Protein Biomarkers for Lung Cancer: A Bioinformatic Approach

A DISSERTATION SUBMITTED TO BRAC UNIVERSITY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF BACHELOR OF SCIENCE IN BIOTECHNOLOGY

Submitted by

Joynob Jeba

Student ID: 12136010

Biotechnology Program

Department of Mathematics and Natural Sciences

BRAC University

Bangladesh

August 2016

DECLARATION

I hereby declare that the research work embodying the results reported in this thesis entitled “Protein Biomarkers for Lung Cancer: A Bioinformatic Approach” submitted by the undersigned has been carried out under the supervision of Ms. Sadia Sayed, 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 has not been submitted to any other institution for any degree or diploma.

Candidate:

_____

(Joynob Jeba)

Certified:

_____

(Ms. Sadia Sayed)

Supervisor

Lecturer,

Biotechnology program

Department of Mathematics and Natural Sciences

BRAC University, Dhaka

17, August, 2016

ACKNOWLEDGEMENTS

First of all, I want to express my gratitude to the Almighty for blessing my life with everything required. Also, I want to express my profound gratitude to Professor Naiyyum Choudhury, former coordinator of the Biotechnology and Microbiology program, Department of Mathematics and Natural Sciences, BRAC University, Chairman, Bangladesh Atomic Energy Regulatory Authority (BAERA); and Professor A. A. Ziauddin Ahmad, Chairperson, Department of Mathematics and Natural Sciences, BRAC University, for the being the inspiration and enabling all the opportunities for betterment of the study.

This project was done under the supervision of my Supervisor Sadia Sayed, Biotechnology program, Lecturer of BRAC University. Her proper guidance, positivity and mental support throughout the study made it possible for me to complete the work properly. I want to wholeheartedly thank my supervisor for being the roof above my head.

I want to especially express my obligation to respected Dr. Aparna Islam, Associate Professor and Coordinator of Postgraduate Biotechnology program; Ms. Jebunnesa Chowdhury, Assistant Professor and Ms. Romanna Siddiqui, Senior Lecturer and current Coordinators of Undergraduate Biotechnology program for sharing their tremendous knowledge from the very beginning of my program.

Lastly, I would like to thank my friends and family for supporting me throughout the study.

Joynob Jeba

ABSTRACT

Lung cancer is one of the most occurring cancers around the world with increasing mortality rate and Bangladesh has a very high rate of lung cancer patient. Detection of lung cancer is not very easy process and quite undetectable until it reaches to the metastatic stage. Protein biomarkers are already being studied to diagnose disease type and stage along with the proper treatment planning for several types of cancer including breast, pancreatic and colon cancer. In this study, 14 proteins (MDK, MMP2, TFPI, TIMP-1, OPN, BIRC6, CEA, CDK4, HSPA5, HSP90α, EFGR, and ACTN4) were chosen that are expressed in lung cancer patient and analyzed using different bioinformatics tools (MEME, SWISS MODEL WORKPLACE). Since sequence motif can be used as a tool for predicting protein function for each of the protein, after assessing the homology modeling and sequence motifs of the 14 proteins, 3(OPN, EGFR, TIMP1) of them were identified for further analysis. The expression level of these lung cancer proteins were checked for better understanding. Additional wet-lab based study of such protein biomarkers maybe used as a panel for further researches that are designed to find out the regulatory properties, rationality and therapeutic agents, for diagnosis of the disease and the best suited treatment system for a lung cancer patient in future.

Table of contents

List of figures

List of Tables

List of abbreviation

NSCLC	Non-Small cell Lung Cancer
SCLC	Small cell lung cancer
IARC	International Agency for research
IHC	ImmunoHistoChemistry
NIDRH	National Institute of cancer research & hospital
LDCT	Low dose computed tomography
EBUS	Endobronchial ultrasound bronchoscopy
VATS	Video assister troracoscopic
CAD	Computer aided detection
SABRT	Stereotactic Ablative Radiation Therapy

PET	Positron emission tomography
EUS	Transesophageal ultrasound
MDK	Midkine
NEGF2	Neurite growth-promoting factor 2
MMP2	Matrix metalloproteinase-2
TFPI	Tissue Factor Pathway Inhibitor
TIMP	Metallopeptidase inhibitor
OPN	Osteopontin
BIRC6	Baculoviral Inhibitor of apoptosis repeat-containing protein 6
CEA	Carcinogenic antigen
CDK4	Cyclin-dependent kinase 4
Hsp90	Heat shock protein 90
HSPA5	Heat shock 70 kDa protein 5
BiP	Binding immunoglobulin protein
EGFR	Epidermal growth factor receptor
ACTN4	Alpha-actinin 4 protein
SIB	Swiss Institute of Bioinformatics
PIR	Protein Information Resourse
EMBL	European Molecular Biology Laboratory
EBI	European Bioinformatics Institute
NLM	National Library of Medicine
NIH	National Institutes of Health
NCBI	National Center for Biotechnology Information

CHAPTER 1:

INTRODUCTION

Humanity and disease share a long and eventful history. As we emerged and evolved, so did the diseases that blight our lives. In the new millennium it is becoming increasingly clear that the biomedical sciences are entering the most exciting phase of their development. Paradoxically, medical practice is also passing through a phase of increasing uncertainty. However, it is difficult to anticipate when the gains of this explosion in scientific knowledge will become available for the prevention and treatment of the major killers of mankind. In this era, the new diseases are more like challenges for since to solve. Science is winning the challenges against the lethal diseases but meanwhile, a large number of people are being prey as to find a cure for any disease it takes years of studies.

Cancer is of such diseases that is washing people's life away globally. Cancer is not just one disease but many diseases. 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). Cancer is a condition where cells in a specific part of the body grow and reproduce uncontrollably. The cancerous cells can invade and destroy surrounding healthy tissue, including organs. According to the World Health Organization (WHO), 7.6 million deaths globally each year are caused by cancer; cancer represents 13% of all global deaths. lung cancer is by far the number one cancer killer.

New scientific researches are being designed to save mankind from the deadly trap of cancers. All different the sections of science working together to find ways to fight back cancers from wiping away a large number of life from earth every year. 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).

This thesis project is inspired from the concept of bioinformatics helping medical science to fight and win the battle against lung cancer. The basic concern is to know about lung cancer and its biomarker- protein, then finding their structures, motifs and expression level that might help demolish lung cancer in an easier way.

1.1Lung cancer

Lung cancer is a form of cancer an out-of-control growth of abnormal cells that starts in the tissues of the lungs or the cells that line the bronchi (tubes that move air into and out of the lungs).

There are 2 main types of lung cancer:

- Non-small cell lung cancer (NSCLC) is the most common type of lung cancer. About 80-85 percent of people with lung cancer have non-small cell lung cancer, or NSCLC. This type of cancer tends to spread less quickly.

- Small cell lung cancer (SCLC) makes up about 15-20 percent of all lung cancer cases. SCLC is an aggressive form of lung cancer that typically starts in the bronchi and spreads very quickly to other parts of the body. It's rare for someone who has never smoked to get SCLC.

If the lung cancer is made up of both types, it is called mixed small cell/large cell cancer. If the cancer started somewhere else in the body and spreads to the lungs, it is called metastatic cancer to the lung.

Lung cancer is most often diagnosed in people ages 65 to 74. About two out of three people diagnosed with lung cancer are 65 or older. About 13 percent of all new cancer cases — and 27 percent of all cancer deaths — are caused by lung cancer. Roughly 221,200 people diagnosed with lung cancer in 2015, and 158,040 people died of the disease.

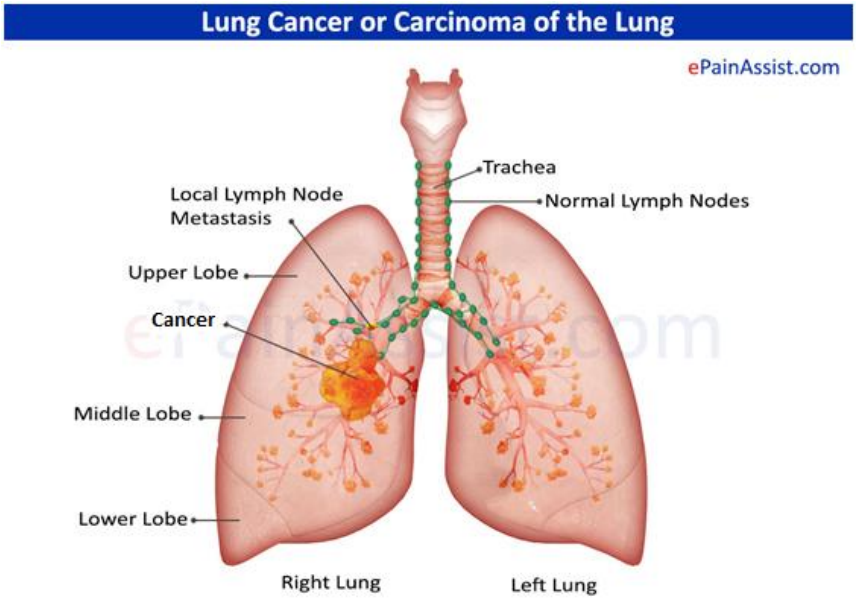

Figure 1.1: Lung Cancer anatomy

1.1.1 Symptoms

Early lung cancer may not cause any symptoms.

- Chest pain
- Cough that does not go away
- Coughing up blood
- Fatigue
- Losing weight without trying
- Loss of appetite
- Shortness of breath
- Wheezing
- Bone pain or tenderness
- Eyelid drooping
- Facial paralysis
- Hoarseness or changing voice
- Joint pain
- Nail problems
- Shoulder pain
- Swallowing difficulty
- Swelling of the face or arms
- Weakness
- Blood clots
- Bleeding

1.1.2 Risk Factors

- Smoking tobacco
- Second-hand smoke
- Radon
- Asbestos
- Outdoor air pollution
- Occupational exposure to chemical carcinogens
- Personal or family history of lung cancer
- Arsenic
- Previous lung disease
- Exposure to radiation
- Indoor burning of coal
- Weakened immune system

- Smoking marijuana
- Indoor burning of wood
- High-temperature frying
- Diet
- Physical inactivity
- Occupational exposure to certain chemicals
- Removal of both ovaries
- Family history of lung cancer.

Table 1. Risk factors for lung cancer

Factors	Tumor risk		
	Low	Middle	High
Size of pulmonary nodules (mm, in diameter)	<8	8-20	>20
Age (years)	<45	45-60	>60
Tumor history	Without tumor history		With tumor history
Smoking history	Never	Smoking, <1 pack per day	Smoking, ≥1 pack per day
History of smoking cessation	Having quit smoking for ≥7 years	Having quit smoking for <7 years	Never quite smoking
Chronic obstructive pulmonary disease	No	Yes	
History of asbestos exposure	No		Yes
Nodule characteristics	Smooth	Lobulated	Burr-like

Table 1.1: Risk factors of lung cancer(Roushney,2014)

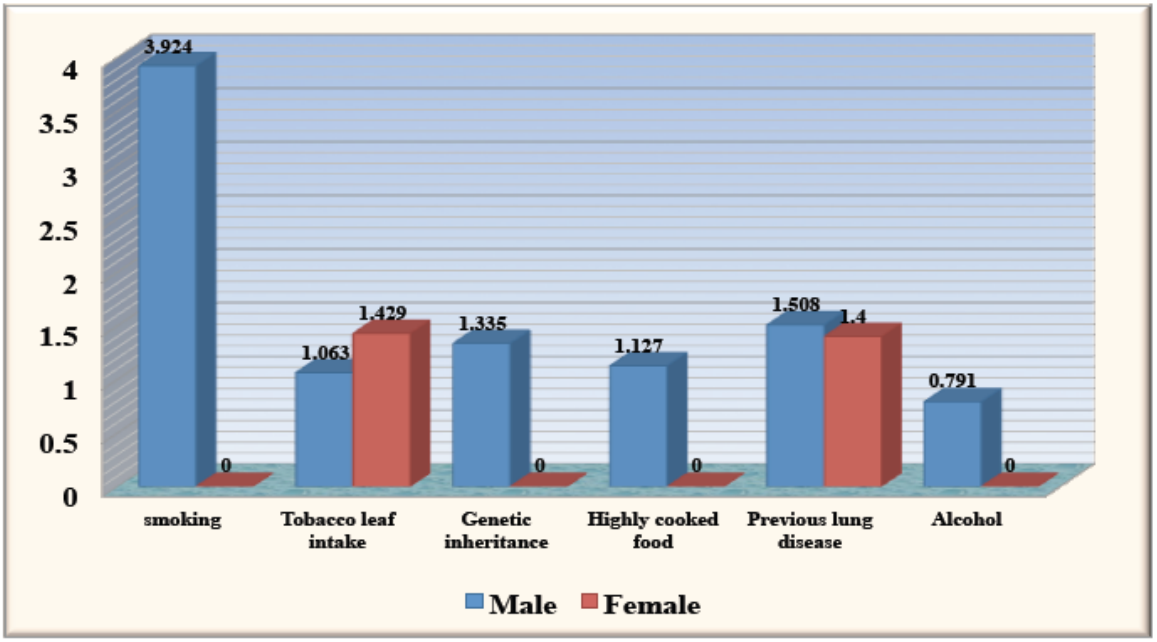

Table 1.2**:** Assessment of Score Based Risk Factors of Lung Cancer for Bangladeshi People(Roushney,2014)

1.1.3 Classifications of lung cancer

In 2004, the World Health Organization (WHO) and the International Agency for Research on Cancer (IARC) published a revised classification of lung cancer. This classification incorporated a number of developments, including recognition of lung cancer heterogeneity, the introduction of diagnostic immunohistochemistry(IHC), and the recognition of newly described entities such as fetal adenocarcinoma, cystic mucinous tumors, and large cell neuroendocrine carcinoma.

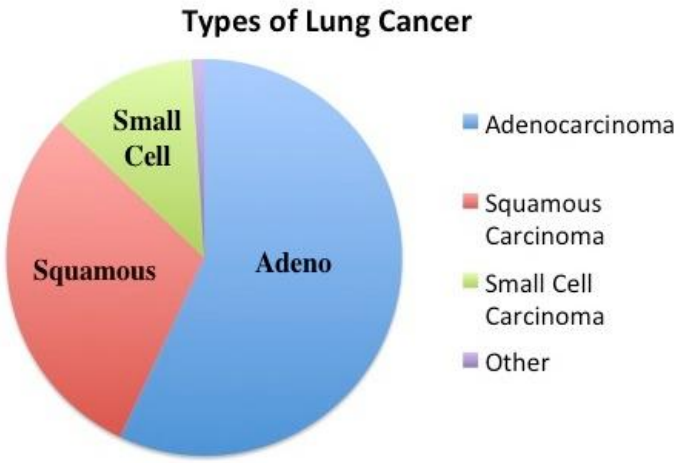

Figure 1.2: Primary type of lung cancer

The 2004 WHO Classification of Invasive Lung Cancer:

- Squamous cell carcinoma

  Variants: papillary, clear cell, small cell, basaloid

- Adenocarcinoma : Adenocarcinoma, mixed subtype
  - Acinar adenocarcinoma
  - Papillary adenocarcinoma
  - Bronchioloalveolar carcinoma
- Large cell carcinoma
  Variants: large cell neuroendocrine carcinoma, combined large cell neuroendocrine carcinoma, basaloid carcinoma, lymphoepithelioma-like carcinoma, clear cell carcinoma
- Adenosquamous carcinoma
  Variants: pleomorphic carcinoma, spindle cell carcinoma, giant cell carcinoma, carcinosarcoma.
- Carcinoid tumor
  Variants: typical carcinoid, atypical carcinoid

- Salivary gland tumors
  Variants: mucoepidermoid carcinoma, adenoid cystic carcinoma, epithelial-myoepithelial carcinoma.

In 2011, the International Association for the Study of Lung Cancer (IASLC), the American Thoracic Society (ATS) and the European Respiratory Society (ERS) made revisions to the pathologic (diagnostic) classification of lung cancer. Most importantly, revisions have been made to the classification of adenocarcinoma of the lung.

- Non-Small-Cell Lung Cancer (NSCLC)
- Adenocarcinoma
- Squamous (Cell) Carcinoma
- Small Cell / 'Oat' Cell / Neuroendocrine Carcinoma

1.1.4 Lung cancer and the world

Lung cancer has been the most common cancer in the world for several decades. There are estimated to be 1.8 million new cases in 2012 (12.9% of the total), 58% of which occurred in the less developed regions. The disease remains as the most common cancer in men worldwide (1.2 million, 16.7% of the total) with the highest estimated age-standardized incidence rates in Central and Eastern Europe (53.5 per 100,000) and Eastern Asia (50.4 per 100,000). Notably low incidence rates are observed in Middle and Western Africa (2.0 and 1.7 per 100,000 respectively). In women, the incidence rates are generally lower and the geographical pattern is a little different, mainly reflecting different historical exposure to tobacco smoking. Thus the highest estimated rates are in Northern America (33.8) and Northern Europe (23.7) with a relatively high rate in Eastern Asia (19.2) and the lowest rates again in Western and Middle Africa (1.1 and 0.8 respectively).

Lung cancer is the most common cause of death from cancer worldwide, estimated to be responsible for nearly one in five (1.59 million deaths, 19.4% of the total). Because of its high fatality (the overall ratio of mortality to incidence is 0.87) and the relative lack of variability in survival in different world regions, the geographical patterns in mortality closely follow those in incidence.

There were an estimated 14.1 million cancer cases around the world in 2012, of these 7.4 million cases were in men and 6.7 million in women. This number is expected to increase to 24 million by 2035.

This growing cancer burden, within the overall context of non-communicable diseases (NCDs), was a key focus of the September 2011 UN High Level Meeting on NCDs. SEER data for 2014 estimates that new lung cancer cases in the U.S. were 224,210 with lung cancer representing 13.5 % of all new cancer cases and 27.2 % of all cancer deaths.

Lung Cancer

Estimated Incidence, Mortality and Prevalence Worldwide in 2012

Estimated numbers (thousands)	Men			Women			Both sexes		
	Cases	Deaths	5-year prev.	Cases	Deaths	5-year prev.	Cases	Deaths	5-year prev.
World	1242	1099	1267	583	491	626	1825	1590	1893
More developed regions	490	417	593	268	210	341	758	627	933
Less developed regions	751	682	674	315	281	286	1066	963	960
WHO Africa region (AFRO)	12	11	10	6	6	5	18	16	15
WHO Americas region (PAHO)	178	149	208	146	113	175	324	262	383
WHO East Mediterranean region (EMRO)	26	23	22	7	6	6	33	29	28
WHO Europe region (EURO)	323	283	343	126	105	133	449	388	476
WHO South-East Asia region (SEARO)	116	104	79	46	42	34	162	146	113
WHO Western Pacific region (WPRO)	588	528	605	251	220	273	839	748	878
IARC membership (24 countries)	514	438	582	279	219	343	794	657	925
United States of America	112	92	140	102	76	128	214	168	269
China	459	422	431	193	175	179	653	597	610
India	54	49	24	17	15	8	70	64	32
European Union (EU-28)	214	186	234	99	82	106	313	268	340

Table 1.3: Estimated Incidence, Mortality & Prevalence for lung cancer worldwide(2012)

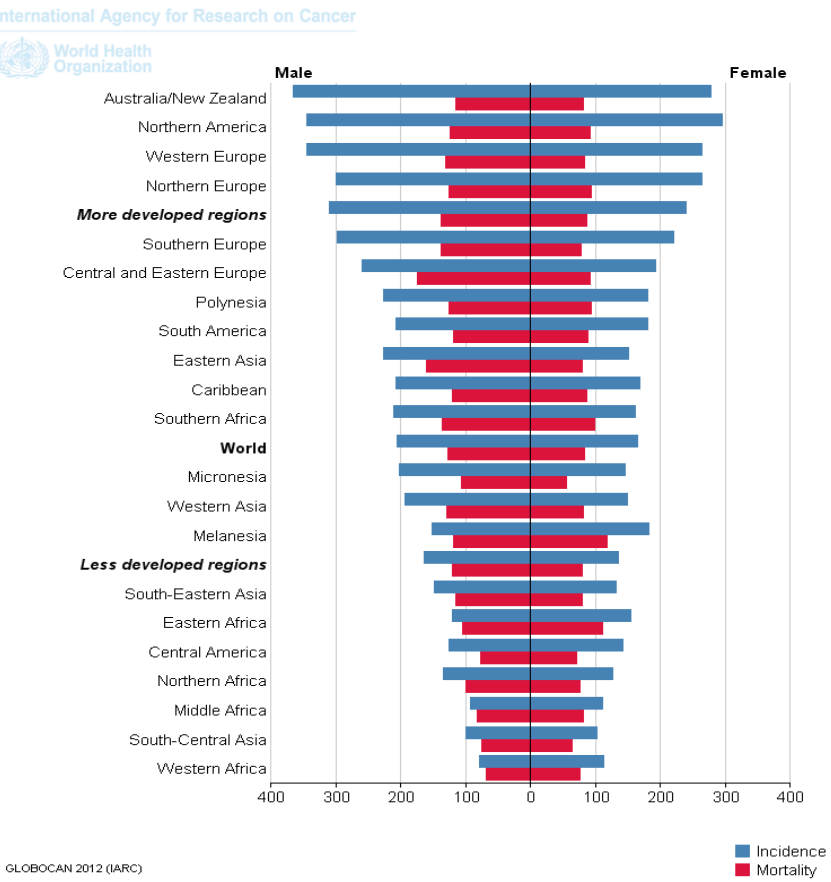

Figure 1.3: Gender based incidence & mortality rate comparison of lung cancer worldwide(2012)

1.1.5 Lung cancer in Bangladesh

Lung cancer has been the most common cancer in the world since 1985 and the leading cause of cancer death(parveen,2011). Worldwide it is by far the most common cancer of men and increasingly being recognized in Bangladesh. In Bangladesh Cancer Registry Report (2005-2007)5, most frequent cancer was lung cancer followed by carcinoma cervix and breast. The lung cancer is number one cancer problem comprising of 16.7% of all cancers of Bangladeshi people(syed,2013). Bangladeshi men had the highest age- and socioeconomic deprivation-standardized incidence rates of lung cancer compared with other ethnic groups. In male, lung cancer was found to be the most common cancer (15.75%) (Ragia,2015). WHO study (2013) estimates that there are 196 000 lung cancer cases in Bangladesh among those aged 30 years. Sex-specific top ten malignancies for three years are given in which males lung cancer topped the list (25.5%) that far exceeds the proportion of others. In NICRH which is the first national cancer registry Bangladesh, about 60% of the male lung cancer patients were smokers whereas 5% female lung cancer patients were smokers.

