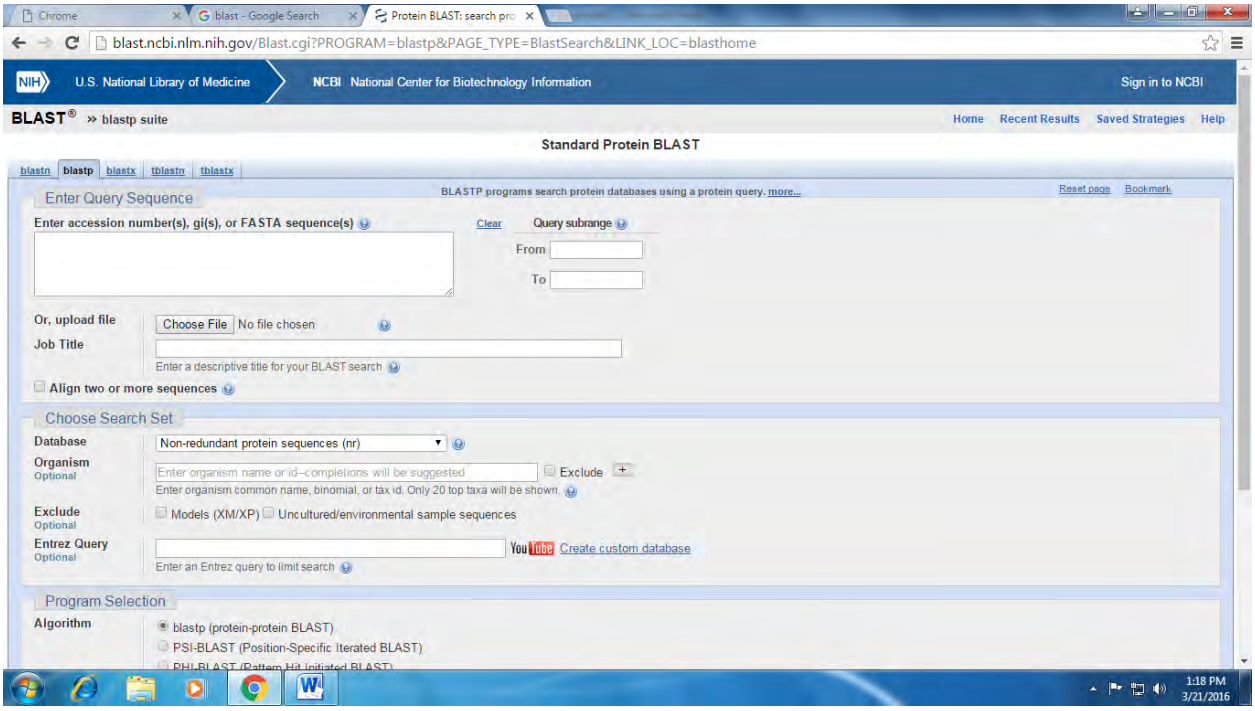


Figure 2.9: BlastP query entry page.

2.2.1.3 Clustal Omega

Clustal Omega is the new version of the old Clustal W series tools which does multiple sequence alignment using seeded guide trees and HMM profile-profile techniques. Usually it can work with three or more sequences at a time. This is a part of EMBL-EBI as a project of processing big data to know biology better and find new information.(Nucleic Acid Research 43, 2015). URL Link: http://www.ebi.ac.uk/Tools/msa/clustalo/

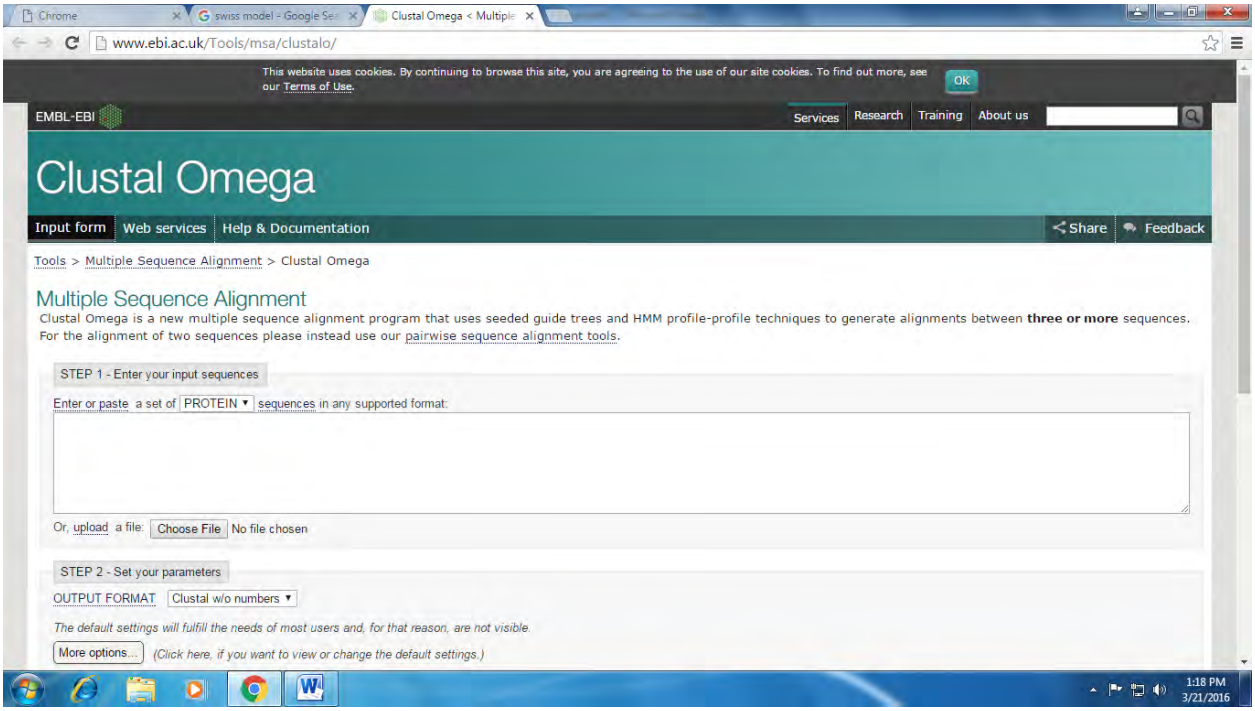


Figure 2.10: Clustal Omega homepage.

2.2.1.4 SWISS MODEL

SWISS MODEL Workspace is a tool that provides with an environment having automated comparative homology modelling platform. SWISS-MODEL was first stated working in 1993 by Manuel Peitsch Nicolas Guex and TorstenSchwede, and further developed at GWER - Glaxo Well come Experimental Research in Geneva and the SIB - Swiss Institute of Bioinformatics. The SWISS, (which is a relational database of annotated three-dimensional comparative protein structure models), was established in 2004. In 2005, SWISS-MODEL service was extended by SWISS-MODEL Workspace, a web-based work bench for protein homology modelling and assessment of the result. This workspace provides three work mode to work with it – automated mode, alignment mode and project mode. Biozentrum (University Basel) and the Advanced Biomedical Computing Center (NCI Frederick, USA) provides with the computational services for this software. (Arnold, 2015). URL Link: http://swissmodel.expasy.org/workspace/

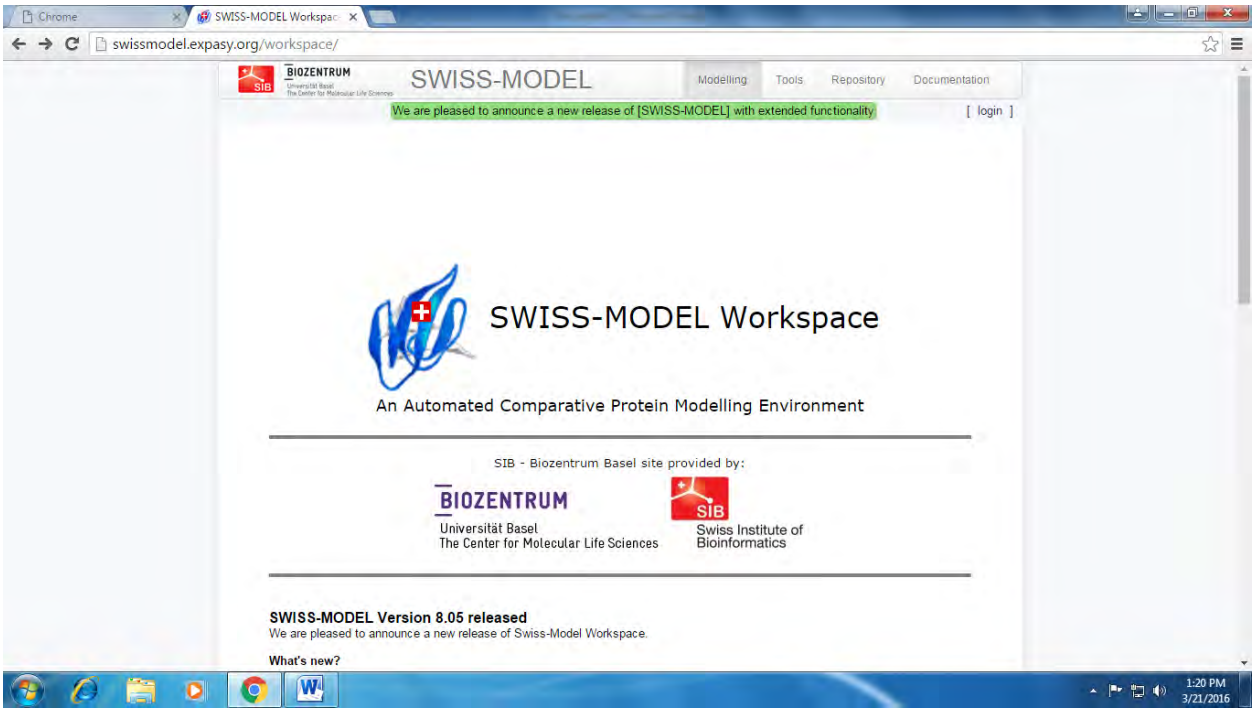


Figure 2.11: SWISS-MODEL Workspace home page.

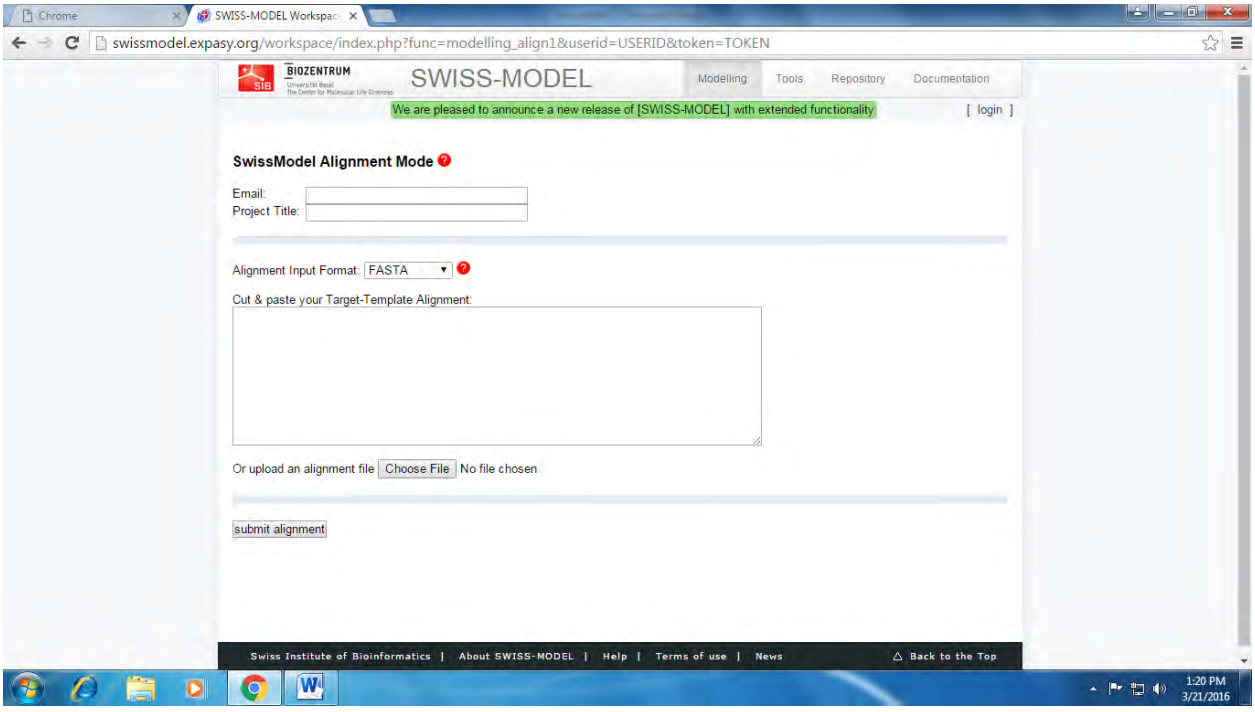


Figure 2.12: SWISS- MODEL Workspace alignment submitting page.

