12-2015

Virtual Screening of Phytochemicals for Cancer Therapy

Asma Ahmad Al Haddad

UAEU

جامعة الإمارات العربية المتحدة
United Arab Emirates University

United Arab Emirates University

College of Science

Department of Biology

VIRTUAL SCREENING OF PHYTOCHEMICALS FOR CANCER THERAPY

Asma Ahmad Al Haddad

This thesis is submitted in partial fulfillment of the requirements for the degree of Master of Science in Molecular Biology and Biotechnology

Under the Supervision of Dr. Ranjit Vijayan

December 2015

Declaration of Original Work

I, Asma Ahmad Al Haddad, the undersigned, a graduate student at the United Arab Emirates University (UAEU), and the author of this thesis entitled "*Virtual Screening of Phytochemicals for Cancer Therapy*", hereby, solemnly declare that this thesis is my own original research work that has been done and prepared by me under the supervision of Dr. Ranjit Vijayan, in the College of Science at UAEU. This work has not previously been presented or published, or formed the basis for the award of any academic degree, diploma or a similar title at this or any other university. Any materials borrowed from other sources (whether published or unpublished) and relied upon or included in my thesis have been properly cited and acknowledged in accordance with appropriate academic conventions. I further declare that there is no potential conflict of interest with respect to the research, data collection, authorship, presentation and/or publication of this thesis

Student's Signature  Date 17/2/2016

Approval of the Master Thesis

This Master Thesis is approved by the following Examining Committee Members:

1) Advisor (Committee Chair): Dr. Ranjit Vijayan
   Title: Assistant Professor
   Department of Biology
   College of Science

Signature _____ Date 14/12/2015

2) Member: Dr. Rabah Iratni
   Title: Associate Professor
   Department of Biology
   College of Science

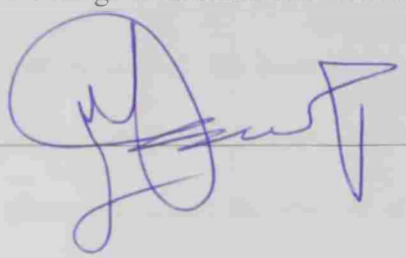
Signature _____ Date 14/12/15

3) Member (External Examiner): Prof. Firdos Alam Khan
   Title: Professor and Chairperson
   School of Life Sciences
   Manipal University Dubai

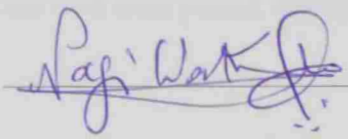
Signature _____ Date 14/12/2015

This Master Thesis is accepted by:

Dean of the College of Science: Dr. Ahmed Murad

Signature  Date 6/3/2016

Dean of the College of the Graduate Studies: NAGI WAKIM

Signature  Date 24/3/2016

Copy 3 of 6

Abstract

Cancer is one of the leading causes of deaths in the world. Vast majority of contemporary cancer therapies have side effects. Dietary plants and plant-based products have been shown to reduce the severity of the condition with minimal or no side effects. However, the precise molecular interactions that lead to this are largely unknown. In this study, several target proteins in cancer pathways were chosen and a collection of chemicals derived from plants, or phytochemicals, were virtually screened *in silico* using molecular docking techniques. Furthermore, the docking modes were subjected to extended molecular dynamics (MD) simulations. The pharmacokinetic properties of these phytochemicals were also checked using *in silico* prediction techniques. A short list of phytochemicals were identified, including ellagic acid, macrocarpal A, epigallocatechin gallate and quercetin, which appeared to bind to the protein Rab5A, a GTPase involved in intracellular membrane trafficking. MD simulations also produced stable interactions between the protein and the ligands. These molecules also showed an acceptable drug score and could therefore be subjected to further validation.

Keywords: *in silico*, Molecular Docking, MD Simulation, EGFR, Rab5A.

Title and Abstract (in Arabic)

الفحص الافتراضي للمواد الكيميانية النباتية لعلاج السرطان

الملخص

أشارت الدراسات العلمية المختصة بالأوبئة إلى أن الاستهلاك اليومي من المواد الكيميائية المستخرجة من النباتات، يمكن أن تسهم في تقليل الإصابة بأنواع عديدة من السرطان.

إن دراسة هذه المواد الكيميائية المستخلصة من النبات التي لها تأثير في كبح مسار البروتينات المرتبطة بالسرطانهي طريقة واعدة في تطوير العقاقير لعلاج السرطان. فائدة إضافية هي أن الغالبية العظمى من هذه المواد الكيميائية تكون طبيعية وبالتالي يكون لها آثار جانبية ضئيلة من ناحية السمّية.

إنالهدف الرئيس هو الحصول على قائمة من المواد الكيميائية النباتية التي أثبتت الدراسات أن لها تأثيراً مباشراً على السرطان، ثم اختيار قائمة من البروتينات التي يمكن استخدامها كأهداف لتلك المواد الكيميائية لتستخدم كعقار للسرطان.تتناول هذه الدراسة القيام بتقنية الفحص الافتراضيوالإقحام الجزيئي بالتنبؤ بشكل ثلاثي الأبعاد عن كيفية تفاعل المواد الكيميائية أمام البروتين الهدف. مقارنة بطرق الفحص الأخرى .

الفحص الافتراضي له العديد من المزايا، من السهولة التحكم به آليا ،ويقلل عدد المركبات التي يتم اختبارها تجريبيا وبالتالي توفر الوقت والمال ويتطلب الحد الأدنى من المعدات والموارد وتم استخدام المحاكاة بالديناميك الجزيئيلإثبات النتائج الظاهرة في الإقحام الجزيئي، وتأكيداً إذا كانت مستمرة بالارتباط في الموقع النشط للبروتين طوال فترة المحاكاة للحصول على نتائج أكثر دقة.

أهم نتائج هذه الدراسة أثبتت أن المواد الكيميائية النباتية: حمض الإيلاجيك، ماكروكاربل أو إي جي سي جي استطاعت أن تثبطبروتين راب5 أ، وهو إنزيم من نوع GTPase يساهم في تنظيم المرور داخل الخلايا، أما المحاكاة بالديناميك الجزيئي (MD Simulation) فقد أظهر ثبات الروابط بين البروتين والمواد الكيميائية. أيضا أظهرت أن هذه المواد مقبولة من ناحية استخدامه كدواء آمن مما يحفزنا في متابعة اكتشاف المزيد عن هذه المواد.

مفاهيم البحث الرئيسية: الفحص الافتراضي، الإقحام الجزيئي، المحاكاة بالديناميك الجزيئة، إي جي إف آر، راب5 أ.

Acknowledgements

This project would not have been possible without the support of many people. A special thanks to Dr. Ranjit Vijayan, my advisor, for his countless hours of reflecting, reading, encouraging, and most of all patience throughout the entire process. Thanks to the United Arab Emirates University for providing me with the financial means to complete this project. I would also like to thank my colleague Ms Priya Antony for her help and support.

I would like to thank the Department of Biology, especially those members of my master's committee for their input, valuable discussions and accessibility. In particular, I would like to thank Dr. Rabah Iratni, Dr. Synan Abu Qamar and Dr. Khaled Tarabily for their hard work, expertise and patience.

Dedication

I dedicate this thesis to my beloved father and mother for their endless love and support, and to my family and friends.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ADME	Absorption, Distribution, Metabolism, Excretion
AKT1	Murine Thymoma Viral Oncogene Homolog 1
AMBER	Assisted Model Building and Energy Refinement
AP-1	Activator Protein 1
APL	Abelson murine leukemia
ASP	Astex Statistical Potential
BAD	Bcl-2-Associated Death
BCR	Breakpoint Cluster Region
CAK	CDK-activating kinase
CCDC	Cambridge Crystallographic Data Centre
CDK2	Cyclin Dependent Kinase 2
CDK4	Cyclin Dependent Kinase 4
CDK6	Cyclin Dependent Kinase 6
CHARMM	Chemistry at HARvard Molecular Mechanics
ChemPLP	Combined ChemScore and piecewise linear potential
ERK1/2	Extracellular signal-Regulated Kinase 1 and 2
EGFR	Epidermal Growth Factor Receptor
ERBB2	Erb-B2 Receptor Tyrosine Kinase 2
FOS	FBJ Murine Osteosarcoma Viral Oncogene Homolog
G1, 2	Gap 1, 2
GA	Genetic search Algorithm
GAFF	General Amber Force Field
GEF	Guanine nucleotide Exchanges Factors
GOLD	Genetic Optimization for Ligand Docking
GROMOS	GROningen MOlecular Simulation

GDP	Guanosine-5'-diphosphate
GTP	Guanosine-5'-triphosphate
HIV	Human Immunodeficiency Virus
HTS	High-Throughput Screening
JAK1	Janus Kinase 1
JAK2	Janus Kinase 2
JAK3	Janus Kinase 3
JUN	Jun Proto-Oncogene
M	Mitosis
MAP	Mitogen Activated Protein
MAPK	Mitogen Activated Protein Kinase
MAPKK1	Mitogen Activated Protein Kinase Kinase 1
MD	Molecular Dynamic
mTOR	mammalian Target of Rapamycin
NMR	Nuclear Magnetic Resonance
NF- κB	Nuclear Factor-kappa B
OPLS-AA	Optimized Potentials for Liquid Simulations- All Atom
PDB	Protein Data Bank
PDGFR	Platelet-Derived Growth Factor Receptor
PI3K	Phosphoinositide 3-kinase
PH	Pleckstrin Homology
PyRx	Python Prescription
RAS	Rat Sarcoma
Rab5A	Ras-related protein 5A
RMSD	Root Mean Square Deviation
S DNA	DNA Synthesis
SDF	Spatial Data File

SEQRES	SEQuence RESidue
SMILES	Simplified Molecular Input Line Entry Specification
SOS	Son of Sevenless
STAT	Signal Transducer and Activator of Transcription
TKI	Tyrosine-Kinase Inhibitor
TPSA	Topological Polar Surface Area
TYK2	Tyrosine Kinase 2
VDW	Van der Waals
VMD	Visual Molecular Dynamics

Chapter 1: Introduction

1.1 What is cancer?

A fine balance between cell proliferation and cell death maintains organisms. If this homeostasis is disturbed by stress, decrease in cell death or increase in proliferation rate, a tumor might occur, which can further progress into a cancer.

Tumor development is most commonly described as natural selection followed by clonal expansion, resulting in monoclonal tumors originating from the progeny of a single cell (Nowell, 1976).

Aberrations that confer growth advantages to the cell will accumulate during the clonal selection process. These changes are consequences of different processes: 1) activation of proto-oncogenes rendering the gene constitutively active or active under conditions in which the wild type gene is not, 2) inactivation of tumor suppressor genes, reducing or abolishing the activity of the gene product, and 3) alteration of repair genes, which normally keep genetic alterations to a minimum. Genomic analyses focusing on structural and numerical aberrations of chromosomes have long suggested that cancer could have a genetic component (Vogelstein & Kinzler, 2004).

Evidence indicates that tumors of various types often derive directly from the normal tissues in which they are first discovered. However, tumors are capable of moving within the confines of the human body. In many patients multiple tumors were discovered at anatomical sites quite distant from where the disease first began, a consequence of the tendency of cancers to spread thought the body and to establish new colonies of cancer cells, a process called metastasis (Figure 1.1). Metastasis is often

traceable directly back to the site where the disease had begun (the primary tumor). (Weinberg, 2013)

Figure 1.1: Metastases of cancer cells in the lung of two mice

1.1.1 Causes of cancer

Cancer is the result of cells that uncontrollably grow and do not die. Normal cells in the body follow an orderly path of growth, division, and death. Apoptosis is programmed cell death and when this process breaks down, cancer begins to form. Unlike regular cells, cancer cells do not experience programmatic death and instead continue to grow and divide. This leads to a mass of abnormal cells that grows out of control. Cancer can be the result of many causes including genetic predisposition that is inherited from family members. It is possible to be born with certain genetic mutations or a fault in a gene that makes an individual statistically more likely to develop cancer later in life.

Another cause is the age. There is an increase in the number of possible cancer-causing mutations in our DNA as one ages. This makes age an important risk factor for cancer. Several viruses have also been linked to cancer such as: human papillomavirus,

hepatitis B and C, and Epstein-Barr virus. Human immunodeficiency virus (HIV) and anything else that suppresses or weakens the immune system, because it inhibits the body's ability to fight infections, increases the chance of developing cancer.

Carcinogens are substances that are directly responsible for damaging DNA, promoting or aiding cancer. Tobacco, asbestos, arsenic, radiation such as gamma and X-rays, the sun, and compounds in car exhaust fumes are all examples of carcinogens. When our bodies are exposed to carcinogens, free radicals are formed that removes electrons from other molecules in the body. Theses free radicals damage cells and affect their ability to function normally.

1.1.2 Medicinal Plant and Cancer

Medicinal plants have been used for the treatment of various diseases from ancient times. The anticancer properties of plants have been recognized for centuries. Plant-based drugs with potent anticancer effects should add to the efforts to find a cheap drug with limited clinical side effects.

The National Cancer Institute (NCI) has screened approximately 35,000 plant species for potential anticancer activities. Among them, about 3,000 plant species have demonstrated reproducible anticancer activity (data available at http://www.ars-grin.gov/duke/).

Many studies have focused on the chemoprotective properties of plants such as anticarcinogenic properties of *Abrus precatorius* on Yoshida sarcoma in rats, fibrosarcoma in mice and ascites tumor *cells* (Reddy et al, 1969). Similarly, Dhar and colleagues have examined the anticancer properties of *Albizzia lebbeck* on sarcoma in

mice and *Alstonia scholaries*on benzo[a]pyrene-induced forestomach carcinoma in humans (Dhar et al, 1968).

Other plants that have shown anticarcinogenic properties include *Anacardium occidentale* in hepatoma, *Asparagus racemosa* in human epidermoid carcinoma, *Boswellia serrata* in human epidermal carcinoma of the nasopharynx, *Erthyrina suberosa* in sarcoma, *Euphorbia hirta* in Freund virus leukemia, *Gynandropis pentaphylla* in hepatoma, *Nigella sativa* in Lewis lung carcinoma, *Peaderia foetida* in human epidermoid carcinoma of the nasopharynx, *Picrorrhiza kurroa* in hepatic cancers, and *Withania somnifera* tumors (Dhar et al, 1968).

The anticancer characteristics of a number of plants are still being actively researched and some have shown promising results. Some plants and plant products that have shown promise as anticancer agents are discussed in detail in the following sections. Although many bioactive plant chemicals have been identified very little is known about the precise molecular mechanisms in the prevention of cancer.

1.2 Problem statement

The purpose of this study was to identify how a collection of plant chemicals (p

precise mechanism of how these compounds interact with their target could be of benefit in developing novel cancer therapies. Moreover, the results could lead to the development of specific phytochemical based treatments that have fewer side effects for patients. Identification of potential drugs could also have significant commercial benefits.

1.3 Objectives

In this study, a collection of natural molecules was screened *in silico* to identify if they could interact with major proteins in different cancer pathways. The main objectives of this study were:

1. To select a list of phytochemicals that has been shown to have an effect on cancer.

2. To obtain a list of proteins that can be used as drug targets.

3. To perform *in silico* screening to obtain a short list of candidate molecules that could be used for cancer therapy.

These objectives were addressed *in silico* using molecular docking and molecular dynamic simulation study. This approach has the potential to generate a wealth of information on effects of these compounds on cancer pathways, which will help to explain the mechanisms of action.

This thesis focused on sixty-seven plant compounds that are known to be able to protect against cancer development. Seven different target proteins, which are currently targets for cancer treatment, were chosen. This included serine-threonine protein kinase1 (AKT1), cyclin dependent kinase 2 (CDK2), epidermal growth factor receptor (EGFR), extracellular signal regulated kinase 2 (ERK2), janus kinase 1 (JAK1), mitogen activated protein kinase kinase 1 (MAPKK1) and Ras-related protein 5A (Rab5A).

Chapter 2: Review of literature

Cancer is the third leading cause of death in the UAE. A retrospective analysis of patients admitted to Tawam Hospital during years 1980 to 1984 indicated the presence of 1,357 cases of cancer. These included breast cancer (9%), head and neck cancer (9%), lung cancer (7%), non-Hodgkins lymphoma (6%), acute leukemia (5%), cancer of the cervix (5%), stomach cancer (5%), Hodgkins lymphoma (4%), cancers of the colon and rectum (4%), thyroid cancer (4%), and others. (Holt, 1985)

Most cancers involve several important pathways that regulate different cellular processes such as cell fate, cell survival and genome maintenance. The driver genes, which are responsible for the formation of tumors, function through these signaling pathways.

2.1 Cancer Pathways

Several signaling molecules that play a critical role in cancer pathways have been identified and are now recognized as potential therapeutic target. One of the most important pathways is the Epidermal Growth Factor Receptor (EGFR) pathway. This pathway exists on the cell surface and normally helps the cells to grow and divide. Upon stimulation of EGFR, an extensive network of signal transduction pathways is activated, results in autophosphorylation of key tyrosine residues. These tyrosine phosphorylated sites lead to the activation of downstream signaling cascades including the RAS/ERK pathway, MAPK/ERK pathway, the phosphatidylinositol 3-kinase (PI3K)/ AKT pathway and the Janus kinase/Signal transducer and activator of transcription (JAK/ STAT) pathway. These pathways act in a coordinated manner to promote cell survival.

These transmembrane proteins are activated following the binding with peptide growth factors of the EGF-family of proteins. The ability of cells to sense their environment and decide to survive or die is dependent largely upon growth factors.

2.1.1 EGFR Pathway

EGFR is a key growth factor regulating cell survival found on the surface of cells. EGFR play an important role in epithelial cell function. In breast cancer, the EGF receptor, ERBB2 can be amplified and over-expressed and this is associated with poor prognosis and drug resistance (Henson et al., 2007).

EGFR has been shown to be frequently overexpressed or hyper activated in a number of epithelial tumors. Figure 2.1 shows downstream signaling pathways regulated by EGFR (Nyati et al., 2006).

The downstream signaling effects of these aberrations lead to impaired apoptosis, and/or enhanced proliferation, angiogenesis, necrosis, and treatment refractoriness, suggesting a causative relationship between receptor dysregulation and the pathobiology of many cancers (Taylor et al., 2012).

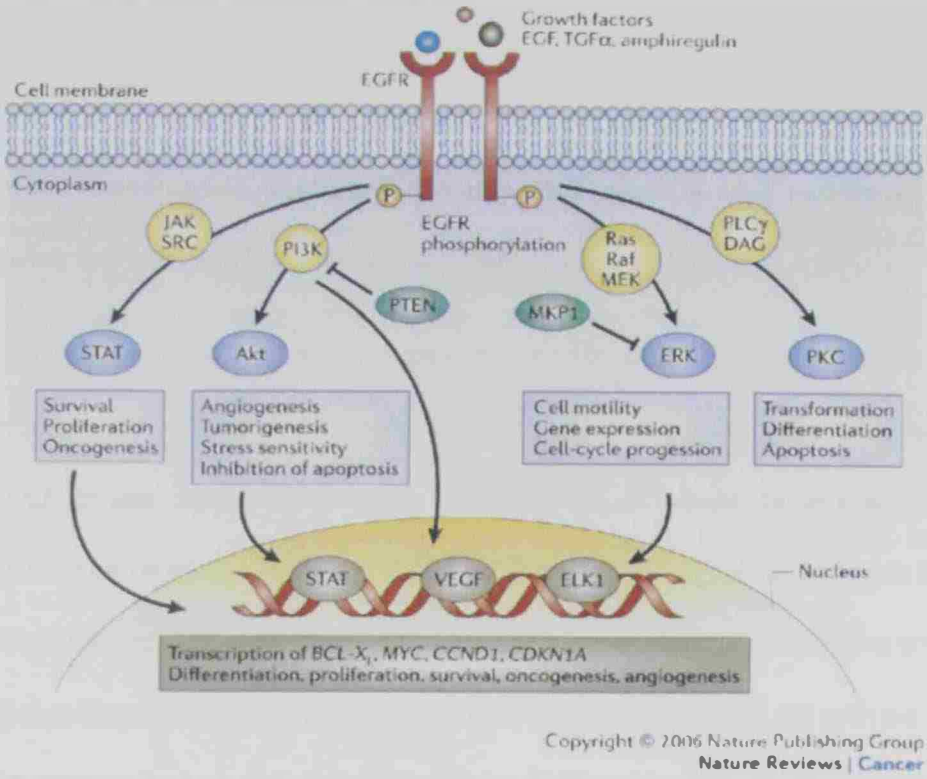


Figure 2.1: The main downstream signaling pathways regulated by EGFR

2.1.2 Ras/ERK Pathway

EGF activates the ERK pathway through the binding of Grb2 or Shc to phosphorylated ErbB receptors, which in turn results in the recruitment of the son of sevenless (SOS) to the activated receptor dimer (Henson and Gibson, 2006). SOS then activates RAS leading to the activation of RAF-1. RAF-1 subsequently phosphorylates MEK1 and MEK2, which activate respectively ERK1 and ERK2.

This pathway results in cell proliferation and in the increased transcription of Bcl-2 family members and inhibitor of apoptosis proteins (IAPs), thereby promoting cell survival. Growth factors and mitogens use the Ras/Raf/MEK/ERK signaling cascade to transmit signals from their receptors to regulate gene expression and prevent apoptosis. Abnormal activation of this pathway occurs in human cancer due to mutations at upstream

membrane receptors and Ras and B-Raf as well as genes in other pathways, which serve to regulate Raf activity (McCubrey et al, 2007).

The Ras pathway is a signal transduction cascade. Ras (rat sarcoma) proteins are one of the most well studied oncoproteins and are a major component of the SOS-Ras-Raf-MAP kinase mitogenic cascade. About 25% of human tumors and 90% of all pancreatic tumors have a mutant form of the intracellular signaling molecule Ras (a small GTP binding signaling molecule). Ras proteins bind to GDP in an inactive state but upon stimulation through GEFs (guanine nucleotide exchanges factors) are activated by binding to GTP (Figure 2.1). Active Ras functions as GTPase through GAPs (GTPase-activating proteins), which hydrolyse the GTP and reach the inactive state (Hejmadi, 2010).

RAS oncogenes have been originally discovered as retroviral oncogenes, and ever since constitutively activating RAS mutations have been identified in human tumors, they are the focus of intense research (Rajalingam et al., 2007).

In tumors, typically, a single point mutation transforms the normal Ras into a Ras oncoprotein, in which the GTPase activity is lost, leading to a constantly active Ras and a cancer cell that cannot turn off its proliferation status (Hejmadi, 2010). Alternatively, Ras activation pathways may be disturbed either due to oncoprotein or mutations in regulatory pathways (as shown in Figure 2.2).

membrane receptors and Ras and B-Raf as well as genes in other pathways, which serve to regulate Raf activity (McCubrey et al, 2007).

The Ras pathway is a signal transduction cascade. Ras (rat sarcoma) proteins are one of the most well studied oncoproteins and are a major component of the SOS-Ras-Raf-MAP kinase mitogenic cascade. About 25% of human tumors and 90% of all pancreatic tumors have a mutant form of the intracellular signaling molecule Ras (a small GTP binding signaling molecule). Ras proteins bind to GDP in an inactive state but upon stimulation through GEFs (guanine nucleotide exchanges factors) are activated by binding to GTP (Figure 2.1). Active Ras functions as GTPase through GAPs (GTPase-activating proteins), which hydrolyse the GTP and reach the inactive state (Hejmadi, 2010).

RAS oncogenes have been originally discovered as retroviral oncogenes, and ever since constitutively activating RAS mutations have been identified in human tumors, they are the focus of intense research (Rajalingam et al., 2007).

In tumors, typically, a single point mutation transforms the normal Ras into a Ras oncoprotein, in which the GTPase activity is lost, leading to a constantly active Ras and a cancer cell that cannot turn off its proliferation status (Hejmadi, 2010). Alternatively, Ras activation pathways may be disturbed either due to oncoprotein or mutations in regulatory pathways (as shown in Figure 2.2).