1.1.6 Present Screening system of Lung Cancer

Screening is looking for cancer before a person has any symptoms. This can help find cancer at an early stage. When abnormal tissue or cancer is found early, it may be easier to treat. By the time symptoms appear, cancer may have begun to spread. Some screening tests are used because they have been shown to be helpful both in finding cancer early and decreasing the chance of dying from these cancers. Other tests are used because they have been shown to find cancer in some people; however, it has not been proven in clinical trials that use of these tests will decrease the risk of dying from cancer.

Three screening tests have been studied to see if they decrease the risk of dying from lung cancer.

The following tests are assigned for lung cancer screening

- Low dose spiral CT scan(LDCT scan): A procedure that uses low-dose radiation to make a series of very detailed pictures of areas inside the body. It uses an x-ray machine that scans the body in a spiral path. The pictures are made by a computer linked to the x-ray machine. This procedure is also called a low-dose helical CT scan.
- Chest X-ray: An x-ray of the organs and bones inside the chest. An x-ray is a type of energy beam that can go through the body and onto film, making a picture of areas inside the body.
- Sputum cytology: Sputum cytology is a procedure in which a sample of sputum (mucus that is coughed up from the lungs) is viewed under a microscope to check for cancer cells.

If a screening test result is abnormal then more tests are done to find out if a patient actually has cancer. These are called diagnostic test

1.1.7 Present Diagnosis System

The early detection of cancer, as for other diseases, is one of the most important and challenging endeavors in clinical medicine.(yuji, 2015) If symptoms suggest lung cancer then more tests including blood work and x-rays or scans will then be ordered. If the chest x-ray or computed tomography (CT) scan shows an abnormal growth that could be a tumor, additional testing is performed to make a diagnosis. Usually, a small piece will

need to be removed from the chest and examined with a microscope. This procedure is called a biopsy. Importantly, the decision to perform a biopsy does not mean that cancer is present. Biopsies are routinely performed to check for both cancer as well as many other diseases.

A biopsy can be done in one of several ways:

- Bronchoscopy is a procedure where a flexible tube with a camera and other small instruments is inserted through mouth or nose and then into the windpipe.
- Endobronchial ultrasound bronchoscopy or EBUS is a technique that combines flexible bronchoscopy with ultrasound to first see lymph nodes in the chest and then to take biopsies from enlarged lymph nodes.
- CT-guided fine needle biopsy is performed by locating the tumor with a CT scan and inserting a thin needle through the skin to remove a tiny sample of tissue.
- Needle aspiration is performed by inserting a needle into lumps or lymph nodes that can be felt under the skin or seen with an ultrasound.
- Thoracentesis is insertion of a needle and small catheter into fluid collections in the chest to remove the fluid and look at it with a microscope.
- Surgery may be needed to remove the tumor entirely if it is small or if other biopsy procedures have not made a definitive diagnosis. The most common surgical procedures are mediastinoscopy, which is used to biopsy lymph nodes in the center of the chest; video-assisted thoracoscopic surgery (VATS), which is a less invasive way to biopsy lung tissue; and thoracotomy, which is a larger surgery to remove larger portions of lung tissue or tumors.

CADe(Computer-aided detection) and CADx systems for the detection and diagnosis of lung cancer have been important areas of research in recent decades. However, these areas are being worked on separately. CADe systems do not present the radiological characteristics of tumors, and CADx systems do not detect nodules and do not have good levels of automation. As a result, these systems are not yet widely used in clinical settings (Macedo,2015).

1.1.8 Present Treatment System

The treatment options for non-small cell lung cancer (NSCLC) are based mainly on the stage (extent) of the cancer, but other factors, such as a person's overall health and lung function, as well as certain traits of the cancer itself, are also important.

Depending on the stage of the cancer and other factors, treatment options for people with NSCLC can include:

- Surgery

- Radiofrequency ablation (RFA)
- Radiation therapy
- Chemotherapy
- Targeted therapies
- Immunotherapy

Palliative treatments can also be used to help with symptoms. In many cases, more than one of type of treatment is used.

- Treating stage 0 NSCLC

Stage 0 NSCLC is limited to the lining layer of airways and has not invaded deeper into the lung tissue or other areas, it is usually curable by surgery alone. No chemotherapy or radiation therapy is needed.

- Treating stage I NSCLC

For stage I NSCLC, surgery may be the only treatment needed. This may be done either by taking out the lobe of the lung containing the tumor (lobectomy) or by taking out a smaller piece of the lung (sleeve resection, segmentectomy, or wedge resection). At least some lymph nodes within the lung and in the space between the lungs will also be removed and checked for cancer cells.

- Treating stage II NSCLC

People who have stage II NSCLC and are healthy enough for surgery usually have the cancer removed by lobotomy or sleeve resection. Sometimes removing the whole lung (pneumonectomy) is needed. Any lymph nodes likely to have cancer in them are also removed. The extent of lymph node involvement and whether or not cancer cells are found at the edges of the removed tissues are important factors when planning the next step of treatment.

In some cases, chemotherapy (often along with radiation) may be recommend before surgery to try to shrink the tumor to make the operation easier.

- Treating stage IIIA NSCLC

Treatment for stage IIIA NSCLC may include some combination of radiation therapy, chemotherapy (chemo), and/or surgery. For this reason, planning treatment for stage IIIA NSCLC often requires input from a medical oncologist, radiation oncologist, and a thoracic surgeon. Treatment options depend on the size of the tumor, where it is in lung, which lymph nodes it has spread to, overall health and treatment impact.

- Treating stage IIIB NSCLC

Stage IIIB NSCLC has spread to lymph nodes that are near the other lung or in the neck, and may also have grown into important structures in the chest. These cancers can’t be

removed completely by surgery. As with other stages of lung cancer, treatment depends on the patient's overall health. In case of fairly good health chemotherapy (chemo) is combined with radiation therapy. Some people can even be cured with this treatment. Patients who are not healthy enough for this combination are often treated with radiation therapy alone, or, less often, chemo alone. These cancers can be hard to treat, so taking part in a clinical trial of newer treatments may be a good option.

- Treating stage IV NSCLC

Stage IV NSCLC is widespread when it is diagnosed. Because these cancers have spread to distant sites, they are very hard to cure. Treatment options depend on where the cancer has spread, the number of tumors, and overall health. For healthy patient treatments such as surgery, chemotherapy (chemo), therapy, immunotherapy, and radiation therapy may help live longer and feel better by relieving symptoms, even though these treatments aren't likely to completely cure patients.

1.2 The Challenge

In the last 15 years, the use of Stereotactic Ablative Radiation Therapy (SABRT) in the management of small peripheral lung tumors has developed considerably, so that it currently represents a standard of care for inoperable stage I non-small cell lung cancer (NSCLC), offering a survival advantage over traditional radiotherapy(Almudena,2014).

Diagnostics have considerably evolved in the last 20 years and positron emission tomography (PET) scan, endobronchial ultrasound (EBUS), and transesophageal ultrasound (EUS) are now part of an accurate preoperative assessment of potentially operable patients in most referral centers. An explosion of new targets have been observed in the last decade, several of which having already led to the development of new targeted agents (EGFR, ELM4–ALK in particular). The pace and sophistication with which new oncological drivers of lung cancer are currently identified, particularly for adenocarcinoma and the speed with which corresponding investigational treatments are subsequently developed, are steadily outgrowing the inflexible machinery of traditional clinical trials.

The greatest challenge of the coming years will be to use and combine all these new techniques and therapeutic modalities, mostly focused on the tumor at the moment, in each individual patient. This will require an enormous effort of multidisciplinary approach for each patient, taking into account a lot of clinical and biological parameters in addition to more and more genetic characteristics which are presently ignored in almost all cases (Thierry, 2011).

1.3 Objective

To find out an easier treatment and early detection of the disease has to be made sure. And for that the screening system must give proper result as it is about the disease certain stage in certain tissue. If detection is done early and in right time it would be easier to treat the disease as the effect of the disease would still be in starting stage giving opportunity to find a cure before harming much.

Biomarkers would be ideal molecules to work with. Biomarkers are being used for both detection of the disease along with finding out its treatment. Biomarker testing is the first step in personalizing cancer treatment because it helps doctors know more about an individual person's tumor. The results of biomarker testing can help doctors predict the likelihood that cancer will return (recurrence) or the response to treatment. This means that better and more informed decisions about treatment options that are best for patient can be understood.

Biomarkers have many potential applications in oncology, including risk assessment, screening, differential diagnosis, determination of prognosis, prediction of response to treatment, and monitoring of progression of disease. Because of the critical role that biomarkers play at all stages of disease, it is important that they undergo rigorous evaluation, including analytical validation, clinical validation, and assessment of clinical utility, prior to incorporation into routine clinical care.(N ,2012)

Accurate information's are needed for every single biomarker molecules. This project is designed to learn more about the biomarkers for the benefit of further studies. 14 proteins are selected based on their biomarker value for lung cancer. Hopefully learning details about the biomarkers will come out handy for detecting the disease and find out the treatment for lung cancer.

1.4 Biomarker

Biological markers (biomarkers) have been defined by Hulka and colleagues as "cellular, biochemical or molecular alterations that are measurable in biological media such as human tissues, cells, or fluids." More recently, the definition has been broadened to include biological characteristics that can be objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacological responses to a therapeutic intervention. In practice, biomarkers include tools and technologies that can aid in understanding the prediction, cause, diagnosis, progression, regression, or outcome of treatment of disease (Richard,2014).

According to the National Cancer Institute, a biomarker is "a biological molecule found in blood, other body fluids, or tissues that is a sign of a normal or abnormal process, or of a condition or disease,"(NCI) such as cancer. Biomarkers typically differentiate an

affected patient from a person without the disease. The alterations can be due to a number of factors, including germ line or somatic mutations, transcriptional changes, and posttranslational modifications. There is tremendous variety of biomarkers, which can include proteins (e.g., an enzyme or receptor), nucleic acids (e.g., a microRNA or other non-coding RNA), antibodies, and peptides, among other categories. A biomarker can also be a collection of alterations, such as gene expression, proteomic, and metabolomic signatures. Biomarkers can be detected in the circulation (whole blood, serum, or plasma) or excretions or secretions (stool, urine, sputum, or nipple discharge), and thus easily assessed non-invasively and serially, or can be tissue-derived, and require either biopsy or special imaging for evaluation.(N., 2012)

Biomarkers are three type-

- Diagnostic biomarker- Biomarkers that actually declares the presence or absence of the disease is known as diagnostic disease.
- Prognostic biomarker- Biomarkers that actually projects which treatment system is being helpful for the patient.
- Predictive biomarker- Biomarker that predicts about which patient might respond positively under which treatment system.

Properties of potential ideal biomarkers-

- An ideal biomarker has to be easy and safe to measure.
- It should be cost effective to follow up.
- It should be easily modifiable with the treatment.
- It must be consistent in all gender and ethnic groups.

The principles of biomarkers in disease have been applied to the detection, screening, diagnosis, treatment and monitoring of cancer. Traditionally, anti-cancer drugs were agents that killed both cancer cells and healthy cells. However, more targeted therapies have now been developed that can be directed to kill cancer cells only, while sparing healthy cells. The assessment of a typical biomarker in cancer helps in the development of therapies that can target the biomarker. This can minimize the risk of toxicity and reduce the cost of treatment. In cancer research, genetic studies are valuable because genetic abnormalities so often underlie the development of cancer.

1.4.1 Protein as a Biomarker

Protein biomarkers include substances that are either produced by cancer cells themselves or by other cells in response to cancer. Most protein biomarkers related to cancer are used to monitor response and/or detect recurrence or progression during follow-up after

treatment. Some biomarkers are used to predict the outcome or for prognosis. The structure of protein is given below.

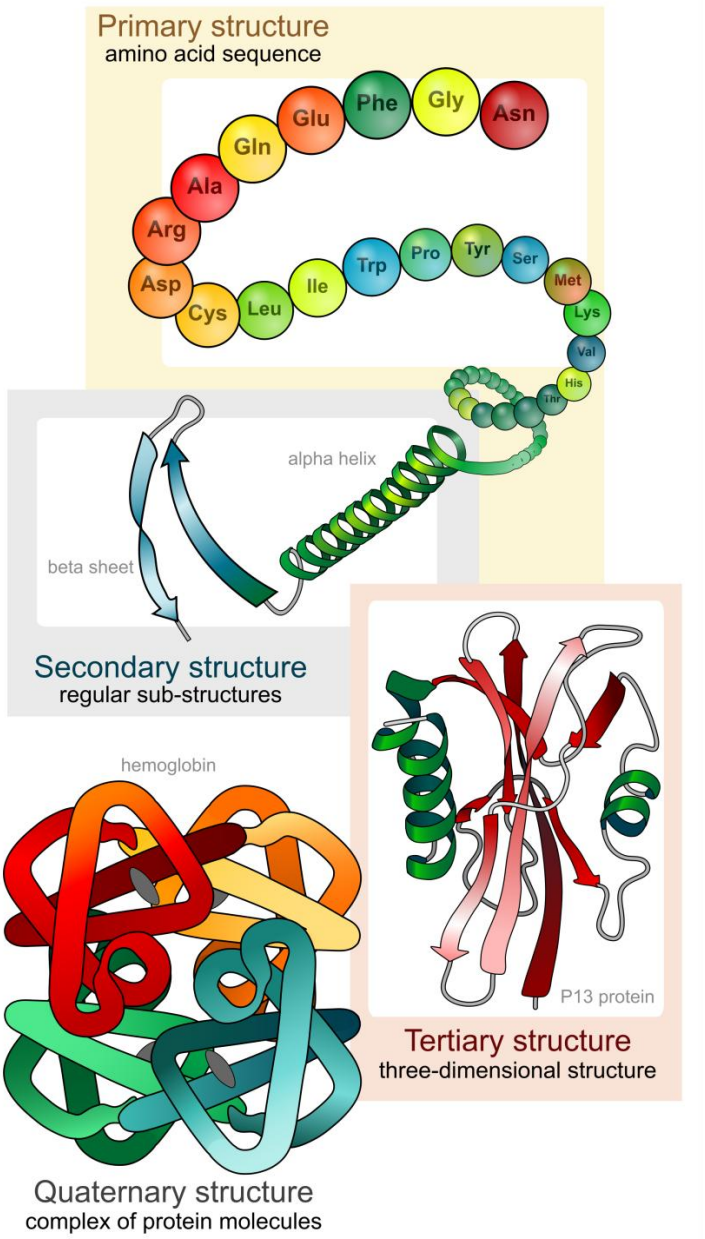

Figure 1.4: Structure of protein molecule.

Structural features of proteins are usually described at four levels of complexity:

- Primary structure: the linear arrangement of amino acids in a protein and the location of covalent linkages such as disulfide bonds between amino acids.

- Secondary structure: areas of folding or coiling within a protein; examples include alpha helices and pleated sheets, which are stabilized by hydrogen bonding.

- Tertiary structure: the final three-dimensional structure of a protein, which results from a large number of non-covalent interactions between amino acids.

- Quaternary structure: non-covalent interactions that bind multiple polypeptides into a single, larger protein. Hemoglobin has quaternary structure due to association of two alpha globin and two beta globin polyproteins.

1.4.2. Selected protein biomarkers

Protein are easier to study since they can be easily traced & evaluated so they are the potential best biomarker. Here in this study 14 protein biomarkers are selected for this thesis project. They are discussed below-

- MDK**:** Stands for Midkine protein encoded by MDK gene, Also known as neurite growth-promoting factor 2 (NEGF2). A basic heparin-binding growth factor & nonglyc-osylated protein. It promotes cellular transformation, angiogenesis and metastasis. The level of MK elevates from the early stage of lung cancer. .(Charles,2015)

- MMP2: 72 kDa type IV collagenase also known as matrix metalloproteinase-2 (MMP-2) and gelatinase A is an enzyme that in humans is encoded by the MMP2 gene. The MMP2 gene is located on chromosome 16 at position 12.2. This protein degrades extracellular matrix, associated with tissue invasion. Cell induced angiogenesis, tumor growth and metastasis. .(Charles,2015)

- TFPI: TFPI (Tissue Factor Pathway Inhibitor) is a Protein Coding gene. The encoded protein is glycosylated and predominantly found in the vascular endothelium and plasma in both free forms and complexed with plasma lipoproteins. It regulates the tissue factor (TF)-dependent pathway of blood coagulation. .(Charles,2015)

- TIMP-1: TIMP metallopeptidase inhibitor 1 is a tissue inhibitor of metalloproteinases. This glycoprotein is a member of the TIMP family. It inactivates metaoproteanase by binding to zinc co-factor, promotes proliferation

& inhibits apoptosis. Increased expression of TIMP-1 has been found to be associated with lung cancer. .(Charles,2015)

- OPN: Stands for osteopontin. Recruits immune cell mediation, wound healing & tissue remodeling. OPN has diverse biologic function in cell adhesion, migration & invasion. Over- expression of level is detected in lung cancer patient. .(Charles,2015)

- BIRC6: Baculoviral Inhibitor of apoptosis repeat-containing protein 6 is a protein that in humans is encoded by the BIRC6 gene. This protein confers apoptosis resistance to cancer cells. Elevated expression of BIRC6 protein in non-small-cell lung cancers is associated with cancer recurrence and chemo resistance..(Xin, 2013)

- CEA: It is a oncoferal glycoprotein, stands for carcinogenic antigen. Found in blood of lung cancer patient (not in healthy blood). CEA plays important role in cell adhesion and intercellular signaling. Increasing level of CEA in blood directs to the possibility of lung cancer.(Charles,2015)

- CDK4: Cyclin-dependent kinase 4 also known as cell division protein kinase 4 is a member of the cyclin-dependent kinase family & the protein encoded by this gene is a member of the Ser/Thr protein kinase family. The expression level of CDK4 protein significantly increases in lung cancer tissues compared to normal tissues. CDK4 mediates cell cycle progression by regulating the expression of p21 expression in lung cancer.(Aibing.2011)

- HSP90: Hsp90 (heat shock protein 90) is a chaperone protein. It assisting other cellular proteins to fold properly, and stabilizes them against oxidative and heat stress, as well as helping with protein degradation. HSP90 can act as a "protector" of unstable protein by-products of DNA mutations. Hsp90 is an important target in cancer therapy. Increased HSP90 expression has been linked to worse prognosis in patients with non–small cell lung cancer.(Edward,2013)

- HSPA5: Binding immunoglobulin protein (BiP) also known as 78 kDa glucose-regulated protein (GRP-78) or heat shock 70 kDa protein 5 (HSPA5) is a protein that in humans is encoded by the HSPA5 gene. Molecular chaperone HSPA5 is a key survival factor in development and cancer. HSPA5 may also be

important for tumor metastasis. It is elevated in metastatic cancer cell lines, lymph node metastasis. HSPA5 inhibits tumor cell invasion and growth and metastasis. (Xianliang, 2015)

- EGFR: Epidermal growth factor receptor (EGFR), a receptor tyrosine kinase. EGFR produces this protein. Receptors play an important role in tumor cell survival and activated phosphorylated EGFR results in the phosphorylation of downstream proteins that cause cell proliferation, invasion, metastasis, and inhibition of apoptosis. It's is frequently over expressed in non small cell lung cancer.(Lee, 2006)

- ACTN4: Alpha-actinin 4 protein is a member of the cytoskeletal protein family; it binds to actin filaments to preserve cytoskeletal structure and cell morphology. ACTN4 is expressed in no muscle cells and is commonly associated with focal adhesion contacts and migrating cells. In cancer tissues the expression level increases. (Ming,2015)

CHAPTER 2:

METHOD

2.1 Protein

2.1.1 Structure Prediction

When it comes to study protein structure they are pretty complicated for detailed study. For the structure prediction in this thesis project homology modeling was followed. The steps that were taken to find out the structure is given below.