2.2.2 Discovering Sequence Motif

For finding motifs in the protein sequences two steps were followed. First, protein sequences were collected from database in their fasta format and then put into a web tool as input to get the result. Here NCBI is used as the database and MEME was used as the web tool to find sequence motif. Details about these two are given below:

2.2.2.1 NCBI

NCBI stands for National Center for Biotechnology Information. This is a great source for biomedical and genomic information. The late Senator Claude Pepper established the National Center for Biotechnology Information (NCBI) on November 4, 1988, as a division of the National Library of Medicine (NLM) at the National Institutes of Health (NIH). NLM was chosen for its experience in creating and maintaining biomedical databases and NIH because of its largest biomedical research facility in the world. As a national resource for molecular biology information, NCBI's mission is to develop new information technologies to help understanding the fundamental molecular and genetic processes that are responsible for good health and diseases. More specifically, the NCBI has been charged with creating automated systems for storing and analyzing knowledge about molecular biology, biochemistry, and genetics; facilitating the use of such databases and software by the research and medical community; coordinating efforts to gather biotechnology information both nationally and internationally; and performing research into advanced methods of computer-based information processing for analyzing the structure and function of biologically important molecules.. URL Link: http://www.ncbi.nlm.nih.gov/

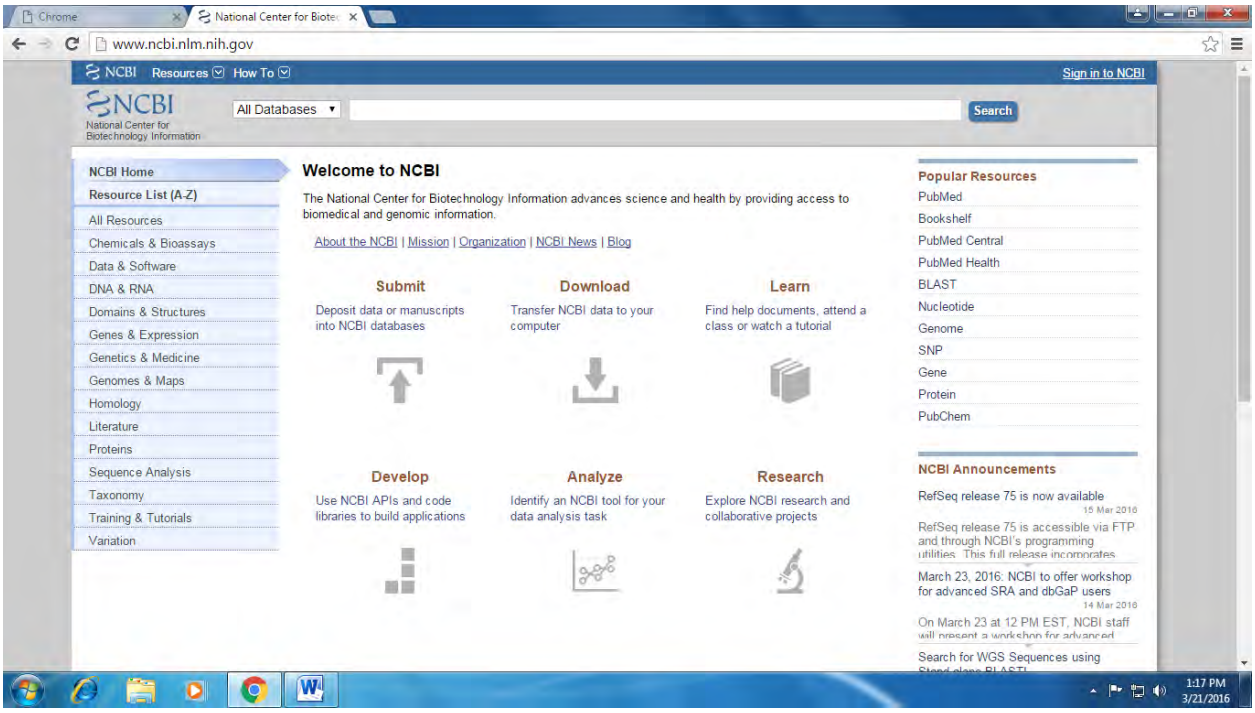


Figure 2.13: NCBI home page.

2.2.2.2 MEME

MEME stands for Multiple Expectation maximization for Motif Elicitation. MEME discovers novel, ungapped motifs (recurring, fixed-length patterns) in nucleic acid or protein sequences (sample output from sequences).

MEME splits variable-length patterns into two or more separate motifs. MEME represents motifs as position-dependent letter-probability matrices which describes the probability of each possible letter at each position in the pattern. Individual MEME motifs do not contain gaps. Patterns with variable-length gaps are split by MEME into two or more separate motifs.

MEME takes as input a sequence or a group of sequences and outputs as many motifs as requested. This tool can choose the best width, number of occurrences, and description for each motif by the help of statistical modelling. MEME on the web can take a second (control) set of input sequences and then discovers motifs that are enriched in the primary set relative to the control set. This discovery is called discriminative motif discovery.The MEME Suite was developed byTimothy Bailey at the Institute for Molecular Bioscience at the University of Queensland and William Stafford Noble in the Department of Genome Sciences at the University of Washington. . This Suite have previously been supported byColumbia University, the Computational Biology Research Center at the National Institute of Advanced Industrial Science and Technology, the National Biomedical Computation Resource, and the San Diego Supercomputer Center. Maintenance and development of the MEME Suite is funded by

the National Institutes of Health.it also receives support from Amazon and Google. (Timothy, 2009). URL Link: http://meme-suite.org/tools/meme

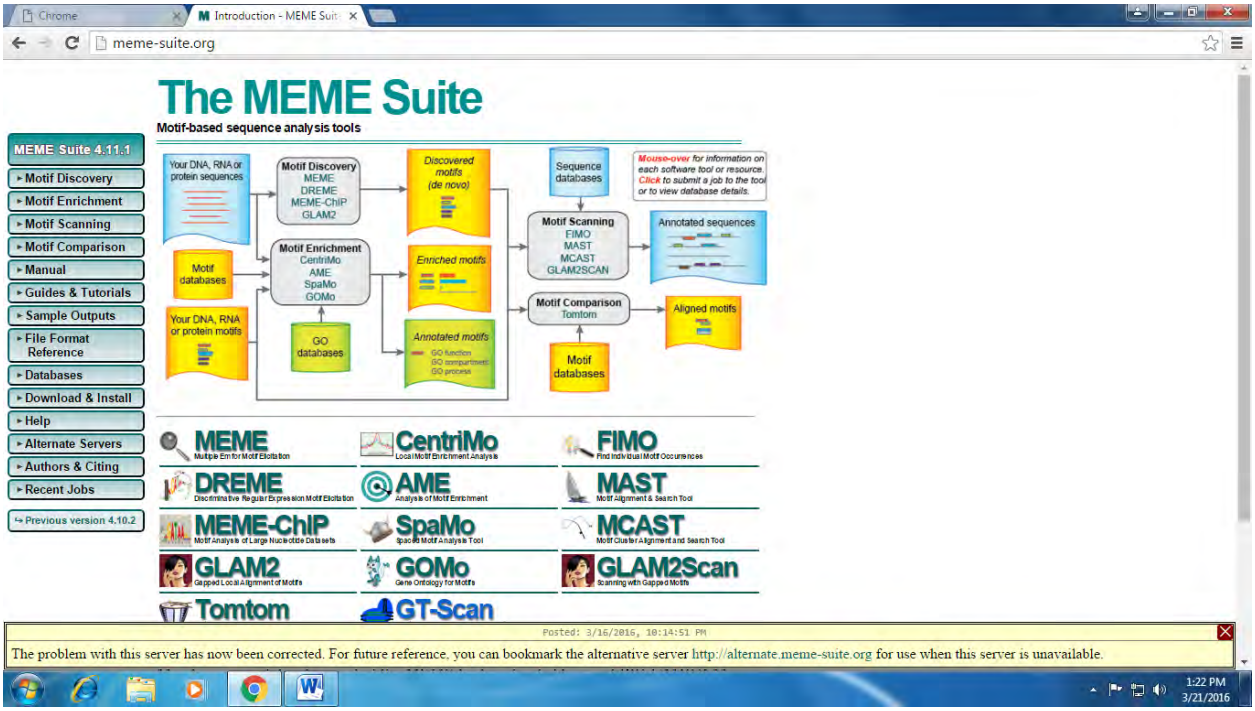


Figure 2.14: MEME Suite home page.

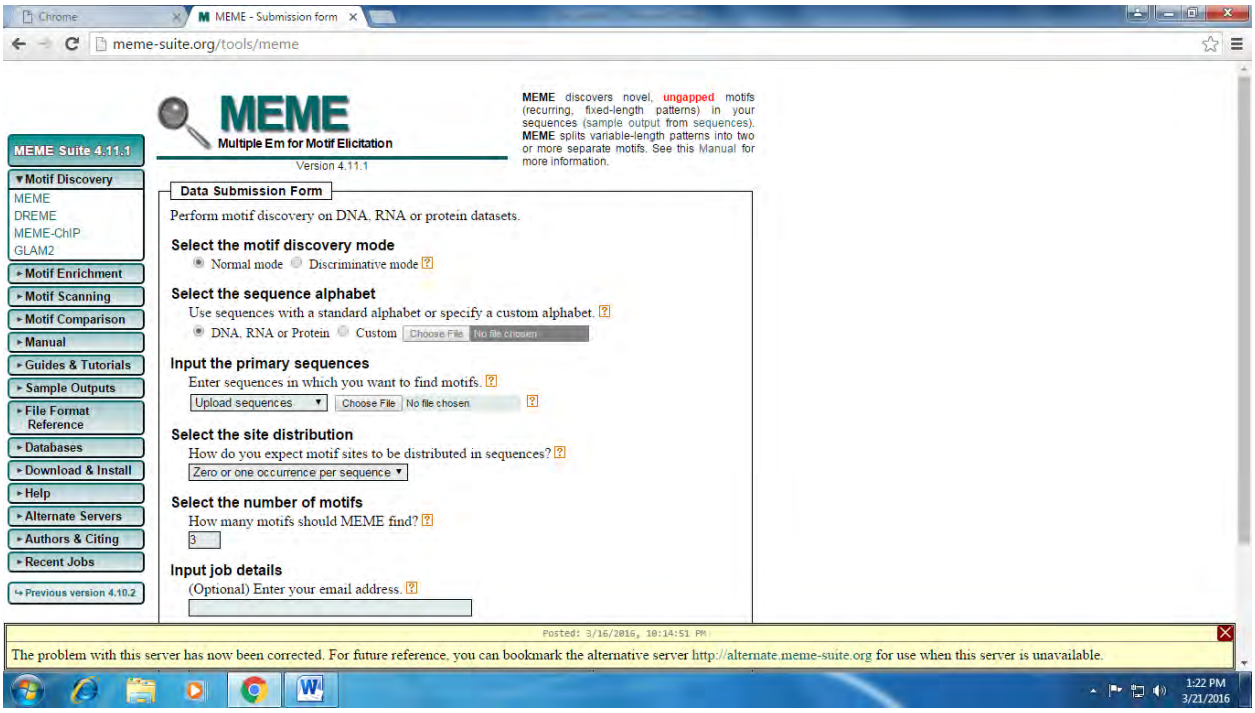


Figure 2.15: MEME data submission form.

2.2.3 Finding Expression Level

To find out the expression level of these proteins in different conditions of breast cancer GEO tool of NCBI was used. Just the name of the protein name and little clue “breast cancer” was given as input and possible options appear.

2.2.3.1 GEO Profile

GEO represents to Gene Expression Omnibus. The GEO Profiles database stores gene expression profiles derived from curated GEO Datasets. Each Profile is presented as a chart that displays the expression level of one gene across all Samples within the Dataset. Experimental parameter is provided in the bars along the bottom of the charts to see whether a gene is differentially expressed across different experimental conditions. every Profile have various types of links including internal links that connect genes that exhibit similar behavior, and external links to relevant records in other NCBI databases.GEO Profiles can be searched using many different attributes including keywords, gene symbols, gene names, GenBank accession numbers, or Profiles flagged as being differentially expressed.(Barrett, 2013). URL Link: https://www.ncbi.nlm.nih.gov/geoprofiles/

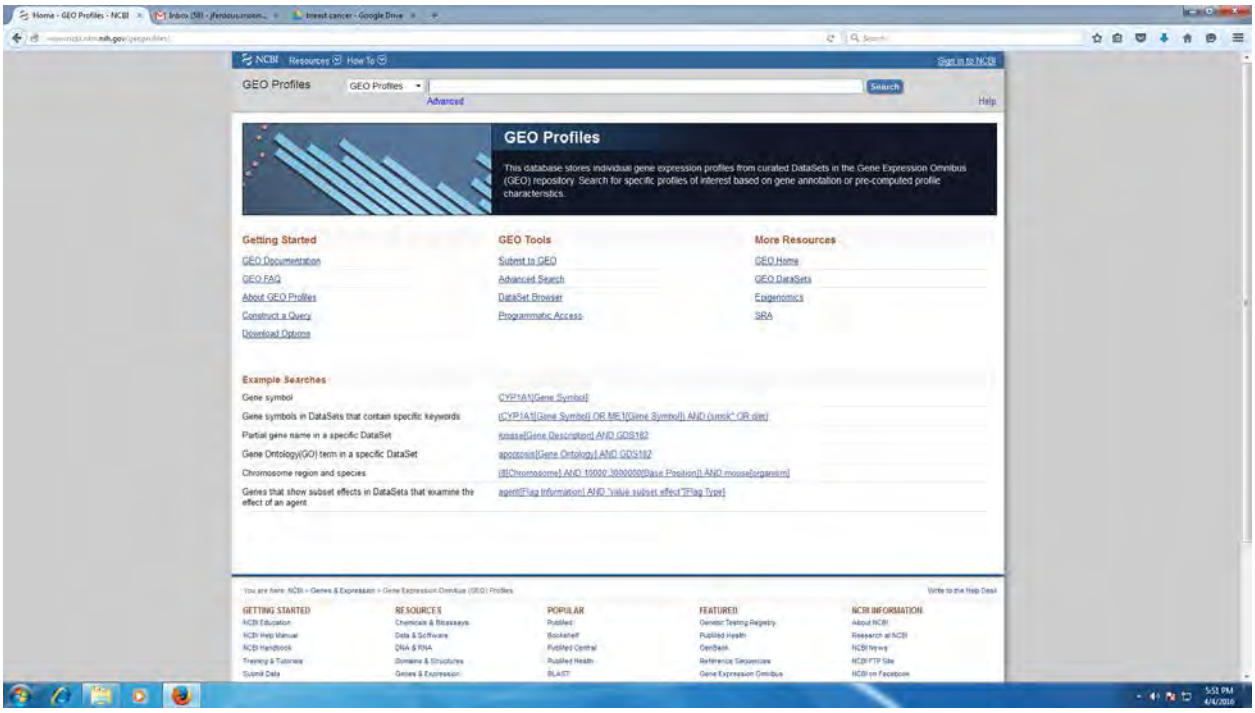


Figure2.16: GEO Profile home page.

CHAPTER: 3

RESULT

Since this thesis was focused on two different types of biomarker molecules, the result section is also divided on two sections, named by these two different biomarker molecules.