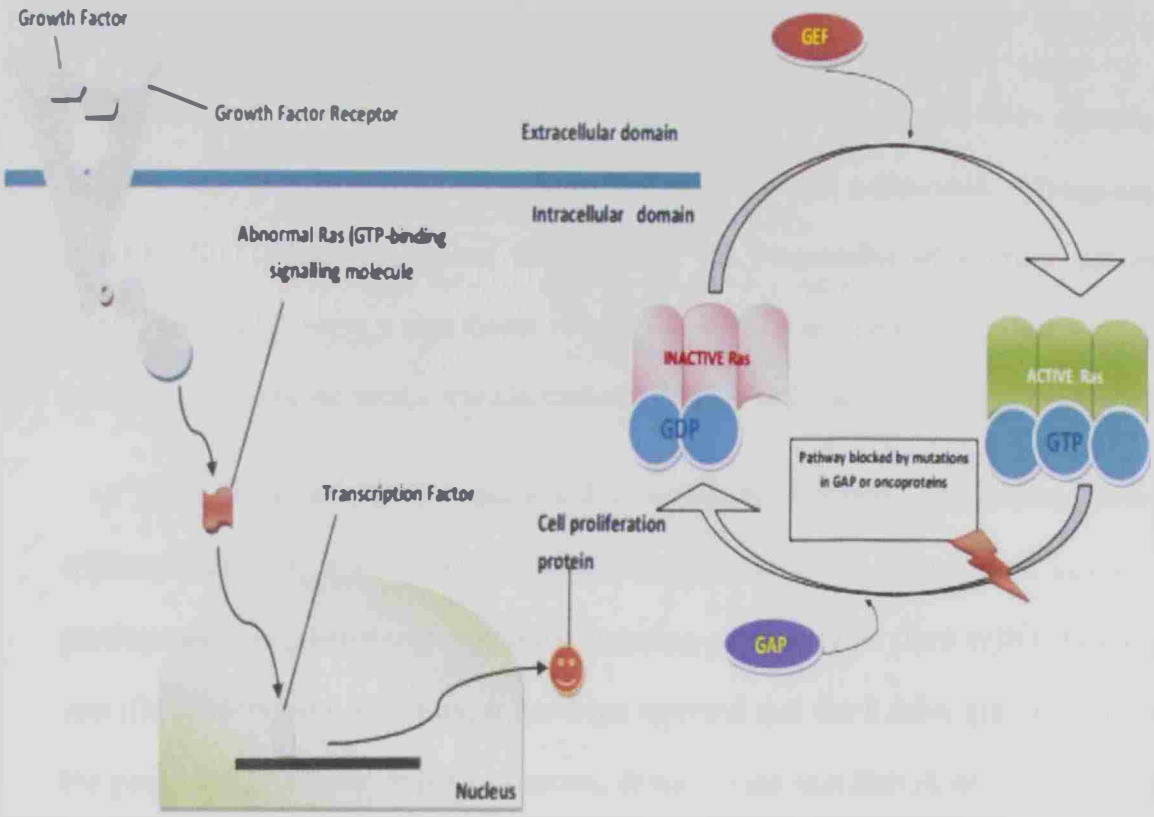


Figure 2.2: Changes in Ras signaling

Ras is vital in several signaling pathways and it acts as a binary switch. Binding GDP (guanine nucleotide diphospate) in an inactive state and GTP (guanine nucleotide triphospate) in its active state when it stimulates GF signaling. Inactivate Ras is activated by GEF (guanine nucleotide exchange factor) by exchanging its GDP for GTP. This step is short lived because Ras has an intrinsic GTPase activity, which means that it hydrolyses GTP to GDP. Mutations or oncoproteins that block the GTPase function restrict Ras in a permanently active, signaling state. (Hejmadi, 2010)

The Rab family is part of the Ras superfamily of small GTPases. There are at least 60 Rab genes in the human genome. The different Rab GTPases are localized to the cytosolic face of specific intracellular membranes, where they function as regulators of

distinct steps in membrane traffic pathways. Their cellular oncogenes were identified in humans and their mutations were found in some human carcinomas. (Hutagalung & Novick, 2011). They have been implicated in the progression of multiple cancers as membrane traffic plays a significant role in cancer biology, primarily in the loss of cell polarity and in the metastatic transformation of tumor cells (Hutagalung & Novick, 2011).

Rab5A is one of the important Rab subfamily members, a regulatory guanosine triphosphatase that is associated with the transport and fusion of endocytic vesicles, and participates in regulation of intracellular signaling pathways that allow cells to adapt to the specific environment. Recently, it has been reported that the Rab5A gene is involved in the progression of many types of cancers. It was found that Rab5A acts as an oncogene and it has been correlated with lung, stomach, and hepatocellular carcinomas (Zhao et al., 2010).

2.1.3 JAK/STAT Pathway

EGFR activates Janus kinase (JAK), which is a family of intracellular non-receptor tyrosine kinasethat transduces cytokine-mediated signals via the JAK-STAT pathway. The JAK-STAT pathway is critical in normal hematopoiesis, survival, proliferation and oncogenesis (Figure 2.1). The JAK family of kinases is involved in a variety of disease states, including inflammatory conditions, hematologic malignancies, and solid tumors (Quintas-Cardama & Verstovsek, 2013).

2.1.4 PI3K/ AKT Pathway

Activation of the EGF receptor results in autophosphorylation of key tyrosine residues. As shown in Figure 2.1 EGFR activates phosphatidylinositol 3-kinase (PI3K) that phosphorylates phosphatidylinositol 4,5-biphosphate to form phosphatidylinositol

3,4,5-triphosphate, which then activates Akt by binding at its pleckstrin homology (PH) domain (Nyati et al., 2006)

AKT1 kinase is a serine/threonine kinase known to increase cell survival by regulation of the intrinsic pathway for apoptosis. Alteration in the AKT pathway can lead to abnormal cell signaling, leading to cell proliferation, differentiation, survival and migration. The result of such uncontrolled cell signaling promotes the acquisition of a cancerous phenotype. Phosphorylated Akt has several effects, both in the cytoplasm and in the nucleus, which include the inhibition of proapoptotic factors such as BAD. Akt-mediated activation of mammalian target of rapamycin (mTOR) is also important in stimulating cell proliferation (Nyati et al., 2006).

2.1.5 MAPK/ERK Pathway

EGFR also activates signal transduction networks to MAPK/ERK pathway, which allows cells to perceive changes in the extracellular environment and plays important role in cell motility, gene expression and cell cycle progression (Figure 2.1). Mitogen-activated protein kinase (MAPK) cascades is among the most thoroughly studied of signal transduction systems and have been shown to participate in a diverse array of cellular programs, including cell differentiation, cell movement, cell division, and cell death (Schaeffer & Weber, 1999).

In mammalian systems, MAPK modulates Extracellular signal-Regulated Kinase 1 and 2 (ERK1/2) cascade, which regulates cell growth and differentiation and is involved in the regulation of meiosis and mitosis. Aberrant activation of the ERK pathway has been shown to be an essential feature common to many types of human tumors (Smith

&Schmid, 2006). Thus, inhibition of the ERK pathway is an attractive strategy for cancer treatment.

2.1.6 CDK Pathway

Many of our cells can be triggered to divide upon the proper mutagenic stimulus. The cell cycle must copy the parent cell's chromosomes and seal them in a safe daughter cell with all of the essential components needed for the daughter cell to function on its own. The human cell cycle accomplishes this in approximately 24 hours through four major phases: G1 (gap 1), S (DNA Synthesis), G2 (gap 2), and M (mitosis). Each phase serves a specific function to ensure proper cell division (Will, 2002).

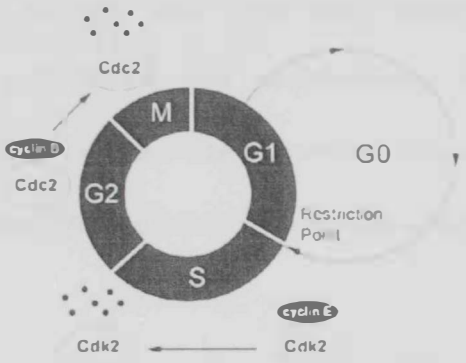


Figure 2.3: The Cell Cycle

The cell progresses through its division cycle, G1->S->G2->M, in the presence of growth signals and active CDK/cyclin complexes specific for each cell cycle. Stage G0, a non-dividing but functional phase, occurs in the absence of growth signals (Figure 2.3). Finally, cyclin degradation is illustrated by the S-phase cyclin/Cdk complex (cyclin E/Cdk2) and the G2/M cyclin/Cdk complex (cyclin B/Cdc2), where each must be

synthesized and degraded systematically for proper cell cycle progression into the next phase (Will, 2002).

Intracellular cyclins and cyclin dependent kinases (C DKs) function as central regulators of both the cell cycle and transcription. CDK2 activation depends on phosphorylation by a CDK-activating kinase (CAK) (Liu & Kipreos, 2000).

Accumulating evidence that inhibiting the activity of CDKs may be an effective therapy in cancers, including breast cancer, led to the development of small molecules that specifically target subgroups of CDKs, including CDK4 and CDK6, the kinases activated by cyclin D (Sutherland & Musgrove, 2009).

For the past two decades, there have been intense studies on the signaling pathways involved in cell growth and proliferation. New targets upstream and downstream of the key pathways are being discovered, revised and interlinked with each other, enabling extracellular signals to elicit multiple biological effects.

2.2 Phytochemicals

Phytochemicals (*phyto* meaning plant in Greek) are naturally occurring biologically active chemical compounds found in plants. They have function in protecting plants from disease and damage and contribute to the plant's color, aroma and flavor. Studies have shown ingested phytochemicals show potential for reducing the risk of cancer and cardiovascular disease in humans (Gibson et al., 1998).

Phytochemicals are also available in supplementary forms, but there is inconclusive evidence that they provide the same benefits as dietary phytochemicals (Mathai, 2000). As shown in Table 2.1, there are three major classes of phytochemicals:

1) Phenols, including the subclasses flavonoids and isoflavones; 2) Thiols, including the subclasses indoles, dithiolthiones and isothiocyanates; 3) Terpenes, including the subclasses carotenoids and limonoids. Within each subclass are individual phytochemicals (Meagher & Thomson, 1999).

2.2.1 Functions of Phytochemicals

Many researchers have focused on phytochemicals for their possible role in preventing or treating cancer and heart disease (Mathai, 2000).

In cancer, phytochemicals may detoxify substances that cause cancer. They appear to neutralize free radicals, inhibit enzymes that activate carcinogens, and activate enzymes that detoxify carcinogens. For example genistein, according to data summarized by Meagher and Thomson (1999) prevents the formation of new capillaries that are needed for tumor growth and metastasis (Miller, 2002)

2.2.2 Possible Mechanisms of Action

Chemoprevention of cancer is a cancer control in which the occurrence of this disease is prevented by administration of chemical compounds. They can halt or at least retard the development and progression of precancerous cells into malignant ones. Chemopreventive agents are subdivided into two main categories blocking agents and suppressing agents. Blocking agents prevent carcinogens from reaching the target sites, from undergoing metabolic activation or from subsequently interacting with crucial cellular macromolecules. Suppressing agents, on the other hand, inhibit the malignant trans formation of initiated cells, in either the promotion or the progression stage as shown in Figure 2.4 (Surh, 2003).

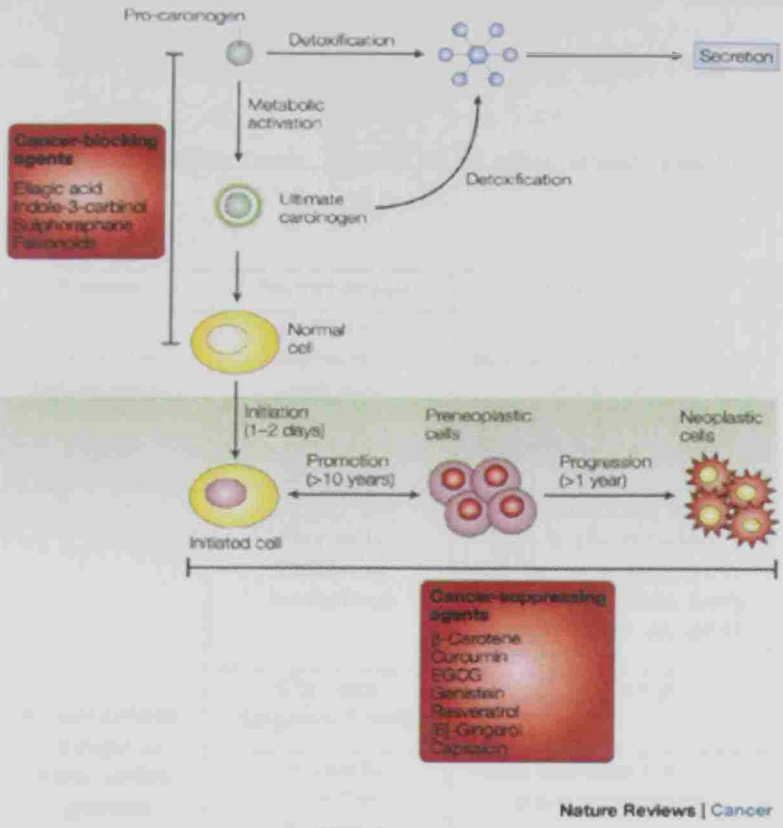


Figure 2.4: Dietary phytochemicals that block or suppress multistage carcinogenesis

Carcinogenesis is initiated with the transformation of the normal cell into a cancer cell. These cells undergo tumor promotion into pre-neoplastic cells, which progress to neoplastic cells. Phytochemicals can interfere with different steps of this process. Some chemopreventive phytochemicals inhibit metabolic activation of the pro-carcinogens to their ultimate electrophilic species, or their subsequent interaction with DNA. These agents therefore block tumor initiation. Alternatively, dietary-blocking agents can stimulate the detoxification of carcinogens, leading to their secretion from the body. Other

phytochemicals suppress the later steps of multistage carcinogenesis. Some phytochemicals can act as both blocking and suppressing agents (Surh, 2003).

Table 2.1:Common phytochemicals, identified by class, source and function

Class	Subclass	Phytochemical	Plant Source	Possible Function
1 Phenols	Flavonoids Anthocyanins (red, blue and purple pigments)	Cyanidin, delphinidin, malvidin, nasunin, pelargonidan, peonidin	Berries (esp. bilberries and blueberries), kale, wine, beans, red cabbage, eggplant	•Antioxidant. •May stimulate detoxification enzymes and protect liver. •Strengthens capillaries. •Blocks inflammation. •May inhibit tumor formation.
	Anthoxanthins (colorless, or whiteto-yellow pigments)	•Flavonols: Quercetin, kaempferol	Olives, onions, kale, lettuce, cherry tomato, broccoli, apples, yellow/green beans, turnip greens, endive tea, apple juice	
		•Flavones: Apigenin, luteolin	Celery, olives	
		•Flavanols: Catechin, epicatechin	Pears, red wine, green tea, white wine, apples	
		•Isoflavones: (phytoestrogens) Diadzein, genistein	Soybeans, lentils, dry peas, soy products	•May inhibit estrogen-promoted cancers • Lower serum cholesterol • May limit effects of menopause
		• Lignans: (phytoestrogens) Coumestrol,enterolactone, enterodiol, zearalenone, zearalenol, sesamin, sesamolinol	Flaxseed, linseed, wheat, barley, beans, peas, whole grains, fungi, tea, pumpkin seeds, sesame seeds, peanuts, cranberries	•Antioxidant •May protect against cancers. •Antiviral, antibacterial, antifungal. • May lower serum cholesterol
	Phenolic acid	Hydroxycinnamic acids: Caffeic, ferulic, chlorogenic, neochlorogenic acids	Blueberries, cherries, pears, apples, oranges, white potatoes, grapefruit, coffee beans	
		Hydroxybenzoic acids: Ellagic, gallic acids	Raspberries, strawberries, green and black grape juice	

	Polyphenols (tannins)	Catechin, epicatechin polymers	Lentils, black-eyed peas, light and dark grapes, red and white wine, apple juice	
2.Terpenes	Carotenoids	Alpha carotene, beta carotene	Carrots, cantaloupe, butternut squash, pumpkins, sweet potatoes	•Antioxidant • May protect against heart disease and stroke.
		Lycopene (red-orange pigments)	Tomatoes, red grapefruit, guava, dried apricots, watermelon	•Antioxidant. • May protect against cancers of the cervix, stomach, bladder, colon and prostate
		Lutein, zeaxanthin (xanthophylls)	Dark green leafy vegetables, broccoli, kiwi	•Antioxidant. • May protect against heart disease and stroke.
	Limonoids (monoterpenes)	Limonene, sobrerol, perillyl alcohol	Citrus fruit, citrus oil, caraway seeds	• May block action of carcinogens. • May inhibit hormone-related cancers
3.Thiols	Diallylsulfides	Diallyl disulfide, S-allyl cysteine, allicin	Garlic, onions, leeks, shallots, chives, broccoli, cabbage, cauliflower, Brussels sprouts, mustard, horseradish, watercress	• May block carcinogens and suppress carcinogenic changes in cells. • May stimulate detoxification enzymes.
	Indoles	Indole-3-carbidol (I3C)		• May protect against breast cancer.
	Isothiocyanates	Sulforaphane		• May protect against colon cancer • Stimulate detoxification enzymes • Protect against cigarette smoke

2.2.3 Phytochemical Research

Studies have looked at the relationship between cancer risk and phytochemicals It has been suggested that people can reduce their risk of cancer significantly by eating more herbs that contain beneficial phytochemicals. Specific studies show that

phytochemicals from the herb called *Origanum majorana* exhibits anticancer activity by promoting cell cycle arrest and apoptosis in breast cancer cell line (Al Dhaheri et al., 2013). Other studies have shown that saffron also has a significant chemopreventive effect against liver cancer through inhibition of cell proliferation and induction of apoptosis (Amin et al., 2011).

2.2.4 Phytochemical Action on Cancer Pathways

Many intracellular signal-transduction pathways converge with the activation of the transcription factors NF-κB and AP1, which act independently or coordinately to regulate target-gene expression. The chemopreventive phytochemicals that are derived from the diet have been shown to suppress constitutive NF-κB activation in malignant cells (Surh, 2003).

AP1 is another transcription factor that regulates expression of genes that are involved in cellular adaptation, differentiation and proliferation. Functional activation of AP1 is associated with malignant transformation as well as tumor promotion. AP1 consists of either homo- or heterodimers between members of the JUN and FOS families, which interact via a leucine-zipper domain. The MAPK-signaling cascade also regulates this transcription factor. Many of the molecular alterations that are associated with carcinogenesis occur in cell-signaling. One of the central components of the intracellular signaling network that maintains homeostasis is the family of proline-directed serine/threonine kinases the mitogen-activated protein kinases (MAPK) as shown in Figure 2.5 (Surh, 2003). Abnormal or improper activation or silencing of the MAPK pathway or its downstream transcription factors can result in uncontrolled cell growth, leading to malignant transformation.

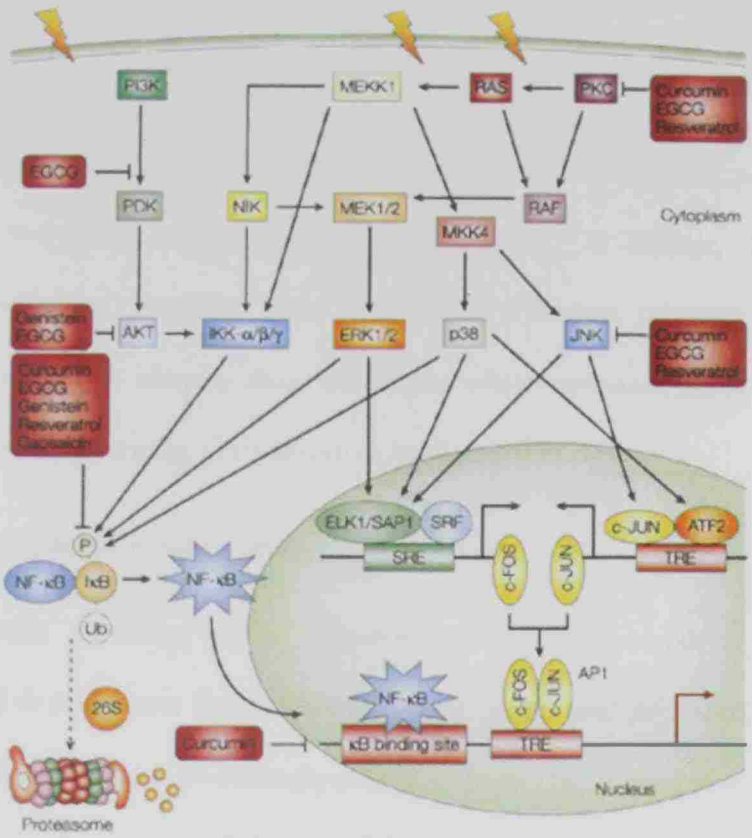


Figure 2.5: Effect of phytochemicals on activation of NF-KB and AP1

On the other hand, some phytochemicals can 'switch on' or 'turn off ' the specific signaling molecules, depending on the nature of the signaling cascade they target, preventing abnormal cell proliferation and growth(Surh, 2003).

To investigate how phytochemicals can be used to treat cancer, *in silico* molecular docking has been adopted in several studies. Phytochemical and target protein structures are collected and molecular docking algorithms are employed to investigate whether a phytochemical is likely to bind to a protein and if yes, the binding site and specific protein-chemical interactions that facilitate the binding.

2.3 *In Silico* Virtual Screening for Drug Discovery

Virtual screening is a central technique in drug discovery today. Many molecules can be tested *in silico* with the aim of selecting the most promising ones and testing them

experimentally which allow researchers to refine their experimental programs to reduce costs and increase research efficiency (Trisilowati & Mallet, 2012). Moreover, computational techniques have been used in drug discovery to search libraries of small molecules in order to identify those structures, which are most likely to bind to a drug target, by virtual screening. (Trisilowati & Mallet, 2012)

2.3.1 Advantages of *In Silico* Study

Computer modeling is an essential component in modern pre-clinical drug discovery and development. Instead of testing each compound in a large compound library experimentally by using high-throughput screening (HTS), virtual screening tools can be used to rank molecules on their probability of binding to the target. This leads to considerable savings in personnel and material costs as only a small number of molecules of the complete library need to be tested experimentally (Trisilowati & Mallet, 2012).

The reason of choosing computational study is that traditional laboratory-based cancer research involves expensive trial and error experimental strategies applied to humans, animals, and their tissues. However *in silico* experiments provides a novel approach to guide the early stages of hypothesis development and experimental design that has the potential to create subsequent efficiencies and cost savings in the laboratory. This computational approach is advantageous because it allows vast numbers of experiments to be carried out that are easily observed at any desired level of detail and can be repeated and controlled at will (Trisilowati & Mallet, 2012).

Such studies involving *in vitro* and *in vivo* animal experiments involve hypothesis generation and testing to determine whether further trials are warranted and are extremely costly both in terms of researchers time and the associated financial investment. Costs,

such as laboratory setup, equipment and space, time spent by academics training others, and the time, equipment, and materials costs involved in repetitive, hands-on experimental work, all contribute to the expense of laboratory-based experimental research (Trisilowati & Mallet, 2012).

Moreover, *in vitro* and *in vivo* models use actual biological materials and/or actual animals to investigate hypotheses and predict effectiveness of treatment strategies. *In silico* models use specifically designed computer programs to mimic these "real" experimental environments and to conduct computational experiments.

2.4 Molecular Docking

Molecular docking is a key tool in structural molecular biology and computer-assisted drug design (Morris and Lim-Wilby, 2008). In combination with virtual screening, molecular docking has become a standard tool in computational biology. This technique is widely applied for studying the interaction between ligands and proteins for potential hit compounds for protein targets. Traditionally, the docking process was thought of in terms of "lock-and-key". Ligands, including small molecules, DNA/RNA, lipids, and even other proteins serve as the "key", while the receptor serves as the "lock".

Once a compound is docked, it is then scored using mathematical models by estimating the binding affinity of the ligand-receptor complex. Different protocols for docking are attempted before determining the correct set of parameters to be used for docking.

2.4.1 Advantage of Molecular Docking

Molecular docking attempts to predict binding interactions and orientations. It is widely used in rational drug design and structure based drug design processes. Another

good reason why many researchers are moving towards docking methods in their research is because it is often difficult to discern the precise atomic level interactions through traditional experimental ways.

The ability of the computer to simulate the interactions in atomic details and with the increasing power of computers, *in silico* methods are able to provide the answer to several difficult research problems.