1. Sequence were taken from a database. Uniprotkb was used for FASTA format.
2. Blast was done for best suited template collection.
3. Alignment was checked with the same sequence and their templates using clustal omega.
4. Alignment was given in Swiss model workspace(homology modeling website)

The databases and softwares used for this project are described below.

2.1.1.1 UniProtKB

The UniProt Knowledge 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 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/

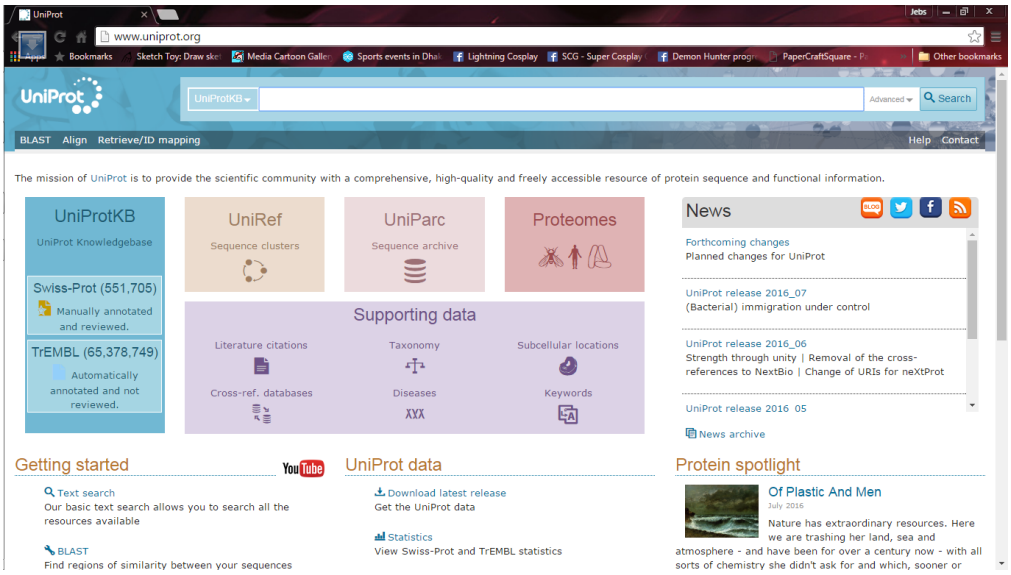

Figure 2.1: UniProtKB homepage.

2.1.1.2 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/

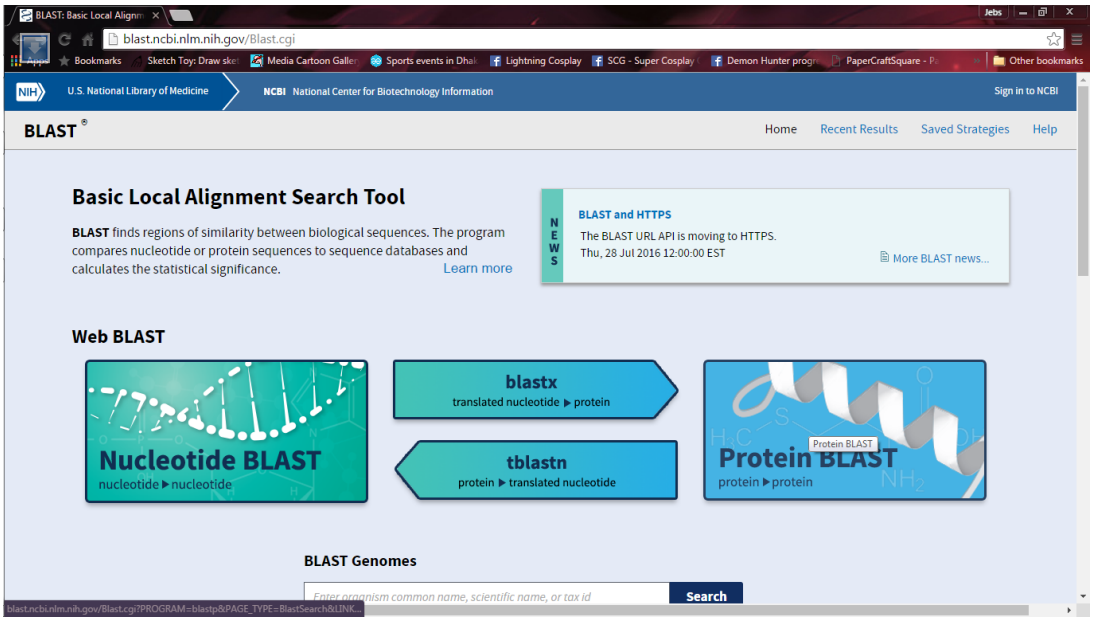

Figure 2.2 : BLAST homepage.

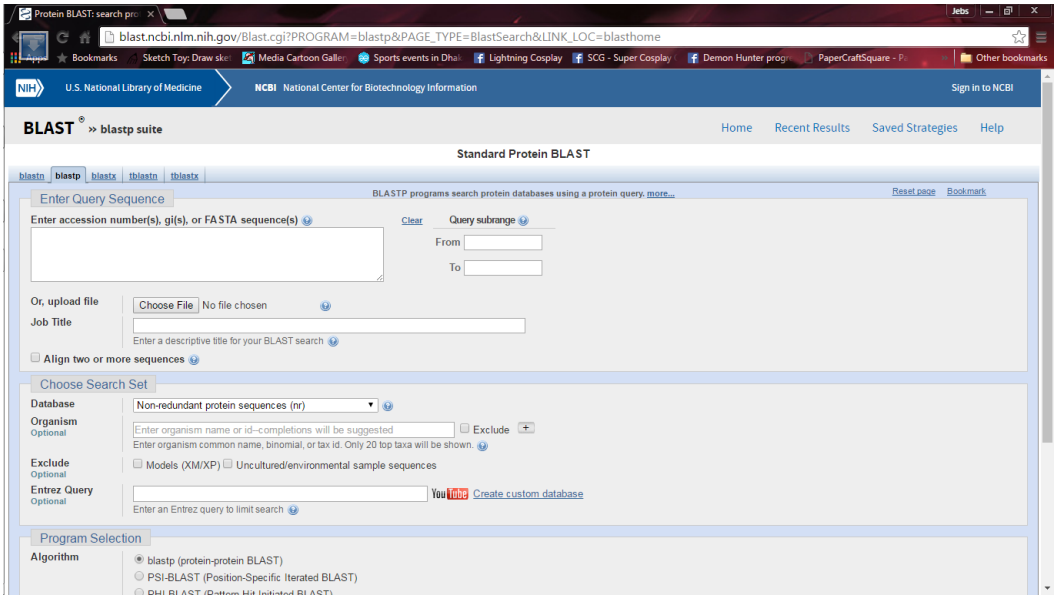

Figure 2.3: BlastP query entry page.

2.1.1.3 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/

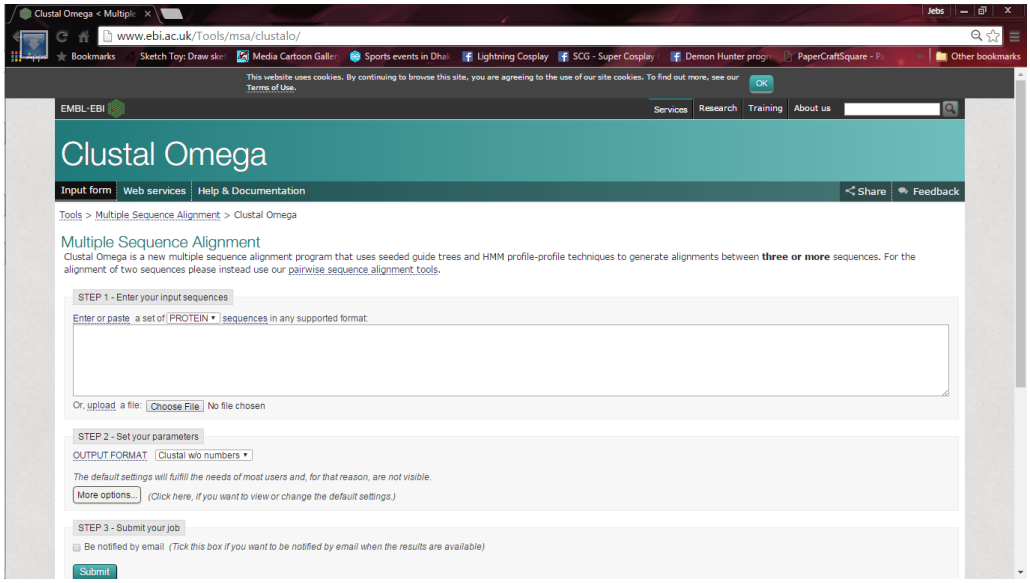

Figure 2.4: Clustal Omega homepage.

2.1.1.4 SWISS MODEL

SWISS MODEL Workspace is a tool that provides with an environment having automated comparative homology modeling platform. SWISS-MODEL was first stated working in 1993 by Manuel Peitsch Nicolas Guex and Torsten Schwede, 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 modeling 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/

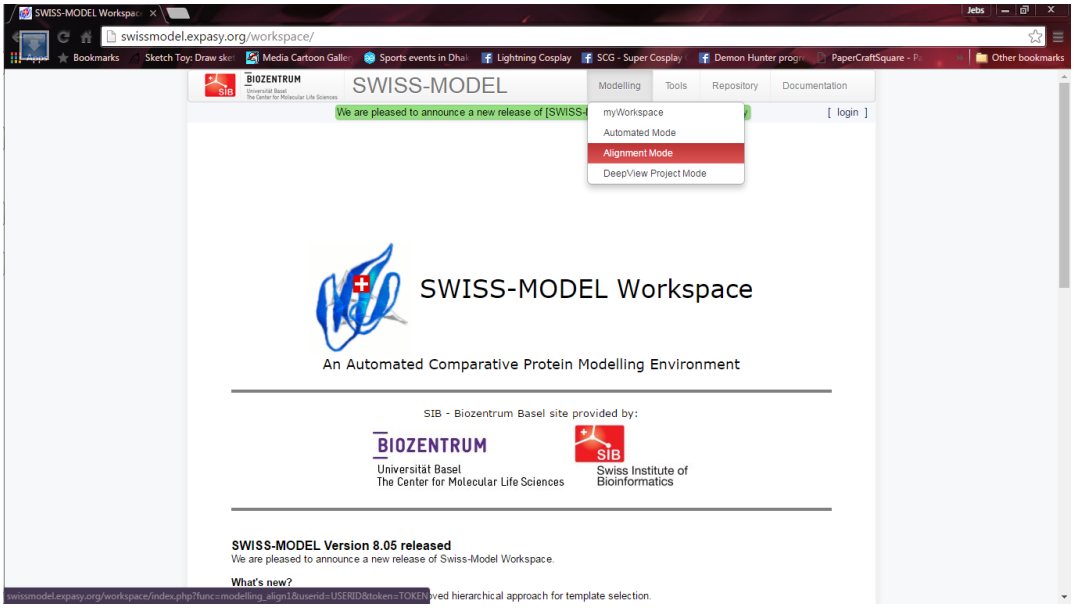

Figure 2.5: SWISS-MODEL Workspace home page.

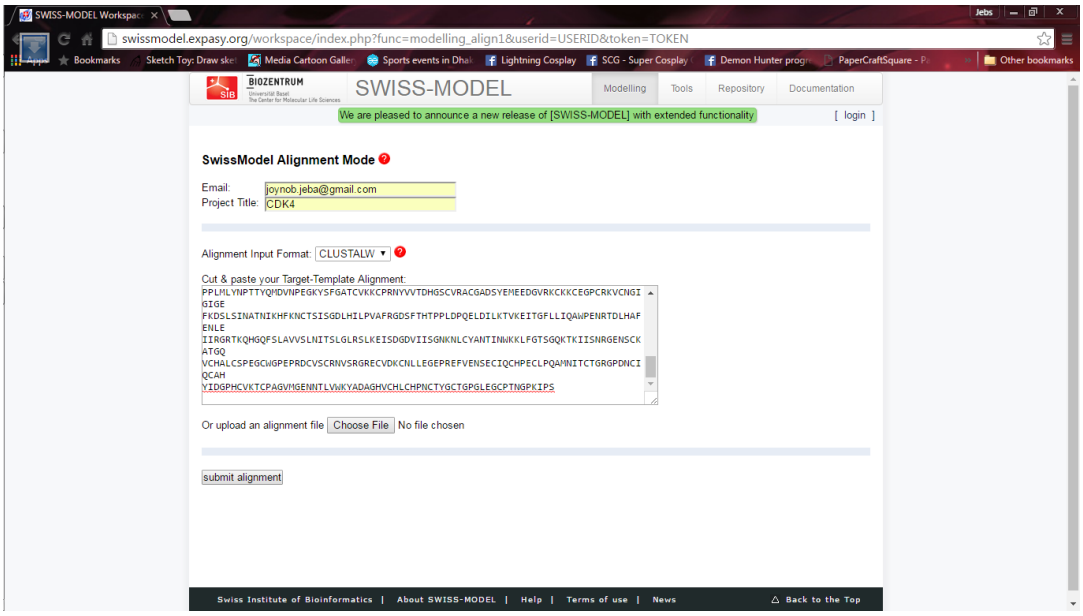

Figure 2.6: SWISS- MODEL Workspace alignment submitting page.

2.1.2 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:

2.1.2.1 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/

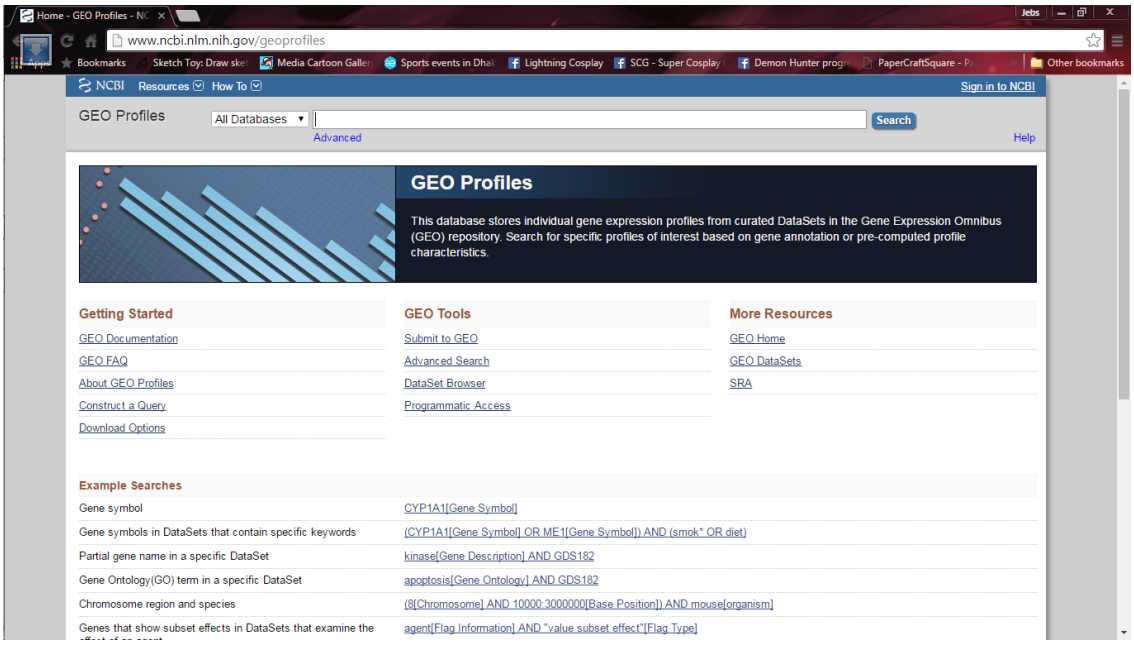

Figure 2.7: NCBI home page.

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 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 modeling. 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 by Timothy 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 by Columbia 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

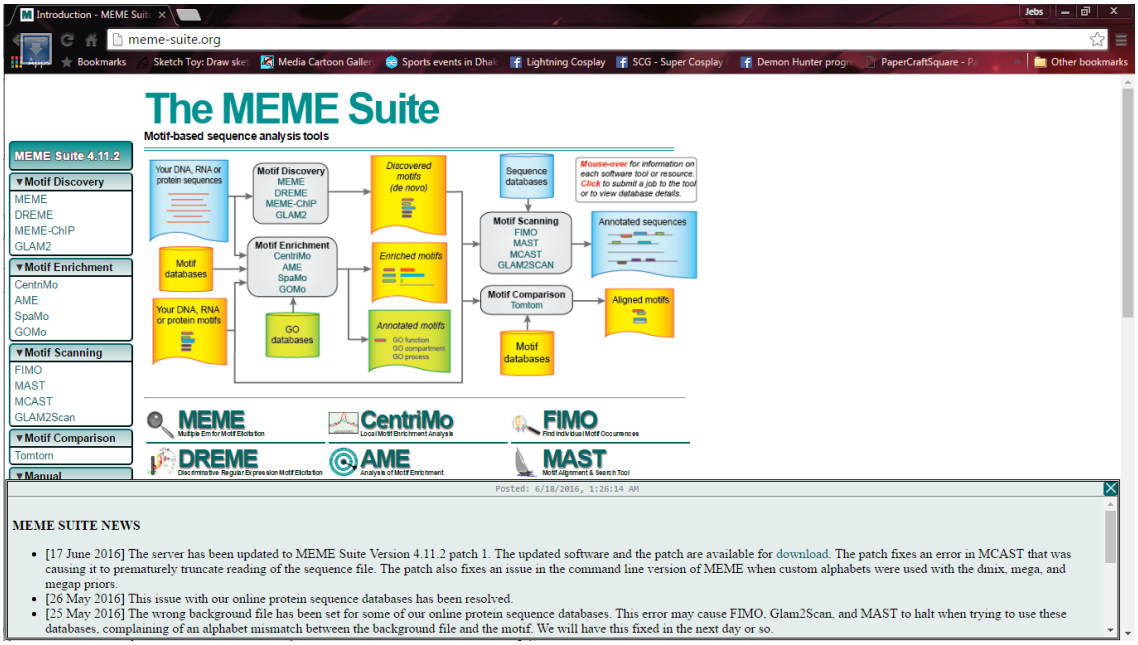

Figure 2.8: MEME Suite home page.