3.1 miRNA

As mentioned earlier, seven promising miRNA molecule were selected to be studied upon in this thesis. For each selected miRNA, three different properties were looked into. One is their structure (secondary, hair pin like), second one is motifs in their sequences and the third one is their differentiating expression profiles. The selected miRNAs are -

- miR10b
- miR21
- miR145
- miR155
- miR191
- miR382
- miR425

Results for those observations are given below:

3.1.1 Structure

Basically in this study, finding the secondary hairpin like structure of the miRNAs was the main focus. By hairpin structure, it means, a secondary structure and it looks like a U shaped pin. These type of structure is formed when a miRNA strand folds and turns and binds with another miRNA strand. This is also known as stem loop structure. Now knowing this kind of structure specifically is important because the folding in a certain way leads to the solution of different unanswered question in binding and functioning of that miRNA molecule.

Figure3.1.: Hairpin structure of miR10B

Figure3.2: Hairpin structure of miR21

Figure3.3: Hairpin stricture of miR145

Figure3.4: Hairrpin structure of miR155

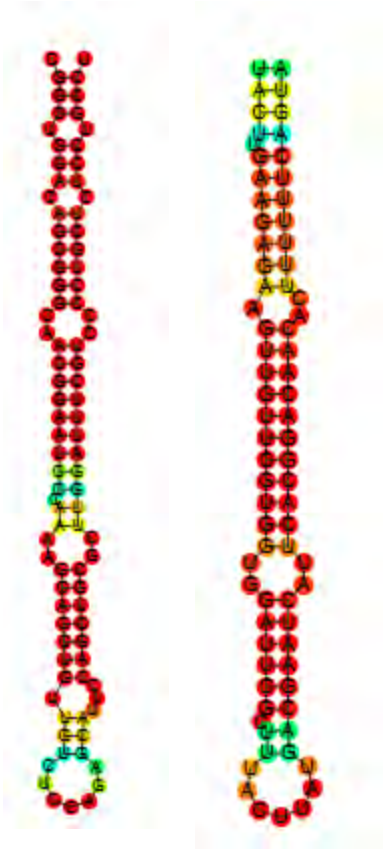


Figure3.5: Hairpin structure of miR191

Figure3.6: Hairpin structure of miR382

Figure3.7: Hairpin structure of miR425

3.1.2 Motif

For seven selected miRNAs, motif was observed. Motif means a sequence that can have special importance biologically or functionally. Now for each miRNA five motifs have been observed. Knowing motifs are important because they are recurrent and they indicate binding sites or functional sequence of that molecule.

3.1.2.1miRNA 10B:

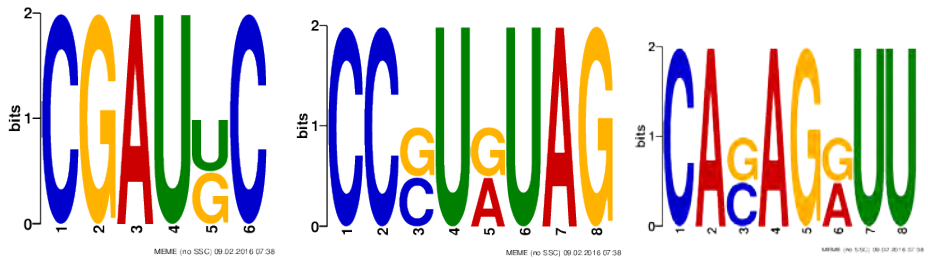


Figure3.8: Motif 1

Figure3.9: Motif 2

Figure 3.10: Motif 3

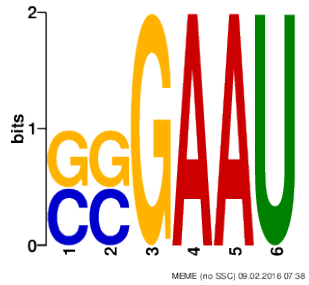


Figure 3.11: Motif 4

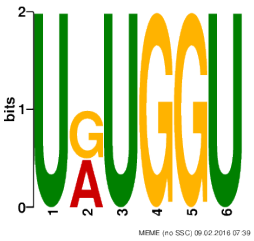


Figure 3.12: Motif 5

3.1.2.2.miRNA 21:

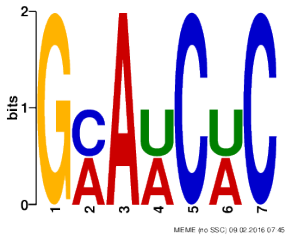


Figure3.13 : Motif 1

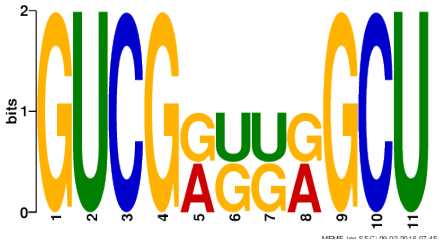


Figure3.14: Motif 2

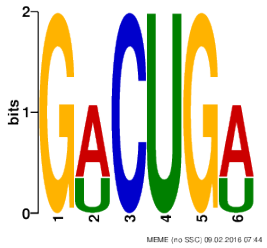


Figure3.15: Motif 3

3.1.2.3miRNA 145:

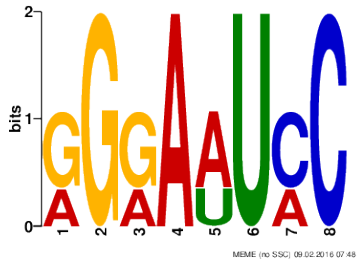


Figure 3.16: Motif 1

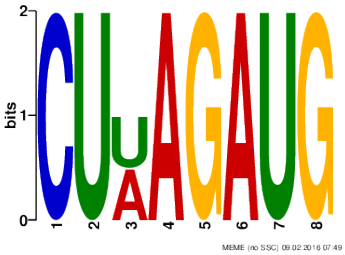


Figure3.17: Motif 2

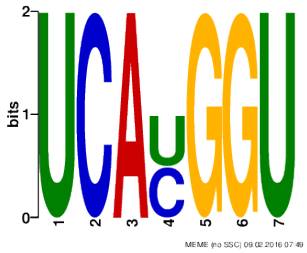


Figure 3.18: Motif 3

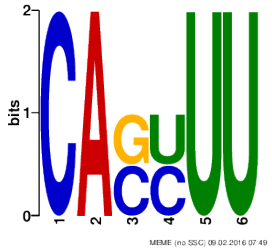


Figure 3.19: Motif 4

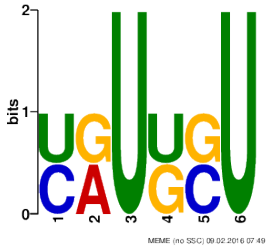


Figure 3.20: Motif 5

3.1.2.4 miRNA155:

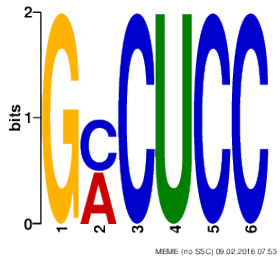


Figure3.21: Motif 1

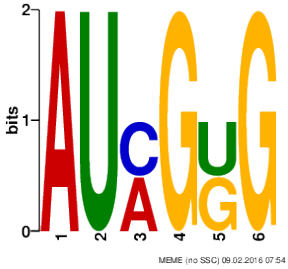


Figure3.22: Motif 2

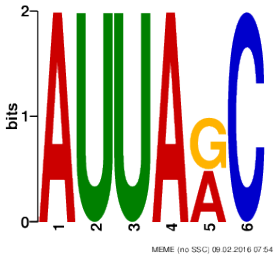
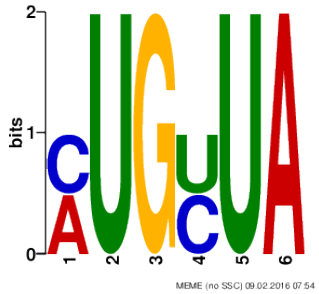


Figure3.23. Motif 3



Figur3.24: Motif 4

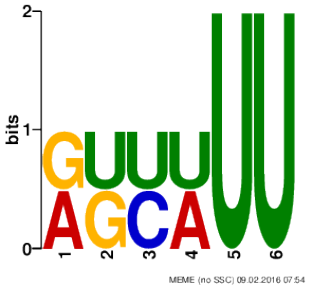


Figure3.25: Motif 5

3.1.2.5 miRNA191:

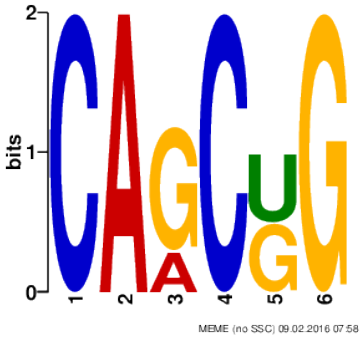


Figure 3.26: Motif 1

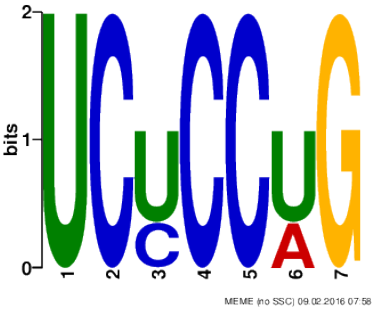


Figure 3.27: Motif 2

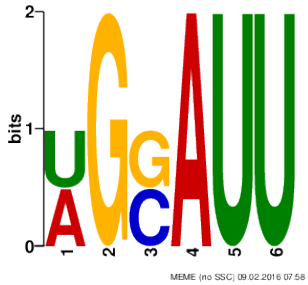


Figure 3.28: Motif 3

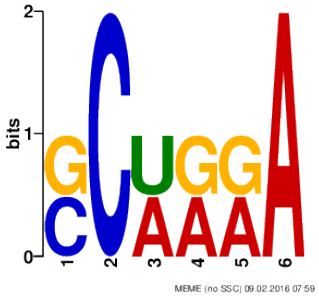


Figure 3.29: Motif 4

3.1.2.6miRNA382:

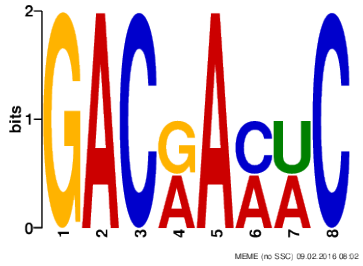


Figure 3.30: Motif 1

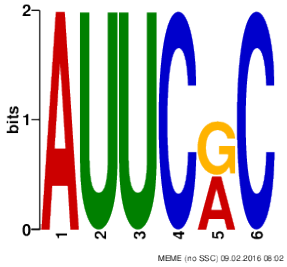


Figure 3.31: Motif 2

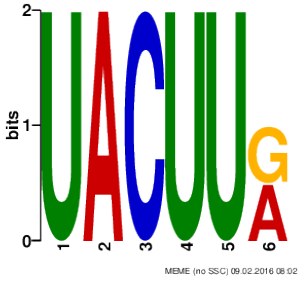


Figure3.32: Motif 3

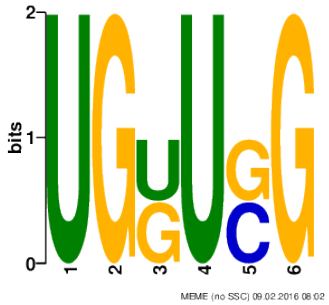


Figure 3.33: Motif 4

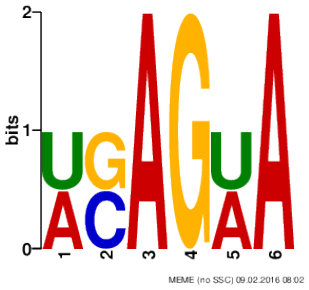


Figure 3.34: Motif 5

3.1.2.7 miRNA425:

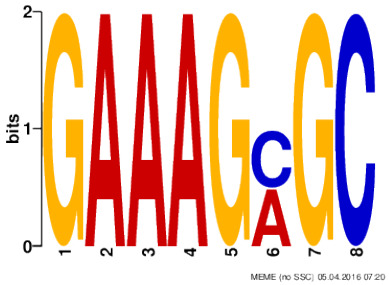


Figure3.35: Motif 1

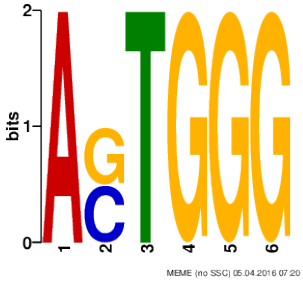


Figure3.36: Motif 2

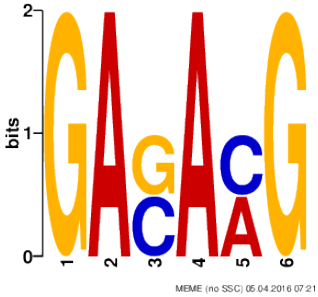


Figure3.37: Motif 3

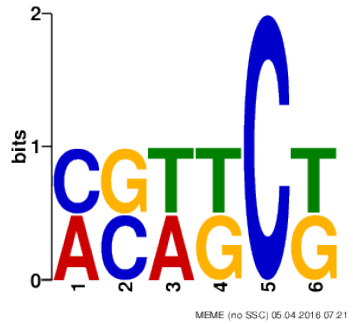


Figure3.38: Motif 4

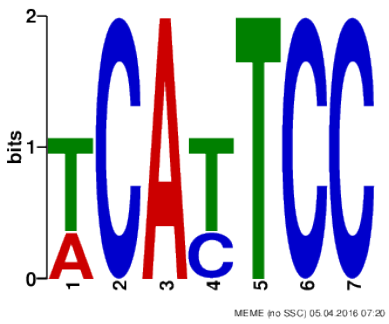


Figure3.39: Motif 5

3.1.3 Expression Level:

The third trait to observe is expression level. This is a tricky one to observe because this involves a multiple step to be expressed and also because this is an actual indicating property that makes the miRNA molecules a biomarker. For this no practical assays were performed but rather GEO tool was used and results of other assays that had been performed on this molecule and stored here, had been taken.