Evaluation of the docked and scored complexes leads to the selection of a short list of compounds that can be moved to the laboratory for *in vitro*, *ex vivo*, and *in vivo* testing. Thus, molecular docking plays an important role in the rational design of drugs.

2.4.2 Classification of Docking Programs

The docking process involves explicitly predicting the binding geometry and interactions between ligand and receptor and estimating the binding affinity of the ligand-receptor complex (scoring). Flexible docking programs can be broadly divided into two classes based on the conformational searching algorithms applied, including system methods like incremental construction, conformational search and stochastic/random methods such as Monte Carlo, genetic algorithms (Kontoyianni et al., 2004). Widely used docking programs, such as FlexX (Kramer et al., 1999) belong to the first class; while GOLD (Verdonk et al., 2003) and AutoDock (Goodsell et al., 1996) belong to the second one. GLIDE (Friesner et al., 2004) program uses a combination of conformational searching and stochastic algorithms. The two major components of any docking tool are the docking algorithm and the scoring function.

In this thesis the GOLD docking tools were used. GOLD is a docking program provided by the Cambridge Crystallographic Data Centre (CCDC). It uses a genetic

search algorithm (GA) based on Mendelian genetics for finding possible binding modes and to explore the full range of ligand conformational flexibility with partial flexibility of the protein, and satisfies the fundamental requirement that the ligand must displace loosely bound water on binding (Jones et al., 1997).

GOLD allows water molecules to be switched 'on' and 'off', displaced and to be rotated around their three principal axes to optimize hydrogen bonding during docking. It can automatically determine whether a water molecule should be bound or displaced by the ligand during docking. The orientation of the water hydrogen atoms can also be optimized by GOLD during docking. GOLD explores the conformational space of a ligand using genetic search algorithms in combination with various scoring functions.

2.4.3 Genetic Algorithm

Predicting which interactions a ligand makes with a protein in pose prediction step depends on the docking software used. The most popular options include genetic algorithms. Genetic algorithm is implemented in the widely used GOLD (Genetic Optimization for Ligand Docking) docking tool.

Genetic algorithms are applied to finding possible conformers. A conformer is encoded as a string of values (torsion angles) called chromosome. Details of different implementations differ but the general idea is the same (Tiikkainen, 2010).

The process starts with a number of random conformers, which are encoded in chromosomes. Each conformer (set of torsion angles) is evaluated with a fitness function, which quantifies the quality of the solution. The population of chromosomes is then subjected to a modification step where they can go through cross-over (two chromosomes swapping pieces) or point mutations (random change in one or more of

properties of the chromosome). Afterwards the resulting chromosomes are evaluated and the top scoring individuals are taken for another round of selection. The process is terminated after a given number of loops (Tiikkainen, 2010).

2.4.4 Scoring Functions

To quickly estimate the binding free energy of a given docking pose, mathematical functions, commonly known as scoring functions, are used. These can be categorized into three main groups: knowledge-based, empirical and molecular mechanics force field scoring functions (Khan et al., 2014).

In order to save computational time scoring functions only consider the direct contact interactions between the ligand and the protein, thus disregarding any long-range interactions. Scoring estimates the chemical interactions such as binding strength and energy state between the ligand and protein to assist in ranking the efficacy of the compound being scored. The score gives a guide as to how good the pose is; the higher the score, the better the docking result is likely to be. Ranking of docking is divided into internal and external ranking (Khan et al., 2014).

Internal ranking, as a molecule is docked to a site, several different poses are generated. Each of the poses is evaluated and scored using the selected scoring function. Thus, an internal ranking between all poses of the molecule in the site is created. The top ranked pose is the pose that the docking program suggests is the lowest energy "the best" pose of the molecule in that specific site. In theory, this would be the same pose the molecule would display in a crystal structure of the complex. However, external ranking, just as the different poses of one molecule in the same site are evaluated with scoring functions, the highest scores obtained for each molecule can then be compared to create

an external ranking, thus determining which molecule is the better binder (Khan et al., 2014).

2.5 Molecular Dynamics Simulation

Molecular docking is performed as an initial step followed by molecular dynamics (MD) simulations. Although molecular docking is a fast and efficient way to quickly estimate the binding mode of a ligand into a binding pocket, it is not the most accurate method for determining interaction energies. Molecular docking programs calculate their scores from a single ligand pose fitted into a snapshot of the protein. Although this method saves computational time, it can also be a source of inaccuracy. In molecular dynamics (MD) simulations the system is simulated under thermal fluctuations, such as it is in real biological systems (Khan et al., 2014).

Molecular dynamics is currently the most popular simulation approach. It carries out simulations on the computer that are difficult or impossible in the laboratory like working at extremes of temperature or pressure. This technique allows the study of biological and chemical systems at the atomistic level on timescales from femtoseconds to milliseconds. It complements experiment while also offering a way to follow processes difficult to discern with experimental techniques. Numerous software packages exist for conducting MD simulations of which one of the widely used one is AMBER (Salomon-Ferrer et al., 2013).

AMBER includes the collection of numerous programs that work together to setup, perform, and analyze MD simulations, from the preparation of the necessary input files, to the analysis of the results. The name AMBER also refers to a series of classical molecular mechanics force fields, primarily designed for the simulation of biomolecules.

This includes the amino acid and nucleic acid parameter sets, termed, for example, ff94, ff99SB, and ff12SB; carbohydrates termed Glycam; phospholipids termed Lipid; nucleic acids; and general organic molecules termed GAFF (Salomon-Ferrer et al., 2013).

All atoms are treated as spheres with attributes such as mass, partial charges and van der Waals radii allowing the calculation of the potential energy at each configuration. In MD, atoms are moved according to Newton's laws of motion and the force of each atom i at time t is calculated from the gradient (∇) of the force field potential energy function. The acceleration of each atom is calculated using Newton's second law and the velocity and time can be approximated from it using a truncated Taylor expansion (Khan et al., 2014).

$$a_i(t) = \frac{F_i}{m_i} = -\frac{1}{m_i}\nabla i\ U_{pot}$$

2.5.1 Molecular Dynamic Simulation Steps

The steps for a typical MD simulation of a protein, starts from a structure file obtained from the PDB before proceeding through minimization, equilibration, and production simulations to obtain meaningful results. The goal of the equilibration stage of an MD simulation is to separate the trajectory into two portions: one containing non-equilibrium fluctuations due to the initial non equilibrium structure, and one which is free of those fluctuations

Logically, a robust equilibration method has important characteristics. Method should provide an equilibration time that is independent of the duration of the simulation but solely on the physics of the system. Interestingly, equilibration procedures for structural studies of proteins have not changed significantly as computational resources have increased. When the RMSD reaches a plateau value, the system has reached a basin

in the potential energy surface. This energetic basin is assumed to correspond with the equilibrium phase space, and the system is deemed suitable for the production stage of molecular dynamics simulation (Walton & VanVliet, 2006).

2.6 Toxicity Data

In silico techniques appear to be the best choice to screen compounds with unknown toxic potential. However, *in silico* modeling of toxicity data is extremely challenging due to the complexity and considerable variability of the data itself.

The major challenge in toxicity modeling is that all molecules are toxic at some level. It is therefore necessary that an *in silico* method can predict both the type of toxicity, as well as the level of toxicity of a compound. Since there are numerous different ways in which toxicity can arise, the prediction of the absolute toxic potential of a compound, either from animal models, or via *in vitro* or *in silico* methods, is extremely challenging (Gleeson et al., 2012).

In addition, the level of toxicity for a given toxicity mechanism may also depend on the level and distribution of drug in the body (ADME processes), making accurate toxicity predictions even more difficult. The Osiris Property Explorer was used in this study (http://www.openmolecules.org/propertyexplorer/applet.html). The OSIRIS Property Explorer can be used to draw chemical structures and calculate on-the-fly various drug-relevant properties (cLogP, solubility, Molecular Weight, Toxicity Risk Assessment (including mutagenicity, tumorogenicity, irritant and reproductive effect, overall Drug-Score, etc) whenever a structure is valid to ensure that the drug is sufficient to achieve proof of concept, efficacy and safety in clinical trials.

Prediction results are valued and color-coded. Properties with high risks of undesired effects like mutagenicity or a poor intestinal absorption are shown in red whereas a green color indicates drug-like behavior.

2.7 *In Silico* Cancer Research

Virtual screening has been used to investigate anti-breast cancer activity of low molecular weight compounds present in wild mushroom (Goodsell et al., 1996). This study confirmed that cinnamic acid derivatives esterified with quinic acid caused significant inhibition of the different proteins involved in overproduction of estrone and estradiol.

Another study created a flavonoid compound database that is found in fruits to gain further understanding of the anticancer and therapeutic effects exerted by these phytochemicals. Computational methods were used to select chemicals that target a particular protein or enzyme and to determine potential protein targets for well characterized as well as novel phytochemicals. They found that isorhamnetin an O-methylated flavonol in herbal medicinal plants could inhibit the kinase activity of MEK1 or PI3-K and the inhibition was due to isorhamnetin's direct binding with these kinases (Chen et al., 2012).

In silico docking techniques are being used to investigate the complementarity at the molecular level of a ligand and a protein target. In a study, nine ligands, which have been isolated and identified from the ethanolic extract of the fruits of *Cucumis trigonus Roxb*, and the standard hepatoprotective agent silymarin, were docked with two novel hepatocarcinoma receptors, Hepatitis B X and Heme Oxygenase 1 (Gopalakrishnan et al, 2015). Out of the nine phytochemical constituents isolated and identified from the

ethanolic extract of the fruits of *Cucumis trigonus*, demeclocycline had the best fitness score compared with the standard drug, silymarin. This suggests that demeclocycline could be an effective potential inhibitor against Hepatitis B X receptor and can be evaluated as hepatoprotective drug molecule using molecular docking studies. These effective properties may be due to the presence of carbonyl and alcoholic-OH functional groups present in the ligand molecule. (Gopalakrishnan et al, 2015).

Among other proteins, leucine-rich alpha-2-glycoprotein (LRG1) plays an important role in the creation of bad blood vessels. So, targeting this protein may be effective in the treatment of cancer. A recent *in silico* study done by Das and colleagues used the phytochemicals beta-carotene, anthocyanins, genistein, elargic acid, limonene, allicin, curcumin, which were docked with this protein. It was identified that all these phytochemicals have good binding affinity and Ki value with a high H-bond energy. Curcumin, genestein and carotinoid showed a high docked energy and intermolecular energy. The H-bond shows a strong interaction with phytochemicals as all were found within 3 Å (Das & Jagadeb, 2015).

Developing molecular drug candidates against a given disease needs proper understanding of the metabolic pathways. Therefore, identification and development of potential ligands or phytochemical specifically for a protein target forms the primary goal in drug discovery process. Most of these studies are at a small scale using a few chemicals and a few proteins. In this project, several phytochemicals and protein targets were studied in an *in silico* molecular docking and molecular dynamics study and ADMET properties were also predicted for these compounds.

Chapter 3: Methodology

Computational techniques provide efficient and inexpensive tools to gain further understanding of the anticancer and therapeutic effects of phytochemicals. This chapter describes the main steps involved starting from obtaining and preparing ligands and target proteins. Molecular docking was done using the GOLD package to find the best-docked pose. Molecular dynamics simulations using AMBER was performed to study the stability of the docked pose and finally toxicity testing was undertaken using the Osiris Property Explorer to identify any side effect of using these ligands as drugs.

3.1 Obtaining and Preparation of Target Proteins

Protein Data Bank (PDB) (see Figure 3.1) is a structural database that contains information about experimentally determined structures of proteins and other macromolecules. Structures of the seven proteins identified as targets in cancer pathways were obtained from the PDB. This included Akt murine thymoma viral oncogene homolog 1 (AKT1), Cyclin Dependent Kinase 2 (CDK2), Epidermal growth factor receptor (EGFR), Extracellular regulated 2 (ERK2), Janus Kinase 1 (JAK1), Mitogen Activated protein kinase kinase 1 (MAPKK1) and Ras-related protein (Rab5A).

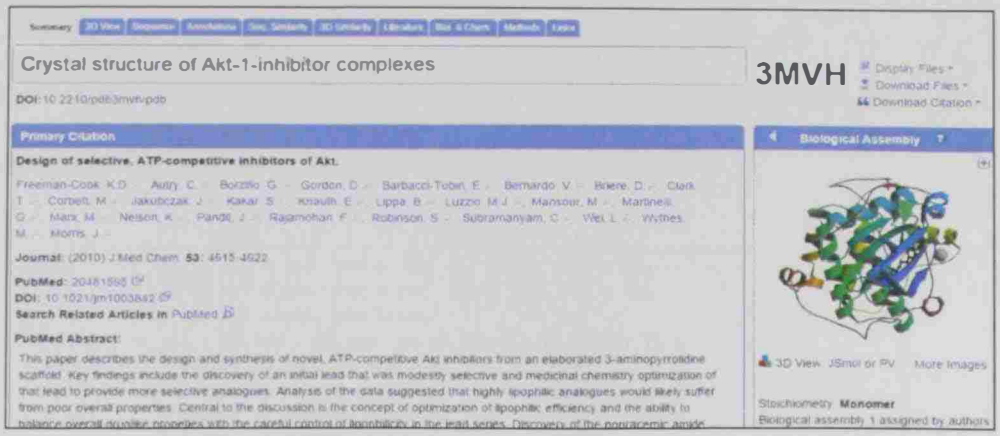


Figure 3.1: Protein Data Bank (PDB) showing the crystal structure of AKT1

High-resolution structures of the 7 proteins from humans (*Homo sapiens*) were selected with a resolution less than 2.5 Å. The PDB ID and the related X-ray data are shown in Table 3.1.

No	Protein name	PDB ID	Resolution (Å)
1	AKT1	3MVH	2.01
2	CDK2	4ACM	1.63
3	EGFR	3W2S	1.9
4	ERK2	2OJG	2
5	JAK1	4EHZ	2.17
6	MAPKK1	4AN3	2.1
7	Rab5A	1N6H	1.5

Table 3.1: List of protein structures obtained from PDB

After downloading appropriate protein structures from the PDB, for docking, one protein chain from each PDB file was used. Each file was processed to remove metadata information like HEADER, TITLE and AUTHOR, SITE records, etc. as shown in Figure 3.2. All lines that do not begin with ATOM keyword, which represents the coordinates of

the atoms in the structure, in PDB files were removed keeping only the protein chain that was necessary.

```
SITE     3 AC2 18 GLU A 228  TYR A 229  ALA A 230  GLU A 234
SITE     4 AC2 18 MET A 281  THR A 291  ASP A 292  PHE A 438
SITE     5 AC2 18 HOH A1058  HOH A1149
CRYST1   42.354   55.261   92.314  90.00 102.64  90.00 P 1 21 1      2
ORIGX1      1.000000  0.000000  0.000000        0.00000
ORIGX2      0.000000  1.000000  0.000000        0.00000
ORIGX3      0.000000  0.000000  1.000000        0.00000
SCALE1      0.023611  0.000000  0.005294        0.00000
SCALE2      0.000000  0.018096  0.000000        0.00000
SCALE3      0.000000  0.000000  0.011102        0.00000
ATOM      1  N   VAL A 145      10.652   4.980  47.052  1.00 53.88           N
ATOM      2  CA  VAL A 145      11.526   4.057  46.227  1.00 52.50           C
ATOM      3  C   VAL A 145      12.984   4.530  46.004  1.00 52.00           C
ATOM      4  O   VAL A 145      13.219   5.623  45.501  1.00 52.53           O
ATOM      5  CB  VAL A 145      10.815   3.665  44.885  1.00 52.73           C
ATOM      6  CG1 VAL A 145      11.810   3.168  43.877  1.00 50.64           C
ATOM      7  CG2 VAL A 145       9.781   2.577  45.157  1.00 51.81           C
ATOM      8  N   THR A 146      13.935   3.665  46.370  1.00 51.46           N
ATOM      9  CA  THR A 146      15.391   3.905  46.358  1.00 50.94           C
ATOM     10  C   THR A 146      16.060   2.589  45.923  1.00 51.24           C
ATOM     11  O   THR A 146      15.348   1.564  45.779  1.00 50.68           O
ATOM     12  CB  THR A 146      15.947   4.285  47.884  1.00 51.52           C
ATOM     13  OG1 THR A 146      15.890   3.146  48.686  1.00 50.07           O
ATOM     14  CG2 THR A 146      15.167   5.441  48.445  1.00 51.02           C
ATOM     15  N   MET A 147      17.396   2.571  45.751  1.00 51.03           N
ATOM     16  CA  MET A 147      18.140   1.318  45.461  1.00 52.19           C
ATOM     17  C   MET A 147      17.837   0.151  46.432  1.00 52.73           C
ATOM     18  O   MET A 147      18.012  -1.020  46.059  1.00 52.34           O
ATOM     19  CB  MET A 147      19.672   1.479  45.453  1.00 52.15           C
ATOM     20  CG  MET A 147      20.328   2.514  44.529  1.00 53.58           C
ATOM     21  SD  MET A 147      20.296   2.189  42.765  1.00 55.63           S
ATOM     22  CE  MET A 147      21.258   0.694  42.561  1.00 50.40           C
ATOM     23  N   ASN A 148      17.421   0.483  47.667  1.00 52.92           N
ATOM     24  CA  ASN A 148      17.112  -0.518  48.712  1.00 53.22           C
ATOM     25  C   ASN A 148      15.762  -1.199  48.493  1.00 51.82           C
ATOM     26  O   ASN A 148      15.452  -2.197  49.166  1.00 52.16           O
ATOM     27  CB  ASN A 148      17.121   0.102  50.151  1.00 53.98           C
ATOM     28  CG  ASN A 148      18.410   0.869  50.470  1.00 58.46           C
ATOM     29  OD1 ASN A 148      19.532   0.345  50.402  1.00 62.34           O
ATOM     30  ND2 ASN A 148      18.250   2.185  50.819  1.00 61.18           N
```

Figure 3.2: AKT1 PDB file with all the records shown

Figure 3.2 shows the coordinates of the atoms that are part of the protein. Some of the header records are also shown. ATOM records describe the protein atoms. In this example the first residue of peptide chain A, is a valine residue with a residue number 145. The second column represents a running counter for atoms in the PDB file and the third column represents the atom type in the residue. The first three floating-point numbers are the x, y and z coordinates of the atom and the unit is Angstroms (Å). The

next three columns are the occupancy, temperature factor, and the element name, respectively.

```
REMARK 465 MISSING RESIDUES
REMARK 465 THE FOLLOWING RESIDUES WERE NOT LOCATED IN THE
REMARK 465 EXPERIMENT. (M=MODEL NUMBER; RES=RESIDUE NAME; C=CHAIN
REMARK 465 IDENTIFIER; SSSEQ=SEQUENCE NUMBER; I=INSERTION CODE.)
REMARK 465
REMARK 465   M RES C SSSEQI
REMARK 465     GLY A   139
REMARK 465     ALA A   140
REMARK 465     MET A   141
REMARK 465     ASP A   142
REMARK 465     PRO A   143
REMARK 465     ARG A   144
REMARK 465     MET A   446
REMARK 465     ILE A   447
REMARK 465     THR A   448
REMARK 465     ILE A   449
REMARK 465     THR A   450
REMARK 465     PRO A   451
REMARK 465     PRO A   452
REMARK 465     ASP A   453
REMARK 465     GLN A   454
REMARK 465     ASP A   455
REMARK 465     ASP A   456
REMARK 465     SER A   457
REMARK 465     MET A   458
REMARK 465     GLU A   459
REMARK 465     CYS A   460
REMARK 465     VAL A   461
REMARK 465     ASP A   462
REMARK 465     SER A   463
REMARK 465     GLU A   464
REMARK 465     ARG A   465
REMARK 465     SER A   477
REMARK 465     SER A   478
REMARK 465     THR A   479
REMARK 465     ALA A   480
```

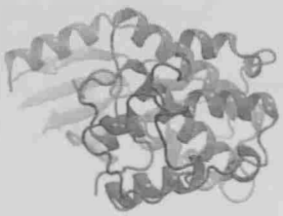
Figure 3.3: PDB file of AKT1 (ID: 3MVH)

Some of the PDB files obtained had amino acids residues missing in the protein structure. Figure 3.3 shows the missing sequence residues of AKT1 protein from 139-144 and 446-480.

Figure 3.4: WHAT IF server

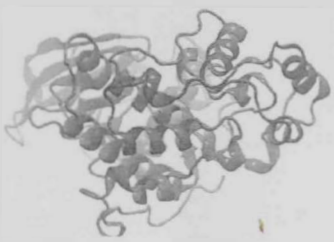
This may happen when a side chain or back bone is highly mobile or disordered. To fix this problem the missing side chains were completed by using 'WHAT IF' (http://swift.cmbi.ru.nl/servers/html/index.html) that is capable of building a model of the missing chains as shown in Figure 3.4. The final structure of the proteins viewed in Visual Molecular Dynamics (VMD) tool is shown in Figure 3.5.

A.AKT1(3MVH)

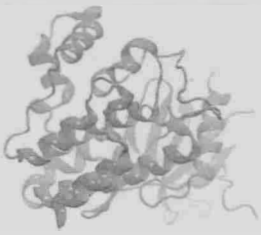


B. CDK2 (4ACM)

C.EGFR (3W2S)



D. ERK2 (2OJG)



E. JAK1 (4EHZ)

F. MAPKK1(4AN3)

F. MAPKK1(4AN3)

Figure 3.5: Fixed chain proteins viewed using VMD

3.2 Obtaining and Preparation of Ligands

Phytochemical structures were obtained from PubChem (https://pubchem.ncbi.nlm.nih.gov), which is a database of chemical molecules, their properties and activity (Figure 3.6)

Figure 3.6: Chemical structure of alpha carotene from PubChem

Research over the last decade has shown that several phytochemicals in fruits and vegetables could have an effect in cancer. Phytochemical groups such as carotenoids, flavonoids and indoles have been chosen in this work based on their anticancer effect. Sixty-seven chemicals were obtained in this project as shown in Table 3.2. Details about the plant source and chemical group of each of the phytochemical are provided in this table.