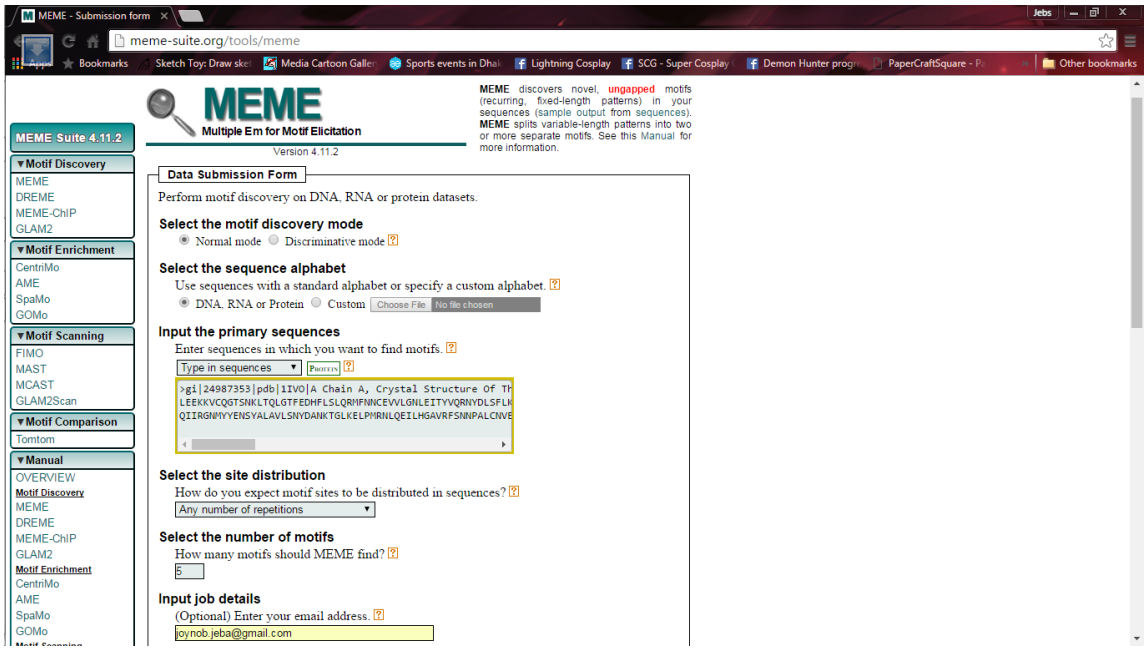

Figure 2.9: MEME data submission form.

2.1.3 Finding Expression Level

To find out the expression level of these proteins in different conditions of lung cancer GEO tool of NCBI was used. Just the name of the protein name and little clue “lung 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 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/

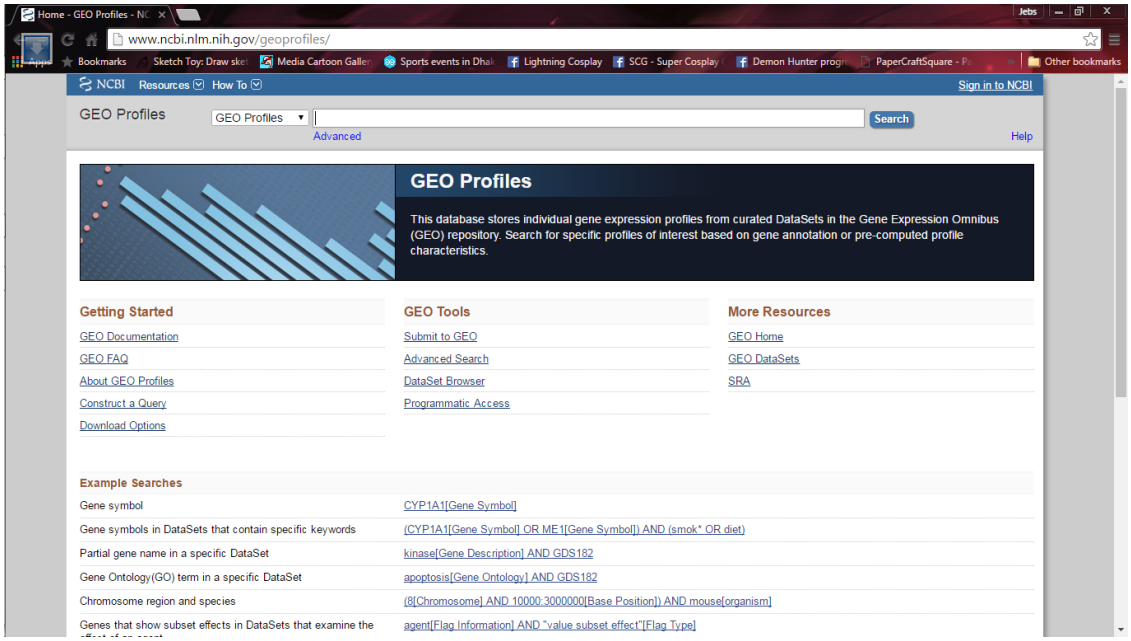

Figure 2.10: GEO Profile home page.

CHAPTER: 3

RESULT

3.1 Protein

Protein is one of the most promising potential biomarkers for lung cancer. For this study, lung cancer specific selected 14 protein's 3D structures, sequence motifs and expression level have been checked. The names of the selected proteins are given below-

- MDK
- MMP2
- TFPI
- TIMP-1
- OPN
- BIRC6
- CEA
- CDK4
- HSPA5
- HSP90Alpha
- EFGR
- ACTN4

Results are given below:

3.1.1 Structure

The 3d structures of these 14 proteins were predicted. There are different classifications of protein structure prediction. For this study homology modeling was done. Tertiary level of structure of a protein is called homology modeling. Its mainly comparative modeling system. In this modeling homologous proteins are used as templates. SWISS MODEL Workspace was used for structure prediction. To determine the structural motif along with their site directed mutagenesis homology modeling can be done which is what makes them probable candidate for lung cancer biomarkers.

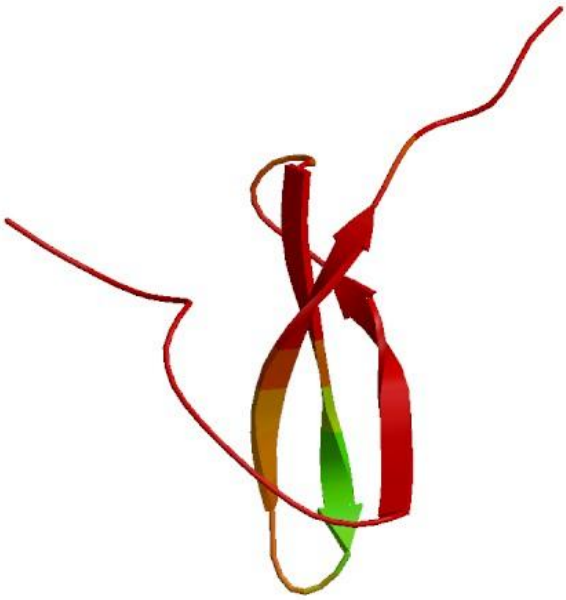

Figure3.1: Homology model of Midkine protein

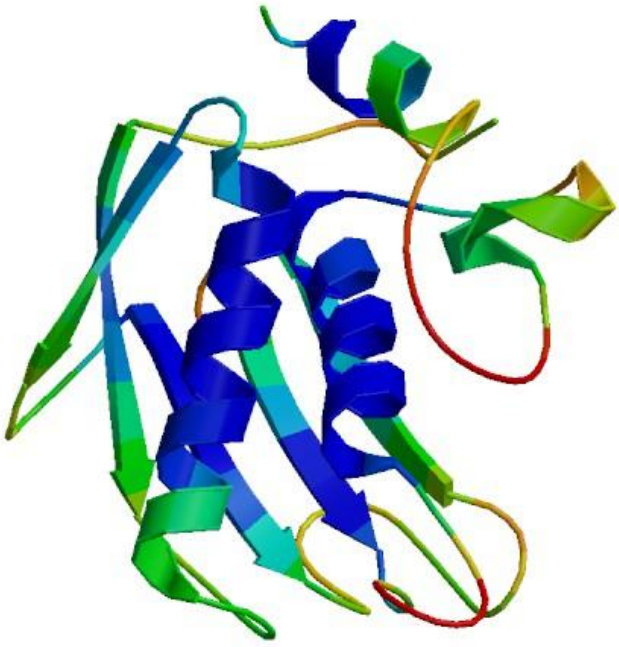

Figure 3.2: Homology model of MMP2 protein

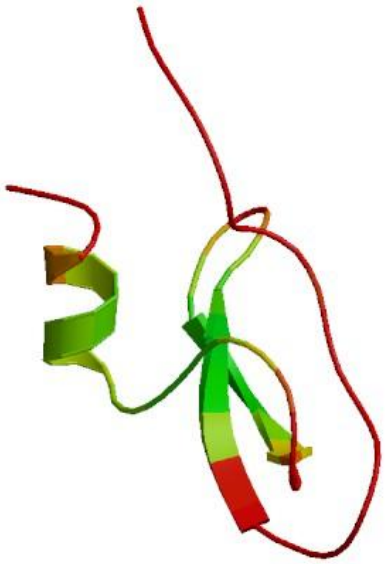

Figure 3.3: Homology modeling of TFPI protein

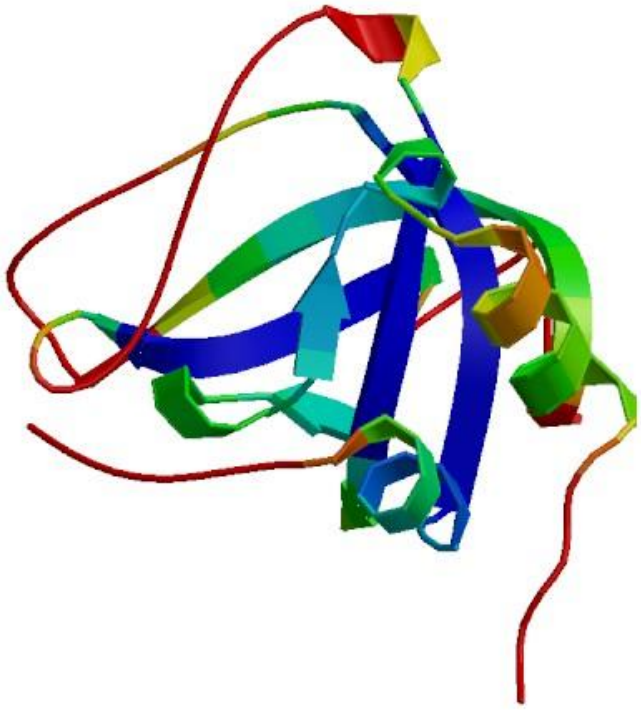

Figure 3.4: Homology modeling of TIMP-1 protein

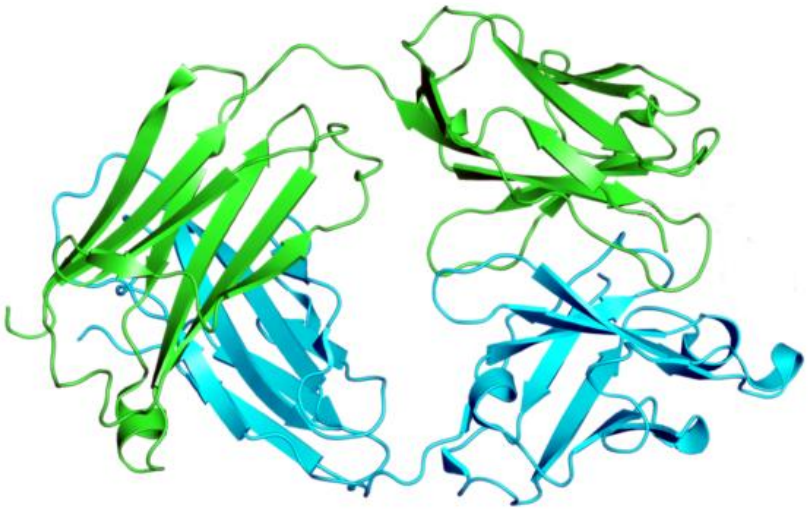

Figure 3.5: Homology modeling of OPN protein

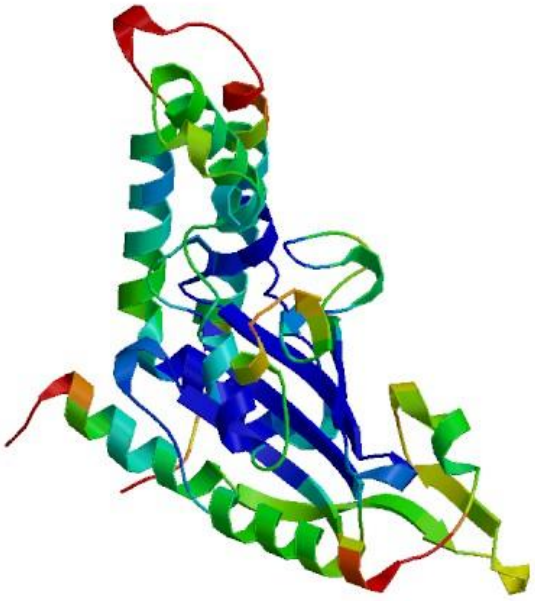

Figure 3.6: Homology modeling of BIRC6 protein

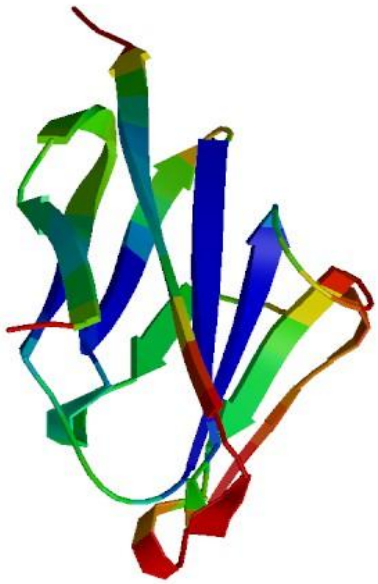

Figure 3.7: Homology model of CEA protein

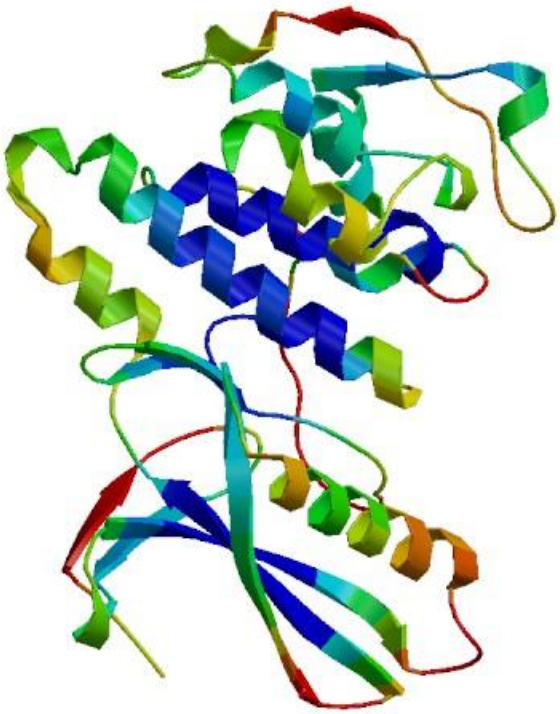

Figure3.8: Homology modeling of CDK4 protein

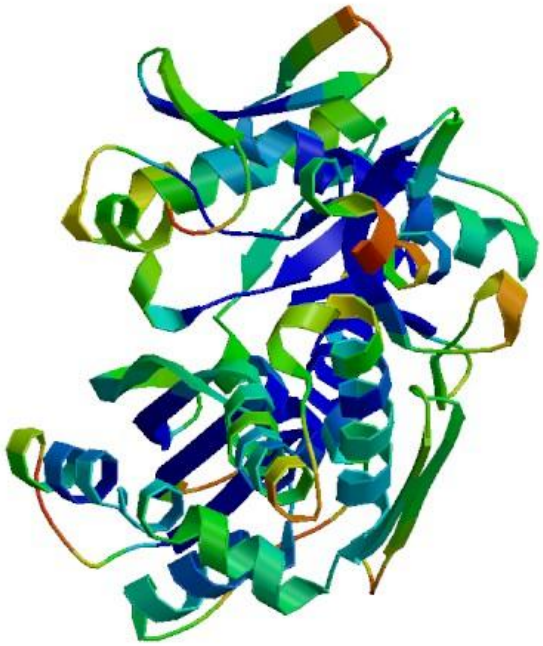

Figure3.9: Homology modeling of HSPA5 protein

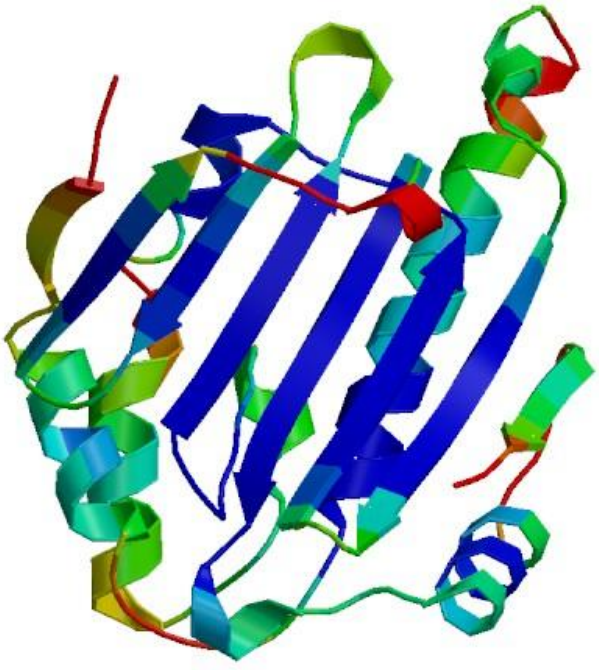

Figure 3.10: Homology modeling of HSP90 protein

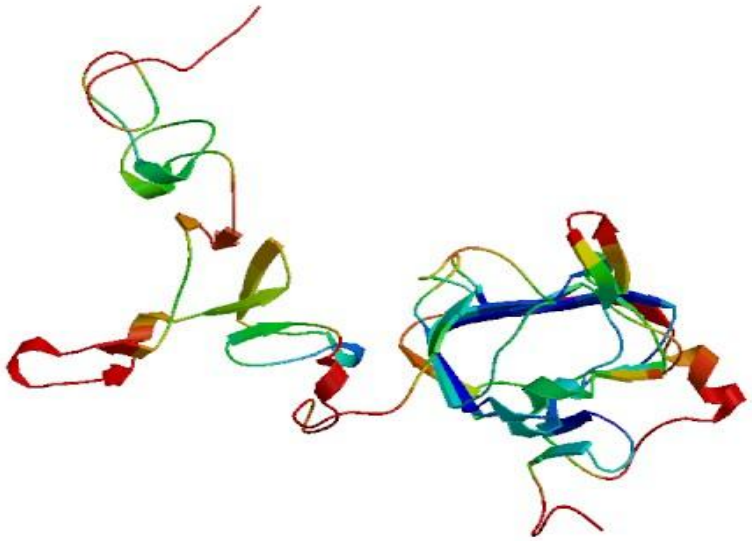

Figure3.11: Homology modeling of EFGR protein

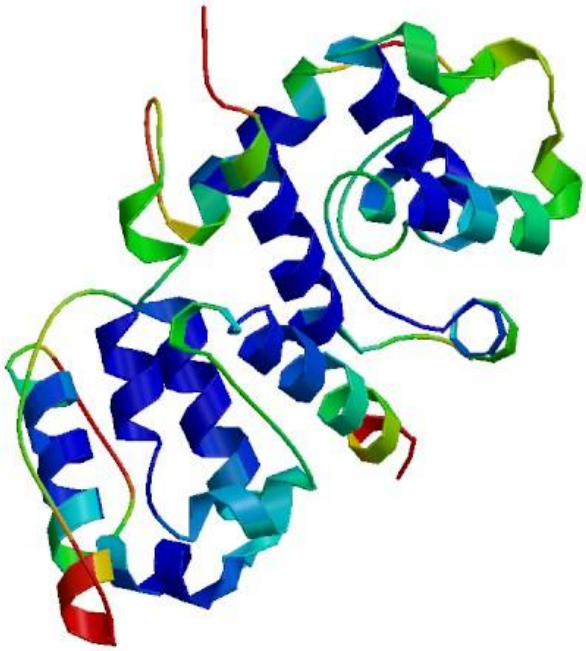

Figure 3.12: Homology modeling of ACTN4 protein

3.2.1 SEQUENCE MOTIF

A sequence motif is a nucleotide or amino-acid sequence pattern that is widespread and has, or is conjectured to have, a biological significance, functional value and has highly specific activity, sequence variation along with protein-protein interactions are also understandable by a sequence motif . Here, sequence motif was observed using MEME software. 5 motifs per proteins were taken.