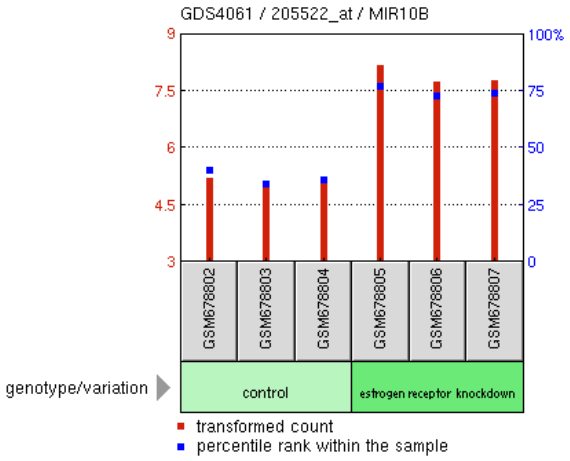


Figure3.40: Expression level of miRNA 10B

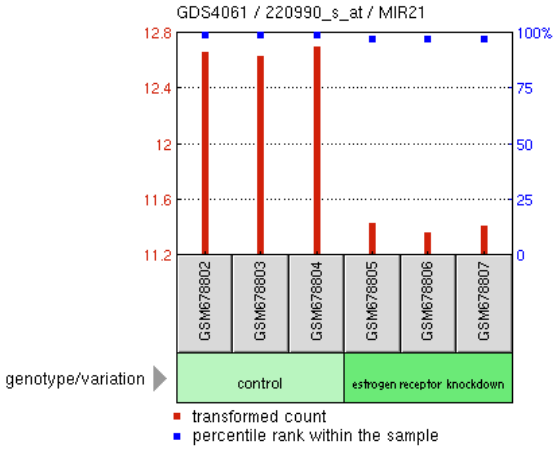


Figure3.41: Expression level of miRNA 21

Here the expression level of miR10B and miR21 is shown. The expression level was measured in normal breast cancer patient and in patient whom were treated with ER mutation. It is seen that miR10B expression level is higher in the ER mutated cells and miR21 is low expressed in the ER mutated cells.

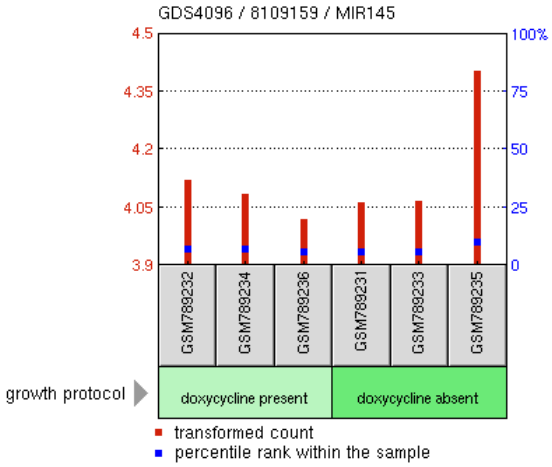


Figure3.42: Expression level of miR145

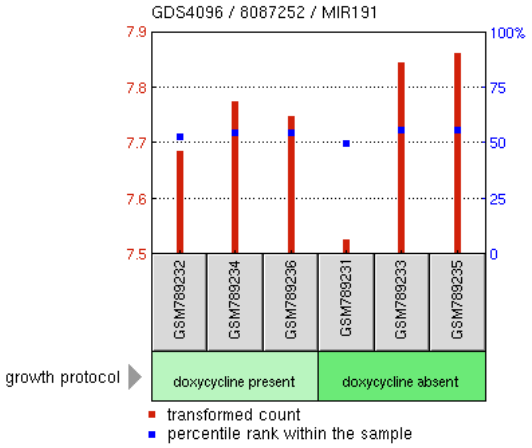


Figure3.43: Expression level of miR191

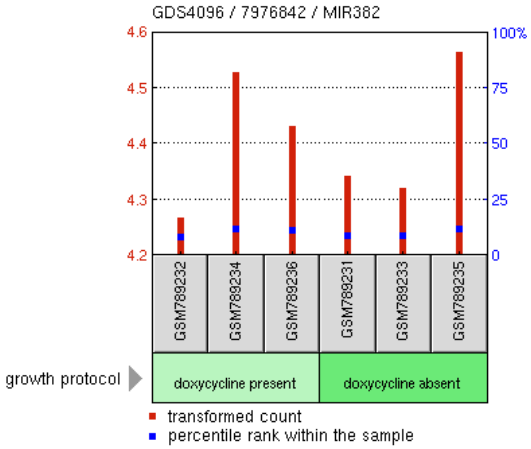


Figure3.44: Expression level of miR382

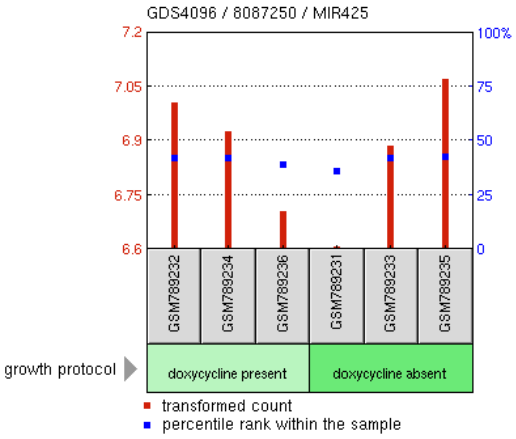


Figure3.45: Expression level of miR425

In these images the expression profile of miR145, miR191, miR382 and miR425 are observed. All of these are observed in breast cancer patients who are treated with doxycycline and patients who are not. Now among four of the miRNAs none of them shows a specific response to this treatment.

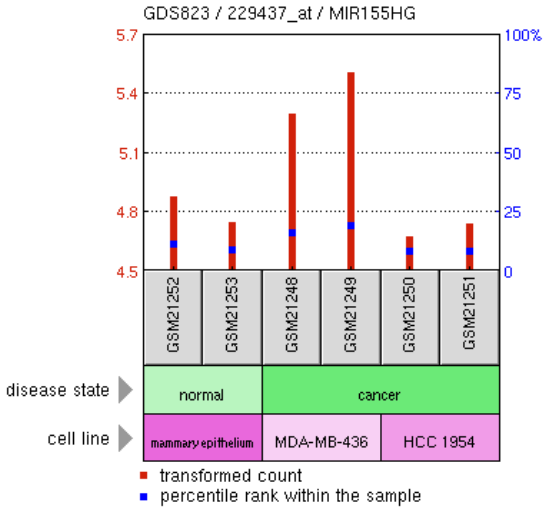


Figure3.46: Expression level of miR155 in breast cancer cell lines

This expression profile shows miR155 expression in two different breast cancer cell lines along with a control. It can be seen that in both cancerous cell lines (MDA-MB-436 and HCC 1954) miR155 is expressed differently than the normal cell line. It is expressed more in the MDA-MB-436 cell line and less in the HCC 1954 cell line.

3.2 Protein

Protein was another biomolecule of this study that was found promising as potential biomarkers for breast cancers. So around 11 proteins 3D structures, sequence motifs and expression level have been checked. The names of the selected proteins are-

- CEA
- P53
- Ki67
- Cyclin D1
- Cyclin E
- TTR
- HSP90
- Her2
- ER
- PR
- ER- beta

Results for the observations are given below:

3.2.1 Structure

For the 11 protein, their 3d structure was predicted. Hence proteins 3D structure prediction involves a lot of steps and different classification, here only homology modelling was focused on. Homology modelling is the tertiary level of structure of a protein. It means comparative modeling where other homologous proteins are used as templates. For homology model prediction SWISS MODEL Workspace was used. Predicting the homology model of these molecules can help in determining the structural motifs as well as site directed mutagenesis that might make them a candidate for biomarker panel of breast cancer.

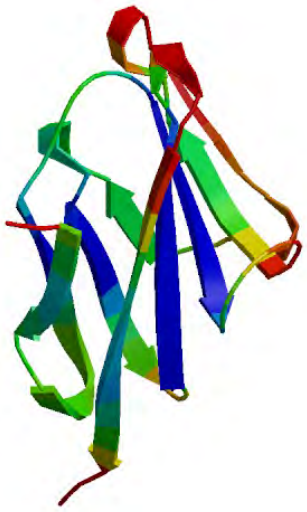


Figure3.47: Homology model of CEA

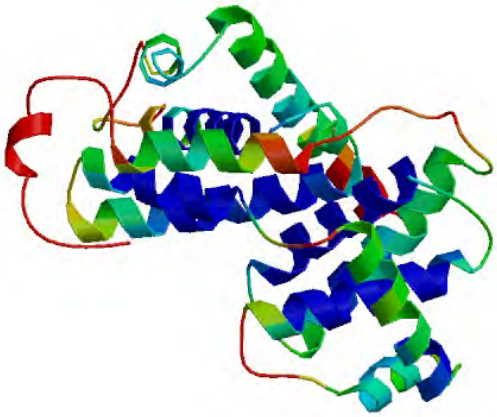


Figure3.48: Homology model of Cyclin D1

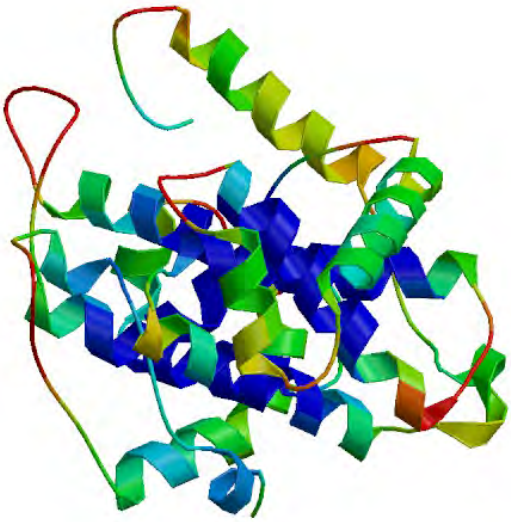


Figure3.49: Homology model of Cyclin E

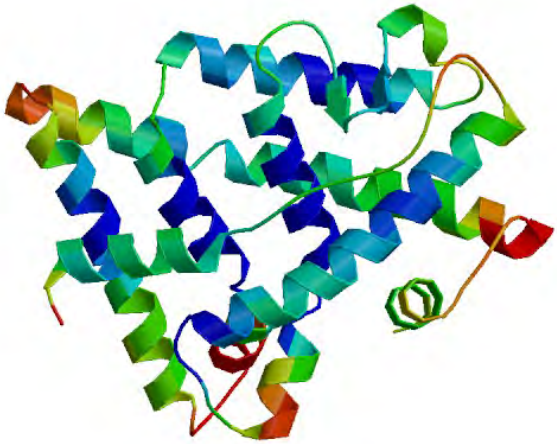


Figure3.50: Homology model of ER

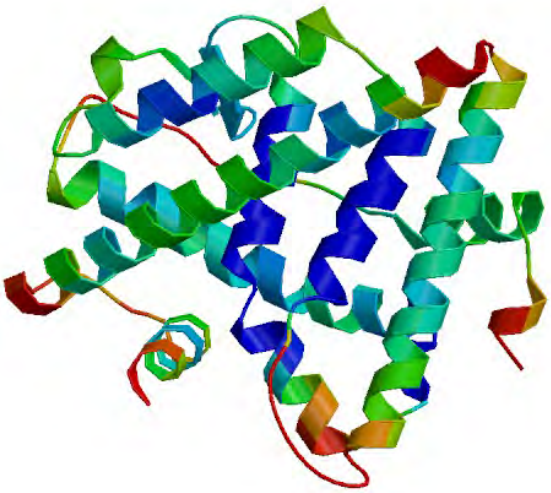


Figure3.51: Homology model of ER Beta

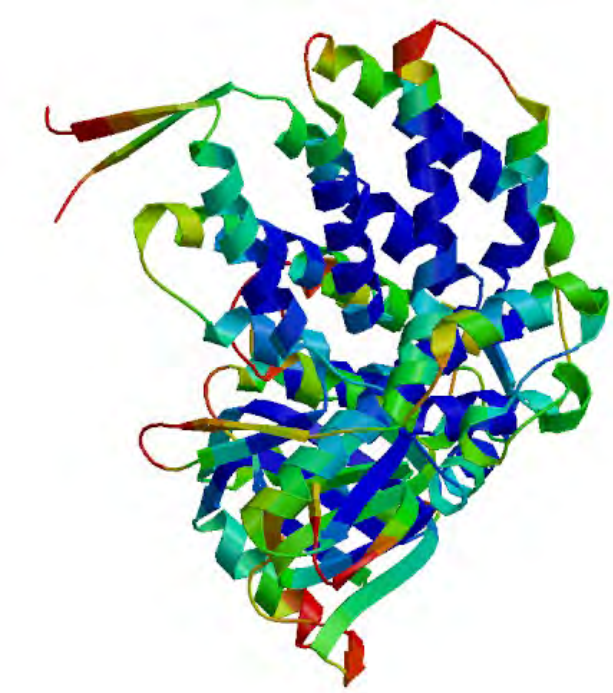


Figure3.52: Homology model of HSP60

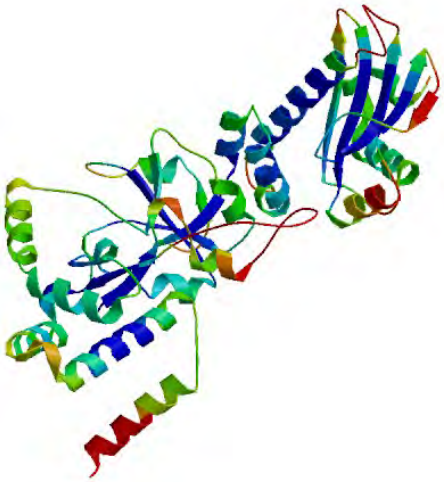


Figure3.53: Homology model of Her2

Figure3.54: Homology model of Ki67

Figure 3.55: Homology model of P53

Figure3.56: Homology model of PR

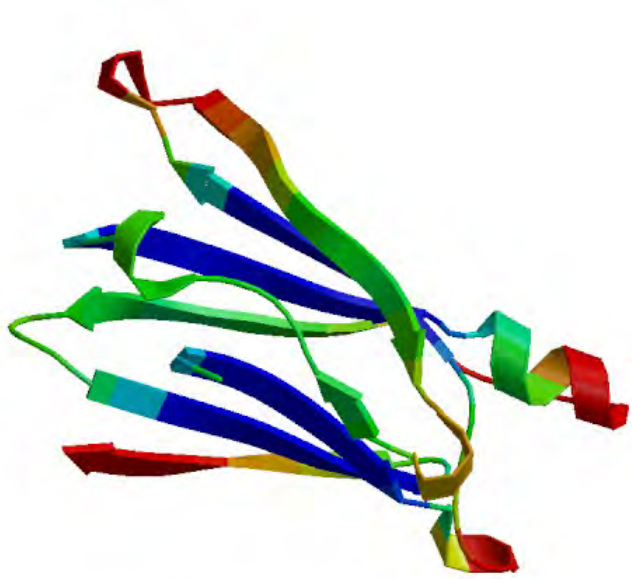


Figure3.57: Homology model of TTR

3.2.2 MOTIF

Like it was said in the miRNA section that motifs are sequences having biological and functional value and they are important for finding answers to their specific activity, motifs for protein were also observed. For finding the motif of proteins, MEME software was used and for each protein at least five motif was observed. Knowing protein motifs are important because they give a clear information about the effects of sequence variation, protein protein interaction etc.