No.	Chemical Name	Molecular Formula	MW (g/mol)	PubChem ID	Plant name	Chemical Group
1	Apigenin	$C_{15}H_{10}O_5$	270.2	5280443	parsley, celery and chamomile	flavone
2	Alpha Carotene	$C_{40}H_{56}$	536.8	4369188	Pumpkin, Carrots and Broccoli	Carotenoids
3	Beta Sitosterol	$C_{29}H_{50}O$	414.7	222284	Corn, soybean and Nuts	Phytosterols
4	Betulinic Acid	$C_{30}H_{48}O_3$	456.7	64971	White birch tree	Triterpene
5	Boswellic Acid	$C_{30}H_{48}O_3$	456.7	168928	Gum resin	Triterpene
6	Canthaxanthin	$C_{40}H_{52}O_2$	564.8	5281227	Carrots, pumpkinsand sweet potatoes	terpenoids
7	Capsaicin	$C_{18}H_{27}NO_3$	305.4	1548943	Chilli pepper	Capsaicinoid
8	Carnosol	$C_{20}H_{26}O_4$	330.4	2579	Rosemary	Phenolic Diterpene
9	Chlorogenic Acid	$C_{16}H_{18}O_9$	354.3	1794427	Coffee bean, bamboo and potatoes	Hydroxycinnamic acids
10	Cryptoxanthin	$C_{40}H_{56}O$	552.8	5281235	Citrus fruit, mangos and papaya	Carotenoids
11	Curcumin	$C_{21}H_{20}O_6$	368.3	969516	Turmeric plant	Diarylheptanoid
12	Daidzein	$C_{15}H_{10}O_4$	254.2	5281708	Kwao Krua and Kudzu	Isoflavones
13	Delphinidin	$C_{15}H_{11}O_7^+$	303.2	128853	Pomegranates, grape and Cranberries	Anthocyanidin
14	Delta Carotene	$C_{40}H_{56}$	536.8	5281230	Pumpkin, Carrots and Broccoli	Carotenoids
15	EGCG	$C_{22}H_{18}O_{11}$	458.3	65064	White tea, green tea	catechin
16	Ellagic Acid	$C_{14}H_6O_8$	302.2	5281855	Raspberries, Strawberries, Cranberries	hexahydroxydiphenic acid.
17	3,3'-Diindolylmethane	$C_{17}H_{14}N_2$	246.3	3071	Cruciferous vegetables	Indole-3-carbinol
18	Acacetin	$C_{16}H_{12}O_5$	284.2	5280442	Black locust (Robinia pseudoacacia) and damiana (Turnera diffusa)	O-methylated flavone
19	Anacardic Acid	$C_{22}H_{36}O_3$	348.5	167551	Cashew nut (Anacardium occidentale)	Phenolic lipids,
20	Chrysin	$C_{15}H_{10}O_4$	254.2	5281607	Passion flowers (Passiflora caerulea)	Flavone
21	Cyanidin	$C_{15}H_{11}O_6^+$	287.2	128861	Red berries, red cabbage and red onion	anthocyanidin
22	Diosmetin	$C_{16}H_{12}O_6$	300.2	5281612	Caucasian vetch	O-methylated flavone
23	Emodin	$C_{15}H_{10}O_5$	270.2	3220	Rhubarb, buckthorn and Japanese knotweed	Resin
24	Gamma Linolenic Acid	$C_{18}H_{30}O_2$	278.4	5280933	Oenothera	Carboxylic acid

25	Gamma Carotene	$C_{40}H_{56}$	536.8	5280791	Pumpkin, Carrots and Broccoli	Carotenoids
26	Garcinol	$C_{38}H_{50}O_6$	602 8	5490884	Garcinia indica	polyisoprenylated benzophenone
27	Genistein	$C_{15}H_{10}O_5$	270.2	5280961	Dyer's broom and *Genista tinctoria*	Isoflavones
28	Gingerol	$C_{17}H_{26}O_4$	294.3	3473	Ginger, chilli peppers and black pepper	ketone, phenol
29	Glycitein	$C_{16}H_{12}O_5$	284.2	5317750	Soy	O-methylated isoflavone
30	Isorhamnetin	$C_{16}H_{12}O_7$	316.2	5281654	*Tagetes lucida*	O-methylated flavonol,
31	Myricetin	$C_{15}H_{10}O_8$	318.2	5281672	Bayberry	flavonol
32	Oleanolic Acid	$C_{30}H_{48}O_3$	456.7	10494	Olive oil, American pokeweed	triterpenoid
33	Oleocanthal	$C_{17}H_{20}O_5$	304.3	11652416	Olive oil	phenylethanoid
34	Peonidin	$C_{16}H_{13}O_6^+$	301.2	441773	Morning glory and peony flowers	O-methylated anthocyanidin
35	Piceatannol	$C_{14}H_{12}O_4$	244.2	667639	Norway spruces and *Aiphanes horrida*	stilbenoid
36	Pinoresinol	$C_{20}H_{22}O_6$	358.3	234817	Sesame seed, Brassica vegetables and Forsythia suspensa	lignan
37	Pterostilbene	$C_{16}H_{16}O_3$	256.2	5281727	Blueberries and grapes	stilbenoid
38	Resveratrol	$C_{14}H_{12}O_3$	228.2	445154	Red grapes and Japanese knotweed (Polygonum cuspidatum)	stilbenoid
39	Stigmasterol	$C_{29}H_{48}O$	412.6	5280794	American ginseng, soybean, calabar bean, and rape seed	phytosterol
40	Tangeritin	$C_{20}H_{20}O_7$	372.3	68077	Citrus peels	O-polymethoxylated flavone
41	Theaflavin	$C_{29}H_{24}O_{12}$	564.4	114777	Black tea	Antioxidant polyphenols
42	Guggulsterone	$C_{21}H_{28}O_2$	312 4	6439929	Guggul plant	Steroid
43	Hesperidin	$C_{28}H_{34}O_{15}$	610.5	10621	Citrus fruits (oranges, lemons or pummelo fruits)	flavanone glycoside
44	Indicaxanthin	$C_{14}H_{16}N_2O_6$	308.2	71308222	Beets, *Mirabilis jalapa* flowers, prickly pears and red dragon fruit	betaxanthin
45	Isobetanin	$C_{24}H_{26}N_2O_{13}$	550.4	6325438	Beetroot and Amaranthus spp	Betacyanin
46	Kaempferol	$C_{15}H_{10}O_6$	286.2	5280863	Tea, broccoli, Delphinium, Witch-hazel and grapefruit,	flavonol
47	Lariciresinol	$C_{20}H_{24}O_6$	360.4	332427	Sesame seeds and in Brassica vegetables	phenylpropanoids.

48	Lutein	$C_{40}H_{56}O_2$	568.8	5368396	Spinach, kale and yellow carrots.	Xanthophyll
49	Luteolin	$C_{15}H_{10}O_6$	286.2	5280445	*Terminalia chebula*, rinds, barks, clover blossom, and ragweed pollen.	Flavone
50	Macrocarpal A	$C_{28}H_{40}O_6$	472.6	454457	*Eucalyptus globulus Labill*	Phloroglucinol Dialdehyde, Diterpene
51	Matairesinol	$C_{20}H_{22}O_6$	358.3	119205	Oil seeds, whole grains, vegetables, and fruits	lignan
52	Myrciaphenone B	$C_{21}H_{22}O_{13}$	482.3	183139	*Myrcia multiflora*	Phenolic glycoside
53	Naringenin	$C_{15}H_{12}O_5$	272.2	932	Grapefruit and tomato	flavanone
54	Pelargonidin	$C_{15}H_{11}O_5^+$	271.2	440832	Red geraniums, berries and pomegranates.	anthocyanidin
55	Quercetin	$C_{15}H_{10}O_7$	302.2	5280343	*Sophora japonica*	flavonol
56	Vulgaxanthin	$C_{14}H_{17}N_3O_7$	339.3	5281217	*Mirabilis jalapa* and swiss chard	betaxanthins
57	Ursolic Acid	$C_{30}H_{48}O_3$	456.7	64945	*Mirabilis jalapa*	triterpene acid
58	Trametenolic Acid	C30H48O3	456.7	56841132	*Trametes lactinea*	Tetracyclic Triterpene
59	Rosmarinyl Glucoside	$C_{24}H_{26}O_{13}$	522.4	11606086	Rosmarinic acid of rosemary, sage, Spanish sage, oregano,thyme, and mint	Polyphenol
60	Rubixanthin	$C_{40}H_{56}O$	552.8	5281252	Rose hips	Xanthophyll
61	Rutin	$C_{27}H_{30}O_{16}$	610.5	5280805	*Carpobrotus edulis*and buckwheat	Flavonoid glycoside
62	Scutellarein	$C_{15}H_{10}O_6$	286.2	5281697	*Scutellaria lateriflora*	Flavone
63	Silibinin	$C_{25}H_{22}O_{10}$	482.4	31553	Milk thistle	Flavonoid
64	Silybin B	$C_{25}H_{22}O_{10}$	482.4	1548994	Milk thistle	Flavonoid
65	Subaphylline	$C_{14}H_{20}N_2O_3$	264.3	5281796	Corn fiber	Alkaloids
66	Zeaxanthin	$C_{40}H_{56}O_2$	568.8	5280899	Corn, saffron, wolfberries,	Xanthophyll
67	Yakuchinone	$C_{20}H_{24}O_3$	312.4	133145	*Alpinia oxyphylla*	diarylheptanone glucoside

Table 3.2: List of phytochemical used in Molecular Docking study

2D structures of the 67 chemicals were obtained from PubChem in SDF format. This was converted to 3D format and its energy was minimized using PyRx

(http://pyrx.sourceforge.net). Energy minimization of the ligand structures is performed to prevent atoms from being unfavorably close to each other as shown in Figure 3.7. Ligand files were then converted to MOL2 format using OpenBabel (https://pypi.python.org/pypi/openbabel) to prepare it for molecular docking as shown in Figure 3.8.

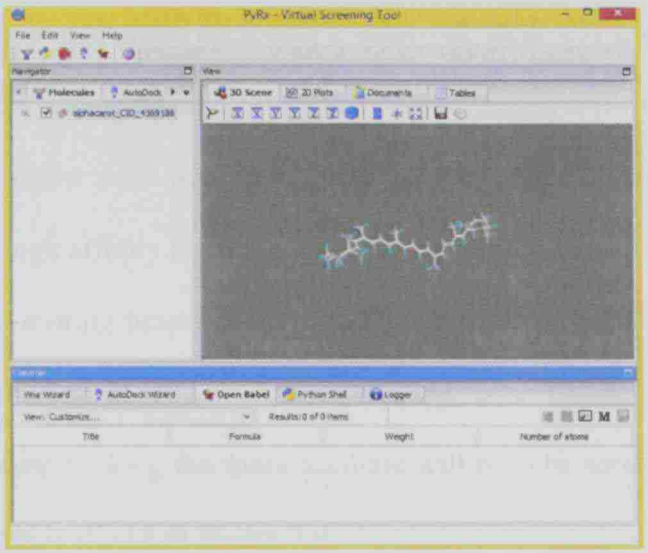


Figure 3.7: Preparing Alpha Carotene by minimizing the structure using PyRx

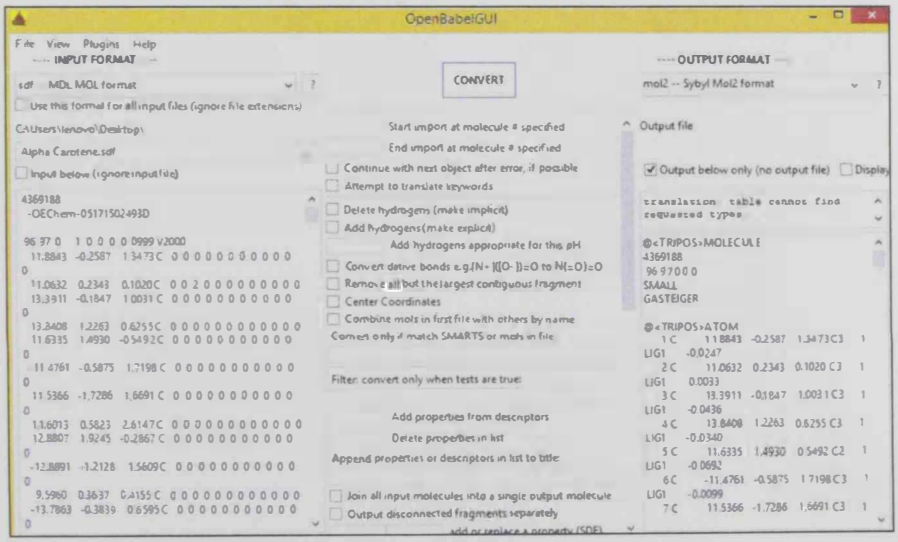
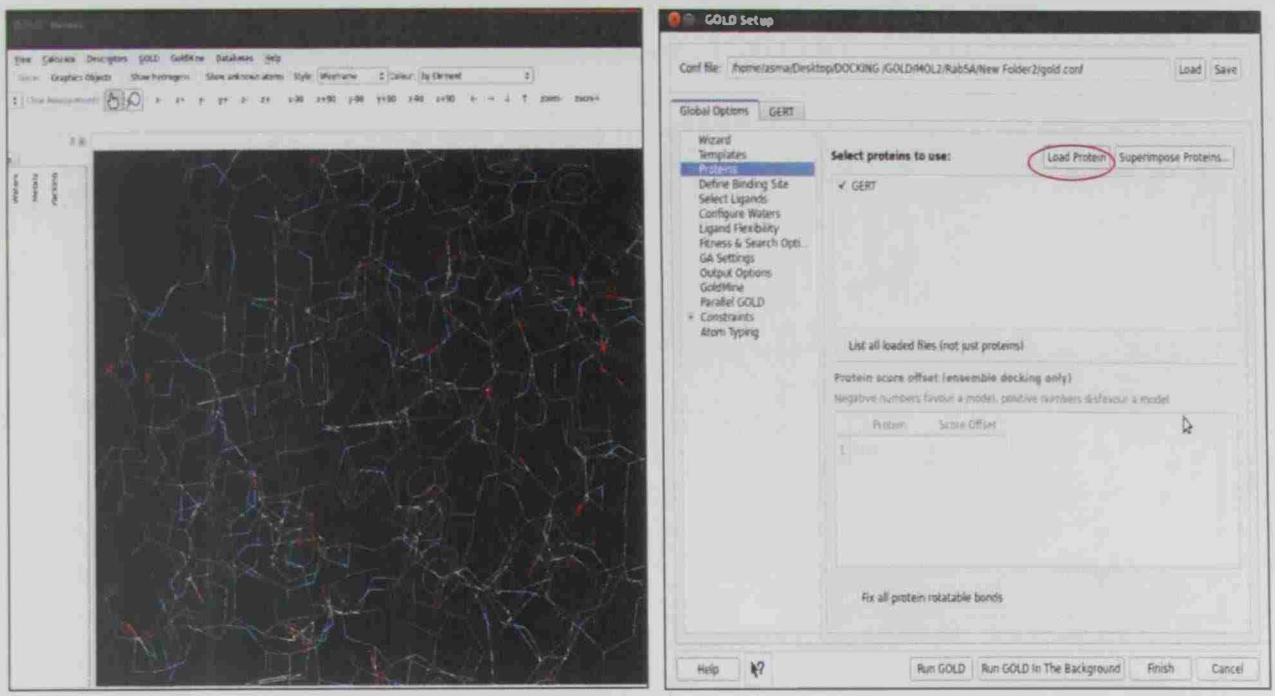


Figure 3.8: Converting sdf format to mol2 using Open Babel

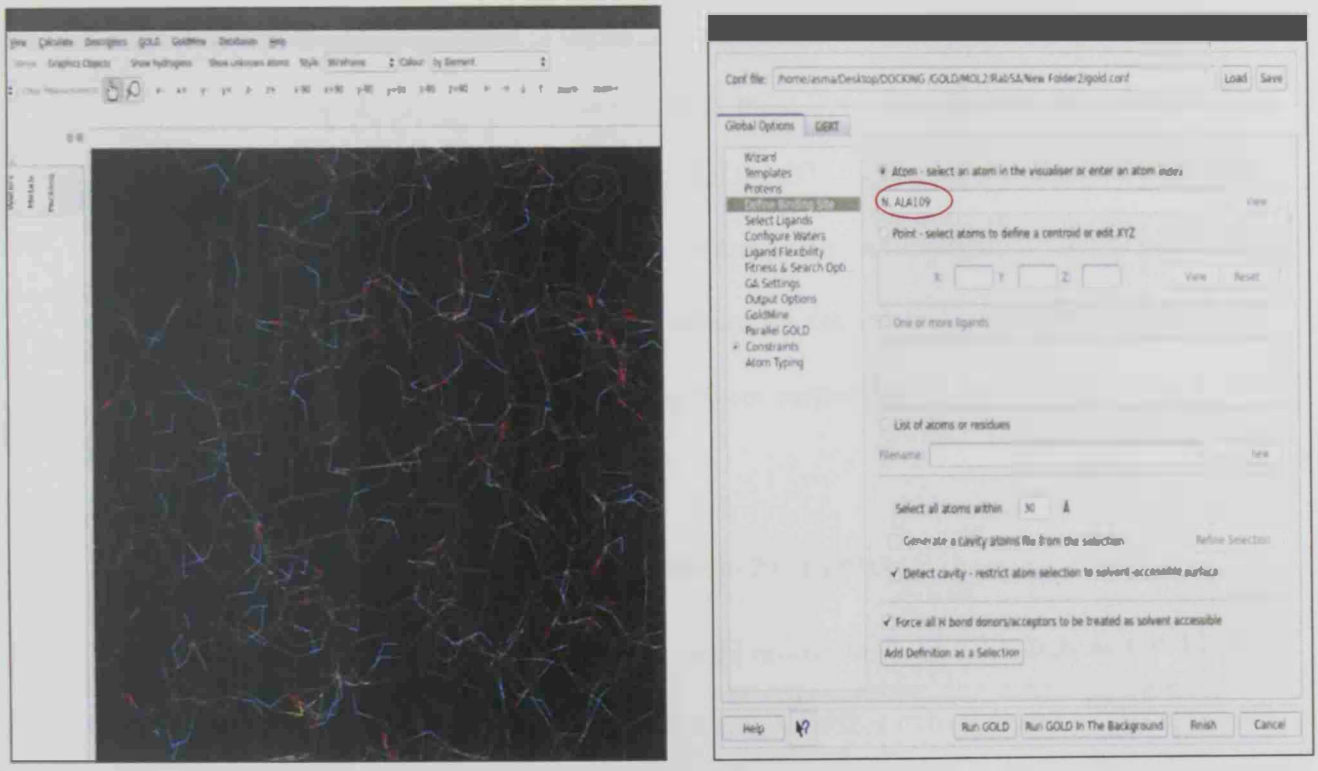
3.3 Molecular Docking of Target Protein with Ligands

Genetic Optimization for Ligand Docking (GOLD) was used to dock the selected ligands to the proteins in a flexible manner. Proteins that were prepared previously were loaded and hydrogen atoms were added to it. Blind docking was used to detect possible binding sites and modes of the ligands by scanning the entire volume of the protein targets. For this a search space enclosing a sphere of 30 Å was used to cover the whole protein. The conformational space of a ligand was explored using the genetic algorithm. 100 genetic algorithm runs were performed in each docking experiment in order to ensure that most high affinity binding modes would be explored.

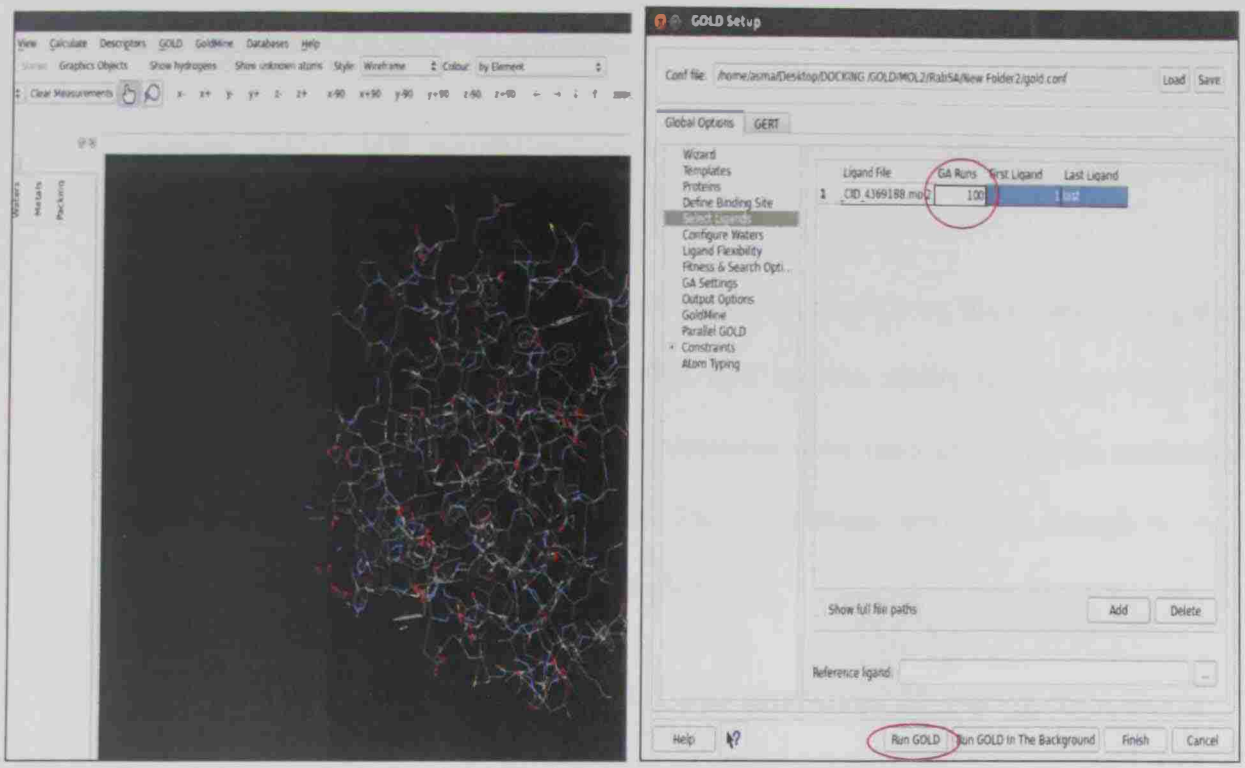
ChemPLP scoring functions was used to quantitatively evaluate the quality of the docking. The time taken to perform a docking is a balance between speed and accuracy, the slower docking the more accurate will be. The series of steps followed to setup a docking run is shown in Figure 3.9.


GOLD Setup
Conf file: /home/asma/Desktop/DOCKING /GOLD/MOL2/Rab5A/New Folder2/gold.conf
Load
Save
Global Options
GERT
Wizard
Templates
Proteins
Define Binding Site
Select Ligands
Configure Waters
Ligand Flexibility
Fitness & Search Opti...
GA Settings
Output Options
GoldMine
Parallel GOLD
Constraints
Atom Typing
Select proteins to use:
Load Protein
Superimpose Proteins...
GERT
List all loaded files (not just proteins)
Protein score offset (ensemble docking only)
Negative numbers favour a model, positive numbers disfavour a model
Protein
Score Offset
Fix all protein rotatable bonds
Help
Run GOLD
Run GOLD In The Background
Finish
Cancel

A.


Conf file: /home/asma/Desktop/DOCKING /GOLD/MOL2/Rab5A/New Folder2/gold.conf
Load
Save
Global Options
GERT
Wizard
Templates
Proteins
Define Binding Site
Select Ligands
Configure Waters
Ligand Flexibility
Fitness & Search Opti
GA Settings
Output Options
GoldMine
Parallel GOLD
Constraints
Atom Typing
Atom - select an atom in the visualiser or enter an atom index
N ALA109
Point - select atoms to define a centroid or edit XYZ
One or more ligands
List of atoms or residues
Filename
Select all atoms within
Generate a cavity atoms file from the selection
Refine Selection
Detect cavity - restrict atom selection to solvent-accessible surface
Force all H bond donors/acceptors to be treated as solvent accessible
Add Definition as a Selection
Help
Run GOLD
Run GOLD In The Background
Finish
Cancel

B.



C.

Figure 3.9: Main steps of GOLD docking.

3.3.1 Analysis Methodology

After the docking run, the best-ranked pose from the 100 runs was chosen. The best pose was analyzed using LIGPLOT (http://www.ebi.ac.uk/thornton-srv/software/LIGPLOT/) and VMD (http://www.ks.uiuc.edu/Research/vmd/). LIGPLOT projects the 3D structure to a 2D image where all the interactions between ligand and protein could be identified. This allows a closer inspection of the hydrogen bonds and hydrophobic interactions.

3.4 Molecular Dynamics Simulation of Docked Complex

Molecular docking predicts how a small molecule will bind to a protein but this is not easy, and no program can guarantee success. A further enhancement to this approach is to measure as accurately as possible the reliability of the dock using molecular

dynamics (MD). MD simulations have to be sufficiently long to draw reliable conclusions.

MD simulations were used for identifying and confirming the binding poses that were found from GOLD docking. AMBER tool has the ability to perform long and accurate MD simulations. AMBER MD simulations were used to predict the molecular motions of the protein-ligand complexes. The simulations were long enough so that representative conformations could be sampled.

The docked files were prepared for MD simulations by placing them first in a cubic box filled with water molecules and salt. Next, energy minimization of the system was performed and then slowly the system was heated to room temperature and pressure. Finally, production MD simulations were performed at the desired temperature and pressure.

3.4.1 Minimization Stage

Before starting MD simulations, the energy of the structures were minimized. Minimization removes any unfavorable interactions that may exist, which might otherwise lead to unstable simulation. The parameters used for the energy minimization stage is shown in Figure 3.10.

```
Minimization to relax initial bad contacts, explicit solvent
 &cntrl
   imin=1,
   ncyc=1000,
   maxcyc=5000,
   ntpr=50,
   cut=8,
   restraintmask='!:WAT&@CA,C,O,N'
   restraint_wt=10.0,
 /
```

Figure 3.10: MD simulation Step 1: minimization stage

As showed in Figure 3.10, to run initial minimization, minimization was turned on using the setting imin = 1. Minimization was switched from steepest descent to conjugate gradient after 1000 steps (ncyc) and the total number of minimization steps was 5000 (maxcyc). The intermediate results was saved every 50 steps (ntpr). A nonbonded cutoff of 8 Å (cut) was used. Water, CA, C, O, N atoms were restrained (restrain mask) and the force constant for restraint was 10 kcal/mol/$Å^2$ (restraint_wt). This ensures that the protein heavy atoms do not deviate significantly from the original structure.