3.2.1.1 MDK:

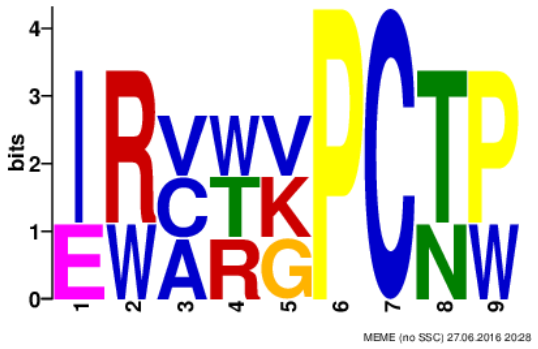

Figure 3.13: Motif 1 of MDK protein

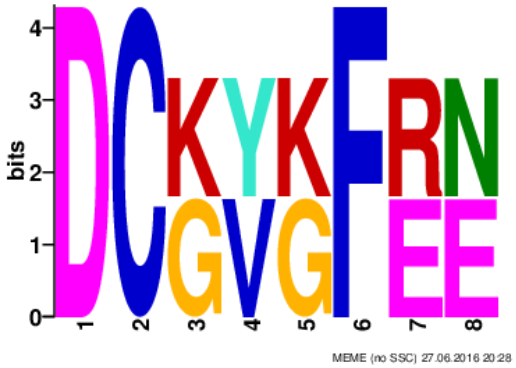

Figure 3.14: Motif 2 of MDK protein

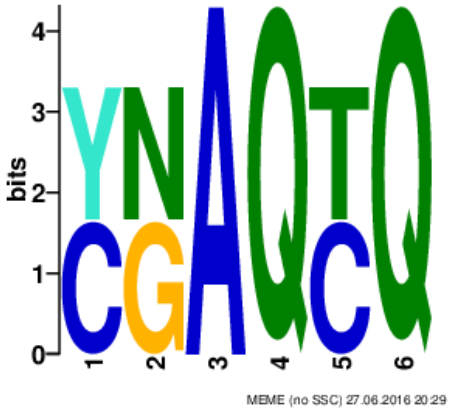

Figure 3.15: Motif 3 of MDK protein

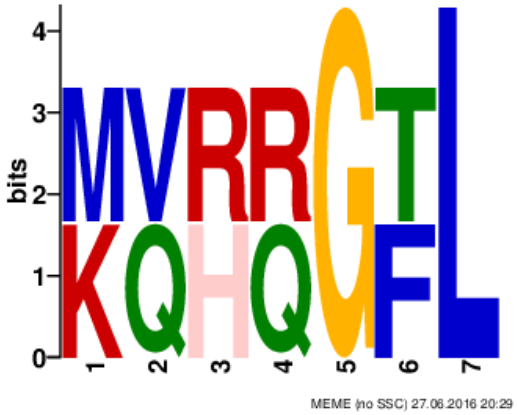

Figure 3.16: Motif 4 of MDK protein

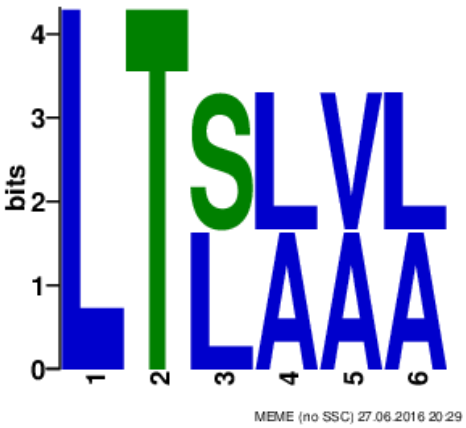

Figure 3.17: Motif 5 in MDK protein

3.2.1.2 MMP2:

Figure 3.18: Motif 1 in MMP2 protein

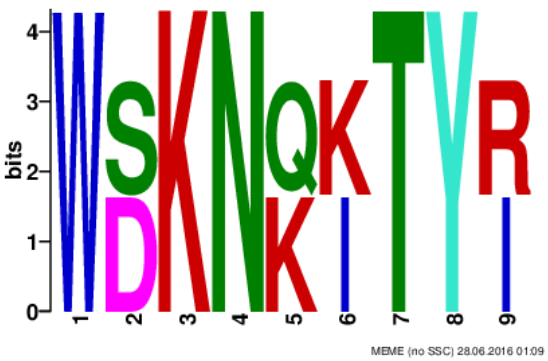

Figure 3.19: Motif 2 in MMP2 protein

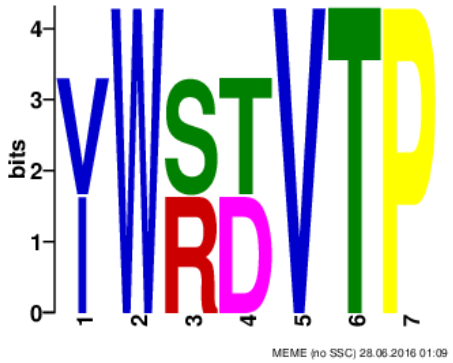

Figure 3.20: Motif 3 in MMP2 protein

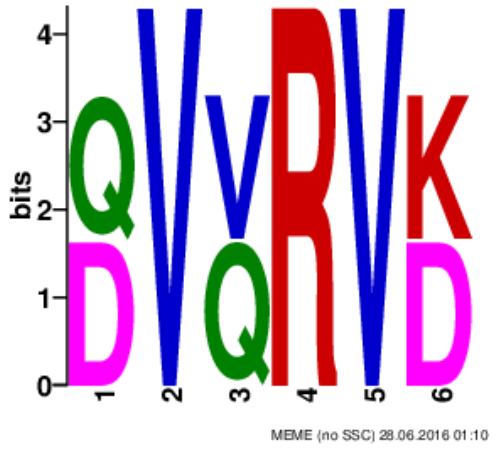

Figure 3.21: Motif 4 in MMP2 protein

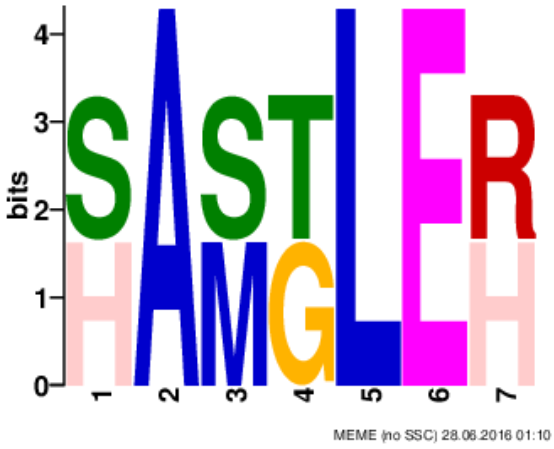

Figure 3.22: Motif 5 in MMP2 protein

3.2.1.3. TFPI

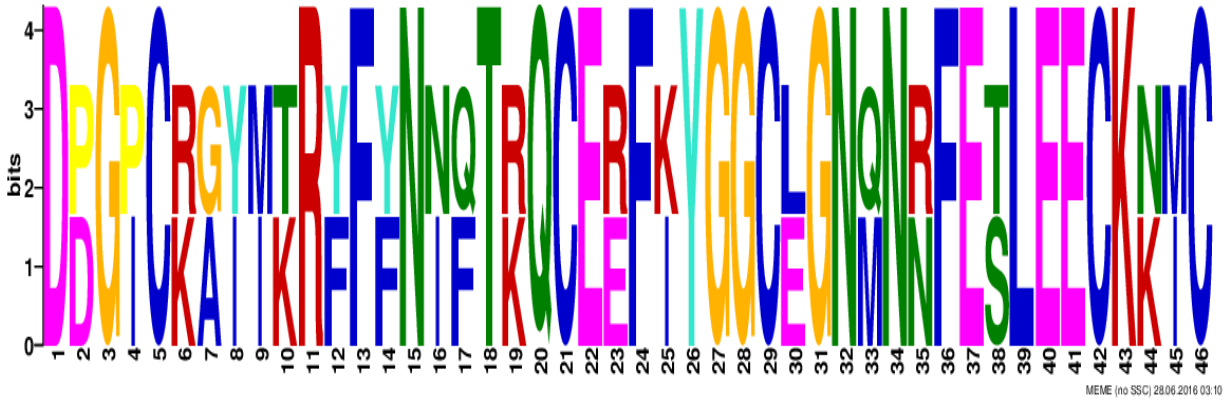

Figure 3.23: Motif 1 of TFPI protein

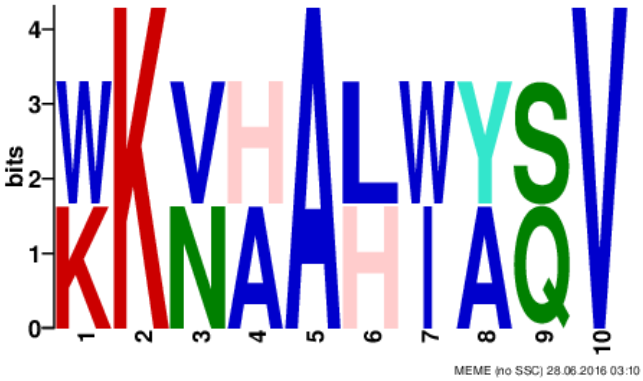

Figure 3.24: Motif 2 of TFPI protein

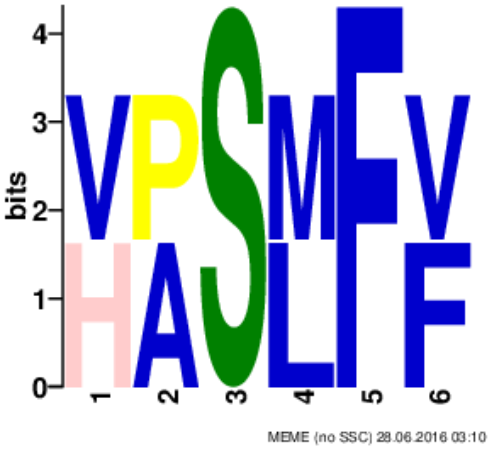

Figure 3.25: Motif 3 of TFPI protein

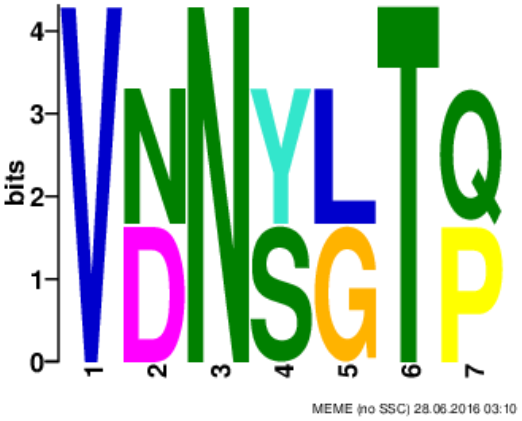

Figure 3.26: Motif 4 of TFPI protein

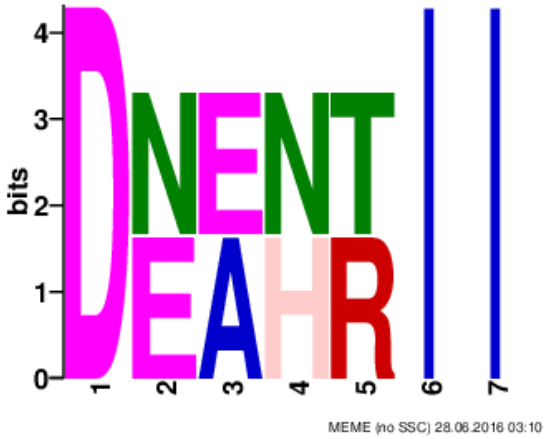

Figure 3.27: Motif 5 of TFPI protein

3.2.1.4TIMP-1

Figure 3.28: Motif 1 of TIMP-1 protein

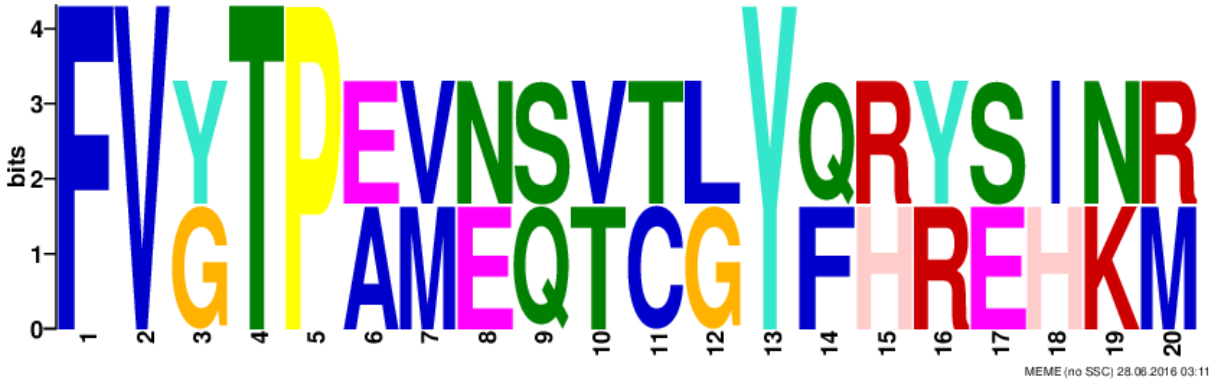

Figure 3.29: Motif 2 of TIMP-1 protein

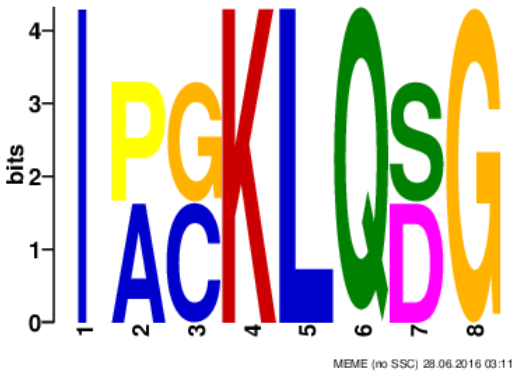

Figure 3.30: Motif 3 of TIMP-1 protein

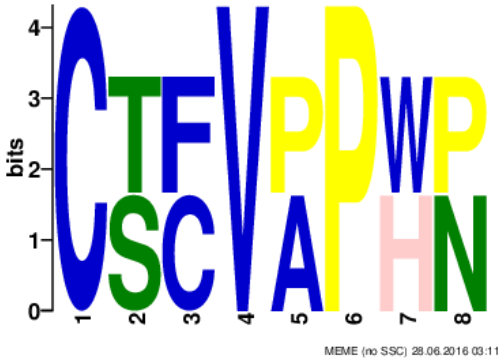

Figure 3.31: Motif 4 f TIMP-1 protein

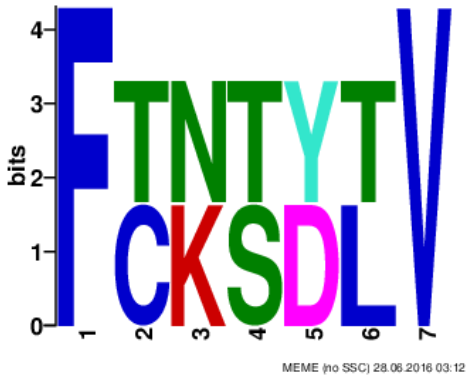

Figure 3.32: Motif 5 of TIMP-1 protein

3.2.1.5 OPN

Figure 3.33: Motif 1 of OPN protein

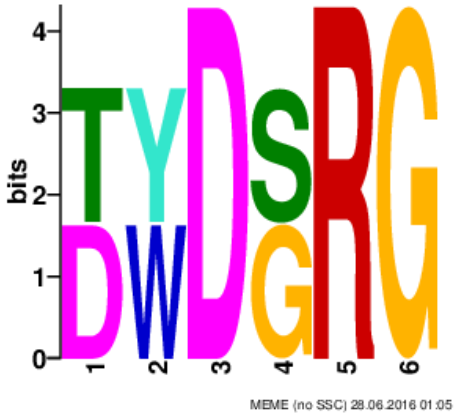

Figure 3.34: Motif 2 of OPN protein

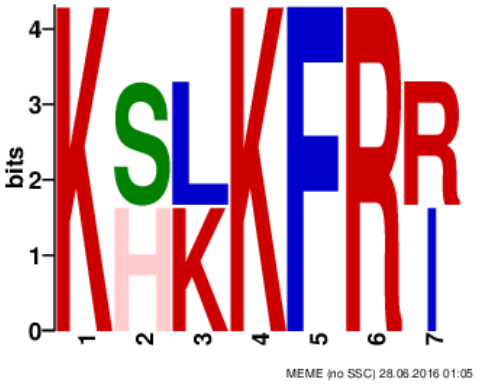

Figure 3.35: Motif 3 of OPN protein

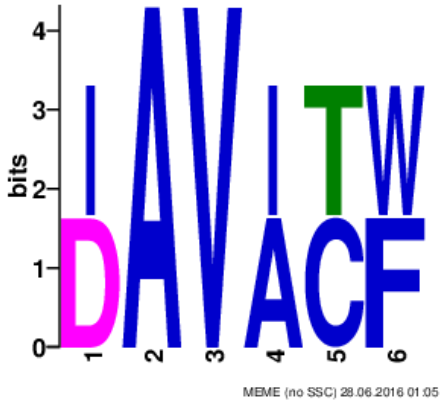

Figure 3.36: Motif 4 of OPN protein

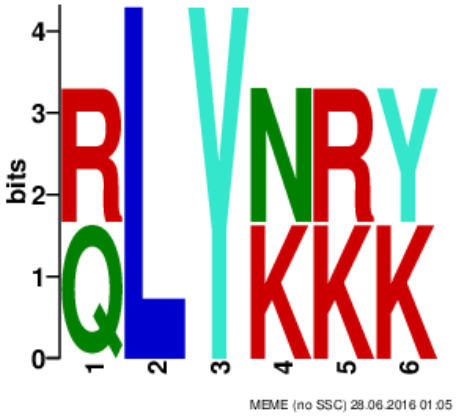

Figure 3.37: Motif 5 of OPN protein

3.2.1.6. BIRC6

Figure 3.38: Motif 1 of BIRC6 protein

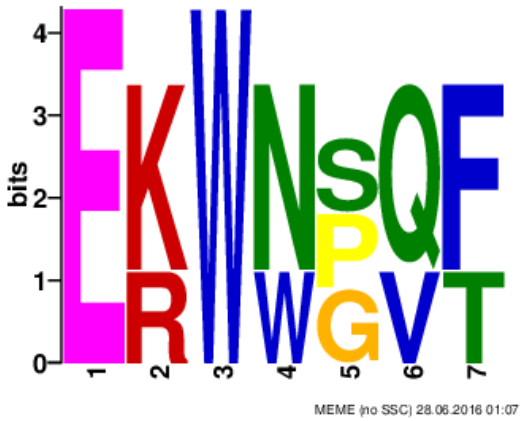

Figure 3.39: Motif 2 of BIRC6 protein

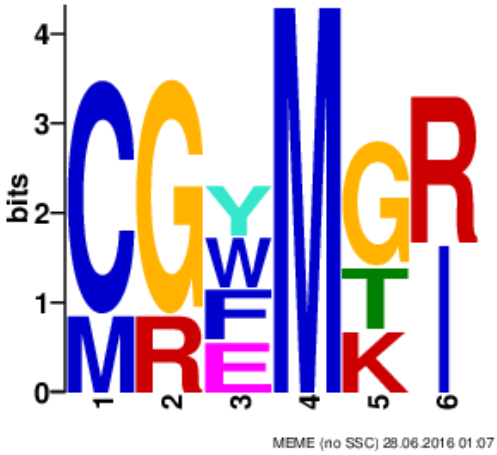

Figure 3.40: Motif 3 of BIRC6 protein

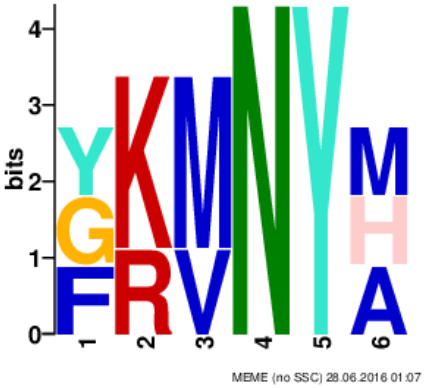

Figure 3.41: Motif 4 of BIRC6 protein

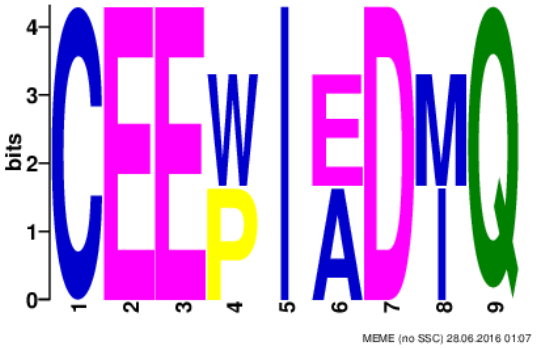

Figure 3.42: Motif 5 of BIRC6 protein

3.2.1.7 CEA:

Figure 3.43: Motif 1 in CEA protein

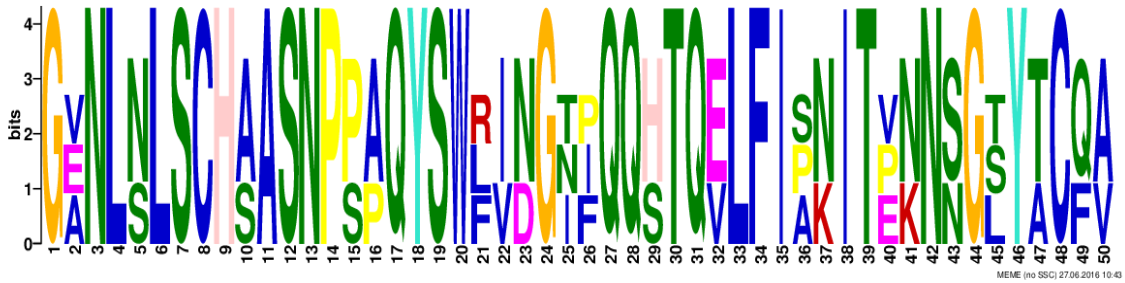

Figure 3.44: Motif 2 in CEA protein.