3.2.2.1CEA:

Figure3.58: Motif 1 in CEA protein

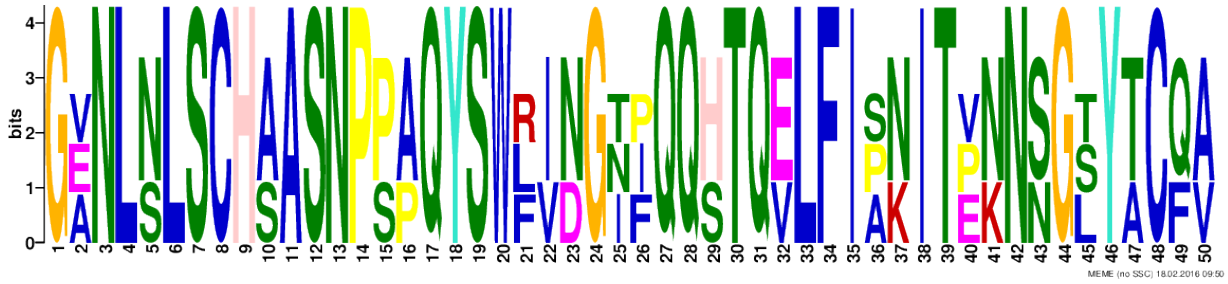


Figure3.59: Motif 2 in CEA protein.

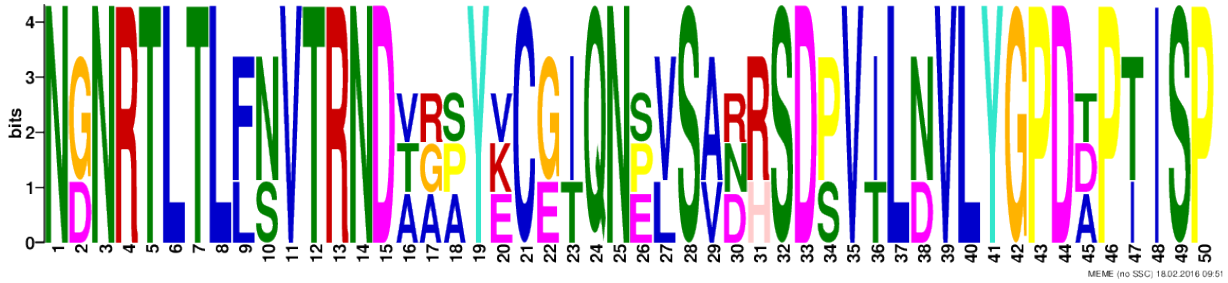


Figure3.60 : Motif 3 in CEA protein.

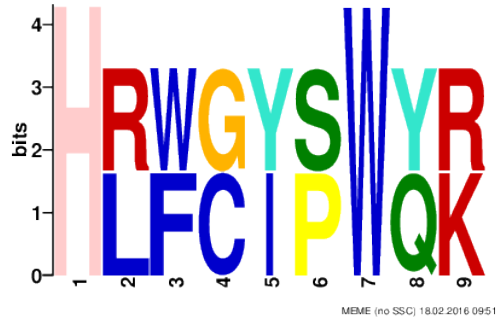


Figure3.61: Motif 4 in CEA protein

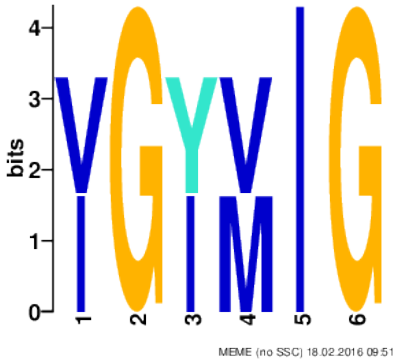


Figure 3.62: Motif 5 in CEA protein

3.2.2.2 Cyclin D1:

Figure3.63: Motif 1 in Cyclin D1 protein

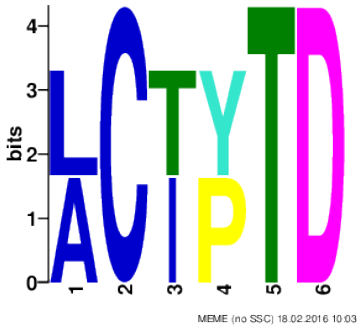


Figure3.64: Motif 2 in Cyclin D1

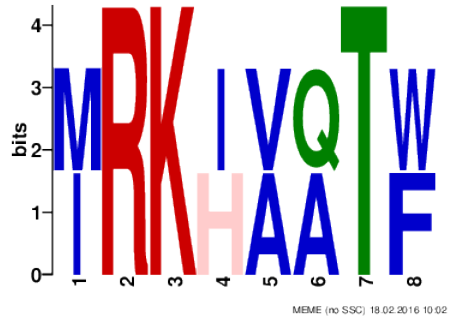


Figure3.65: Motif 3 in Cyclin D1 protein

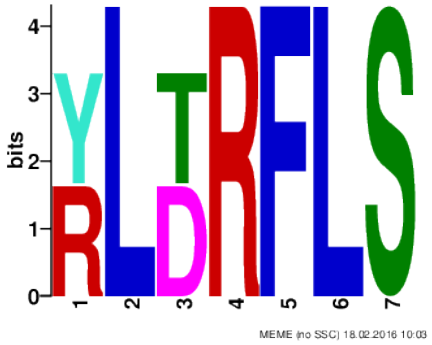


Figure3.66: Motif 4 in Cyclin D1

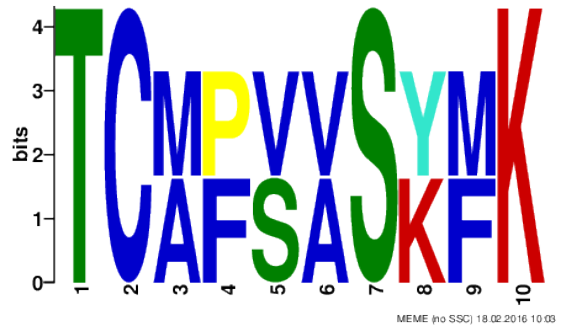


Figure3.67: Motif 5 in Cyclin D1 protein

3.2.2.3 Cyclin E:

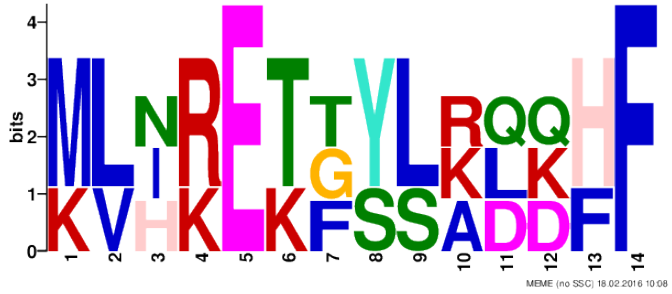


Figure3.68: Motif 1 in Cyclin E protein

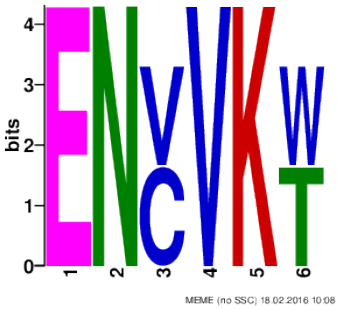


Figure3.69 Motif 2 in Cyclin E

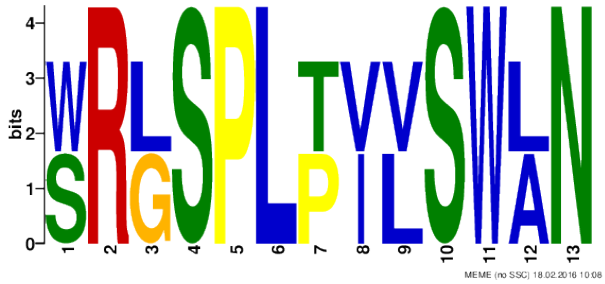


Figure3. 70: Motif 3 in Cyclin E protein

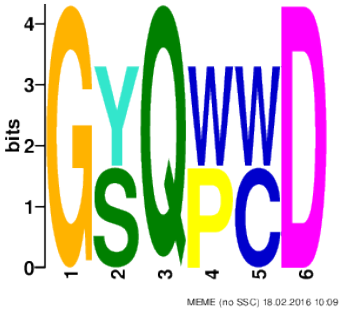


Figure3.71: Motif 4 in Cyclin E protein

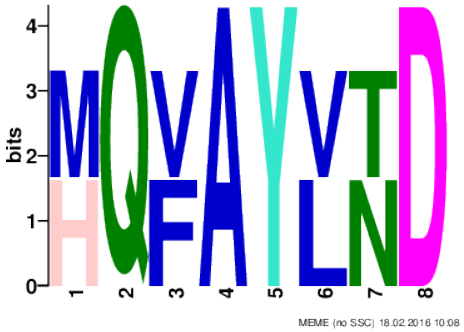


Figure 3.72: Motif 5 in Cyclin E protein

3.2.2.4ER:

Figure3.73: Motif 1 in ER protein

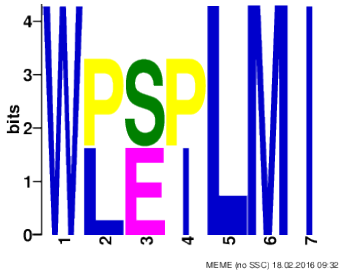


Figure3.74: Motif 2 in ER protein

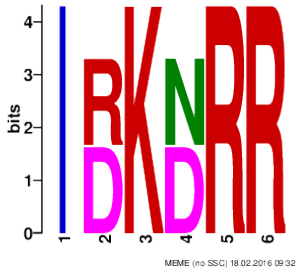


Figure3.75: Motif 3 in ER

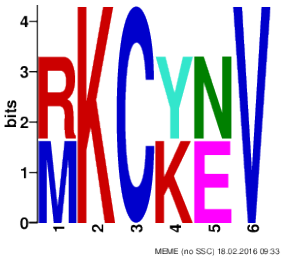


Figure3.76: Motif 4 in ER

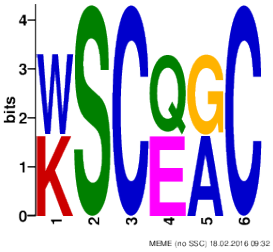


Figure3.77 Motif 5 in ER

3.2.2.5ER Beta:

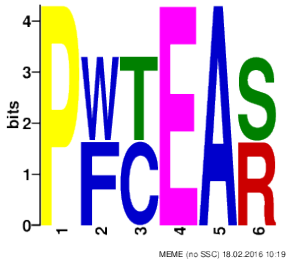


Figure3.78: Motif 1

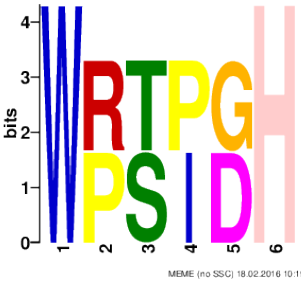


Figure 3.79: Motif 2

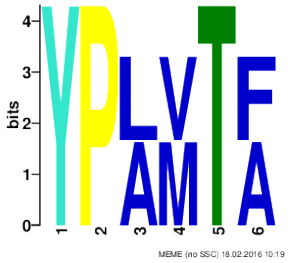


Figure3.80: Motif 3

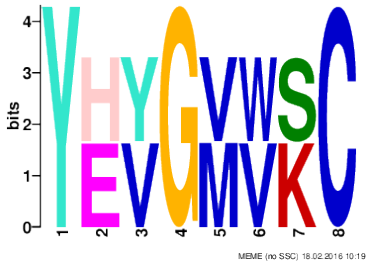


Figure3.81: Motif 4

Figure3.82: Motif 5

3.2.2.6 Her2:

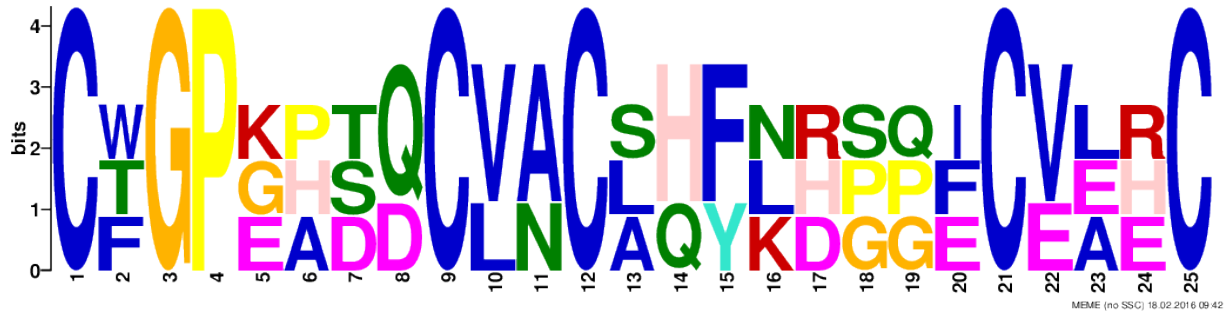


Figure3.83: Motif 1 in Her2 protein

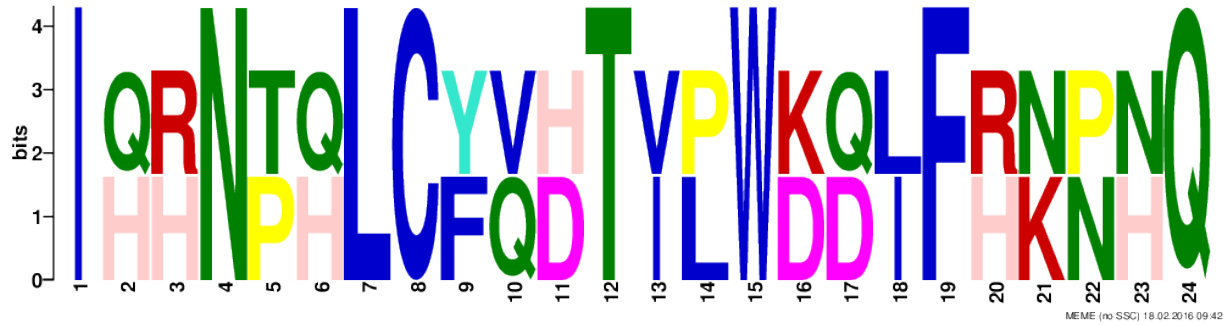


Figure3.84: Motif 2 in Her2 protein

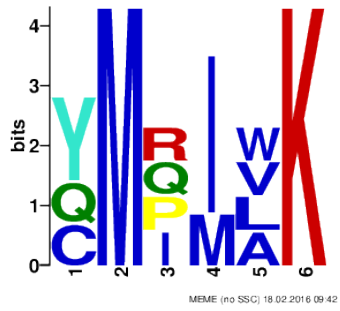


Figure3.85: Motif 3

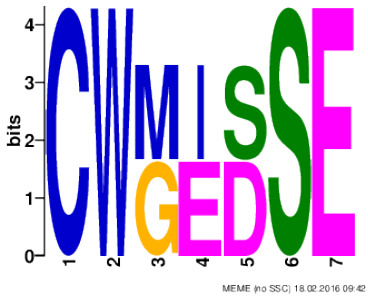


Figure3.86: Motif 4

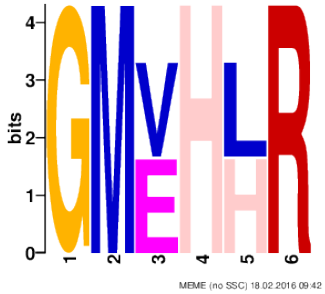


Figure 3.87: Motif 5

3.2.2.7HSP60:

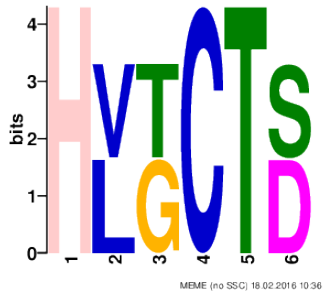


Figure3.88: Motif 1

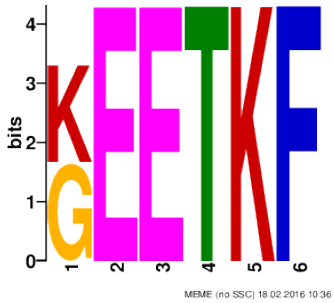


Figure 3.89: Motif 2

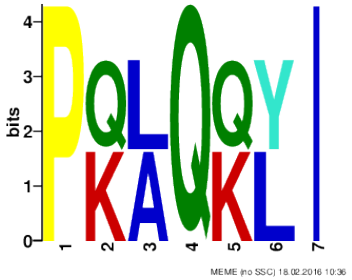


Figure 3.90: Motif 3

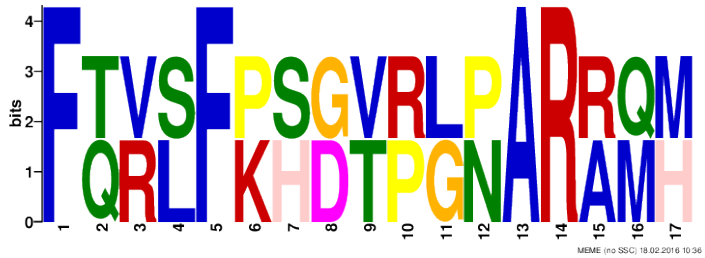


Figure3.91: Motif 4 in HSP60

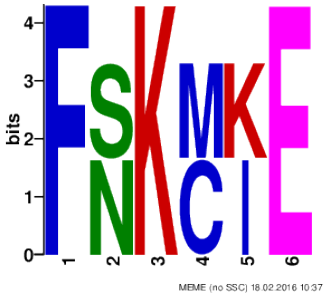


Figure3.92: Motif 5

3.2.2.8 Ki67:

Figure3.93: Motif 1 in Ki67 protein

Figure3.94: Motif 2 in Ki67 protein

Figure3.95: Motif 3 in Ki67 protein

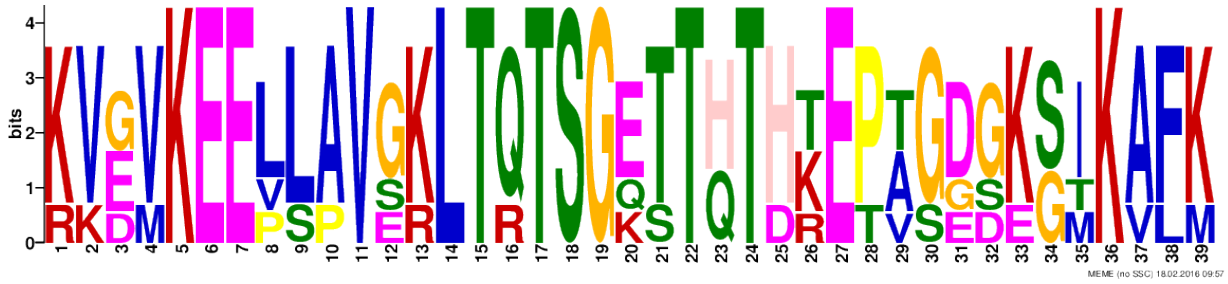


Figure3.96: Motif 4 in Ki67 protein

Figure3.97: Motif 5 in Ki67 protein

3.2.2.9 P53:

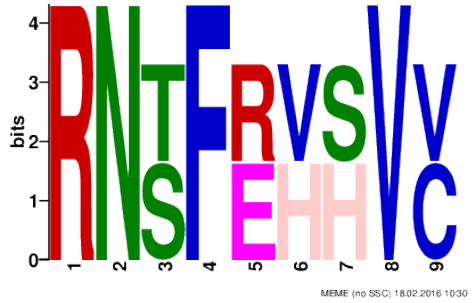


Figure3.98: Motif 1 in P53

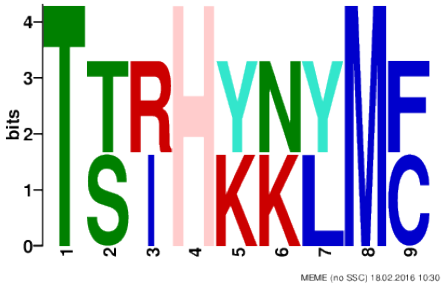


Figure 3.99: Motif 2 in P53

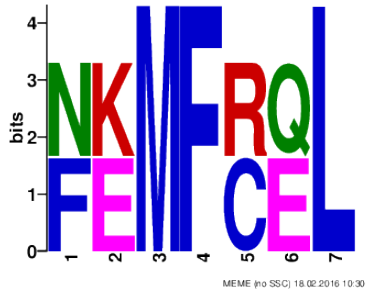


Figure3.100: Motif 3

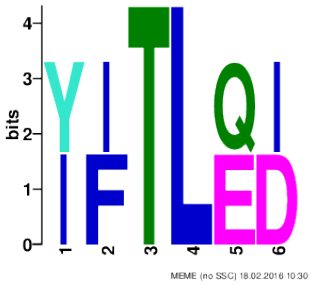


Figure3.101: Motif 4

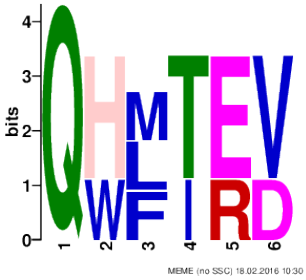


Figure3.102: Motif 5

3.2.2.10 PR:

Figure3.103: Motif 1 in PR Protein

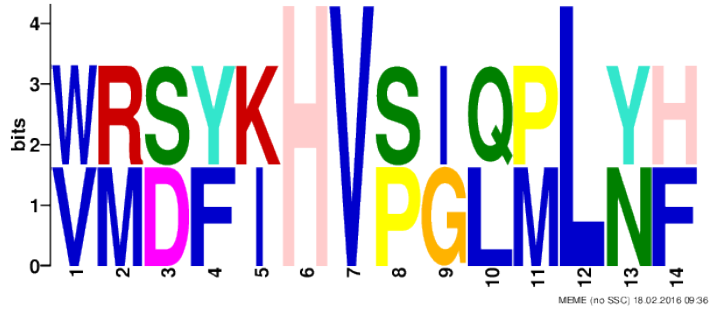


Figure3.104: Motif 2 in PR protein

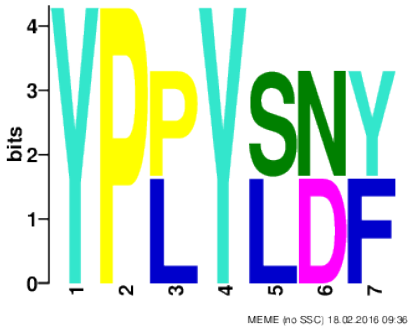


Figure3.105: Motif 3 in PR

Figure3.106: Motif 4 in PR protein

Figure3.107: Motif 5 in PR protein

3.2.2.11.TTR:

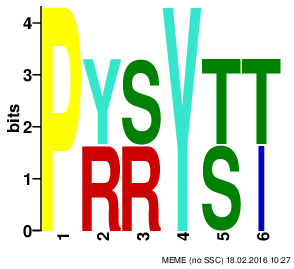


Figure3.108: Motif 1

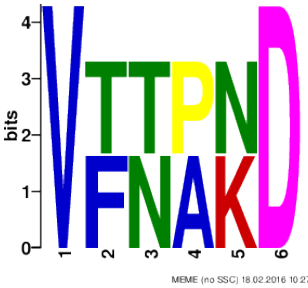


Figure 3.109: Motif 2

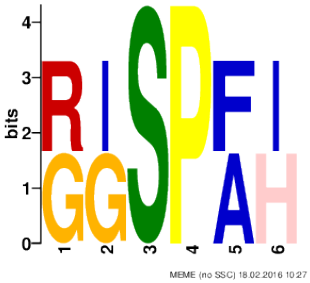


Figure3.110: Motif 3

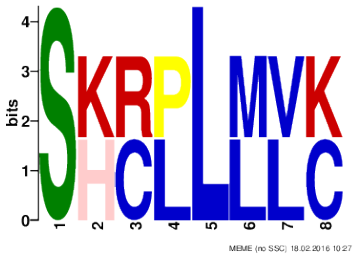


Figure3.111: Motif 4

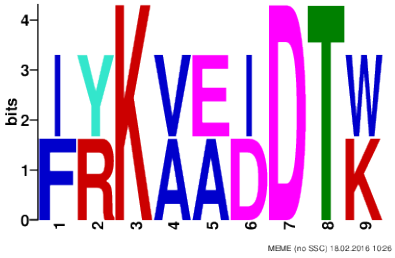


Figure3.112: Motif 5

3.2.3 Expression Level:

Like every biomarker molecule proteins are known as biomarkers only because of their unique expression level in different parameter of breast cancer. For this no practical experiment was

performed rather results were taken from the stored information in GEO Profiles. Each protein molecule is different in their expression profile and that makes them unique biomarker molecule.