3.4.2 Heating the system

After the system was minimized, the next stage was to allow the system to heat up to normal temperature of 300 K.The temperature of the system was raised from 10 K to 300 K gradually using the following parameters. .

```
Explicit solvent initial heating mdin
 &cntrl
   imin=0, irest=0, ntx=1,
   ntpr=1000, ntwx=1000, nstlim=200000,
   dt=0.002, ntt=3, gamma_ln=5.0, ig=-1,
   ntc=2, ntf=2, cut=8, ntb=1,
   iwrap=1, ioutfm=1, nmropt=1,
 /
 &wt
   TYPE= TEMP0 , ISTEP1=0, ISTEP2=150000,
   VALUE1=10.0, VALUE2=300.0,
 /
 &wt TYPE= END /
```

Figure 3.11: MD simulation Step 2: Heating stage

The input file for this phase is shown in Figure 3.11. Minimization was turned off (imin=0) and the initial coordinates were read from the inpcrd file by ntx=1 which is the output file from the previous stage, i.e. minimization. Every 1000 steps (ntpr) the energy readings were written to the output files while the coordinates of atoms were deposited into the trajectory file every 1000 step (ntwx). The total number of MD steps

performedwas 200,000 (nstlim) with a time step of 2 fs per step (dt). The heating was started at 10K (value1) and continued until the desired temperature 300K was reached.

3.4.3 Equilibration and production phase

The final step of the simulation is to run the simulation in "production" phase for the desired time length using the same conditions as equilibration at constant temperature and pressure, which closely matches laboratory conditions. The parameters used in this stage are shown in Figure 3.12.

```
Explicit solvent molecular dynamics constant pressure MD
&cntrl
  imin=0, irest=1, ntx=5,
  ntpr=10000, ntwx=10000, nstlim=50000000,
  dt=0.002, ntt=3, tempi=300,
  temp0=300, gamma_ln=1.0, ig=-1,
  ntp=1, ntc=2, ntf=2, cut=8,
  ntb=2, iwrap=1, ioutfm=1,
/
```

Figure 3.12: MD Simulation Step 3: Production stage

Every 10,000 steps energy informationwas printed to output files (ntpr) and every 10,000 steps the coordinates were written into the trajectory file (ntwx). Number of MD steps in production phase to be performed was 50,000,000 steps (nstlim), which translates to 100 ns of MD simulation time.

3.5 ADME and Toxicity Characteristics of Drugs in Humans

One of the key challenges of drug development is to predict drug responses in humans because drug discovery is a time-consuming and expensive effort and may result in failures in clinical trials. The last step in this project was to predict the ADMET (Absorption, Distribution, Metabolism, Excretion and Toxicity) properties of the phytochemicals used in this study using *in silico* techniques. For this the Osiris Property

Explorer (http://www.openmolecules.org/propertyexplorer/applet.html) was used. To test the ligands the SMILES representation (Canonical SMILES) of the selected ligands were obtained from PubChem. For example the Canonical SMILES of hesperidin (CID: 10621) is showed in Figure 3.13. This was submitted to the Property Explorer to be evaluated. A summary of the analysis produced is shown in Figure 3.14.

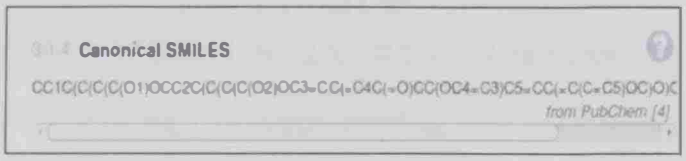


Figure 3.13:Canonical Smiles of Hesperidin (CID: 10621) from Pubchem

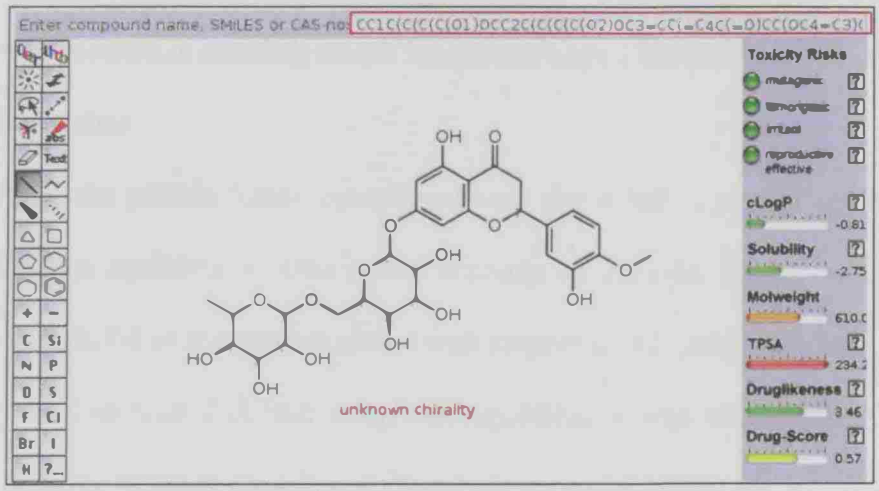


Figure 3.14: Analysis of the ADMET properties using Osiris Property Explorer

Chapter 4: Results

In this study, molecular docking and MD simulations were used to characterize the interaction of phytochemicals with several target proteins.

In order to gets an insight into the binding interactions, intermolecular interactions like hydrogen bonding and hydrophobic interactions were analyzed. These bonds help in determining the binding affinity and drug efficacy.

In GOLD docking, hydrogen bonding and hydrophobic interactions play an important role in stabilizing a ligand energetically in a protein. The number of hydrogen bond for each chemical and protein were calculated for all docking runs. This comparison assisted in deciding which ligand and targets are the most suitable candidates for further studies.

After the protein-ligand complexes were simulated, it is essential to check that simulation has equilibrated. One useful measure to consider is the root mean square deviation (RMSD) of the protein atoms with respect to the initial structure. This should normally be less than 2 Å indicating that equilibration was successful. Analyzing the binding distance between protein and chemicals is needed to test whether the ligand stays tightly bound in the active site. These results produce an atomic-level view of the protein dynamics through multiple MD simulations. The results showed that the architecture of the active site is stabilized by several different hydrogen bonds. Furthermore, the pharmacological properties and toxicity of the candidate chemicals was predicted *in silico* to avoid any failure in drug designing.

4.1 Total Number of Hydrogen Bond in Molecular Docking

Molecular docking results of GOLD were analyzed in LIGPLOT and the total number of hydrogen bond interactions was calculated. The results are summarized in Table 4.1.

No.	Chemical Name	AKT	BCL2	CDK	EGFR	ERK	JAK1	MEK1	Rab5A
1	Apigenin	0	0	0	0	0	0	0	0
2	Alpha Carotene	0	0	0	0	0	0	0	0
3	Beta Sitosterol	0	0	0	0	0	0	0	0
4	Betulinic Acid	0	0	0	0	0	0	0	0
5	Boswellic Acid	0	0	0	0	0	0	0	0
6	Canthaxanthin	0	0	0	0	0	0	0	0
7	Capsaicin	3	3	0	0	0	3	0	3
8	Carnosol	0	0	0	0	3	0	0	0
9	Chlorogenic Acid	3	0	5	4	7	5	6	8
10	Cryptoxanthin	0	0	0	0	0	0	0	0
11	Curcumin	0	0	0	3	3	5	0	3
12	Daidzein	0	0	0	0	0	0	0	0
13	Delphinidin	0	3	0	3	4	0	3	7
14	Delta Carotene	0	0	0	0	0	0	0	0
15	EGCG	4	0	6	3	8	0	4	8
16	Ellagic Acid	4	5	3	4	3	4	5	8
17	3,3'-Diindolylmethane	0	0	0	0	0	0	0	0
18	Acacetin	0	0	0	0	0	0	0	0
19	Anacardic Acid	0	0	6	0	5	4	3	6
20	Chrysin	0	0	0	0	0	0	3	3
21	Cyanidin	0	0	0	3	3	3	3	8

22	Diosmetin	0	0	0	0	3	0	0	3
23	Emodin	0	0	3	0	0	3	0	7
24	Gamma Linolenic Acid	0	0	0	0	4	3	3	4
25	Gamma Carotene	0	0	0	0	0	0	0	0
26	Garcinol	0	0	3	3	5	0	0	0
27	Genistein	0	0	0	0	0	0	0	5
28	Gingerol	0	0	0	0	0	3	3	4
29	Glycitein	0	0	0	0	0	0	4	0
30	Isorhamnetin	3	3	3	4	0	0	3	9
31	Myricetin	6	0	0	4	5	0	4	5
32	Oleanolic Acid	0	0	0	0	0	0	0	0
33	Oleocanthal	0	0	5	3	3	4	3	5
34	Peonidin	0	0	0	0	0	0	0	5
35	Piceatannol	3	0	0	0	3	0	5	3
36	Pinoresinol	0	0	0	0	3	0	3	0
37	Pterostilbene	0	0	0	0	0	0	0	0
38	Resveratrol	0	0	0	0	0	0	0	4
39	Stigmasterol	0	0	0	0	3	0	0	0
40	Tangeritin	0	0	0	0	0	0	0	0
41	Theaflavin	0	6	0	4	5	4	4	8
42	Guggulsterone	0	3	0	0	0	0	0	0
43	Hesperidin	4	3	7	6	7	5	5	4
44	Indicaxanthin	0	3	5	0	3	4	3	4
45	Isobetanin	5	0	3	2	6	4	7	6
46	Kaempferol	0	0	3	3	0	0	0	8
47	Lariciresinol	0	0	0	0	4	3	0	5
48	Lutein	0	0	0	0	0	0	3	0

49	Luteolin	0	3	0	0	0	0	0	5
50	Macrocarpal A	0	0	0	0	4	0	4	6
51	Matairesinol	0	0	0	3	7	0	0	4
52	Myrciaphenone B	0	5	3	8	6	5	4	10
53	Naringenin	0	0	0	0	0	0	0	5
54	Pelargonidin	3	0	0	0	0	0	0	5
55	Quercetin	3	3	3	0	4	0	4	9
56	Vulgaxanthin	0	0	3	4	4	0	6	6
57	Ursolic Acid	0	0	0	0	3	0	0	0
58	Trametenolic Acid	0	0	0	3	0	0	4	3
59	Rosmarinyl Glucoside	5	6	7	6	5	6	6	9
60	Rubixanthin	3	0	0	0	0	0	0	0
61	Rutin	7	5	0	3	6	4	4	0
62	Scutellarein	0	0	0	0	3	0	3	0
63	Silibinin	4	0	3	0	5	0	3	7
64	Silymarin	0	0	3	0	6	0	7	7
65	Subaphylline	3	0	6	5	4	5	4	5
66	Zeaxanthin	0	0	0	0	0	0	0	0
67	Yakuchinone	0	0	3	0	0	0	3	0

Table 4.1:Total numbers of hydrogen bonds between the protein and the docked ligand

Molecular docking results showed that some chemicals did not interact with any of the protein. This is most likely due to the size of ligand, the shape of the binding site, the chemical structure, the polarity, or unfavorable contacts. As a result, some chemicals did not form good interactions with the proteins while others showed strong interaction

with good number of hydrogen bonds. GOLD docking showed that Rab5A has the highest number of interacting chemicals out of the 67 chosen.

4.2 Rab5A as a Promising Target

Rab5A, a member of RAS oncogene family was the most suitable candidate protein based on the results obtained from molecular docking results.

4.2.1 Molecular Docking

In molecular docking, usually ligands bind and interact with the active site of a protein. This active site is the basis of the lock and key model. Ideally candidate drugs must bind strongly with high affinity and specificity to the active site.

Figure 4.1 shows a LIGPLOT diagram of GTP bound to Rab5A as evident from the co-crystallized PDB structure (PDB ID: 1N6L) (Zhu et al., 2003). It can be see that GTP interacts with the same amino acids that interact with the ligands in this study like Ser29, Ser34, Ser35and Thr5, which indicates that the ligands are docking in the same binding site.

Molecular docking results showed that several chemicals could interact strongly with Rab5A (PDB ID: 1N6H) as shown in Figure 4.2 and summarized in Table 4.2. Binding site amino acid residues were analyzed to see whether the chemicals bound to the active site.

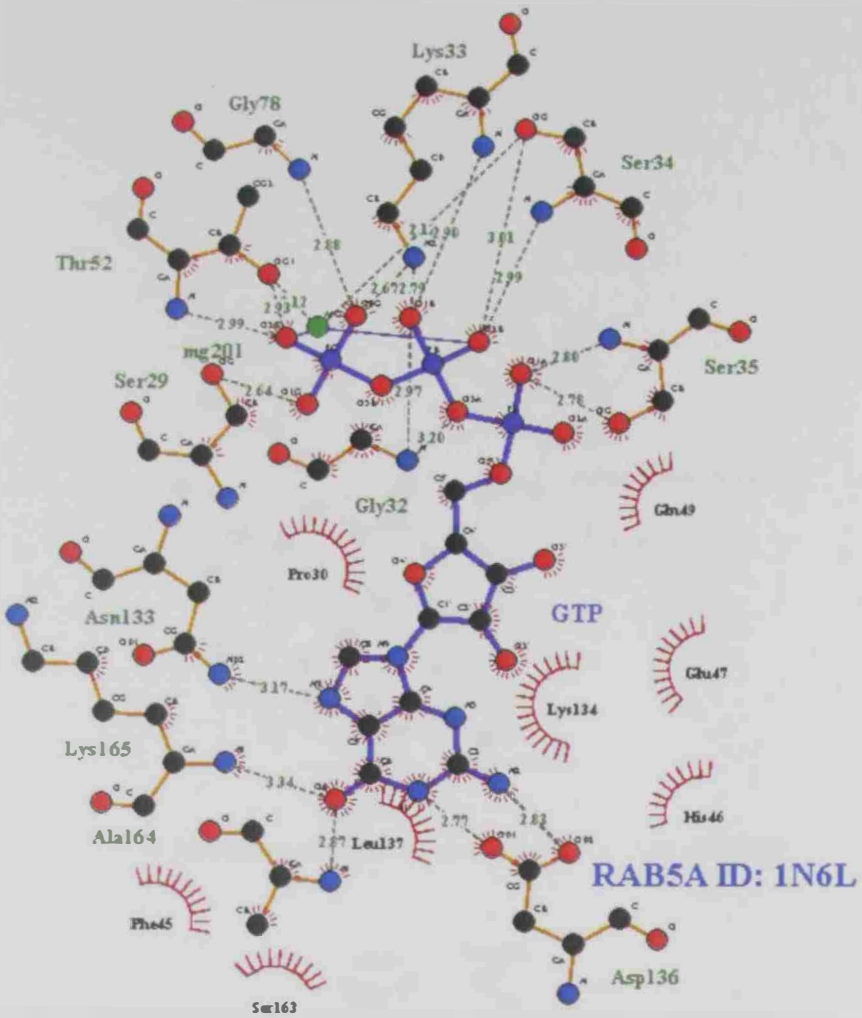
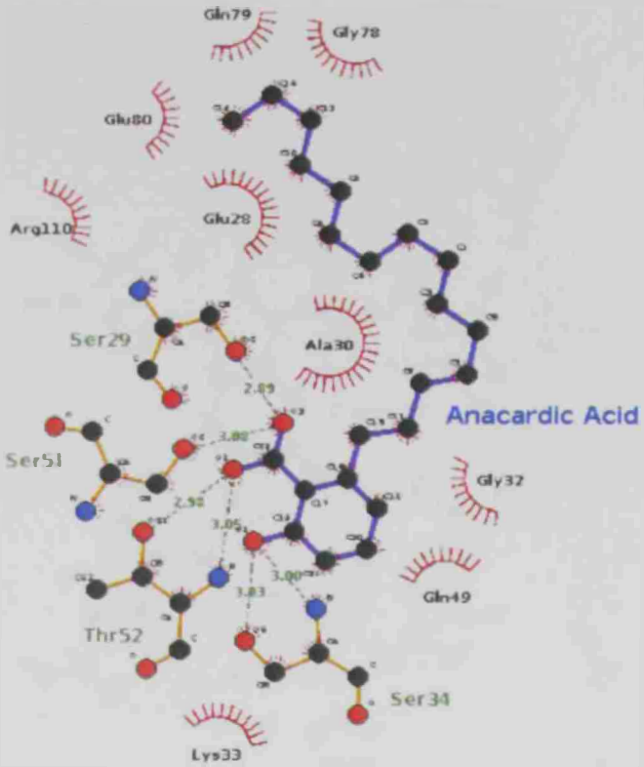
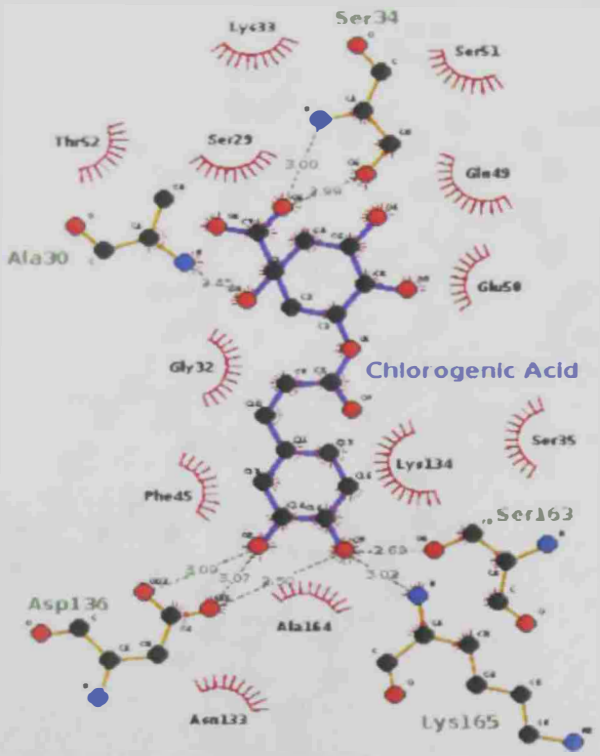


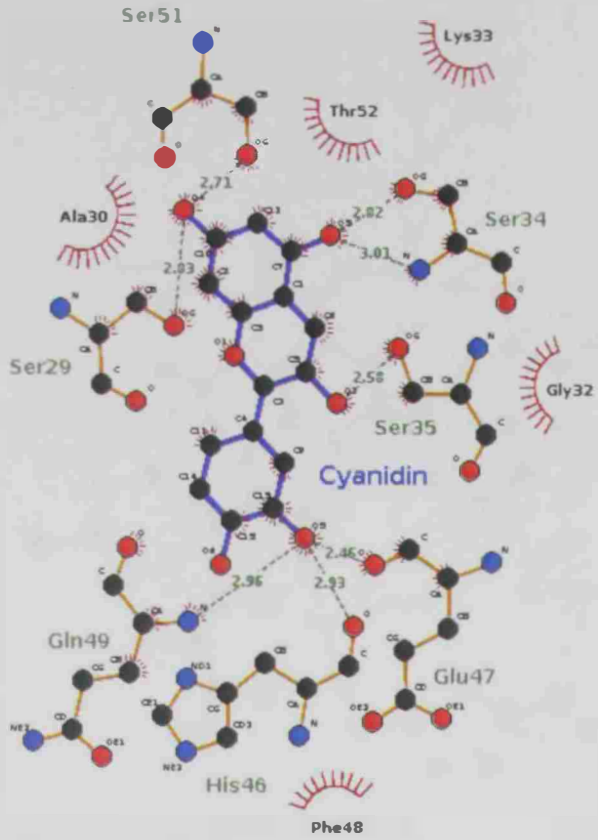
Figure 4.1: LIGPLOT showing hydrophobic interaction and hydrogen bond of GTP and Rab5A (PDB ID: 1N6L)



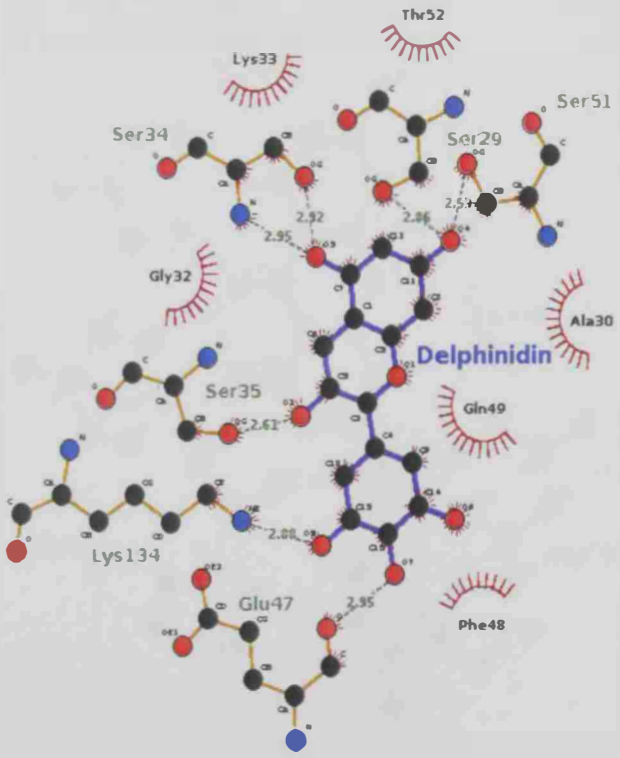
A Interaction of anacardic acid with Rab5A (PDB ID:1N6H)



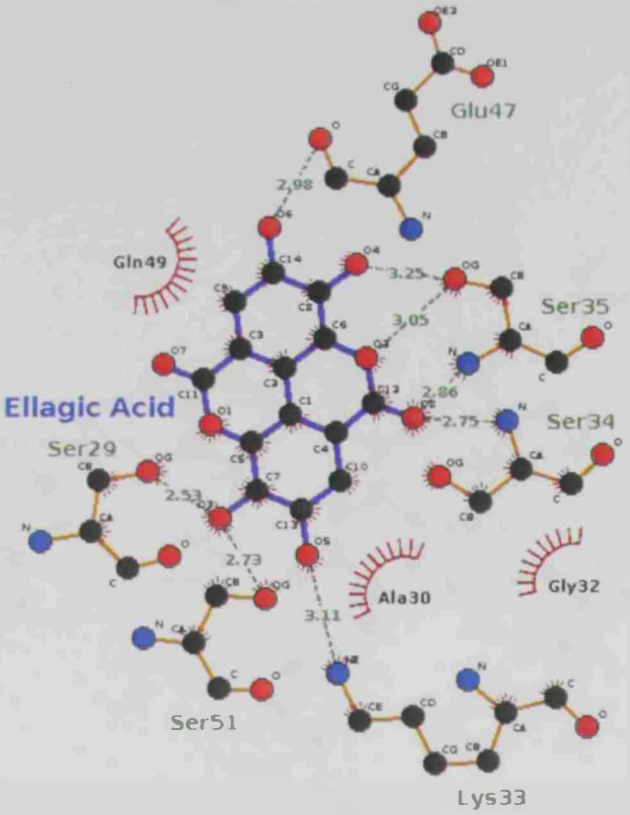
B. Interaction of chlorogenic acid with Rab5A (PDB ID:1N6H)



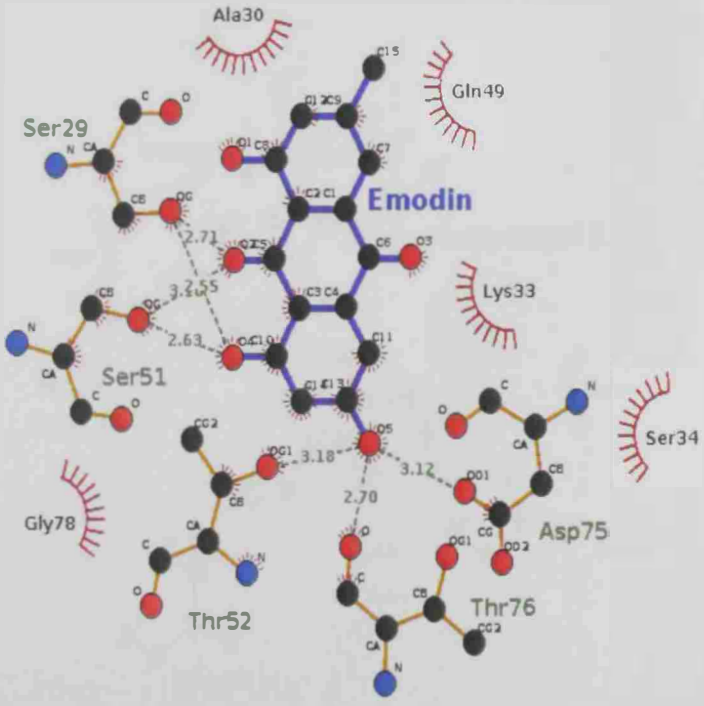
C. Interaction of cyanidin with Rab5A (PDB ID:1N6H)