Figure 3.45: Motif 3 in CEA protein.

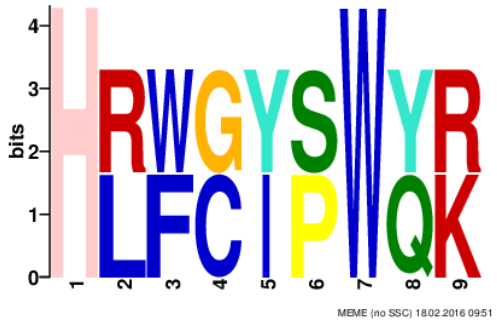

Figure 3.46: Motif 4 in CEA protein

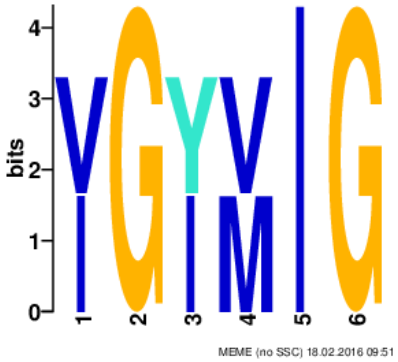

Figure 3.47: Motif 5 in CEA protein

3.2.1.8. CDK4

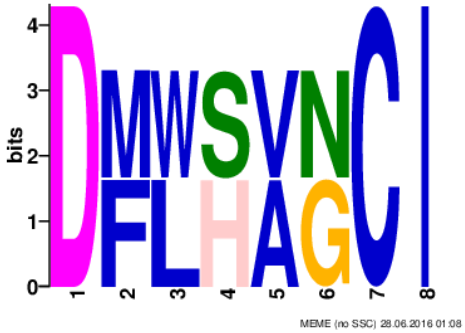

Figure 3.48: Motif 1 of CDK4 protein

Figure 3.49: motif 2 of CDK4 protein

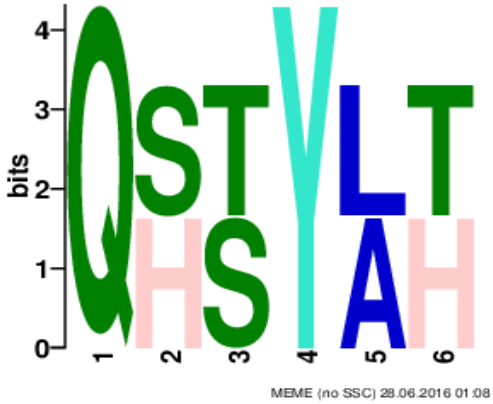

Figure 3.50: Motif 3 of CDK4 protein

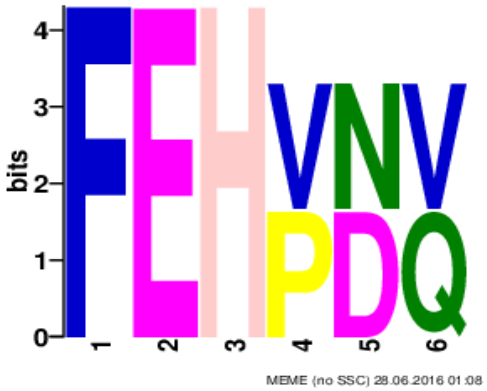

Figure 3.51: Motif 4 of CDK4 protein

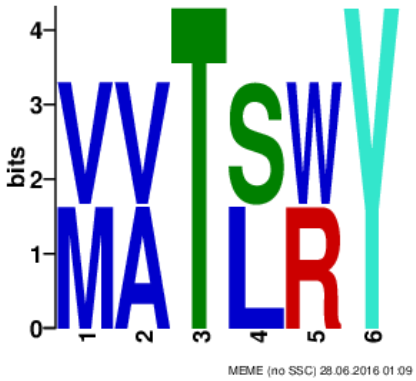

Figure 3.52: Motif 5 of CDK protein

3.2.1.9 HSPA5

Figure 3.53: Motif 1 of HSPA5 protein

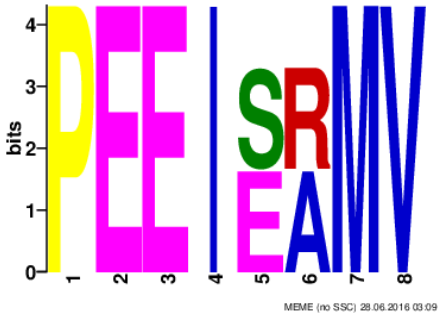

Figure 3.54: Motif 2 of HSPA5 protein

Figure 3.55: Motif 3 of HSPA5 protein

Figure 3.56: Motif 4 of HSPA5 protein

Figure 3.57: Motif 5 of HSPA5 protein

3.2.1.10.1. HSP90- alpha-isoform-1

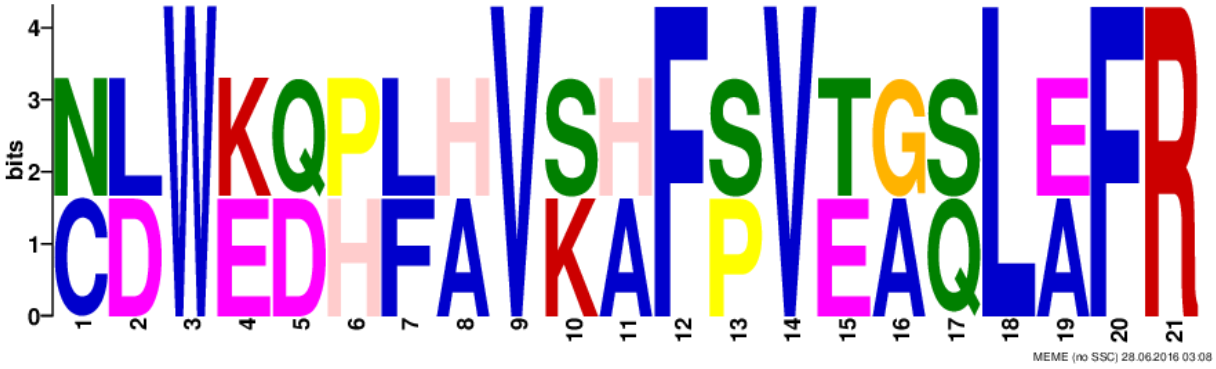

Figure 3.58: Motif 1 of HSP90- α -isoform-1

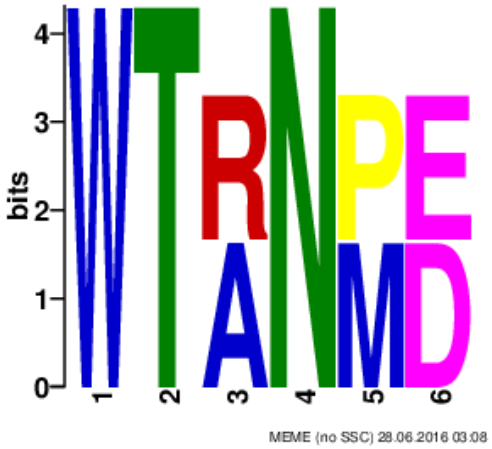

Figure 3.59: Motif 2 of HSP90- α -isoform1

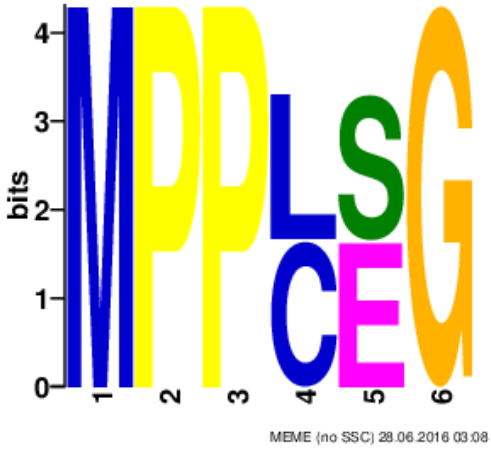

Figure3.60: Motif 3 of HSP90- α- isoform1

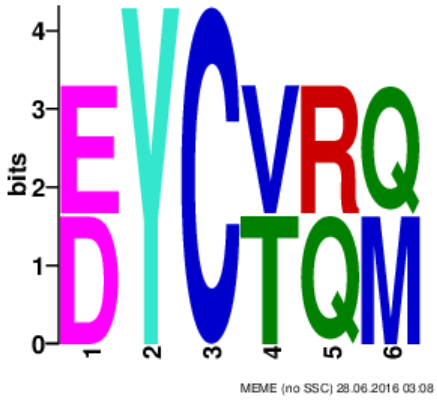

Figure 3.61: Motif 4of HSP90- α -isoform-1

Figure 3.62: Motif 5 of HSP90- α -isoform-1

3.2.1.10.2. HSP90- α -isoform-2

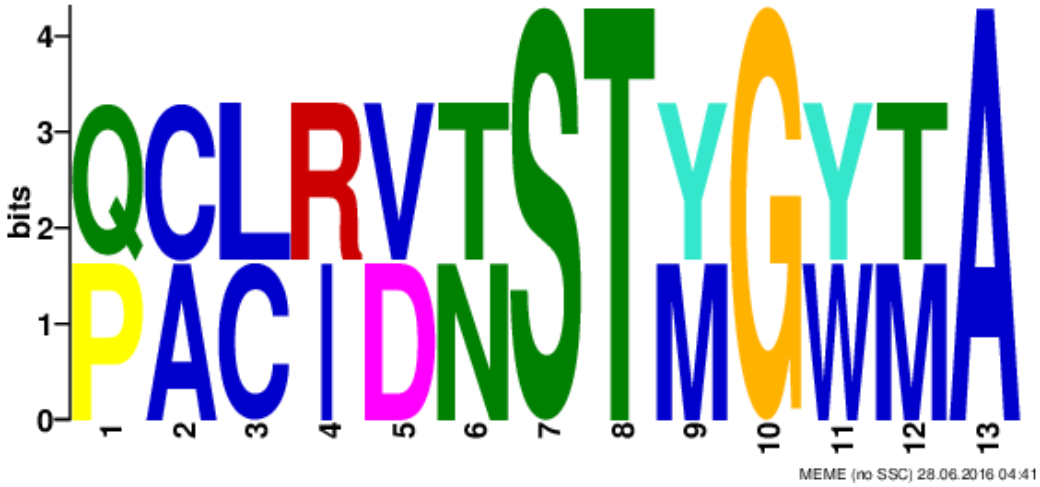

Figure 3.63: Motif 1 of HSP90- α -isoform-1

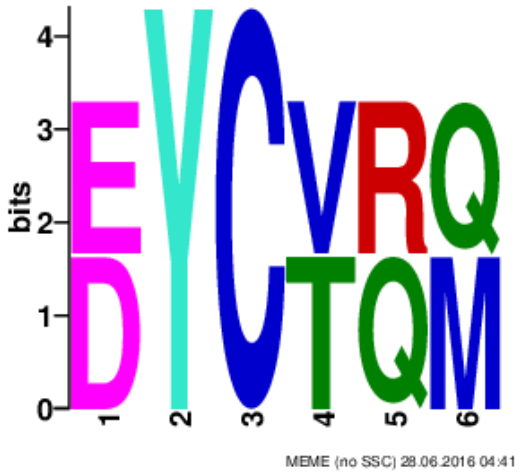

Figure 3.64: Motif 1 of HSP90- α -isoform-2

Figure 3.65: Motif 3 of HSP90- α -isoform-2

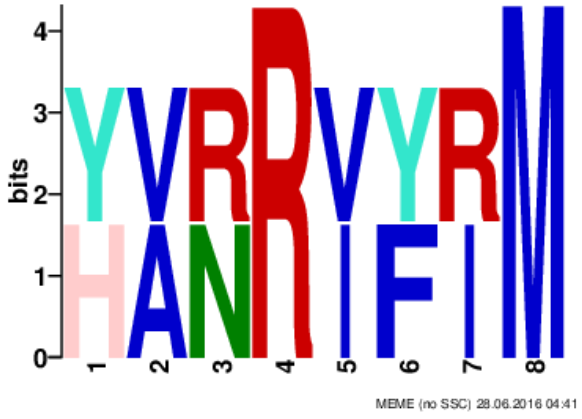

Figure 3.66: Motif 4 of HSP90- α -isoform-2

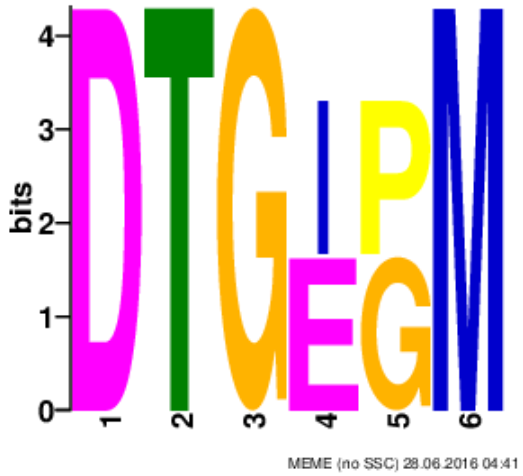

Figure 3.67: Motif 5 of HSP90- α -isoform-2

3.2.1.11 EGFR

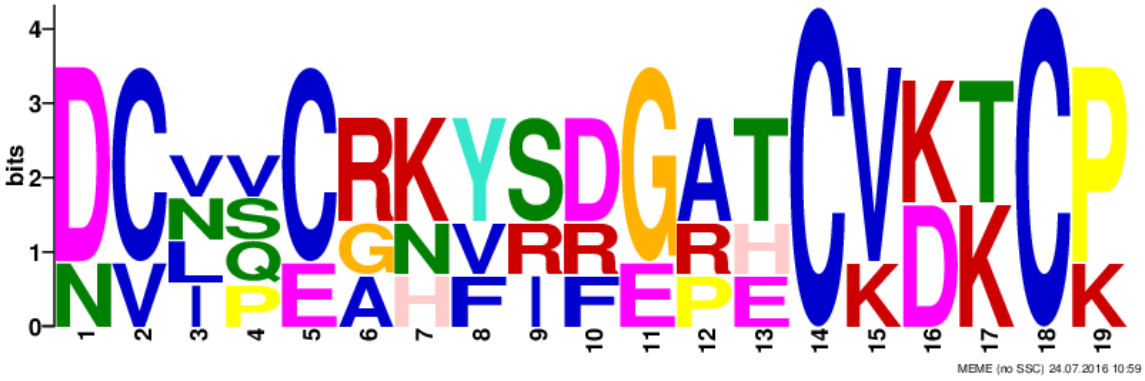

Figure 3.68: Motif 1 of EGFR protein

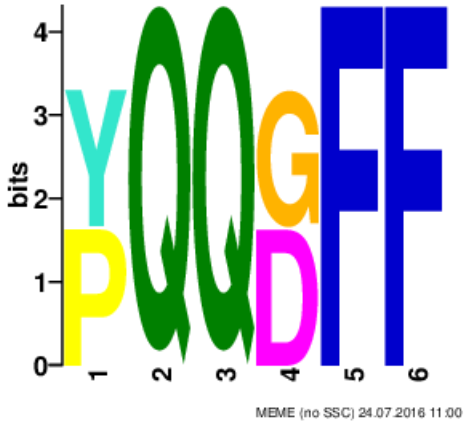

Figure 3.69: motif 2 of EGFR protein

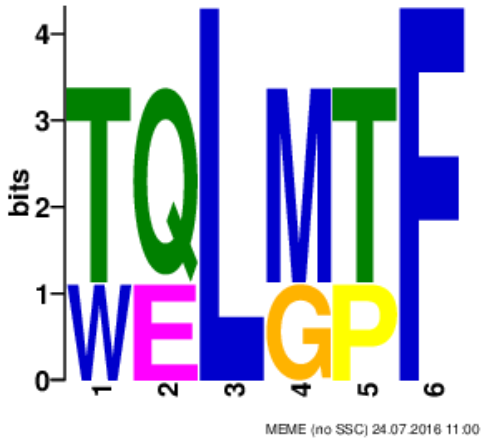

Figure 3.70: Motif 3 of EGFR protein

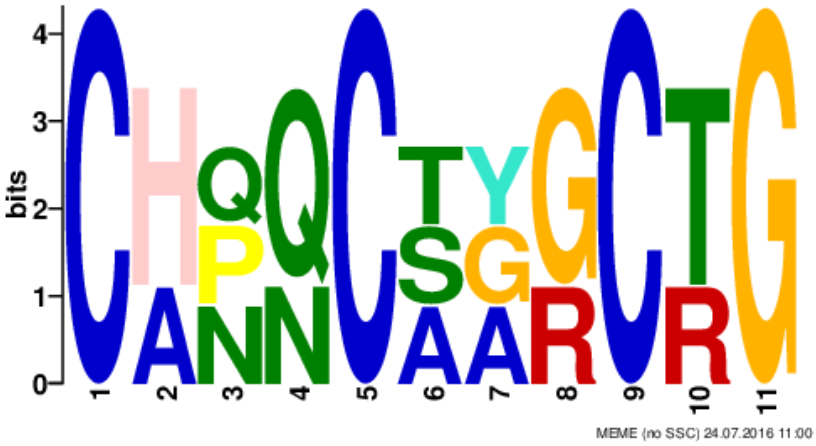

Figure 3.71: Motif 4 Of EGFR protein

Figure 3.72: Motif 5 of EGFR protein

3.2.1.12 ACTN4

Figure 3.73: Motif 1 of ACTN4 protein

Figure 3.74: Motif 2 of ACTN protein

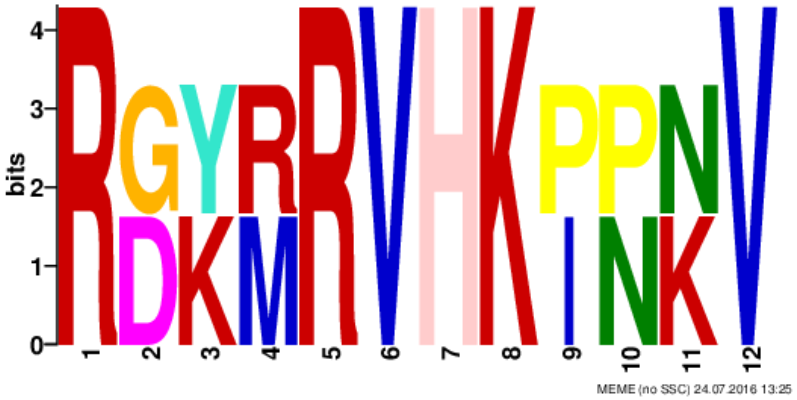

Figure 3.75: Motif 3 of ACTN4 protein

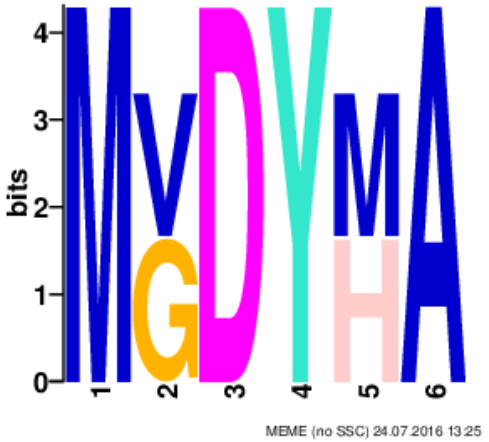

Figure 3.76: Motif 4 of ACTN4 protein

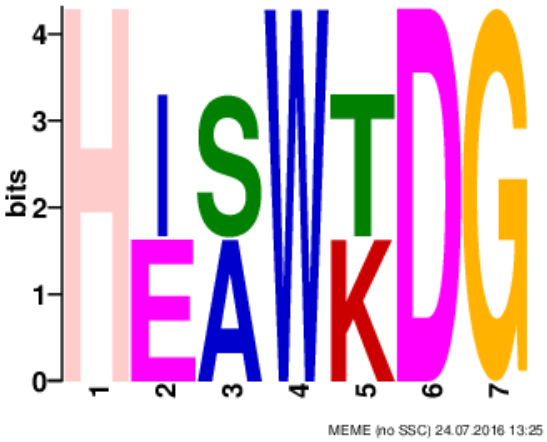

Figure 3.77: Motif 5 of ACTN4 protein

3.2.3 Expression Level:

Every protein that are identified in lung cancer shows different expression level in different stages of cancer. Each protein is very unique according to your structure and expression level which is why they are very unique and ideal to be used as biomarker molecules. To detect the difference of protein expression level between lung cancer and normal tissue for discussing their expression level, datas of different proteins expressed as a result of experiments done before were taken from GEO Profiles. Geo profiles have already stored information for each protein. No practical experiment was done for this part of the project.