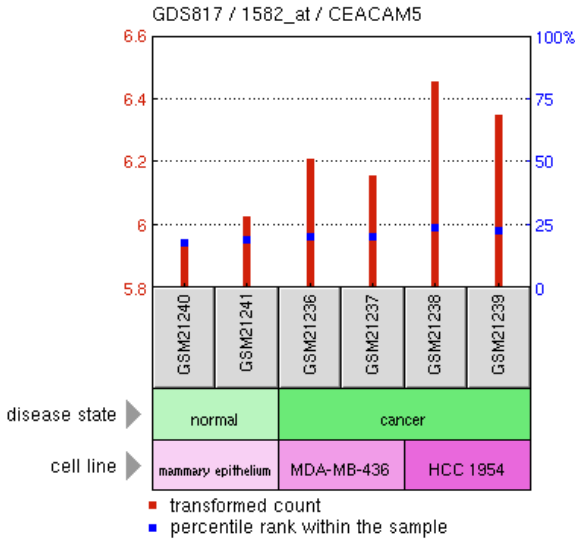


Figure3.113: Expression level of CEA

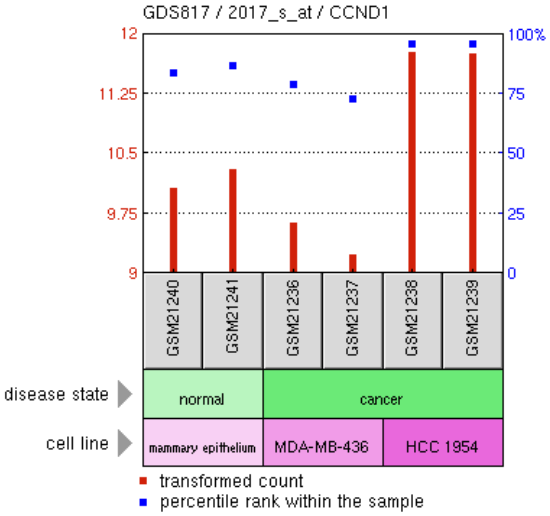


Figure3.114: Expression level of CyclinD1

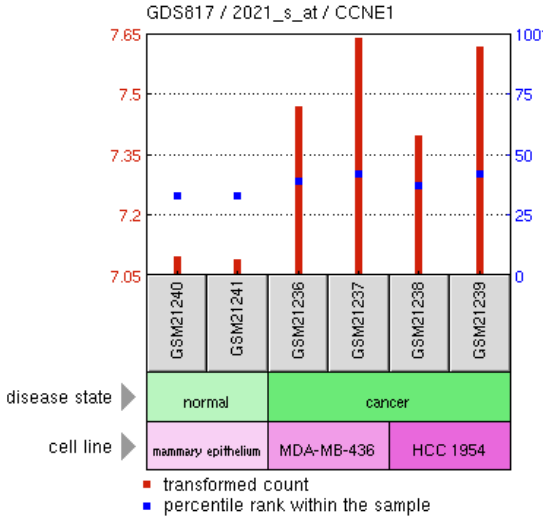


Figure3.115: Expression level of Cyclin E

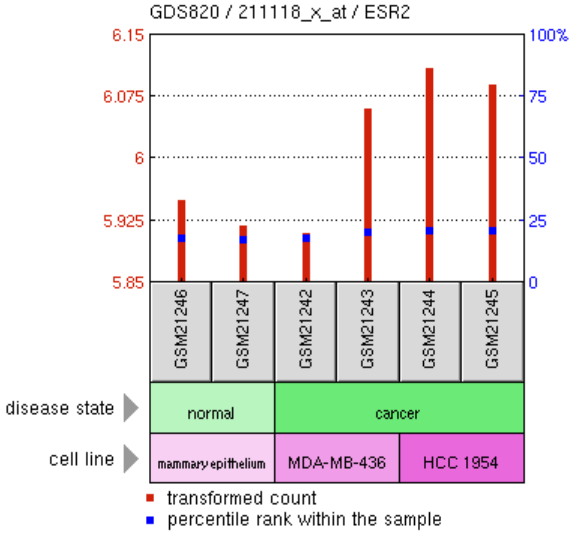


Figure3.116: Expression level of ER beta

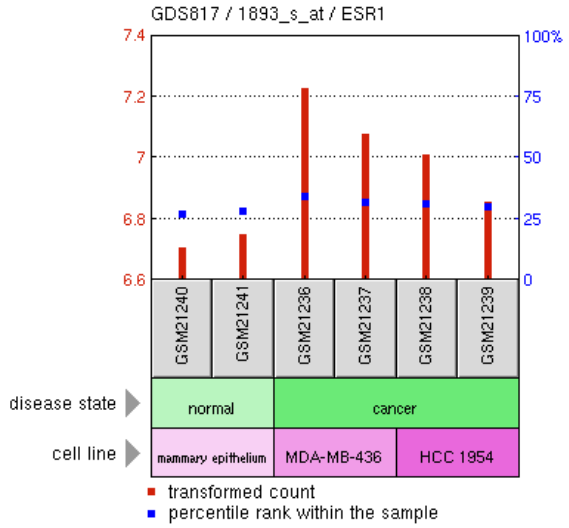


Figure3.117: Expression level of ER

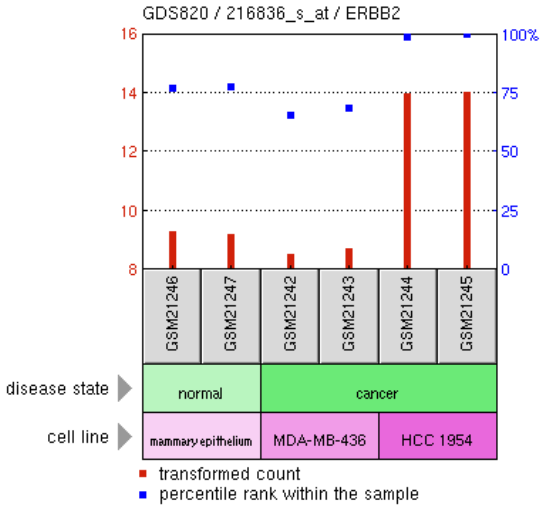


Figure3.118: Expression level of Her2

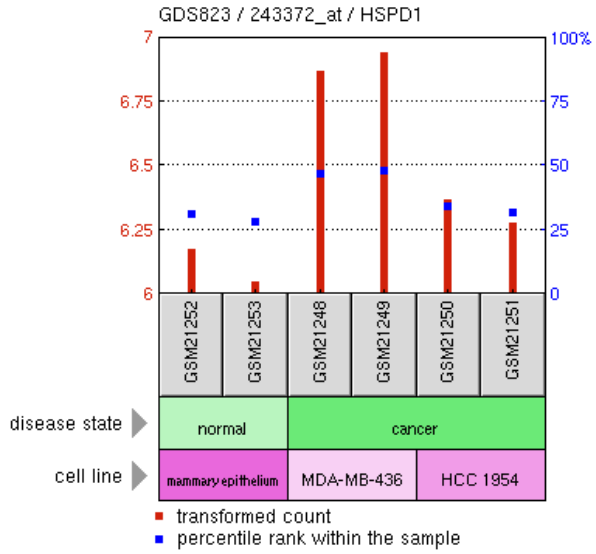


Figure3.119: Expression level of HSP60

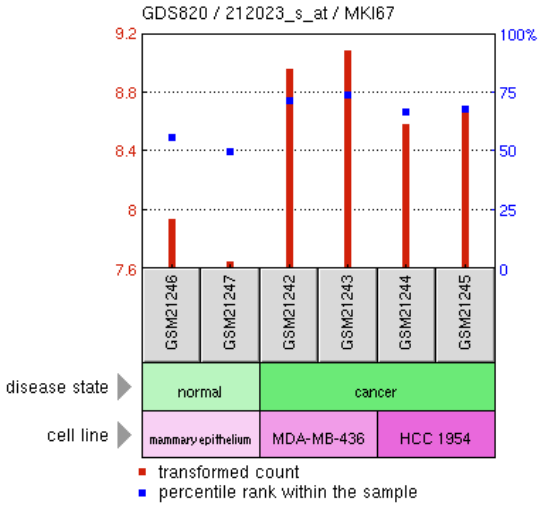


Figure3.120: Expression level of Ki67

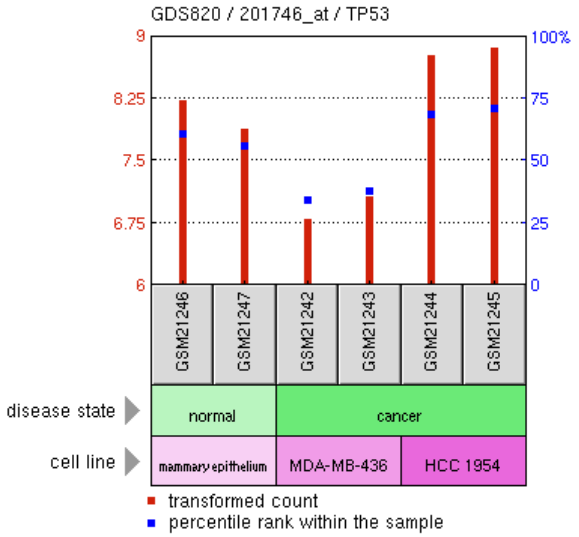


Figure3.121: Expression level of P53

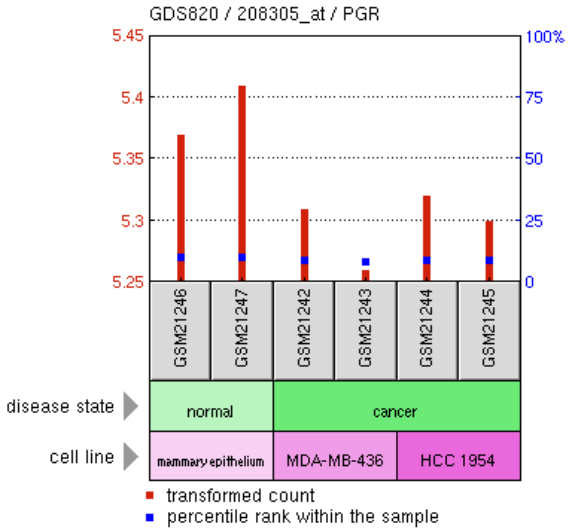


Figure3.122: Expression level of PR

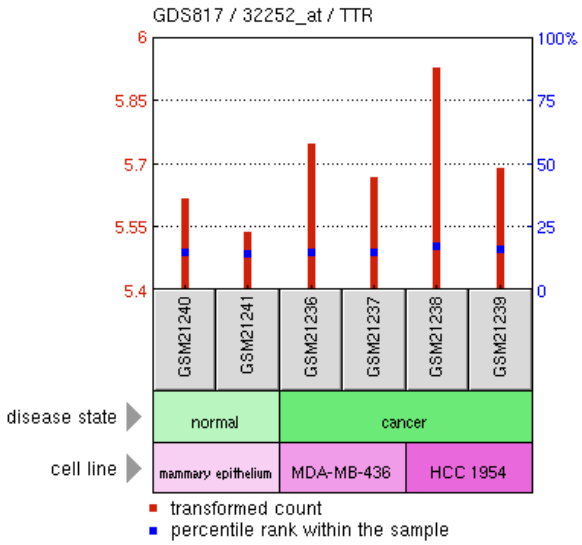


Figure3.123: Expression level of TTR

In these expression profiles the expression level of the selected protein molecules are shown in two different cancerous cell line along with a normal cell line. The expression of CEA, CyclinE, ER, ER Beta, HSP60, Ki67 and TTR are almost similar. All of them are overexpressed in both the breast cancer cell lines (MDB-MB-436 and HCC 1954). On the other side expression level of Cyclin D1, Her2 and P53 are almost similar. They are more expressed in the HCC 1954 cell line and less expressed in the MDB- MB- 436 cell line. Lastly PR is differently expressed than all as it is under expressed in both of the cell lines than the normal one.

CHAPTER 4:
DISCUSSION

According to the studies and found-out results, the results turned out to be as expected. But the result needs to be discussed for clear understanding.

4.1 miRNA Structure

To observe the miRNA structure mfold tool was used. With the accomplishment of the operation there were results shown into two different parameters. One is the Centroid structure (a structure of RNA that represents minimal base pair distance compared to other secondary structures in the Boltzmann ensemble) and another one is the MFE structure (a structure of RNA that contributes minimum of free energy). Among the two options MFE structure was preferred to be presented in the thesis as this form of miRNA structure has been established as standards over two decades. (Zucker, 1981).These two parameters also had subdivisions. Like under MFE there are three options to choose from- normal black and white image, colored based on the base pairing probability and colored by the positional entropy. As this study was about mere structure observation, so the base pair probability based colored image were selected. All the miRNA structures are of MFE structure and colored on base pair probability. The color ranges from blue to red. Blue being the lowest point 0, followed by green yellow and red- the highest point.

4.2 miRNA Motif

To observe the miRNA motif MEME suite software was used. And after being done with the command, the result showing page came with- the motifs, their logos and sites which are available for downloading. Motifs are the sequences with biological and functional importance, logos are the graphical presentation of that motif and sites are the sequences where motifs were found (Timothy, 1994). For each miRNA maximum five motifs were commanded to be found. But there are certain miRNAs- like mir21 and mir191, where five motifs could not be found so they had to be presented with 3 and 4 numbers of motifs. Now if the logos of the motifs are seen carefully it can be seen that they can be of different lengths because the tool automatically selects the standard length from 6 to 50. Most of the motifs are present in two sites, but only a few are present in more than two sites. Those are mir191, mir145 and mir21. These sums up to the fact that these three miRNA has motifs in more than 2 sites in the whole sequence, so they might have the desired motifs that can play an important role in breast cancer screening.

4.3 miRNA Expression Level

Observing the expression level was another parameter in this study of biomarkers and it was done with the help of GEO Profiles, as they secure different assay results on different molecules. So from this huge store book of experiments best suited profiles were taken. As the result is found in charts it is important to know about the terms and organization of the chart. Three different type of expression profile were found here. One is the comparison of normal and breast cancer cell line. Another is the comparison of normal breast cancer cell and cancerous cell with ER mutated. Last type of comparison was between breast cancer patients with or without doxycycline treatment. All these different cell lines are presented in the bottom light pink bars. And the diseased state are shown in the second bottom green bars. Above these two line of bars, ash colored bars show the name of the samples. In the long red lines that represents the transformed count of the expression level. This transformed count is from the actual experiment results as they were performed in affymetrix systems. And the blue squares presents their percentile rank among all the samples. The result of miRNA expression level shows that miR10B and miR21 show their performance better in the ER silenced cell line. They can be used as prognostic biomarker for this type of treatment. miR145, miR191,miR382, miR425 none of them showed any clear result in the doxycycline treatment systems, which makes them weak biomarkers. Lastly miR155 shows different expression in two different breast cancer cell lines rather than the normal condition. This miRNA might not be used as a marker for breast cancer screening but this could be a good biomarker for breast cancer classification.

4.4 Protein Structure

Protein is one tough molecule when it comes to its structure. First of all it has four level of structure. Primary, secondary, tertiary and quaternary. This thesis focused on the tertiary structure which by definition means a special geometric shape of the protein that is formed by the bonding interactions of the different side chains around the main helix. (Murphy, 2009). Now the tertiary structure can be of different types. For example: AB initio, threading, homology modeling etc. Here homology modeling was performed with the help of SWISS MODEL Workspace. This homology modeling means comparing models with another homologous proteins structure. In the homology model the spiral like structure is known as alpha helix and the wide sheet like part is the beta sheet and the thin spread like structure is the single polypeptide chain. The structure is colored and the color shows the residue error, blue being the lowest it goes up to red. All the homology models are validated with QMEAN and the QMEAN score says all of them are more than .77% in the scale of 0 to 1.

4.5 Protein Motif

Protein motif is the sequences that might have biological or functional importance. Just as it was done before with the miRNA molecules, here MEME suite is also used to find proteins sequence motifs. After being done with the command the result showing page came with the motifs, their logos and sites which are available for downloading.Motifs are the sequences with biological and functional importance, logos are the graphical presentation of that motif and sites are the sequences where motifs were found (Timothy, 1994). For each protein molecule limit to the motif search was given 5, which means maximum five motifs were commanded to be found. The logos can be of different lengths because the tool automatically selects the standard length from 6 to 50. Most of the motifs are found in two sites. But there are some motifs in CEA, Cyclin E, Ki67, Her2 and P53 which occurs more than that. This result suggests that these proteins might have the derired motif that can be the ultimate biomarker in the panel.