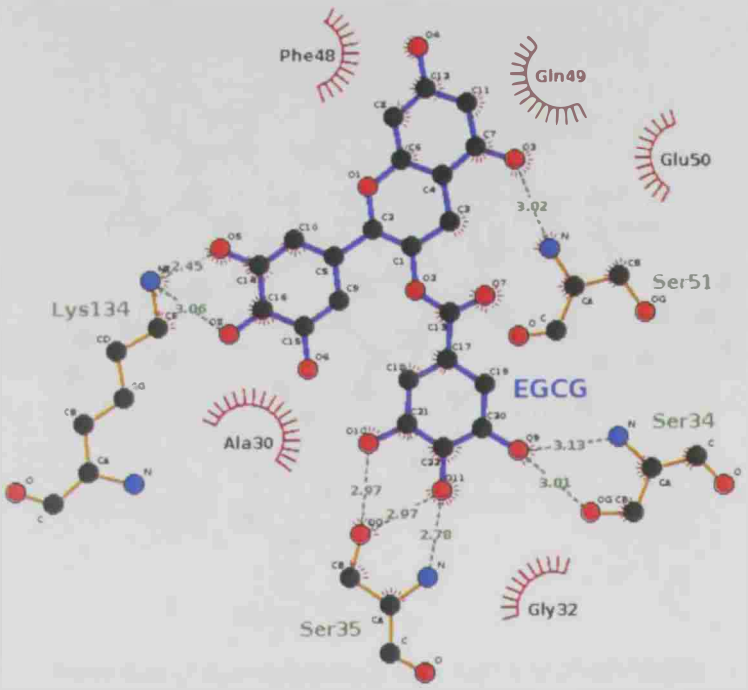
D. Interaction of delphinidin with Rab5A (PDB ID:1N6H)



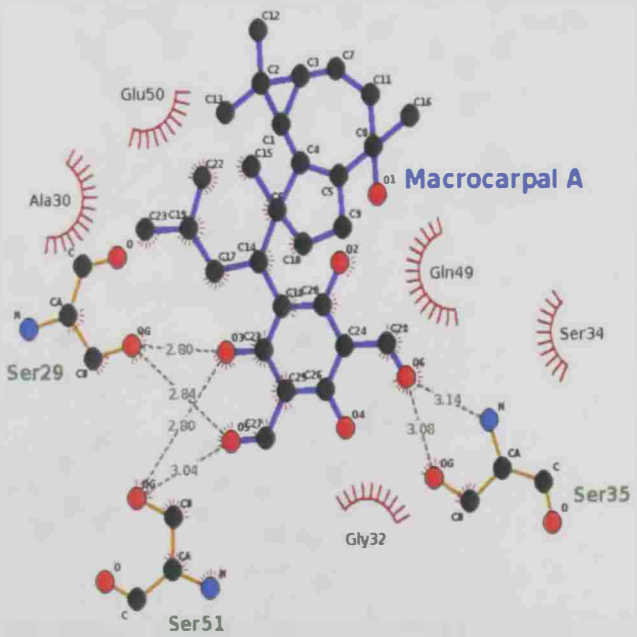
E. Interaction of ellagic acid with Rab5A (PDB ID:1N6H)



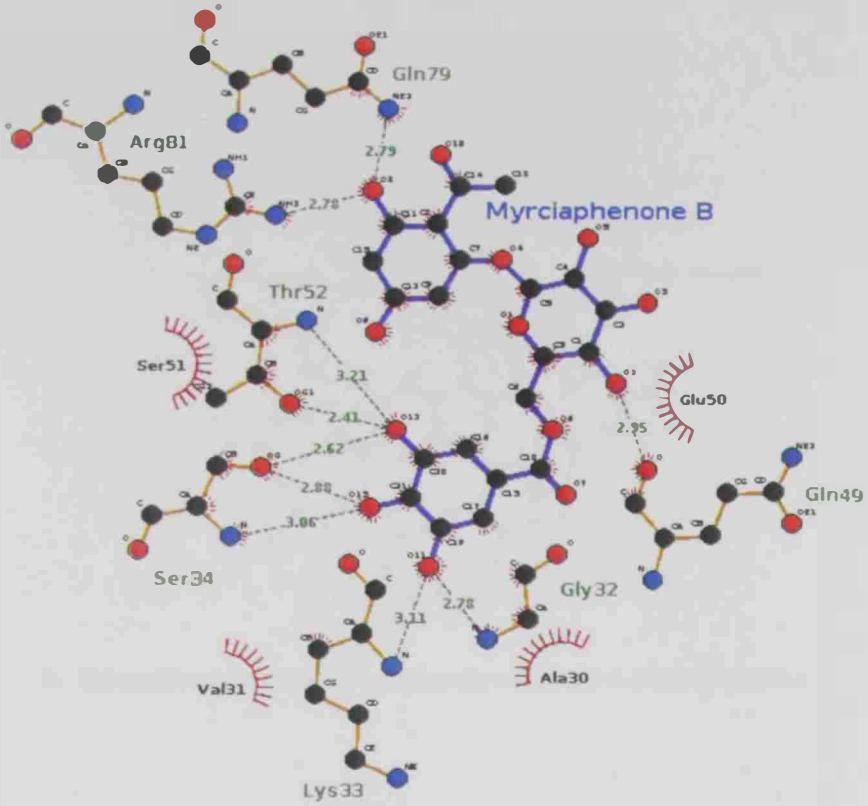
F. Interaction of emodin with Rab5A (PDB ID:1N6H)



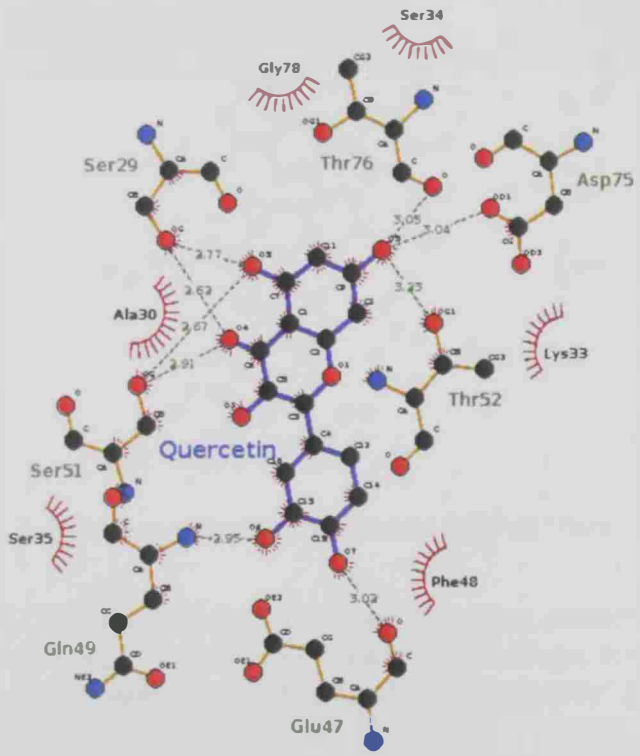
G. Interaction of EGCG with Rab5A (PDB ID:1N6H)



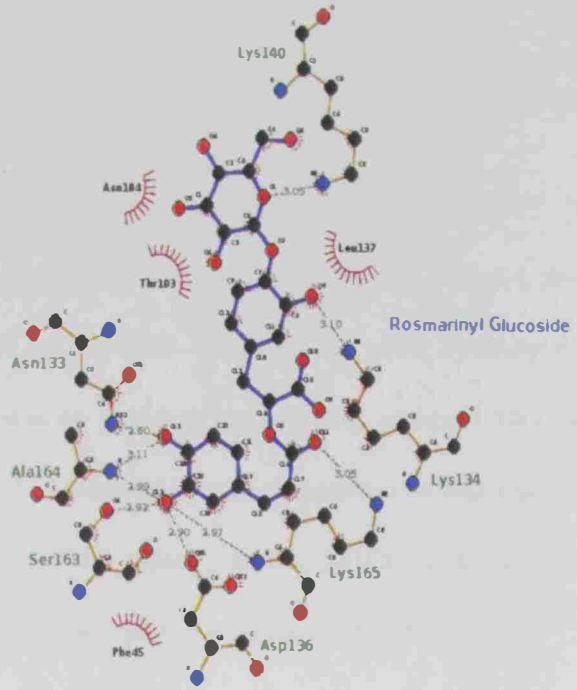
H. Interaction of macrocarpal A with Rab5A (PDB ID:1N6H)



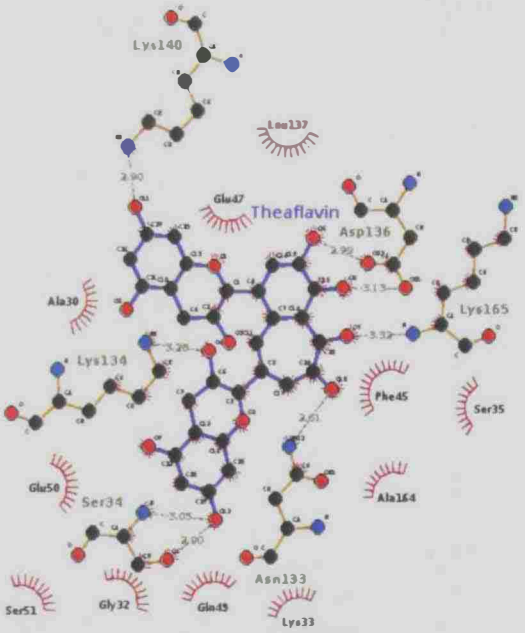
I. Interaction of myrciaphenone B with Rab5A (PDB ID:1N6H)



J. Interaction of quercetin with Rab5A (PDB ID:1N6H)



K. Interaction of rosmarinyl glucoside with Rab5A (PDB ID:1N6H)



L. Interaction of theaflavin with Rab5A (PDB ID:1N6H)

Figure 4.2: LIGPLOT result of Rab5A (ID:1N6H) docking interactions with A. anacardic acid, B. chlorogenic acid, C. cyanidin, D.delphinidin, E. ellagic acid, F.emodin,G.EGCG, H.macrocarpal A I. myrciaphenone B, J. quercetin, K. rosmarinyl glucoside, L.theaflavin.

Based on the Rab5A structure that was complexed with GTP (PDB ID: 1N6L) the active site includes the amino acids: Ser29, Gly32, Lys33, Ser34, Ser35, Thr 52, Gly78, Asn133, Asp136, Ala164 and Lys165. Consequently, the docking results of Rab5A (ID: 1N6H) showed that it shared the same active site residues like Ser29, Ser34 Thr52, Asn133, Ala164 and Lys165 with several chemicals including anacrdic acid, chlorogenic acid,cyanidin, delphinidin, ellagic acid, emodin, EGCG, macrocarpal A, myrciaphenone B, quercetin, rosmarinyl glucosid and theaflavin (Table 4.2).

No	Phytochemical	Amino acids with H bond interaction	Amino acids with Hydrophobic Interaction
1	Anacardic acid	Ser29, Ser34, Ser51, Thr52	Glu28, Ala30, Gly32, Lys33, Gln49, Gly78, Gln79, Glu80, Arg110
2	Chlorogenic acid	Ala30, Ser34, Asp136, Ser163, Lys165	Ser29, Gly32, Lys33, Ser35, Phe45, Gln49, Glu50, Ser51, Thr52, Asn133, Lys134, Ala164
3	Cyanidin	Ser29,Ser34, Ser35, Glu47, Gln49, Ser51	Ala30, Thr52, Lys33, Gly32, Phe48
4	Delphinidin	Ser29, Ser34, Ser35, Glu47, Ser51, Lys134	Ala30, Gly32, Lys33, Phe48, Gln49, Thr52
5	Ellagic acid	Ser29, Lys33, Ser34, Ser35, Ser51	Ala30, Lys33, Ser34, Gln49, Gly78
6	Emodin	Ser29, Ser51, Thr52, Thr76, Asp75	Ala30, Gly32, Phe48, Gln49, Glu50
7	EGCG	Ser34, Ser35, Ser51, Lys134	Ala30, Gly32, Gln49, Phe48, Glu50
8	Macrocarpal A	Ser29, Ser35, Ser51	Ala30, Ser34, Gly32, Gln49, Glu50
9	Myrciaphenone B	Lys33, Gly32, Ser34, Gln49, Thr52, Arg81, Gln79	Ala30, Val31, Glu50, Ser51
10	Quercetin	Ser29, Glu47, Gln49, Ser51, Thr52, Asp75, Thr76	Ala30, Lys33, Ser34, Ser35, Phe48, Gly78
11	Rosmarinyl glucoside	Asn133, Lys134, Asp136, Lys140, Ser163, Ala164, Lys165	Phe45, Thr103, Leu137, Thr103
12	Theaflavin	Ser34, Asn133, Asp136, Lys140, Lys165	Ala30, Gly32, Lys33, Ser35, Phe45, Glu47, Gln49, Glu50, Ser51, Leu137, Ala164

Table 4.2: Amino acids of Rab5A interacting with the ligand based on hydrogen bonds and hydrophobic interaction

4.2.2 Molecular Dynamics Simulation

The MD technique has been assessed as a reliable tool to refine experimental structures. Thus, MD simulations were performed for closely studying the interactions obtained from docking. MD simulation results showed that only four phytochemicals - EGCG, ellagic acid, macrocarpal A and quercetin – bound strongly to Rab5A. The active site of the Rab5A-GTP complex (PDB ID: 1N6L) visualized in 3D using VMD is shown in Figure 4.3.

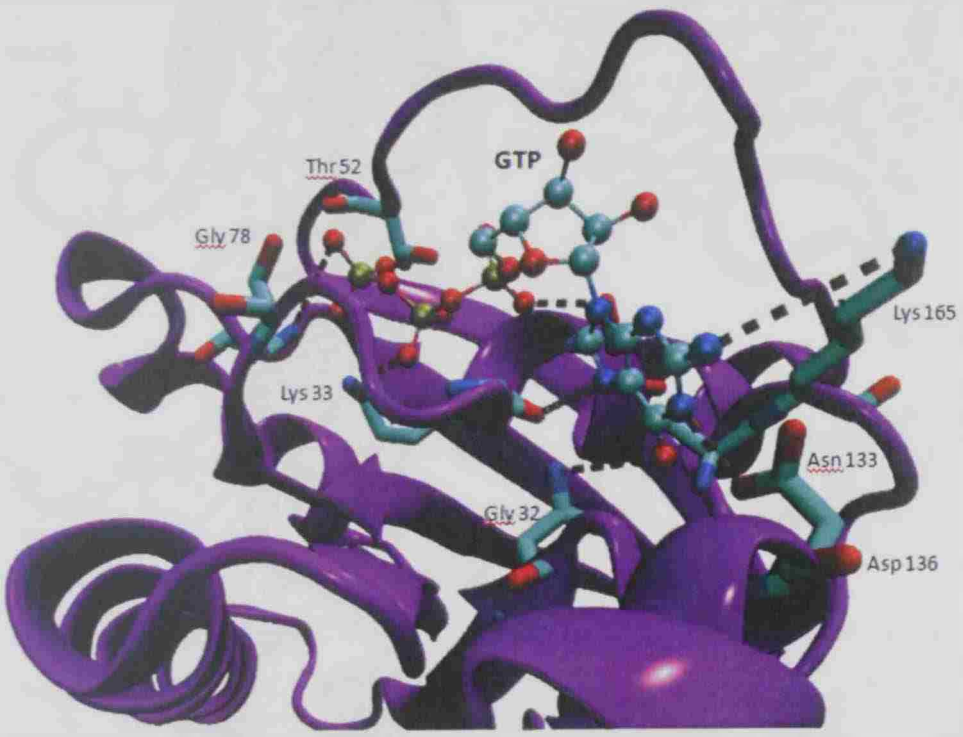


Figure 4.3: GTP bound to the active site of Rab5A showing the hydrogen bonds with several amino acids

Figure 4.4 showed that only four out of 67 phytochemicals that were tested have strong and stable interactions with the target Rab5A protein (PDB ID: 1N6H). EGCG

forms hydrogen bond interactions with three amino acids – Ala30, Ser34, and Gln4 - in the Rab5A active site. Ellagic acid forms two hydrogen bonds with Gly32 and Glu47 and macrocarpal A forms hydrogen bonds with Ala55, Met88, Gly92, and Ala93 while quercetin forms two hydrogen bond with Thr52 and Asp75. These amino acid residues are located in the active site of the Rab5A based on its interactions with GTP.

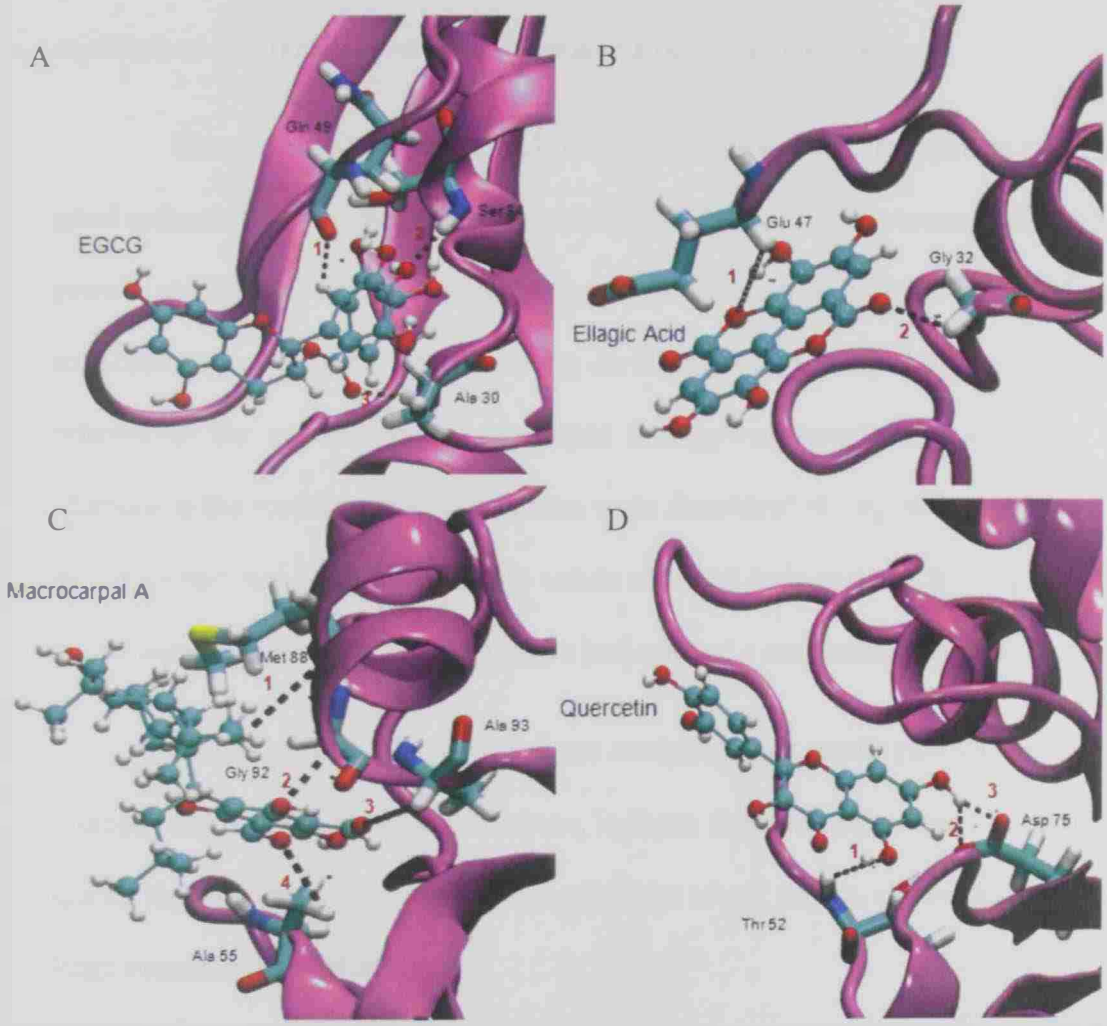


Figure 4.4: EGCG showing hydrogen bond interactions with 3 amino acids (Gln49, Ser34, Ala 30), B. Ellagic acid showing hydrogen bond interactions with 2 amino acids (Glu47 and Gly32), C. Macrocarpal A showing hydrogen bond interactions with 4 amino acids (Met88, Gly92, Ala93 and Ala55), D. Quercetin showing hydrogen bond interactions with 2 amino acids (Thr52 and Asp75).

4.2.3 Root Mean Square Deviation (RMSD) Analysis

The analysis of molecular dynamics (MD) simulation trajectories included physical quantities such as the root mean square deviation (RMSD), binding stability at thermal changes and hydrogen bond interactions showed stability in ligand protein interactions along the MD simulation trajectories. RMSD of various systems reaches equilibrium and fluctuates around the mean value at various times.

Measuring equilibration and protein flexibility from the starting structure is a good indication of structural stability and simulation integrity. To measure how much the protein structure changes over the course of the simulation protein. RMSD during the simulation was measured. By calculating the RMSD values of the backbone heavy atoms relative to the coordinates of the initial (energy-minimized) structures, the overall changes in the model atomic coordinates were monitored during MD simulations. After an initial fast-rise region the RMSD values changed little and reached a more constant value, indicating that the model structure had reached a conformational steady state.

Changes on the order of 1-3 Å are acceptable for small, globular proteins while changes much larger than that, however, indicate that the protein may be misbehaving during the simulation. Flexible regions e.g. large loops, unwound termini, etc. can cause large contributions to RMSD.

In Rab5A, RMSD was calculated for the protein (all atoms of the protein backbone without hydrogen). The overall RMSD finally fluctuated between 1.5 and 2.6 Å (Figure 4.5).It did not exceed 3 Å indicating that the systems had equilibrated well.