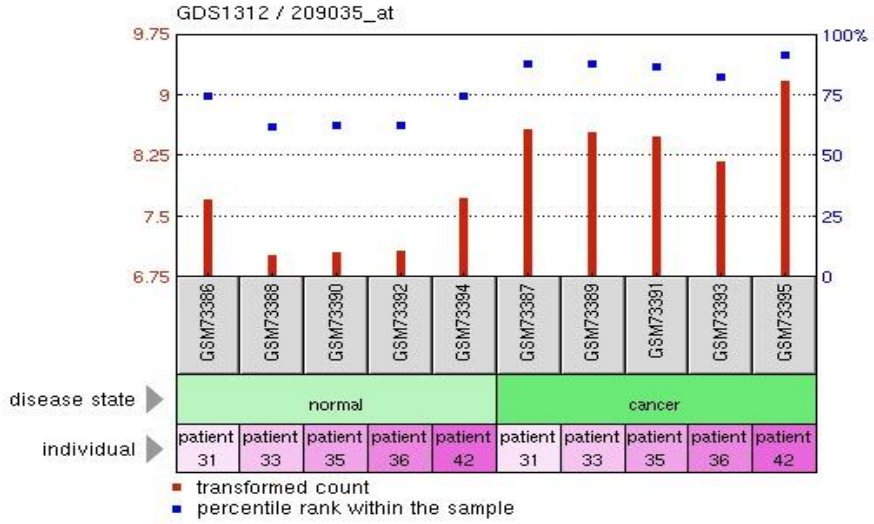

Figure 3.78: Expression level of MDK

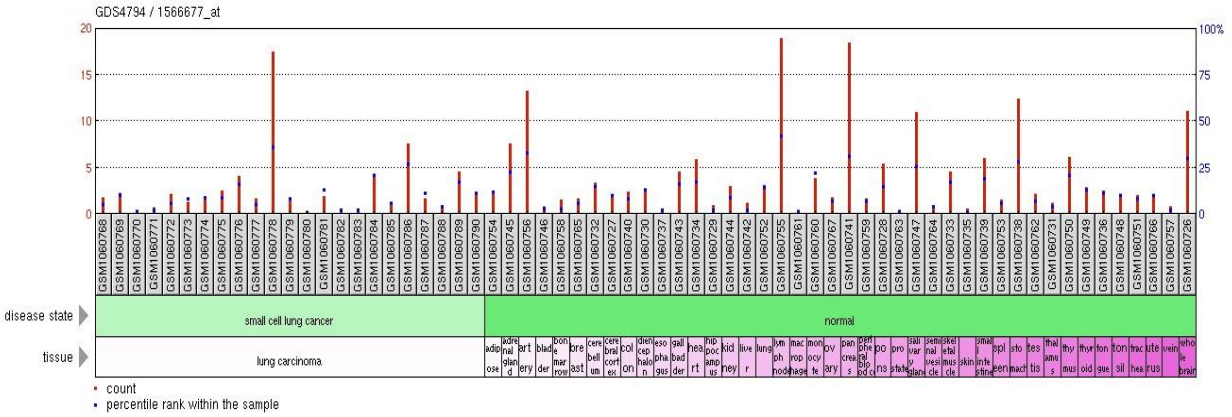

Figure 3.79: Expression level of MMP2

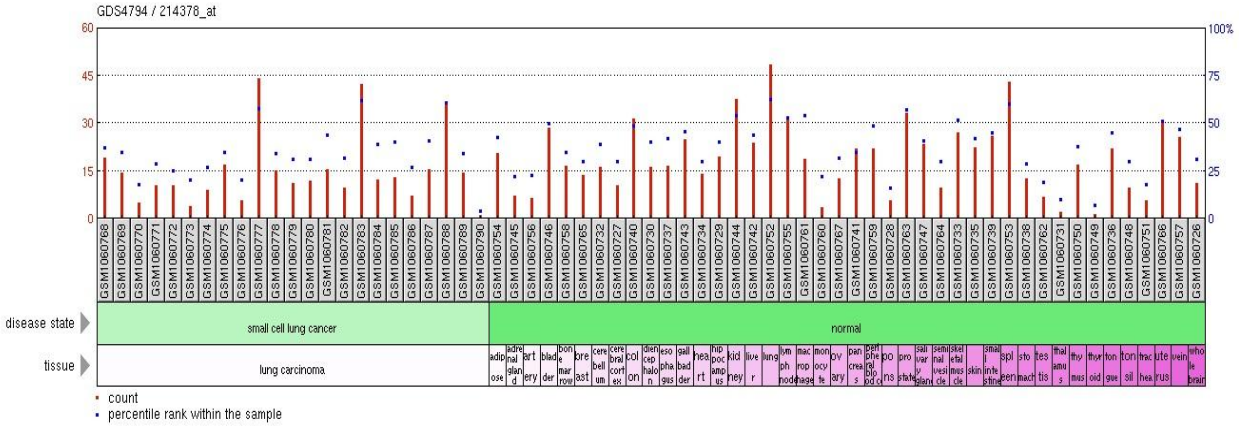

Figure 3.80: Expression level of TFPI

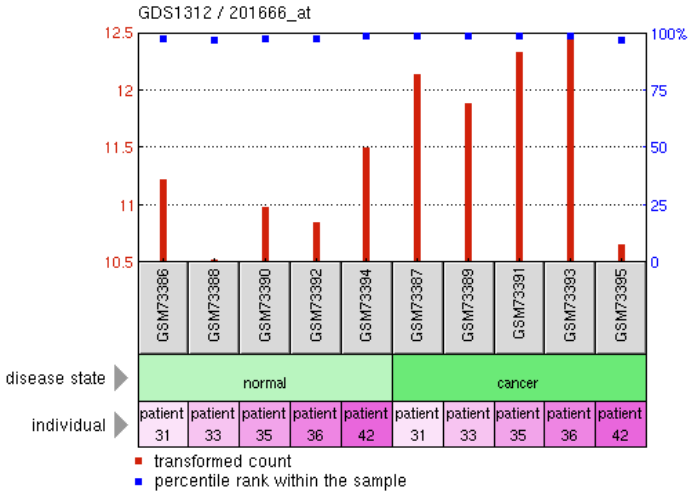

Figure 3.81: Expression level of TIMP1

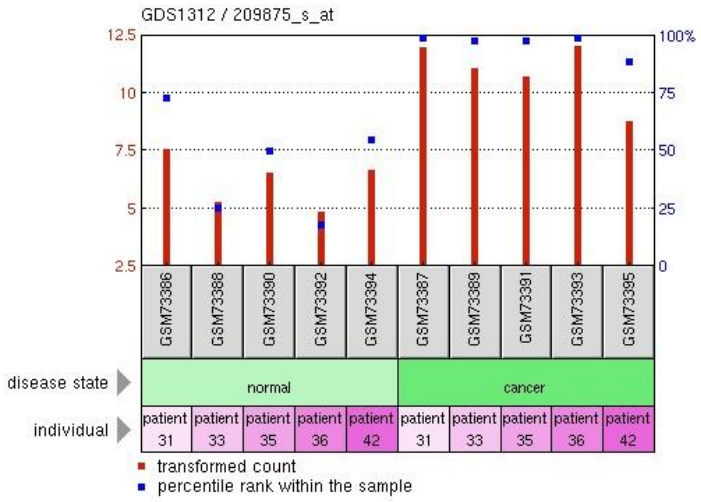

Figure 3.82: Expression level of OPN

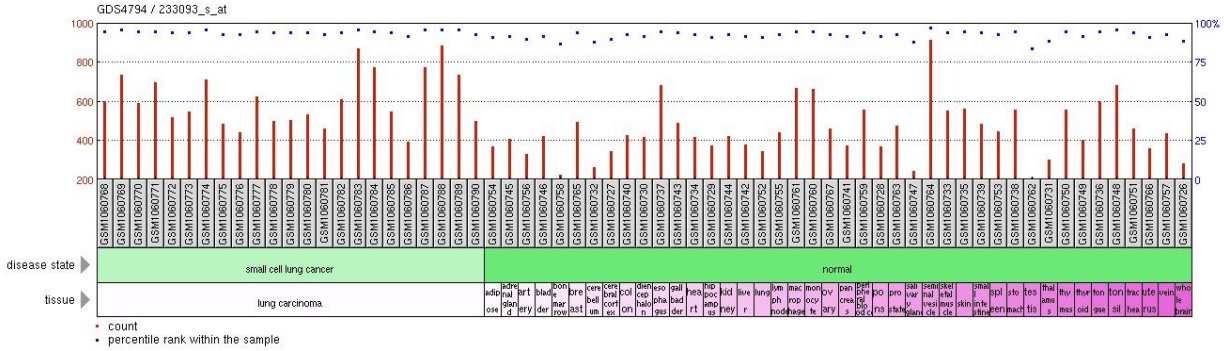

Figure 3.83: Expression level of BIRC6

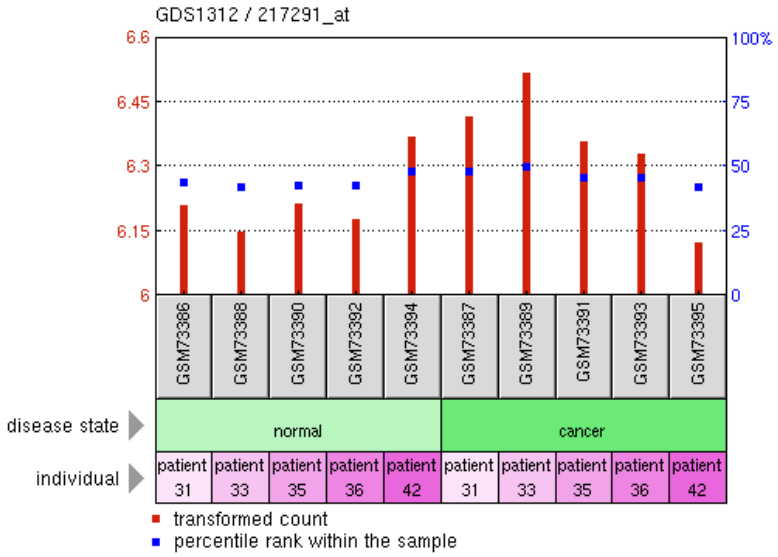

Figure 3.84: Expression level of CEA

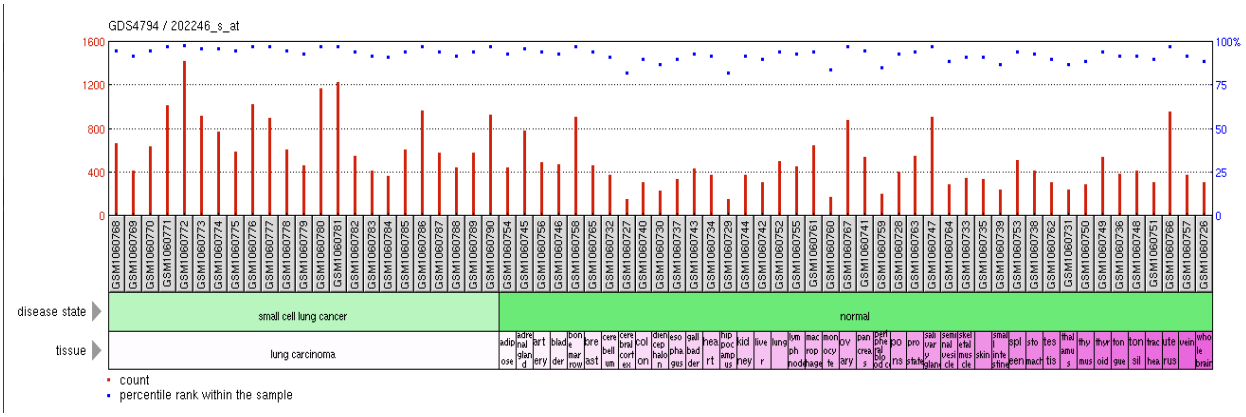

Figure 3.85: Expression level of CDK4

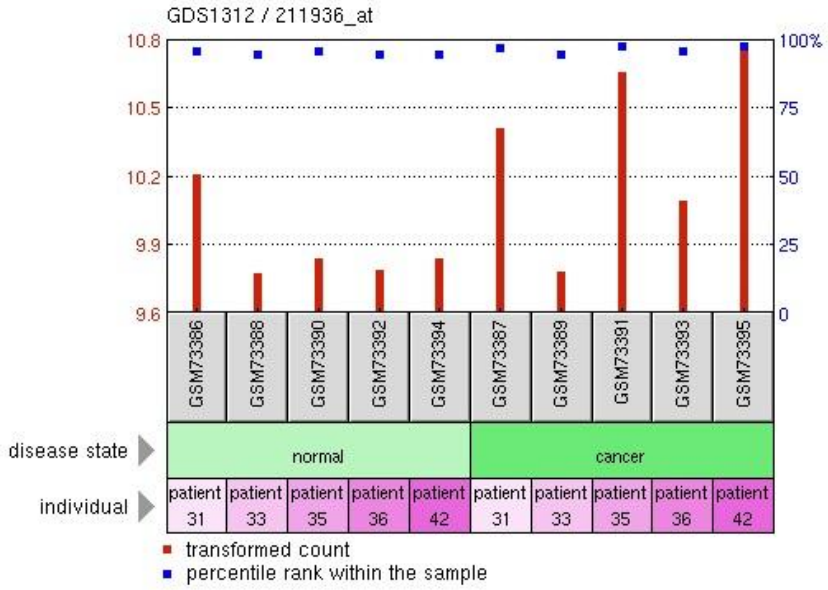

Figure 3.86: Expression level of HSPA5

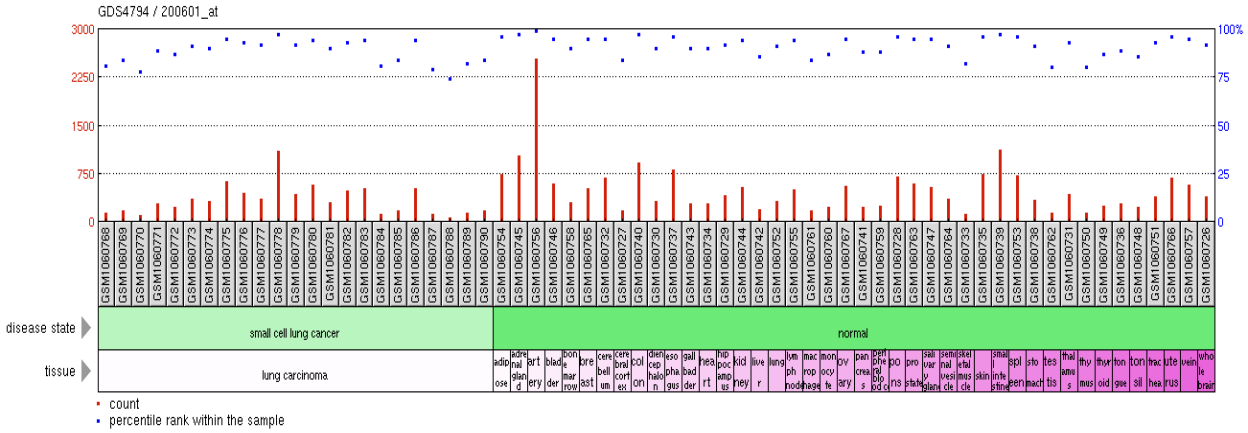

Figure 3.87: Expression level of ACTN4 protein

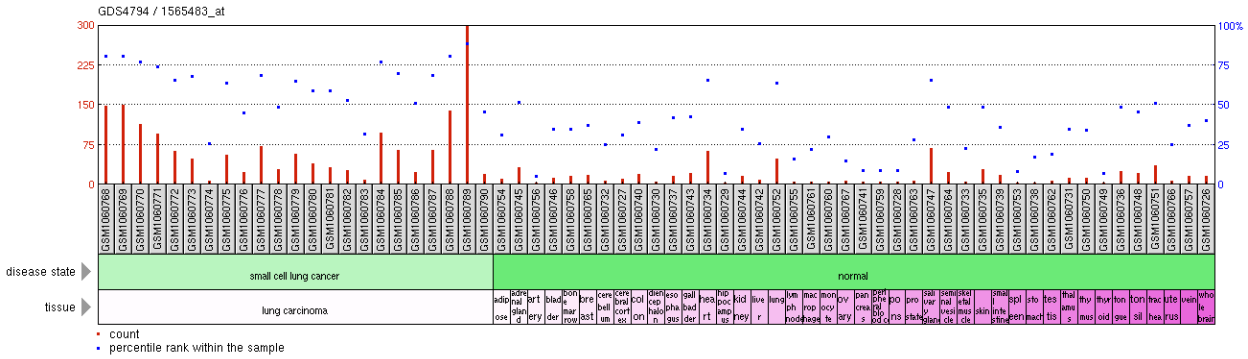

Figure 3.88: Expression level of EGFR

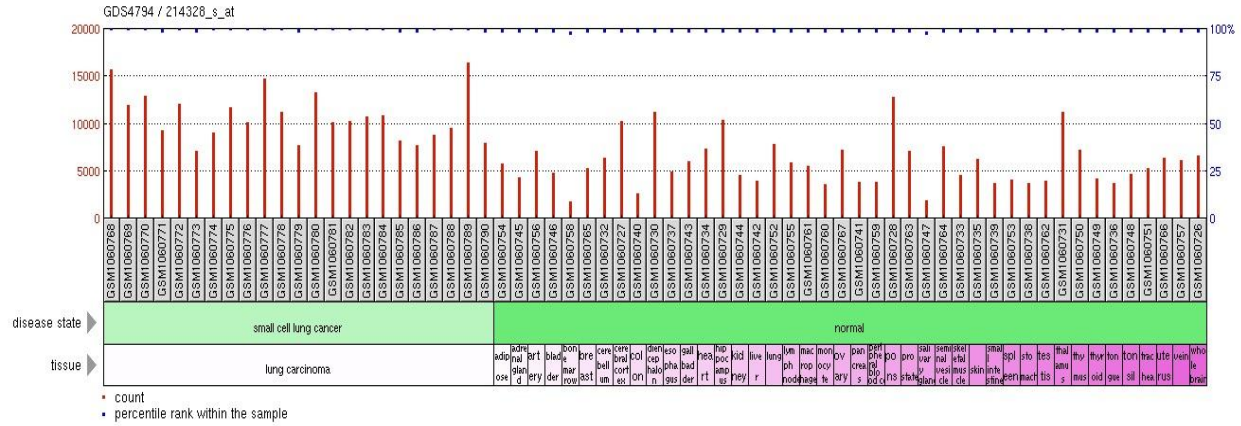

Figure 3.89: Expression level of HSP90AA

Protein expression level of the selected proteins given above in the figures(3.78-3.89) are shown in cancer (patient's) cell along with a normal (patient's) cell to check the difference of selected protein expression level. According to the expression level all of them except MMP2 & TFPI are over expressed in lung cancer patients compared to normal conditions, whereas MMP2 & TFPI are less expressed in lung cancer.