4.6 Protein Expression Level

For observing the expression level of protein GEO Profile is used rather than performing an experiment this very own thesis, as this is not feasible in any condition. So from the collection of various assays performed in various experiments, ones that compliments this thesis were taken to be put in the result section. To explain the charts, three cell lines are compared there to put on a better understanding. One is normal cell line from mammary epithelium, second one is from MDA-MB-436 cell line which have adenocarcinoma and the third one is from HCC 1954 which have stage 3 ductal carcinoma (Barrett, 2013). All these different cell lines are presented in the bottom light pink bars. And the diseased state are shown in the second bottom green bars. Above these two line of bars, ash colored bars show the name of the samples. In the long red lines that represents the transformed count of the expression level. This transformed count is from the actual experiment results. And the blue squares presents their percentile rank among all the samples. Among all the proteins CEA, Cyclin E, ERbeta, ER, HSP60, Ki67, TTR are clearly overexpressed in the diseased person almost as twice as the normal condition. On the other hand PR is under expressed. Clearly this protein molecules can be used as biomarker in the panel for breast cancer screening. The rest of the proteins Cyclin D1, Her2 and P53 show different expressions in different types of breast cancer. These three show better result within two types of breast cancer. It is hoped that they can be a potential biomarker for breast cancer going.

If these results are combined together a better biomarker panel could be decided with CEA, Cyclin E, ER, ER Beta, HSP60, KI67, TTR and PR. On the other hand Cyclin D1, Her2, P53 along with miR155can make a biomarker panel for breast cancer staging. mi10B and miR21 can play biomarker role in the ER silencing treatment systems.

CHAPTER 5:

CONCLUSION

Breast cancer is a global curse. This is the most commonly encountered cancer in our country as well as in the whole world (Baskin, 2010). To fight back with this monster like disease scientists are working all day and night, but still no drastic change is observed in the incidence or mortality rate just because of the diagnosis and treatment systems have not improved enough.

Up until now a lot of study has been performed on breast cancer molecules, but only a minority of them are being used clinically to improve the mortality rate. The reason behind this scenario is inadequate information about the basic properties of these molecules. So clearly once the information about basic properties are gathered experiments based on their clinical use can be designed to find a better solution to this problem. These experiments can lead us to a door to better diagnosis and treatment system of breast cancer.

With an objective to add a little help in the findings of better treatment and diagnosis system this thesis was designed to know more about breast cancer biomarker molecules like protein and miRNA. Among other different biomarker molecules, these molecules were chosen because they can be found or collected in enough amount from patients body (Chung, 2014). The main objective of this research was basically to study about a few basic properties about specific promising breast cancer biomarkers. Among the properties, three were selected- structure, sequence motif and expression level. They were also chosen by their capability, 11 protein and 7 miRNA potential biomarker molecules were selected to be worked with. As the name suggests, the study progressed to find out the 3 basic properties of the selected 11 proteins and 7 miRNA molecules to find the answers about the basic properties, the help of computational tools were taken. In other words it can be said that it was all about a bioinformatics approach to find out the main properties of the potential breast cancer biomarker molecules. This could also be done with experimental approach but bioinformatics makes this process much easier, less time consuming and easy to perform on.

So this thesis project can open a whole new areas of study for these biomarker molecules. The structures of the protein and miRNA can help in future studies of, their binding properties, and target molecules, virtual screening capabilities as well. The sequence motifs can make a contribution in the researches that are designed to find out the regulatory properties, rationality and therapeutic agents. The expression level might come in very handy to detect breast cancer carrier, patient, stage or even types too. All these are only leading to a better panel of biomarkers that can be used for diagnosis of the disease or as therapeutic agents or even a marker panel that can predict the best suited treatment system for a patient.

In the sector of breast cancer disease research, this study can open a field of chances to evaluate the disease, knowing more about the disease and maybe finding a permanent remedy to it. Even though a lot of research and study might be needed to beat this dangerous disease. But this type of approach that includes bioinformatics, might shine a light in the darkness of the unknown. The experimental approach can be used to come to the same conclusion too but bioinformatics approaches are great for finding out information about nucleotide level since they do not miss out on any messages easily or by mistake.

The future of breast cancer therapy lie in the use of biomarkers that offer the potential to identify and treat cancer years before it is either visible or symptomatic. (Bhatt, 2010). So with the hope of new advancements and invention, this study was dedicated as a little contributions towards the invention of a way to fight back breast cancer that includes early diagnosis, better treatment and less mortality rate.

CHAPTER 6:
REFERENCES

1. David P. (2004) MicroRNAs: Genomics, Biogenesis, Mechanism, and Function ,Cell, Vol. 116, 281–297
2. Jacques F., Hai-Rim S., Freddie B., David F., Colin M.and Donald M. (2008) Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008,International Journal of Cancer
3. Story H.,Love R., Salim R., Roberto A., Krieger L.,and Ginsburg G. (2012) Improving Outcomes from Breast Cancer in a Low-Income Country: Lessons from Bangladesh ,International Journal of Breast Cancer Volume 2012, Article ID 423562, 9 pages doi:10.1155/2012/4 23562
4. Rahman M., Ahsan A., Begum F., Rahman K., (2015) Epidemiology, Risk Factors and Tumor Profiles of Breast Cancer in Bangladeshi underprivileged women ,The Gulf Journal of Oncology
5. Long D., Lee R., Williams P., Chan C., Ambros V.& Ding Y. (2007) Potent effect of target structure on microRNA function; doi:10.1038/nsmb1226
6. ArnoldK., Bordoli L., KoppJ. and Schwede T.(2006) The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling
7. Forazy A., Chowdhury B. (2015) Incidence of Breast Cancer in Bangladesh
8. Liu J., Huang W., Yang H. & Luo Y. (2015) Expression and function of miR-155 in breast cancer, Biotechnology & Biotechnological Equipment, 29:5, 840-843, DOI: 10.1080/13102818.2015.1043946
9. Chung L., Moore K., Phillips L., Boyle F., Marsh D.and Baxter R. (2014) Novel serum protein biomarker panel revealed by mass spectrometry and its prognostic value in breast cancer,Breast cancer research
10. Li J., Zhang Z., Rosenzweig J.,Wang Y, and Chan D., (2002),Proteomics and Bioinformatics Approaches forIdentification of Serum Biomarkers to Detect Breast Cancer, ' Clinical Chemistry 48:8 1296–1304
11. Alexander H., Stegner A., Mann C.,Bois G.,Alexander S.,and Sauter E.(2004) ,Proteomic Analysis to Identify Breast Cancer Biomarkers in Nipple aspartate fluid Clinical cancer research ,Vol. 10, 7500–7510
12. Voduc K., Cheang M., Tyldesley S., Gelmon K., Nielsen T., Kennecke H., OncolJ. (2010) Breast Cancer Subtypes and the Risk of Local and Regional Relapse, 28:1684-1691.
13. Li J.,Orlandi R.,White C.,Rosenzweig J.,Zhao J., Seregni E.,Morelli D.,Yu Y.,Meng X.,Zhang Z.,Eric N., Fung T.,and Chan D. (2005) Independent Validation of Candidate Breast Cancer Serum Biomarkers Identified by Mass Spectrometry, , Clinical Chemistry 51:12, 2229–2235.
14. Baskın Y. and Yiitbaı T. (2010) Clinical Proteomics of Breast Cancer , Current Genomics, 11, 528-536

15. Bhatt A., Mathur R., Farooque A., Verma A. &Dwarakanath B.(2010) Cancer biomarkers - Current perspectives ,
16. Duffy M., McGowan P., Harbeck N., Thomssen C.and Schmitt M. (2014) uPA and PAI-1 as biomarkers in breast cancer: validated for clinical use in level-of-evidence-1 studies, ' breast cancer research
17. Kim B., Lee J., Park P.,Shin Y., Lee W., Lee K., Ye S., Hyun H., Kang K., Yeo D., Kim Y., Ohn S., Noh D. and Kim C. (2009), The multiplex bead array approach to identifying serum biomarkers associated with breast cancer, Breast Cancer Research
18. Rai S. (2012) Structural Characterization of Potential Cancer Biomarker Proteins by
19. Kulasingam V. (2008) Identification And Validation Of Candidate Breast Cancer Biomarkers: A Mass Spectrometric Approach
20. Hossain M., Ferdous S., Henrike E., Kos K. (2014) Breast cancer in South Asia: A Bangladesh perspectiveCancer Epidemiology, DOI:10.1016/j.canep.2014.08.004
21. Evangelia , Fourkala O. (2009) Risk factors and novel biomarkers in breast cancer
22. Rothschild S. (2014) microRNA therapies in cancer, , molecular and cellular therapies
23. Aguilar F., Jorge A., Ram´ırez M., Santiago I., Perla K., Silva E., Santuario-Facio S., Ruiz-Flores P., Rodr´ıguez-Padilla P. and Resendez-P´erez D. (2013) Serum circulating microRNA profiling for identification of potential breast cancer biomarkers, Disease Markers 34 ,163–169 163, DOI 10.3233/DMA-120957, IOS Press
24. Zhao H., Shen J., Medico L., Wang D., Ambrosone C. (2010) A Pilot Study of Circulating miRNAs as Potential Biomarkers of Early Stage Breast Cancer. PLoS ONE 5(10): e13735. doi:10.1371/journal.pone.0013735
25. Weige M.and Dowsett M. (2010),Current and emerging biomarkers in breast cancer: prognosis and prediction, Endocrine-Related Cancer
26. Misek D.and Kim E.(2011) Protein Biomarkers for the Early Detection of Breast Cancer, International Journal of Proteomics Volume 2011, Article ID 343582, 9 pages doi:10.1155/2011/343582
27. Kozomara A., Griffiths-Jones S. (2014) miRBase: annotating high confidence microRNAs using deep sequencing data..
28. Griffiths-Jones S., Saini HK., Dongen S., Enright AJ. (2008) miRBase: tools for microRNA genomics
29. Griffiths-Jones S., Grocock RJ., Dongen S., Bateman A., Enright. (2006) miRBase: microRNA sequences, targets and gene nomenclature.
30. Griffiths-Jones S. (2004) The microRNA Registry.
31. M. Zuker.(2003) Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res.31 (13), 3406-3415
32. A. Waugh, P. Gendron, R. Altman, J. W. Brown, D. Case, D. Gautheret, S. C. Harvey, N. Leontis, J. Westbrook, E. Westhof, M. Zuker & F. Major.(2002) RNAML: A standard syntax for exchanging RNA information. RNA8 (6), 707-717

33. M. Zuker & A. B. Jacobson. (1998) Using Reliability Information to Annotate RNA Secondary Structures. RNA4, 669-679
34. Bailey T., Bodén M., Buske F., Frith M., Grant C., Clementi L., Ren J., Li W., Noble W. (2009) "MEME SUITE: tools for motif discovery and searching", Nucleic Acids Research, 37:W202-W208.
35. Barrett T., Wilhite SE., Ledoux P., Evangelista C., Kim IF., Tomashevsky M., Marshall KA., Phillippy KH., Sherman PM., Holko M., Yefanov A., Lee H., Zhang N., Robertson CL., Serova N., Davis S., Soboleva A.(2013) NCBI GEO: archive for functional genomics data sets--update.
36. The UniProt Consortium UniProt: a hub for protein informationNucleic Acids Res. 43: D204-D212 (2015)
37. Altschul S.F., Gish W., Miller W., Myers E.W. & Lipman D.J. (1990) "Basic local alignment search tool." J. Mol. Biol. 215:403-410. PubMed
38. Gish W. & States D.J. (1993) "Identification of protein coding regions by database similarity search." Nature Genet. 3:266-272. PubMed
39. Madden T.L., Tatusov R.L. & Zhang J. (1996) "Applications of network BLAST server" Meth. Enzymol. 266:131-141. PubMed
40. Altschul S.F., Madden T.L., Schäffer A.A., Zhang J., Zhang Z., Miller W. & Lipman D.J. (1997) "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res. 25:3389-3402. PubMed
41. Zhang Z., Schwartz S., Wagner L., & Miller W. (2000), "A greedy algorithm for aligning DNA sequences" J Comput Biol 2000; 7(1-2):203-14. PubMed
42. Zhang J. &Madden T.L. (1997) "PowerBLAST: A new network BLAST application for interactive or automated sequence analysis and annotation." Genome Res. 7:649-656. PubMed
43. Morgulis A., Coulouris G., Raytselis Y., Madden T.L., Agarwala R., & Schäffer A.A. (2008) "Database indexing for production MegaBLAST searches." Bioinformatics 15:1757-1764. PubMed
44. Camacho C., Coulouris G., Avagyan V., Ma N., Papadopoulos J., Bealer K., & Madden T.L. (2008) "BLAST+: architecture and applications." BMC Bioinformatics 10:421. PubMed
45. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. (2011) Molecular systems biology 7 :539 PMID: 21988835
46. The EMBL-EBI bioinformatics web and programmatic tools framework. (2015 July 01) Nucleic acids research 43 (W1) :W580-4 PMID: 25845596
47. Analysis Tool Web Services from the EMBL-EBI. (2013 July) Nucleic acids research 41 (Web Server issue) :W597-600 PMID: 23671338
48. Arnold K., Bordoli L., Kopp J., and Schwede T. (2006). The SWISS-MODEL Workspace: A web-based environment for protein structure homology modelling. Bioinformatics, 22,195-201.

49. Kiefer F., Arnold K., Künzli M., Bordoli L., Schwede T. (2009). The SWISS-MODEL Repository and associated resources. Nucleic Acids Research. 37, D387-D392.
50. Schwede T., Kopp J., Guex N., and Peitsch MC. (2003) SWISS-MODEL: an automated protein homology-modeling server. Nucleic Acids Research 31: 3381-3385.
51. Guex, N. and Peitsch, M. C. (1997) SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modelling. Electrophoresis 18: 2714-2723.
52. Peitsch, M. C. (1995) Protein modeling by E-mail Bio/Technology 13: 658-660.