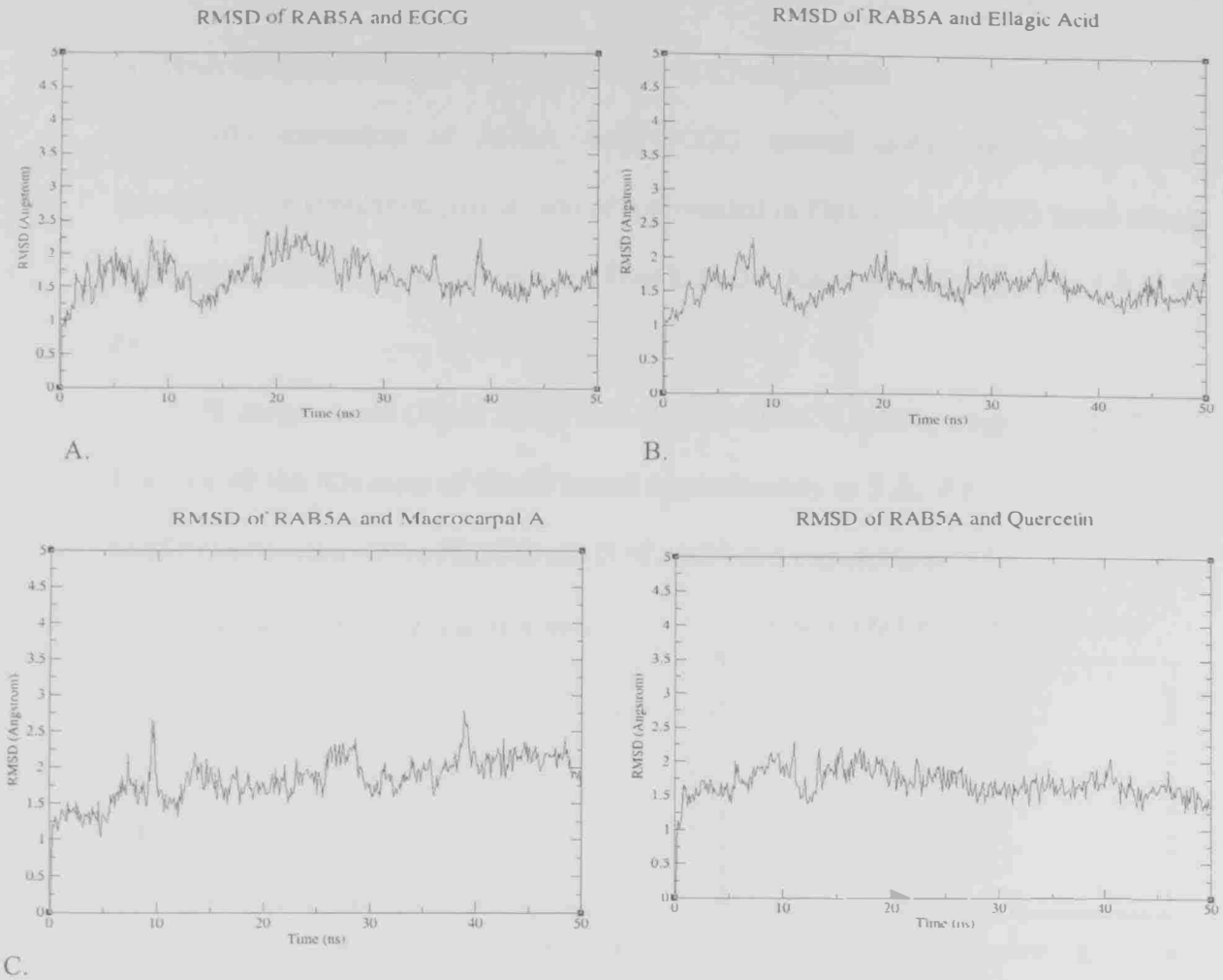


Figure 4.5: RMSDs of the protein backbone heavy atoms with each ligand (EGCG, ellagic acid, macrocarpal A, quercetin).

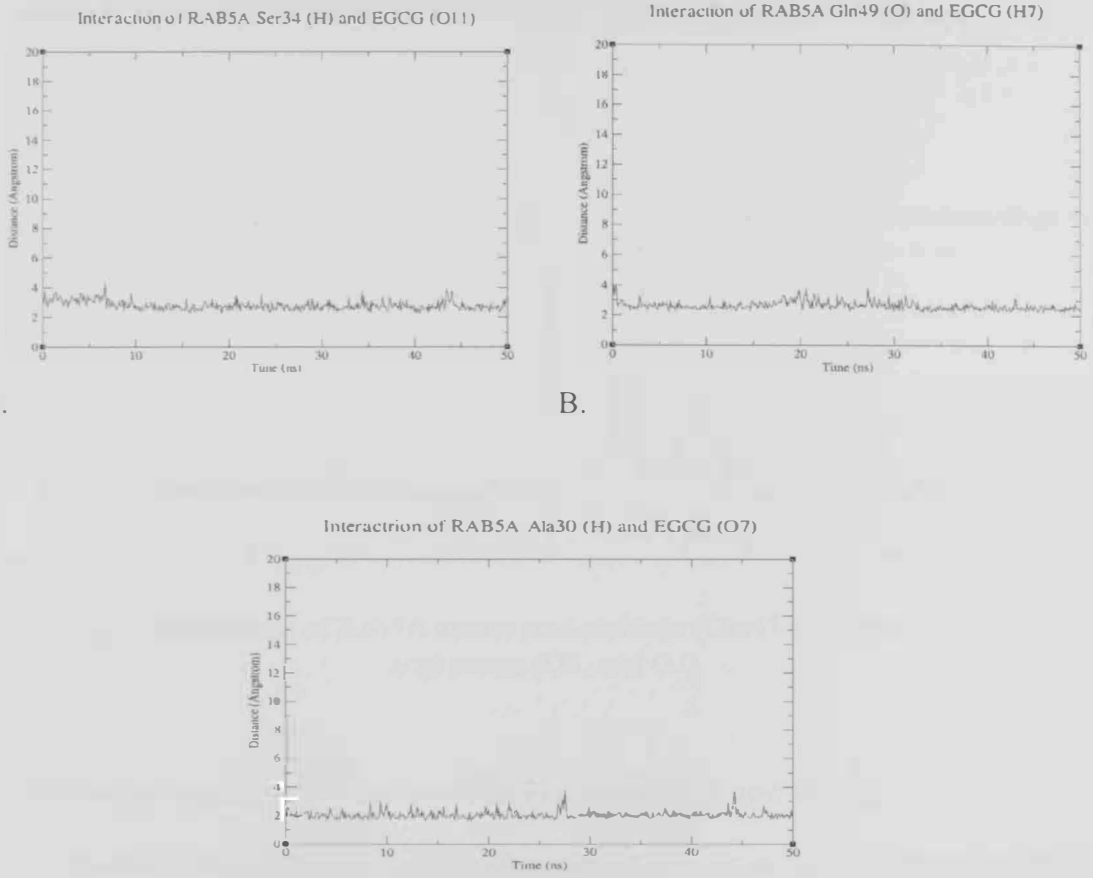
4.2.4 Analysis of Hydrogen Bond Interaction

Hydrogen bond interactions with amino acids play an import role in stabilizing protein-ligand complexes. It plays a key role in both the formation and stabilization of protein structures. They form and break while a protein deforms, for instance during the transition from a non-functional to a functional state. In this work, hydrogen bonding was analysed along the length of the MD simulation trajectories. The time dependence of the hydrogen bonds formed by the protein and ligand during the MD simulations was evaluated.

A. Hydrogen bond interactions between EGCG and Rab5A

MD simulation of Rab5A with EGCG showed stable hydrogen bonding throughout the simulation period, which is presented in Figure 4.6. EGCG forms strong hydrogen bonds with three amino acids Gln49, Ser34, Ala30 with distance 2 to 3 Å in 50 ns.

Hydrogen bond (H) of Ser34 with (O11) of EGCG, and hydrogen bond (H7) of EGCG with the (O) atom of Gln49 bound approximately in 3 Å. While other hydrogen bond form between (O7) of EGCG and H of Ala30 and was stable at 2 Å.



A. B.

C.

Figure 4.6: Interactions of Rab5A amino acid residues A. Gln49, B.Ser34 and C. Ala30 with EGCG atoms (H7, O11 and O7) respectively

B. Hydrogen bond interactions between ellagic acid and Rab5A

Ellagic acid produced more fluctuation when its formed hydrogen bonding interactions with Glu47 and Gly32 of Rab5A. The distance was high and not stable in the first 10 ns due to optimization of interactions with the water solvent. Hydrogen bond interaction distance decreases after 10 ns and was fluctuating around 4 Å for Gly47 and 3 Å for Gly32 until the end of MD simulation trajectory. Figure 4.7 illustrates hydrogen bond distance between (HA) of Glu47 and (O1) of ellagic acid and was fluctuating around 4 Å. However, with Gly32 the hydrogen interaction distance was closer to 3 Å between (HA3) of Gly32 and (O8) of ellagic acid.

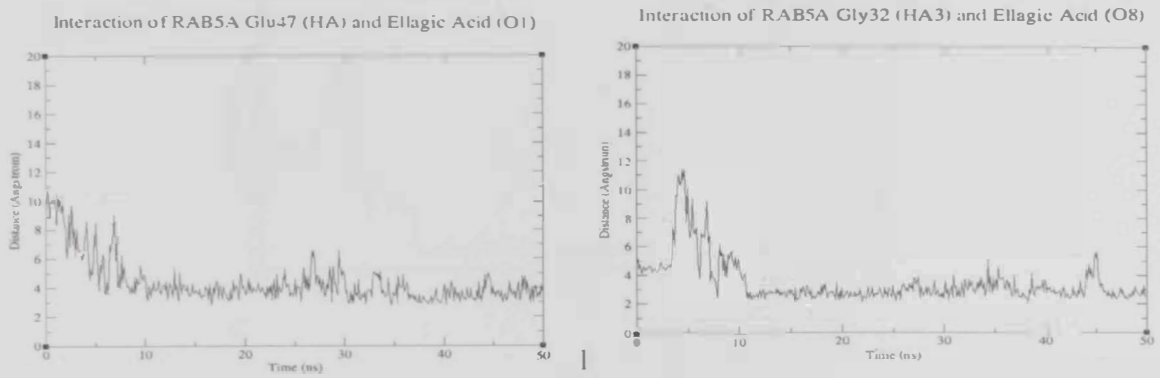


Figure 4.7: Interactions of Rab5A amino acid residues (Glu47 and Gly32) with ellagic acid atoms (O8, and O1)

C. Hydrogen bond interactions between macrocarpal A and Rab5A

Stable hydrogen bonds formed after the first 15 ns in the simulation of Rab5A with macrocarpal A. Figure 4.8 illustrates that (O) of Gly92 form hydrogen bond with (H37) of macrocarpal A, which was constant around 4 Å. The other hydrogen bond formed between (HA) of Ala93 amino acids and O5 of macrocarpal A at around 4 Å in

the MD simulation. Third hydrogen bond formed between O atom of Met88 amino acid andH22 atom of macrocarpal A and was nearly constant around 4 Å in the MD simulation.

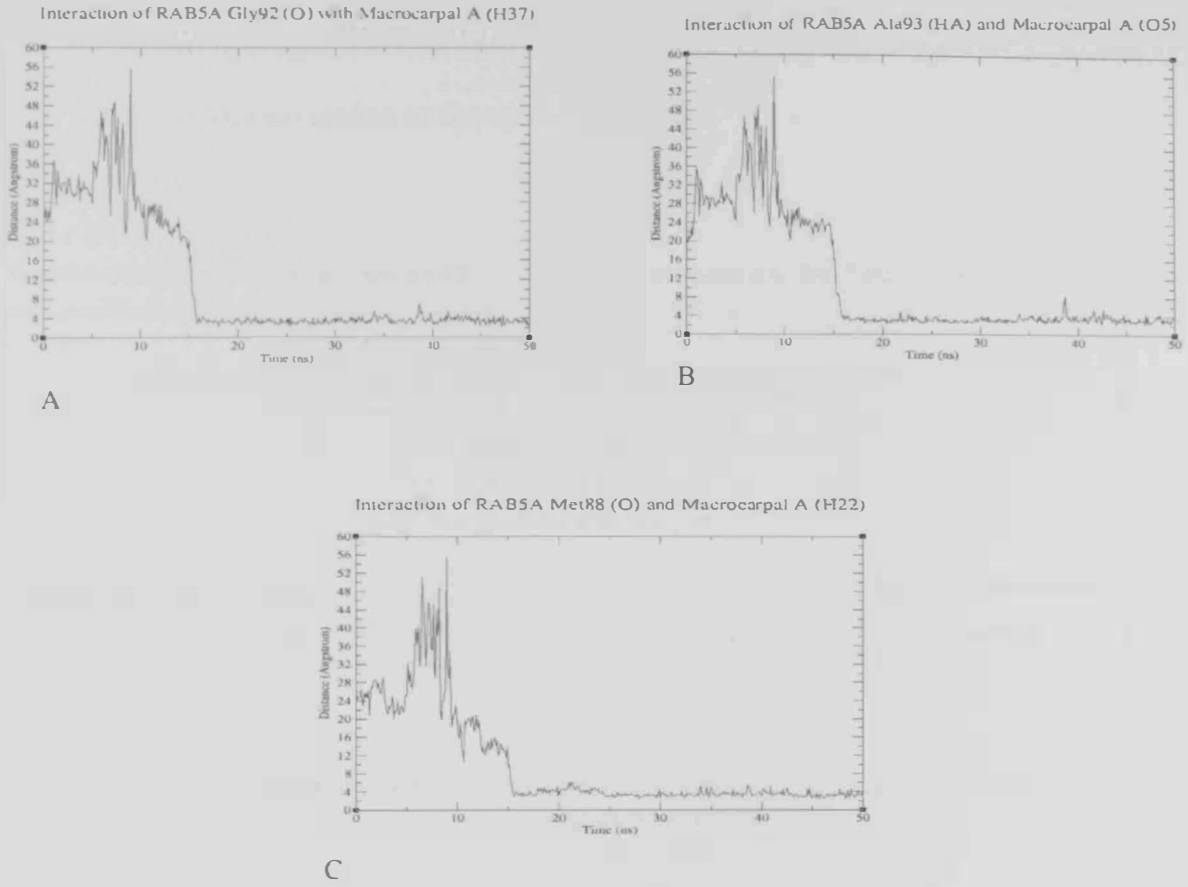


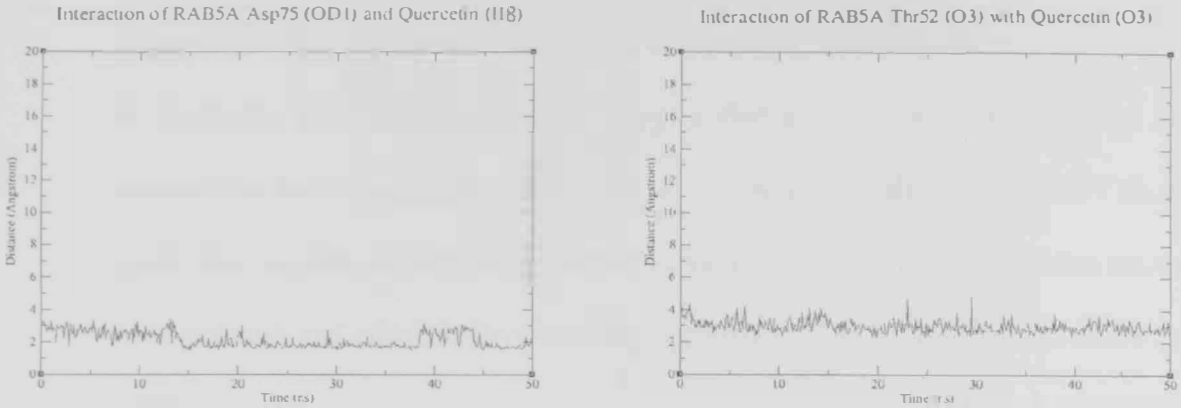
Figure 4.8: Interactions of Rab5A amino acid (Gly92, Ala93 and Met88) with macrocarpal A atoms (H37, O5, H22)

D. Hydrogen bond interactions between quercetin and Rab5A

MD simulation showed that hydrogen bond interaction between quercetin and Rab5A was fluctuating around 2 Å between (OD1) of Asp75 and (H8) of quercetin and

D. Hydrogen bond interactions between quercetin and Rab5A

MD simulation showed that hydrogen bond interaction between quercetin and Rab5A was fluctuating around 2 Å between (OD1) of Asp75 and (H18) of quercetin and the second hydrogen bond between (O3) of Th52 and (O3) of quercetin with a distance of 3 Å in the MD simulation as shown in Figure 4.9



A. B.

Figure 4.9:Interactions of Rab5A amino acid residues (Asp75 and Thr52) with quercetin atoms (H8 and O3)

4.3 EGFR as second promising target

The second most promising target was EGFR in this study (PDB ID: 3W2S). EGFR is involved in the pathogenesis and progression of different carcinoma types. Most ligands did not show any good interactions with EGFR except for one extracted from citrus fruits called hesperidin. It showed strong hydrophobic interaction and good hydrogen bond interactions based on molecular docking and MD simulation. Furthermore, hesperidin didn't show any toxicity risk in ADMET analysis.

4.3.1 Molecular Docking

The analysis of EGFR molecular docking did not produce any good results with most of the phytochemicals; the only one that shows good and stable interaction was hesperidin. LIGPLOT analysis showed that it has interactions with several amino acid residues - Phe723, Gly724, Cys797, Asp837 and Arg841 as shown in Figure 4.10.

Previous studies have found that Cys797 is critical for ligand binding to EGFR (Zhong et al., 2009). Drugs like gefitinib and erlotinib have been used in patients with either wild- type or mutant EGFR. EGFR was found to be more sensitive to erlotinib than to gefitinib. Their docking studies suggest that Cys797 could be the key residue responsible for the enhanced sensitivity of erlotinib. It implies that Cys797 should be taken into account in drug design targeting the EGFR. Two point mutations resistant to both gefitinib and erlotinib have been documented as T790M and D761Y. (Zhong et al., 2009)

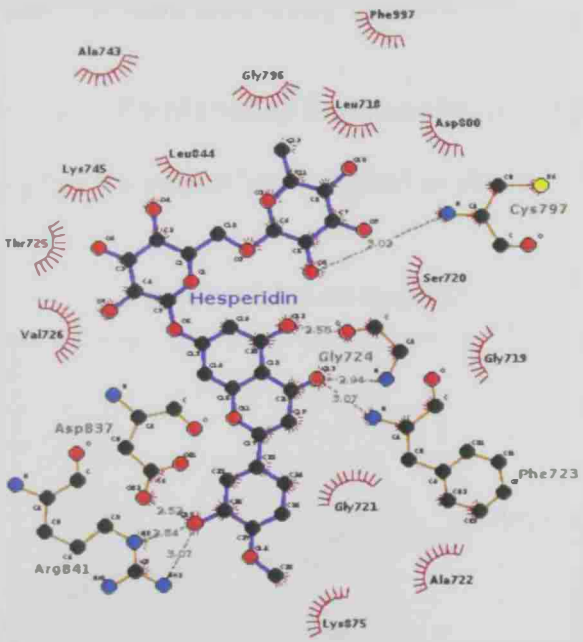


Figure 4.10: LIGPLOT result of EGFR with hesperdin (PDB ID: 3W2S)

4.3.2 Molecular Dynamics Simulation

MD simulation showed that hesperidin remained stably bound in its binding site. Many amino acid were involved in the strong binding to hersperidin including Phe723, Gly724, Gly721, Gly719, Leu718 and Cys797 as shown in Figure 4.11 viewed in VMD.

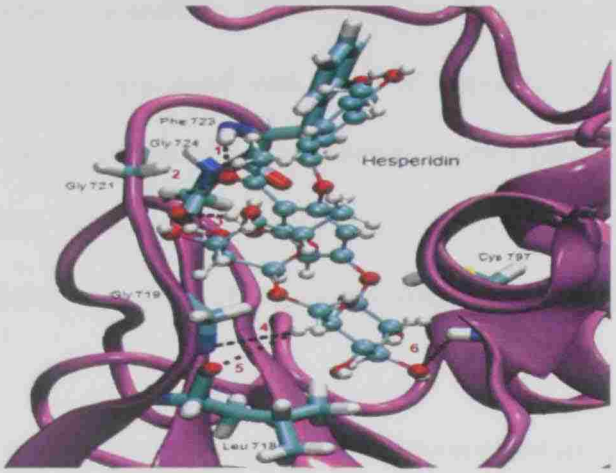


Figure 4.11: Hesperidin bounded to the active sites of EGFR viewed in VMD

4.3.3 Root Mean Square Deviation (RMSD) Analysis

In EGFR, the overall RMSD finally fluctuated between 1.5 and 2.6 Å, and didn't exceed 3 Å, indicating that the system had stabilized as shown in Figure 4.12

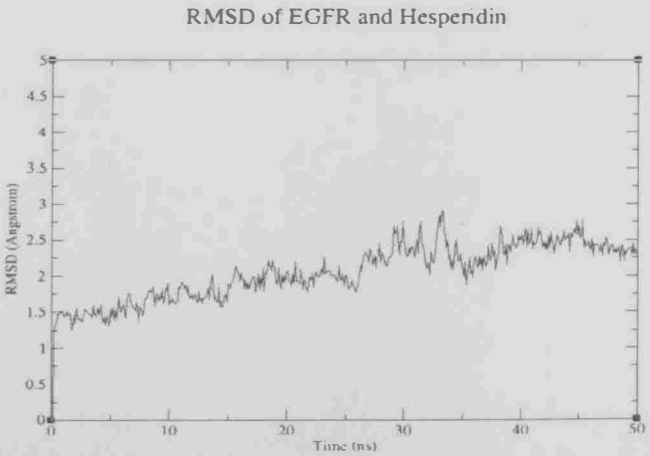
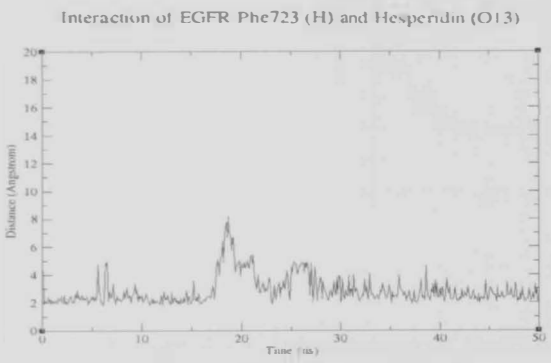


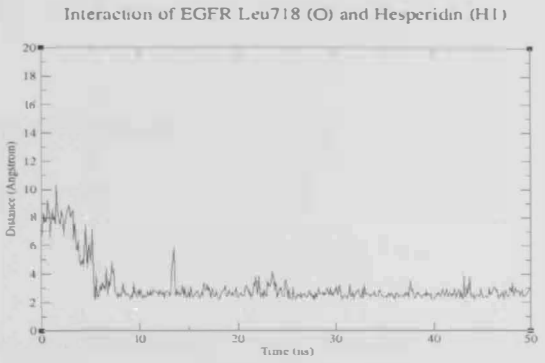
Figure 4.12:RMSDs of Hesperidin core evaluated with respect to the crystal structure along the MD trajectories.

4.3.4 Analysis of Hydrogen Bond Interactions

All protein-ligand atoms hydrogen bond interactions had a distance between 2 to 4 Å as shown in Figure 4.13. Hydrogen atom of residue Phe723 of EGFR and (O13) of hesperidin showed a fluctuating interaction within 2 Å. Leu18 residue formed hydrogen bond with (H1) atom of hesperidin within the first 5 ns and was then stable at 2 Å. Gly724 (O) formed a hydrogen bond with (H29) of hesperidin and was stable at 3 Å for most of the simulation. Hydrogen bond interaction formed between (H) atom of Cys97 and O6 of hesperidin fluctuated between 2 and 4 Å. While nitrogen atom of Gly719 formed hydrogen bond with (H1) of hesperidin after around 10 ns and fluctuated at a distance of 4 Å.



A



B.