CHAPTER 4:

DISCUSSION

The result of the studies proteins, their structure, sequence, motif and expression level is discussed for better understanding in this section.

4.1 Protein Structure prediction

Protein structure prediction is the inference of the three-dimensional structure of a protein from its amino acid sequence that is, the prediction of its folding and its secondary and tertiary structure from its primary structure. The tertiary structure can be of different types. For example: AB initio, threading, homology modeling etc.

For this thesis project homology modeling was performed using SWISS MODEL Workspace. Homology modeling, also known as comparative modeling of protein, refers to constructing an atomic-resolution model of the "target" protein from its amino acid sequence and an experimental three-dimensional structure of a related homologous protein (the "template").The structures are differently colored which shows the residue error, blue being the lowest it goes up to red and 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.

For each structure QMEAN value was given. For TIMP1 0.54, HSPA5 0.87, MDK 0.23, CDK4 0.65, TFPI 0.44, BIRC6 0.79, EFGR 0.64, Alpha Actinin 4 0.86 and for HSP90 & MMP2 0.76 is the value, given the range limit of 0 to 1.

4.2 Protein Sequence Motif

Structural motifs are short segments of protein 3D structure, which are spatially close but not necessarily adjacent in the sequence. Structural motifs may be conserved in a large number of different proteins. Their role may be structural or functional. Here, MEME suite was used. After giving the proper commands along with the required forms of data for each of the proteins the result showing page came up. As a result sequence motifs, their logos and sites were available for downloading. Motifs are the sequences that have biological and functional importance, logos are the graphical presentation of that motif and sites are the sequences where motifs were found. For each protein molecule the limit for motif search was set to 5 hence, 5 motifs for each protein was found. The logos are of different lengths because of the tool. It automatically selects the standard length from 6 to 50.

Most of the motifs has two sites. But there are some motifs in CEA, BIRC6, MDK, MMP2, EGFR, and ACTN which up to 4 sites. This result suggests that these proteins might have the desired motif that can be the ultimate biomarker.

For each motif per protein different E-values were given. Motifs with small E-values (e.g., less than 0.001) are very unlikely to be random sequence artifacts. The E-value of the match of a sequence in a database to a group of motifs is defined as the expected number of sequences in a random database of the same size that would match the motifs as well as the sequence does and is equal to the combined p-value of the sequence times the number of sequences in the database.

4.3 Protein Expression Level

Protein expression level can be used to determine stages of disease and type as well, since proteins are expressed in different level and unique. GEO Profile was used for getting protein expression level. There are many profiles to check protein expression level as this is NCBI database. So only the profiles with desired protein for this thesis were taken. Expression levels of the proteins are compared in normal and cancerous condition. The bottom bar has different individual samples along with lung cancer tissue indication. And the diseased state are shown in the second bottom green bars. Above these two lines 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, MDK, TIMP1, OPN, HSPA5, HSP90AA, EGFR. ACTN4 are clearly over expressed in the cancer patient. On the other hand MMP2 and TFPI are under expressed in diseased condition. So it’s pretty clear that these protein molecules can be used as biomarker in the panel for lung cancer screening. Out of these 14 proteins EGFR, OPN and TIMP1 showed visibly higher expression in lung cancer patient. So these 3 proteins has the potential to be called the ultimate biomarkers for lung cancer for which further wet-lab studies are required.

CHAPTER 5:

CONCLUSION

Worldwide, in 2008, lung cancer was the leading cause of cancer deaths in males and the second leading cause of cancer deaths in females, about 1 400 000, or 18% of all cancer deaths.(Philip,2012). This disease is tough to defeat but the scientist are working hard to find a proper cure that can minimize the mortality rate.

The general prognosis of lung cancer is poor because doctors tend not to find the disease until it is at an advanced stage. Five-year survival is around 54% for early stage lung cancer that is localized to the lungs, but only around 4% in advanced, inoperable lung cancer. For better treatment it needs to be diagnosis at an early stage as it can reduce hardship of treatment.

Since biomarkers are very promising now a days to work with as it has the ability to look at the core detail of the disease. Protein biomarker was chosen for this thesis as protein expression in human body is abundant and easily detectable. Finding out protein biomarker homology was the main focus of this thesis. So protein structure along with their sequence motif and expression level was checked. Now the 14 proteins were chosen because of their capability. All of these 14 proteins are expressed differently in different stages of lung cancer and are quite prominent. Among the properties, EGFR, OPN and TIMP1 were further selected on the basis of structure, sequence motif and expression level. But out of initial 14 proteins, the chosen three proteins (EGFR, OPN, TIMP1) showed high expression level in lung cancer patient. So these three proteins (EGFR, OPN, TIMP1) can be the potential biomarkers for lung cancer.

To study protein biomarkers using different bioinformatics tools was the intention of this thesis project. The structure, sequence motifs and expression levels can help in further studies on these lung cancer biomarkers.

1. Protein structures can help to learn about these proteins binding sites along with their target molecules.
2. Sequence motifs can help study regulatory properties along with their therapeutic agents.
3. The expression level is very important to know in which stage how much expression occurs.

For summery, these three properties have the ability to give better treatment opportunity as the study goes further. All together a bioinformatic approach to learn about the small details of lung cancer enables the broad road of finding out the proper and best treatment to stand against this disease.

CHAPTER 6:

REFERENCES

1. Aibing Wu, Bin Wu, Jinsong Guo, Weiren Luo, Dong Wu, Huiling Yang, Yan Zhen, Xiaoli Yu, Hao Wang, Ying Zhou, Zhen Liu, Weiyi Fang and Zhixiong Yang. Elevated expression of CDK4 in lung cancer. Wu et al. Journal of Translational Medicine 2011.
2. 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
3. Amita Patnaik, Lee S. Rosen, Sara M. Tolaney, Anthony W. Tolcher, Jonathan W. Goldman, Leena Gandhi, Kyriakos P. Papadopoulos1, Muralidhar Beeram1, Drew W. Rasco1, John F. Hilton3, Aejaz Nasir4, Richard P. Beckmann4, Andrew E. Schade4, Angie D. Fulford4, Tuan S. Nguyen4, Ricardo Martinez4, Palaniappan Kulanthaivel4, Lily Q. Li4, Martin Frenzel4, Damien M. Cronier4, Edward M. Chan4, Keith T. Flaherty5, Patrick Y. Wen3, and Geoffrey I. Shapiro
4. Analysis Tool Web Services from the EMBL-EBI. (2013 July) Nucleic acids research 41 (Web Server issue) :W597-600 PMID: 23671338
5. Anastasios Dimou and Vassiliki Papadimitrakopoulou ,* Non-Small Cell Lung Cancer beyond Biomarkers: The Evolving Landscape of Clinical Trial Design. J. Pers. Med. 2014, 4, 386-401; doi:10.3390/jpm4030386
6. 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.
7. Asa Ben-Hur and Douglas Brutlag. Sequence motifs: highly predictive features of protein function.
8. 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
9. Bhatt A., Mathur R., Farooque A., Verma A. & Dwarakanath B.(2010) Cancer biomarkers - Current perspectives ,
10. Bhatt A., Mathur R., Farooque A., Verma A. & Dwarakanath B.(2010) Cancer biomarkers - Current perspectives ,
11. C.M.V. Goparaju, H.I. Pass, J.D. Blasberg, N. Hirsch, and J.S. Donington. Functional Heterogeneity of Osteopontin Isoforms in Non-Small Cell Lung Cancer. Published in final edited form as: J Thorac Oncol. 2010 October ; 5(10): 1516–1523. doi:10.1097/JTO.0b013e3181eba6bd
12. 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
13. Celine Mascaux, Murry W. Wynes, Yasufumi Kato, Cindy Tran, Bernadette Reyna Asuncion,Jason M. Zhao, Mark Gustavson, Jim Ranger-Moore, Fabien Gaire, Jun Matsubayashi, Toshitaka Nagao, Koichi Yoshida, Tatuso Ohira, Norihiko Ikeda, and Fred R. Hirsch. EGFR Protein Expression in Non–Small Cell

Lung CancerPredicts Response to an EGFR Tyrosine Kinase Inhibitor—A Novel Antibody for Immunohistochemistry or AQUA Technology. doi: 10.1158/1078-0432.CCR-11-0209. 2011 American Association for Cancer Research

14. Charles E. Birse, Robert J. Lagier, William FitzHugh, Harvey I. Pass, William N. Rom, Eric S. Edell, Aaron O. Bungum, Fabien Maldonado, James R. Jett, Mehdi Mesri, Erin Sult, Elizabeth Joseloff, Aiqun Li, Jenny Heidbrink, Gulshan Dhariwal, Chad Danis, Jennifer L. Tomic, Robert J. Bruce, Paul A. Moore, Tao He, Marcia E. Lewis and Steve M. Ruben: Blood-based lung cancer biomarkers identified through proteomic discovery in cancer tissues, cell lines and conditioned medium. Birse et al. Clinical Proteomics (2015) , DOI 10.1186/s12014-015-9090-9.
15. Chothia C. and Lesk A.M.(1986).The relation between the divergence of sequence and structure in proteins. EMBO J5:823–6.
16. Edward B. Garon, Richard S. Finn, Habib Hamidi, Judy Dering, Sharon Pitts, Naeimeh Kamranpour, Amrita J. Desai, Wylie Hosmer, Susan Ide, Emin Avsar, Michael Rugaard Jensen, Cornelia Quadt, Manway Liu, Steven M. Dubinett, and Dennis J. Slamon. The HSP90 Inhibitor NVP-AUY922 Potently Inhibit Non–Small Cell Lung Cancer Growth. Mol Cancer Ther; 12(6) June 2013
17. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. (2011) Molecular systems biology 7 :539 PMID: 21988835
18. Guex, N., Peitsch, M.C., Schwede, T. (2009). Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PdbViewer: A historical perspective. Electrophoresis, 30(S1), S162-S173.
19. Henry M. Marshall, Rayleen V. Bowman, Ian A. Yang, Kwun M. Fong, Christine D. Berg, Sep 10, 2013. Screening for lung cancer with low-dose computed tomography: a review of current status. J Thorac Dis 2013;5(S5):S524-S539. doi: 10.3978/j.issn.2072-1439.2013.09.06
20. Ioannis Prassas, Caitlin C Chrystoja, Shalini Makawita and Eleftherios P Diamandis. Bioinformatic identification of proteins with tissue-specific expression for biomarker discovery. Prassas et al. BMC Medicine 2012, 10:39
21. Ioannis Prassas, Caitlin C Chrystoja, Shalini Makawita and Eleftherios P Diamandis. BMC Medicine 2012 Bioinformatic identification of proteins with tissue-specific expression for biomarker discovery.
22. John E. Hale, Valentina Gelfanova, James R. Ludwig and Michael D. Knierman 11th August, 2003. Application of proteomics for discovery of protein biomarkers.
23. Khashayar Esfahani, Victor Cohen, HSP90 as a novel molecular target in non-small-cell lung cancer Lung Cancer: Targets and Therapy 2016:7

24. Kiefer F, Arnold K, Künzli M, Bordoli L, Schwede T (2009). The SWISS-MODEL Repository and associated resources. Nucleic Acids Research. 37, D387-D392.
25. Marco Biasini, Stefan Bienert, Andrew Waterhouse, Konstantin Arnold, Gabriel Studer, Tobias Schmidt, Florian Kiefer, Tiziano Gallo Cassarino, Martino Bertoni, Lorenza Bordoli, Torsten Schwede. (**2014**). SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Research*; (1 July 2014) 42 (W1): W252-W258; doi: 10.1093/nar/gku340.
26. Marti---Renom M.A., Stuart A.C., Fiser A., Sanchez R., Melo F., Sali A. (2000). Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomoltruct 29: 291–325.
27. Michael R. Mehan, Deborah Ayers, Derek Thirstrup, Wei Xiong, Rachel M. Ostroff, Edward N. Brody, Jeffrey J. Walker, Larry Gold1, Thale C. Jarvis, Nebojsa Janjic, Geoffrey S. Baird, Sheri K. Wilcox: Protein Signature of Lung Cancer Tissues. April 2012 | Volume 7 | Issue 4 | e35157
28. Ming-Chuan Wang, PhD, Ying-Hua Chang, PhD, Chih-Chieh Wu, PhD, Yu-Chang Tyan, PhD, Hua-Chien Chang, MS, Yih-Gang Goan, MD,Wu-Wei Lai, MD,Pin-Nan Cheng, MD, and Pao-Chi Liao, PhD. Alpha-Actinin 4 Is Associated with Cancer Cell Motility and Is a Potential Biomarker in Non–Small Cell Lung Cancer Journal of Thoracic Oncology Volume 10, Number 2, February 2015 Functional Properties of the ACTN4 Protein in NSCLC
29. Min-Kyu Han, Young-Hee Oh, Jimin Kang, Young-Pil Kim, Soowon Seo, Jhingook Kim, KeunChil Park and Hak-Sung Kim.(2009) Protein profiling in human sera for identificationof potential lung cancer biomarkers using antibody microarray. DOI 10.1002/pmic.200800777 Proteomics 2009, 9, 5544–5552
30. 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
31. N. Lynn Henrya, *, Daniel F. Hayesb Cancer biomarkers MOLECULAR ONCOLOGY 6 (2012) 140 e146
32. N. Lynn Henrya,, 6 February 2012 Daniel F. Hayesb. Cancer biomarkers.
33. PARVEEN SHAHIDA AKHTAR, ZAFOR MOHAMMAD MASUD, MOHAMMAD TAREK ALAM, MAKSUDA BEGUM. J MEDICINE 2011; 12 : 115-119. Profile of Lung Cancer: A One-Year Report.
34. Peitsch, M. C. (1995) Protein modeling by E-mail Bio/Technology 13: 658-660.
35. Rafia Parveen, Shaikh Shofiur Rahman, Syeda Adib Sultana, Zakir Hossain Habib4Cancer Types and Treatment Modalities in Patients Attending at Delta Medical College Hospital. Delta Med Col J. Jul 2015
36. RH Jack, EA Davies1 and H Møller. Lung cancer incidence and survival in different ethnicgroups in South East England. British Journal of Cancer (2011)

105, 1049 – 1053 & 2011 Cancer Research UK All rights reserved 0007 – 0920/11
37. Richard Mayeux. Biomarkers: Potential Uses and Limitations NeuroRx: Vol. 1, 182–188, April 2004 © The American Society for Experimental Neuro-Therapeutics, Inc.
38. Roushney Fatima Mukti, Pratul Dipta Samadder, Abdullah Al Emran &, Farzana Ahmed &, Iqbal Bin Imran, Anyanna Malaker, Sabina Yeasmin. Score Based Risk Assessment of Lung Cancer and its Evaluation for Bangladeshi People. Asian Pacific Journal of Cancer Prevention, Vol 15, 2014
39. S M Lee Is EGFR expression important in nonsmall cell lung cancer? Committee on Publication Ethics – Seminar 2006, 10th March 2006, BMA House, London, UK
40. Steven M. Dubinett, David Geffen School of Medicine at University of California, Los Angeles, Los Angeles, CA Avrum Spira, Boston University School of Medicine, Boston, MA. Challenge and Opportunity of Targeted Lung Cancer Chemoprevention. *Journal of Clinical Oncology*, Vol 31, No 33 (November 20), 2013: pp 4169-4171.
41. Syed Akram Hussain1, and Richard Sullivan . Cancer Control in Bangladesh.Jpn J Clin Oncol 2013;43(12)1159–1169 doi:10.1093/jjco/hyt140 Advance Access Publication 25 October 2013
42. Syed Akram Hussain1,* and Richard Sullivan2 Cancer Control in Bangladesh Jpn J Clin Oncol 2013;43(12)1159–1169 doi:10.1093/jjco/hyt140 Advance Access Publication 25 October 2013
43. The EMBL-EBI bioinformatics web and programmatic tools framework. (2015 July 01) Nucleic acids research 43 (W1) :W580-4 PMID: 25845596
44. The UniProt Consortium UniProt: a hub for protein information Nucleic Acids Res. 43: D204-D212 (2015)
45. Thierry Le Chevalier1,2* Non-small cell lung cancer: the challenges of the next decade Specialty Grand challenge article published: 30 September 2011 doi: 10.3389/fonc.2011.00029
46. Ting Xiao, Wantao Ying, Lei Li, Zhi Hu, Ying Ma, Liyan Jiao, Jinfang Ma, Yun Cai,Dongmei Lin, Suping Guo, Naijun Han, Xuebing Di, Min Li, Dechao Zhang, Kai Su, Jinsong Yuan, Hongwei Zheng, Meixia Gao, Jie He, Susheng Shi, Wuju Li, Ningzhi Xu, Husheng Zhang, Yan Liu, Kaitai Zhang, Yanning Gao, Xiaohong Qian, and Shujun Chenga, An Approach to Studying Lung Cancer-related Proteins in Human Blood. 2005 by The American Society for Biochemistry and Molecular Biology, Inc.
47. Ting Xiao,a,b Wantao Ying,c,b Lei Li,c,b Zhi Hu,a,b Ying Ma,a Liyan Jiao,c Jinfang Ma,a Yun Cai,c Dongmei Lin,d Suping Guo,a Naijun Han,a Xuebing Di,a Min Li,a Dechao Zhang,e Kai Su,e Jinsong Yuan,a Hongwei Zheng,a Meixia

Gao,a Jie He,a Susheng Shi,d Wuju Li,f Ningzhi Xu,a Husheng Zhang,a Yan Liu,a Kaitai Zhang,c,g Yanning Gao,a,h Xiaohong Qian,c,i and Shujun Chenga,j. An Approach to Studying Lung Cancer-related Proteins in Human Blood 2005 by The American Society for Biochemistry and Molecular Biology, Inc.

48. Wei Liu1 Yong Wu, Libo Wang, Ling Gao, Yingping Wang, Xiaoliang Liu, Kai Zhang, Jena Song, Hongxia Wang, Thomas A Bayer, Laurel Glaser, Yezhou Sun, Weijia Zhang, Michael Cutaia7, David Y Zhang, Fei Ye. Protein signature for non-small cell lung cancer prognosis. Am J Cancer Res 2014;4(3):256-269
49. Xianliang Chen, Xiaoying Guan, Huiyu Zhang, Xiaobin Xie, Hongyan Wang, Jie Long, Tonghui Cai, Shuhua Li, Zhen Liu and Yajie Zhang. DAL-1 attenuates epithelial-to mesenchymal transition in lung cancer Journal of Experimental & Clinical Cancer Research (2015) 34:3 DOI 10.1186/s13046-014-0117-2
50. Xin Dong, MD, Dong Lin, MD, PhD, Chris Low, MSc, Emily A. Vucic, BSc, John C. English, MD, John Yee, MD, Nevin Murray, MD, Wan L. Lam, PhD, Victor Ling, PhD,Stephen Lam, MD, Peter W. Gout, PhD, and Yuzhuo Wang, PhD. Elevated Expression of BIRC6 Protein in Non–Small-Cell Lung Cancers is Associated with Cancer Recurrence and Chemoresistance. *Journal of Thoracic Oncology* • Volume 8, Number 2, February 2013
51. Yuji Imafuku, Gilbert S. Omenn and Samir Hanash∗ Departments of Pediatrics, Medicine, and Human Genetics, University of Michigan, Ann Arbor, MI 48109, USA Proteomics approaches to identify tumor antigen directed autoantibodies as cancer biomarkers, Disease Markers 20 (2004) 149–153 149 , IOS Press.
52. 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
53. 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