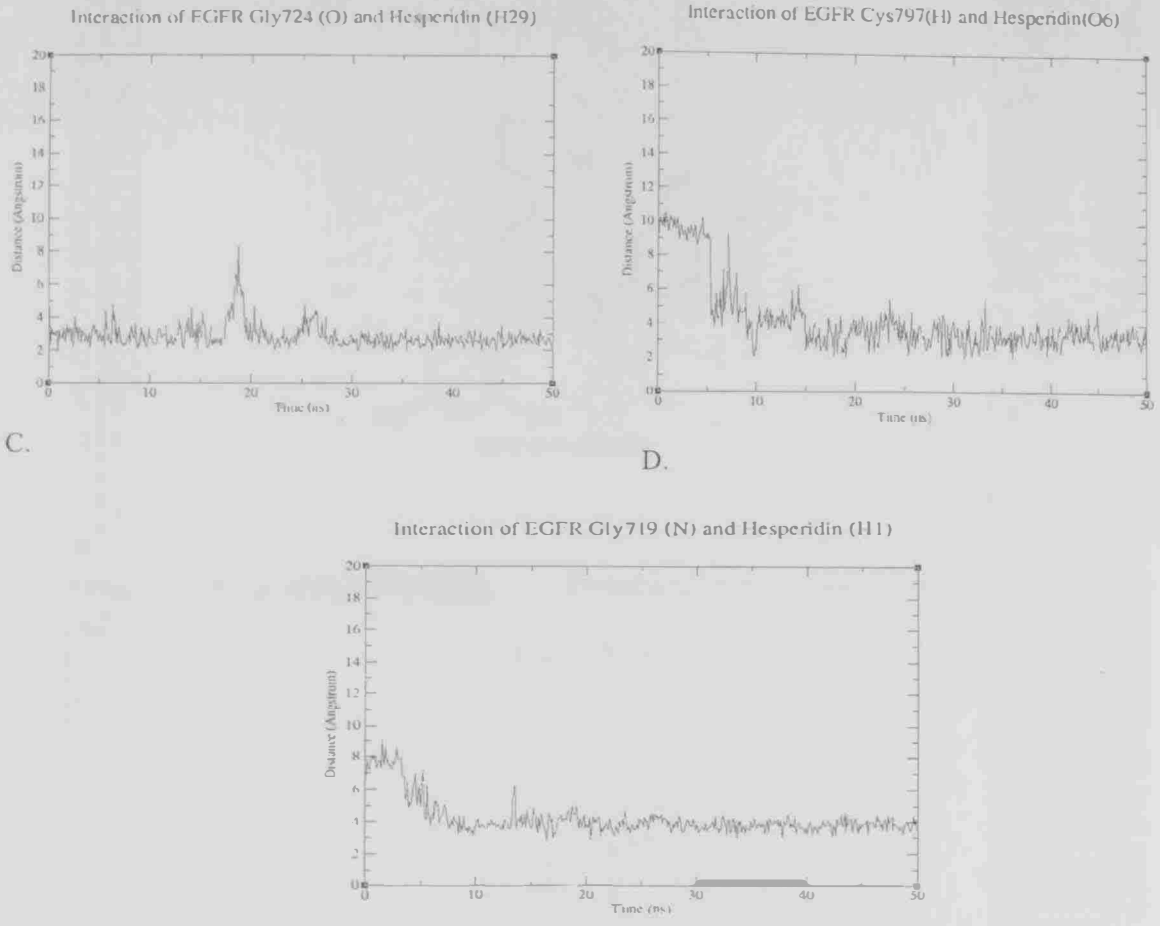


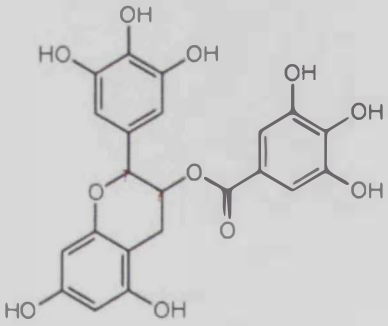
Figure 4.13: Interactions of Rab5A amino acid residues (Phe723, Leu718, Gly724, Cys797 and Gly719) with ligand atoms (O13, H1, H29, O6, and H1)

4.4 Toxicity and Solubility of Ligands

Characterization of Absorption, Distribution, Metabolism, Excretion, and Toxicity (ADMET) properties were predicted using *in silico* methods to know whether the phytochemicals have the potential of adversely affecting humans. This includes properties like mutagenic, tumorogenicity, irritant or reproductive effects as predicted by the Osiris Property Explorer (Figure 4.14).

Enter compound name, SMILES or CAS-no: 2)O)O)C3=CC(=C(C(=C3)O)O)O)OC(=O)C4=CC(=C(C(=C4)O)O)O

A. EGCG



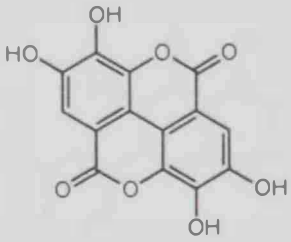
unknown chirality

Toxicity Risks
- mutagenic
- tumorigenic
- irritant
- reproductive effective

cLogP 2.05
Solubility -2.16
Molweight 458.0
TPSA 197.3
Druglikeness 1.58
Drug-Score 0.7

Enter compound name, SMILES or CAS-no: C1=C2C3=C(C(=C1O)O)OC(=O)C4=CC(=C(C(=C43)OC2=O)O)O

B. Ellagic Acid



Toxicity Risks
- mutagenic
- tumorigenic
- irritant
- reproductive effective

cLogP 1.28
Solubility -3.29
Molweight 302.0
TPSA 133.5
Druglikeness -1.6
Drug-Score 0.51

Enter compound name, SMILES or CAS-no: C(C(=C(C(=C1O)C=O)O)C=O)O)C2(CCC3C2C4C(C4(C)C)CCC3(C)O)C

C Macrocarpal A

unknown chirality

Toxicity Risks
mutagenic
tumorigenic
irritant
reproductive effective

cLogP 5.36

Solubility -5.5

Molweight 472

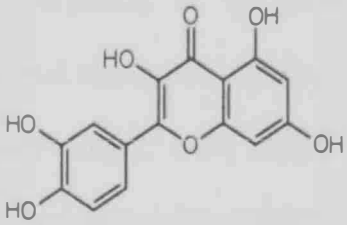
TPSA 115

Druglikeness -4.1

Drug-Score 0.07

Enter compound name, SMILES or CAS-no: C1=CC(=C(C=C1C2=C(C(=O)C3=C(C=C(C=C3O2)O)O)O)O)O

D. Quercetin



Toxicity Risks
mutagenic
tumorigenic
irritant
reproductive effective

cLogP 1.49

Solubility -2.49

Molweight 302.0

TPSA 127.4

Druglikeness 1.6

Drug-Score 0.3

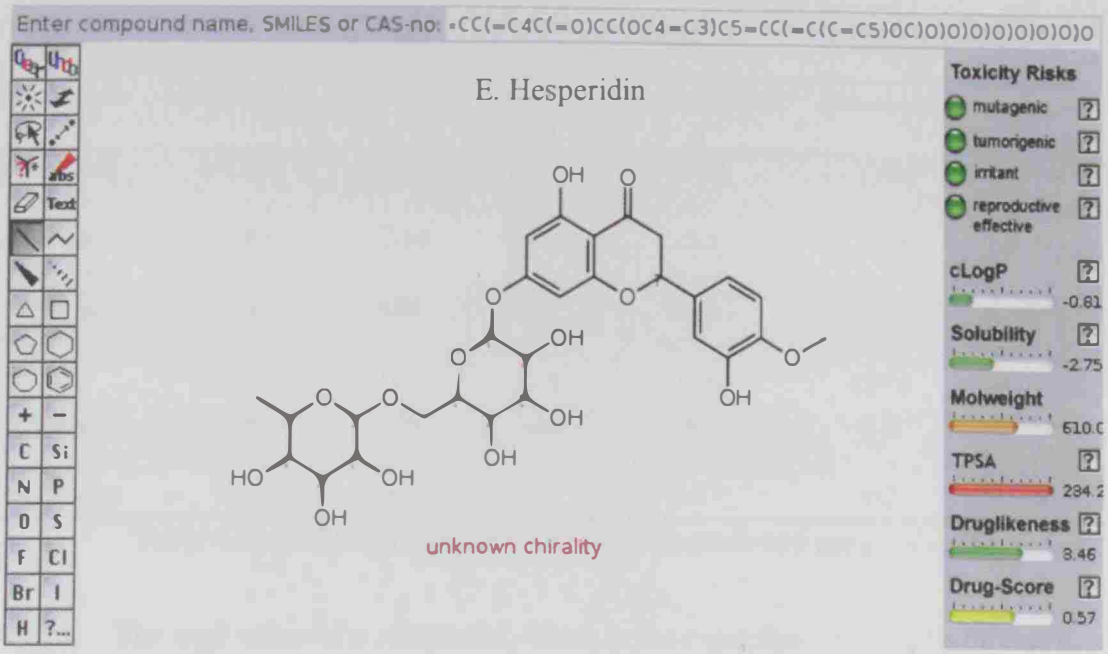


Figure 4.14:ADMET properties showed the toxicity risks of for EGCG, ellagic Acid, macrocarpal A, quercetin and hesperidin

The toxicity risk properties or the ones that that have side effects like mutagenicity or poor intestinal absorption are shown in red. Green color indicates that the properties are acceptable. Ellagic acid, EGCG and hesperidin did not show any toxicity risks (mutagenic, tumorigenicity, irritant or reproductive effects). However macrocarpal A showed high risk effect in tumorigenicity and irritant effect, while mutagenic and tumorigenicity effect was high for Quercetin.

Parameters	Hesperidin	EGCG	Ellagic acid	Macrocarpal A	Quercetin	Acceptable range
Calculated LogP	-0.01	2.05	1.28	5.36	1.49	-0.7 to 5.0 $c_{octanol}/c_{water}$
Solubility (Log S)	-2.75	-2.16	-3.29	-5.5	-2.49	0 to -6 mol/L
Molecular Weight	610	458	302	472	302	< 500 g/mol
TPSA	234.2	197	133.5	115	127.4	20 to 130 Å²
Druglikeness	3.46	1.58	-1.6	-4.1	1.6	>0
Drug Score	0.57	0.7	0.51	0.07	0.3	0 - 1

Table 4.3: ADMET results compared to the preferred range of values

The logP value of a compound, which is the logarithm of its partition coefficient between n-octanol and water ($c_{octanol}/c_{water}$), is a well-established measure of the compound's hydrophilicity. It has been shown for compounds to have a reasonable probability of being well absorb if their logP value must not be greater than 5.0 and smaller than -0.7. Low hydrophilicities and therefore high logP values cause poor absorption or permeation. As shown in Table 4.3 all ligands showed acceptable range, only marcorcarpal A showed a slight out of range value for logP (5.36).

The aqueous solubility of a compound significantly affects its absorption and distribution characteristics. Typically, a low solubility goes along with a bad absorption and therefore the general aim is to avoid poorly soluble compounds. As shown in Table 4.3, macrocarpal A had bad absorption (logP= 5.36) and low solubility (logS= -5.5 mol/L) unlike other ligands.

Molecular weight on the other hand optimizes compounds for high activity on a biological target. This property usually reduces with increased molecular weight. Compounds with higher weights are less likely to be absorbed and therefore to ever reach

the place of action. Thus, keeping molecular weights as low as possible is desirable. Only hesperidin had a slightly high molecular weight in the shortlisted chemicals.

Topological Molecular Polar Surface Area (TPSA) calculates the sum of surface contributions of polar atoms (e.g. oxygens, nitrogens and hydrogens) in a molecule. This correlates well with drug transport properties, such as intestinal absorption, or blood-brain barrier penetration. Hesperidin and EGCG showed a slightly higher TPSA compared to the other ligands.

A positive value for Druglikeness indicates that the molecule contains predominantly fragments, which are frequently present in commercial drugs. A molecule with high druglikeness score may not really qualify for being a drug because of its high lipophilicity. Only ellagic acid and macrocarpal A showed negative score in this case.

The DrugScore combines cLogP, Solubility (logS), molecular weight, Topological Molecular Polar Surface Area (TPSA), druglikeness, drug and toxicity risks to judge the compound's overall potential to qualify as a drug. Table 4.3 shows that the overall score is below 1 for all the short listed compounds indicating that they may be good candidate drugs.

Chapter 5: Discussion and Conclusion

Many drugs are small molecules (ligands) that bind to the active site of specific targets (proteins). Therefore, identification of ligands that bind to these proteins using protein-ligand docking is an important step in structure-based drug design.

The overall research goal of this project was to understand the atomic level interactions between a list of phytochemical ligands and receptor proteins based on their binding mode and to select a few candidate ligands for cancer therapy using *in silico* methods like molecular docking and MD simulation.

5.1 Molecular Docking

7 different proteins were selected and 67 different phytochemicals, known to have mechanism of action as anticancer compounds, were docked with these proteins. Two promising target protein - Rab5A and EGFR - showed good hydrogen and hydrophobic interactions with some of the chemicals chosen.

Rab5A PDB ID: 1N6L) structure complexed with GTP in the active site (Ser 29, Gly 32, Lys 33, Ser 34, Ser 35, Thr 52, Gly 78, Asn 133, Asp 136, Ala 164 and Lys 165) was compared with ligands docked to the apo Rab5A structure (PDB ID: 1N6H). Ligands shared the same active site position (Ser29, Ser34, Thr52, Asn133, Ala164 and Lys165) supporting the results. Several phytochemicals like anacardic acid, chlorogenic acid,cyanidin, delphinidin, ellagic acid, EGCG, emodin, macrocarpal A, myrciaphenone B, quercetin, rosmarinyl glucosid and theaflavin docked to the Rab5A active site. LIGPLOT results showed EGFR bound ellagic acid, hesperidin, isorhamnetin,

myrciaphenoneB and rosmarinyl glucoside. The docking showed that ligands bounded to the known active site, Phe723, Gly724, Cys797, Asp 837.

5.2 MD Simulation

In order to confirm the binding mode of ligands and to study the dynamics of the EGFR and Rab5A structures complexed with phytochemicals, MD simulations were performed using AMBER. Differences in stability and dynamics among the structure of the protein were observed. The result showed that only 4 out of 67 phytochemicals (EGCG, ellagic acid, macrocarpal A and quercetin) tested showed strong and stable interactions with Rab5A (PDB ID: 1N6H) and 1 phytochemical had good interaction with EGFR (hesperidin).

To explore the dynamic stability of protein-ligand complexes and to ensure the rationality of the sampling method, RMSDs of Rab5A and EGFR simulations were found. RMSD is a parameter indicating the stability of the entire protein in time. Calculating the RMSD values of both protein atoms showed that this was between 1.5 and 2.6 Å, which was within the acceptable range. RMSD values from MD simulation proved that the simulation systems had reached equilibrium and that the protein-ligand complexes were stable along the entirety of the simulation.

Hydrogen bond stability was monitored in both docking studies and MD simulations. Hydrogen bond distance of Rab5A fluctuated between 2 and 4 Å for the four ligands (EGCG, ellagic acid, macrocarpal A and quercetin). EGFR also showed good hydrogen bond interactions with hesperidin varying from 2 to 4 Å.

5.3 ADME and Toxicity

All the compounds studied are of natural origin and regularly consumed in human diet. However, both ADME and toxicity risk properties were evaluated. The overall drug score showed that it was within the acceptable range for all the shortlisted ligands. Hence, these compounds could be considered as potent lead compounds against cancer. However, this would need to be confirmed in lab experiments.

5.4 Recommendation for Future Research

The following could be done to extend the results from the current study:

1. Obtaining list of different structures of Rab5A and EGFR to compare molecular docking and MD simulation result with each protein.

2. Molecular modelling results using Rab5A and EGFR proteins could be tested *in vitro* and *in vivo* to confirm the effectiveness of using the chemicals as drugs.

5.5 Limitation of the study

The major limitations of this study are:

1. Due to limited resources, only 7 proteins and 67 phytochemicals could be used in this study.

2. *In silico* studies provided a short list of potential chemicals. However, this would need to be confirmed with wet lab experiments and finally in clinical trials.

5.6 Summary

In this project, a selection of natural molecules was screened *in silico* to identify if they could interact with proteins in different cancer pathways. Virtual screening and simulations were used to obtain a short list of candidate molecules that have potential to be used for cancer therapy.

This study aimed to answer four vital questions:

1. Does the selected chemicals form good interactions with the chosen protein?

2. If it binds, does it bind to the active site of the protein or at random locations?

3. If it was in the active site, does it bind tightly and stably or could it be detached easily?

4. Does the chosen chemicals have any risk making it unsuitable for humans?

Molecular docking and MD simulations were employed to study if the ligands would remain stable in the bound state. A comparison of the docking results with previous studies involving active site amino acid residues confirmed that the phytochemical ligands bound to the active site of the two target proteins – Rab5A and EGFR. Furthermore, simulation results showed that only four ligands remained stable in simulations which were EGCG, ellagic Acid, macrocarpal A and quercetin for Rab5A and hesperidin for EGFR.

Bibliography

A. Journal Article:

Al Dhaheri Y, Attoub S, Arafat K, AbuQamar S, Viallet J, Saleh A, et al. (2013) Anti-Metastatic and Anti-Tumor Growth Effects of Origanum majorana on Highly Metastatic Human Breast Cancer Cells: Inhibition of NFκB Signaling and Reduction of Nitric Oxide Production. *PLoS One* 8(7), e68808

Amin, A., Hamza, A. A., Bajbouj, K., Ashraf, S. S., & Daoud, S. (2011). Saffron: A potential candidate for a novel anticancer drug against hepatocellular carcinoma. *Hepatology*, 54(3), 857-867.

Chen, H., Yao, K., Nadas, J., Bode, A. M., Malakhova, M., Oi, N., Dong, Z. (2012). Prediction of molecular targets of cancer preventing flavonoid compounds using computational methods. *PloS One*, 7(5), e38261.

Das RP, Jagadeb M. 2015. Molecular docking approach towards targeting Leucine-rich alpha-2-glycoprotein (LRG1) through Phytochemicals to prevent bad angiogenesis of cancer. *International Journal of Current Microbiology and Applied Science.* 4(5): 929-938

Dhar ML, Dhar MM, Dhawan BN, Mehrotra BN, Ray C.(1968). *Indian Journal of Experimental Biology*. 6(4), 232–247

Friesner, R. A., Banks, J. L., Murphy, R. B., Halgren, T. A., Klicic, J. J., Mainz, D. T., Perry, J. K. (2004). Glide: A new approach for rapid, accurate docking and

scoring. 1. method and assessment of docking accuracy. *Journal of Medicinal Chemistry*, 47(7), 1739-1749.

Friesner, R. A., Murphy, R. B., Repasky, M. P., Frye, L. L., Greenwood, J. R., Halgren, T. A., Mainz, D. T. (2006). Extra precision glide: Docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *Journal of Medicinal Chemistry*, 49(21), 6177-6196.

Gibson, E., Wardle, J., & Watts, C. (1998). Fruit and vegetable consumption, nutritional knowledge and beliefs in mothers and children. *Appetite*, 31(2), 205-228.

Gleeson, M. Paul; Modi. (2012). The Challenges Involved in Modeling Toxicity Data In Silico: A Review. *Current Pharmaceutical Design*, 18(9), 1266-1291.

Goodsell, D. S., Morris, G. M., & Olson, A. J. (1996). Automated docking of flexible ligands: Applications of AutoDock. *Journal of Molecular Recognition*, 9(1), 1-5.

Gopalakrishnan SB , Kalaiarasi T , Gnanendra S. (2015). Molecular docking studies of phytochemical constituents of the fruits of Cucumis trigonus roxb. against hepatocarcinomo receptors. *International Journal of Research in Pharmacy and Science*, 5(2), 30 – 36

Hahn, N.I. (1998). Are Phytoestrogens Nature's Cure for What Ails Us? A Look at the Research. *Journal of the American Dietetic Association* 98, 974-976.

Henson, E. S., Johnston, J. B., Los, M., & Gibson, S. B. (2007). Clinical activities of the epidermal growth factor receptor family inhibitors in breast cancer. *Biologics: Targets & Therapy*, 1(3), 229-239.

Holt CP. The scope of the UAE cancer problem. (1985). *Emirates Medeical Journal*. 3, 163-6.

Hutagalung, A. H., & Novick, P. J. (2011). Role of rab GTPases in membrane traffic and cell physiology. *Physiological Reviews*, 91(1), 119-149.

Jones, G., Willett, P., Glen, R. C., Leach, A. R., & Taylor, R. (1997). Development and validation of a genetic algorithm for flexible docking. *Journal of Molecular Biology*, 267(3), 727-748.

Khan S.Y., Gutiérrez-de-Terán, H., Boukharta, L., & Åqvist, J. (2014). Toward an optimal docking and free energy calculation scheme in ligand design with application to COX-1 inhibitors. *Journal of Chemical Information and Modeling*, 54(5), 1488-1499.

Kontoyianni, M., McClellan, L. M., & Sokol, G. S. (2004). Evaluation of docking performance: Comparative data on docking algorithms. *Journal of Medicinal Chemistry*, 47(3), 558-565.

Kramer, B., Rarey, M., & Lengauer, T. (1999). Evaluation of the FLEXX incremental construction algorithm for protein–ligand docking. *Proteins: Structure, Function, and Bioinformatics*, 37(2), 228-241.

Kräutler, V.; Oostenbrink, C. 2005. The GROMOS Software for Biomolecular Simulation: GROMOS05. *Journal of Computational Chemistry*. 26, 1719–1751.

Liu, J., & Kipreos, E. T. (2000). Evolution of cyclin-dependent kinases (CDKs) and CDK-activating kinases (CAKs): Differential conservation of CAKs in yeast and metazoa. *Molecular Biology and Evolution*, 17(7), 1061-1074.

MacKerell, A. D., Bashford, D., Bellott, Dunbrack, R. L., Evanseck, J. D., Field, M. J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir, L., Kuczera, K., Lau, F. T. K., Mattos, C., Michnick, S., Ngo, T., Nguyen, D. T., Prodhom, B., Reiher, W. E., Roux, B., Schlenkrich, M., Smith,J. C., Stote, R., Straub, J., Watanabe, M., Wiórkiewicz-Kuczera, J., Yin, D., Karplus, M. All-Atom Empirical Potential for Molecular Modeling and Dynam- ics Studies of Proteins . 1998. *The Journal of Physical Chemistry A. B.* 102, 3586–3616.

McCubrey.J A, Steelman, L, Chappell.W, Abrams,S.L, Wong.WT, Chang, F, Lehmann, B,Terrian, David M; Milella, M,Tafuri, A .(2007).Roles of the Raf/MEK/ERK pathway in cell growth, malignant transformation and drug resistance.*Biochimica et Biophysica Acta (BBA)-Molecular Cell Research.* 1773(8), 1263-1284

Morris, M., Lim-Wilby M. (2008). Molecular Modeling of Proteins. *Methods Molecular Biology.* 443, pp 365-382

Nowell, P. C. (1976). The clonal evolution of tumor cell populations. *Science (New York, N.Y.).* 194(4260), 23-28.

Nyati, M. K., Morgan, M. A., Feng, F. Y., & Lawrence, T. S. (2006). Integration of EGFR inhibitors with radiochemotherapy. Nature Reviews Cancer, 6(11), 876-885.

Pritam Das, R., & Jagadeb, M. (2015). Molecular docking approach towards targeting leucine-rich alpha-2-glycoprotein (LRG1) through phytochemicals to prevent bad angiogenesis of cancer. *International Journal of Current Microbiology and Applied sciences*, 4(5), 929-938

Quintas-Cardama, A., & Verstovsek, S. (2013). Molecular pathways: Jak/STAT pathway: Mutations, inhibitors, and resistance. *Clinical Cancer Research*:19(8), 1933-1940.

Rajalingam, K., Schreck, R., Rapp, U. R., & Albert, Š. (2007). Ras oncogenes and their downstream targets. *Biochimica Et Biophysica Acta (BBA)-Molecular Cell Research*, 1773(8), 1177-1195.

Reddy V.V., Sirsi M. (1969). Effect of Abrus precatorius L. on Experimental Tumors. *Cancer Research*. 29(7), 1447–1451.

Salomon-Ferrer, R., Case, D. A., & Walker, R. C. (2013). An overview of the amber biomolecular simulation package. Wiley Interdisciplinary Reviews: *Computational Molecular Science*, 3(2), 198-210.

Schaeffer, H. J., & Weber, M. J. (1999). Mitogen-activated protein kinases: Specific messages from ubiquitous messengers. *Molecular and Cellular Biology*, 19(4), 2435-2444.

Smith, D. A., & Schmid, E. F. (2006). Drug withdrawals and the lessons within. *Current Opinion in Drug Discovery & Development*, 9(1), 38-46.

Surh, Y. (2003). Cancer chemoprevention with dietary phytochemicals. *Nature Reviews Cancer*, 3(10), 768-780.

Sutherland, R. L., & Musgrove, E. A. (2009). CDK inhibitors as potential breast cancer therapeutics: New evidence for enhanced efficacy in ER disease. *Breast Cancer Research*, 11(6), 112.

Taylor, T. E., Furnari, F. B., & Cavenee, W. K. (2012). Targeting EGFR for treatment of glioblastoma: Molecular basis to overcome resistance. *Current Cancer Drug Targets*, 12(3), 197-209. doi: CCDT-EPUB-20120116-001 [pii]

Trisilowati, & Mallet, D. G. (2012). *In silico* experimental modeling of cancer treatment. *Oncology*, 828701.

Verdonk, Marcel L; Cole, Jason C; Hartshorn, Michael J; Murray, Christopher W; Taylor, Richard D. (2003). Improved protein–ligand docking using GOLD. *Proteins: Structure, Function, and Bioinformatics*, 52 (4), 609-623,

Vogelstein, B., & Kinzler, K. W. (2004). Cancer genes and the pathways they control. *Nature Medicine*, 10(8), 789-799.

Walton, E. B., & VanVliet, K. J. (2006). Equilibration of experimentally determined protein structures for molecular dynamics simulation. *Physical Review E*, 74(6), 061901.

Will, R. (2002). Cancer and the role of cell cycle check points. Reviews in *Undergraduate Research,* 1, 1-7

Zhao, Z., Liu, X., Wu, H., Zou, S., Wang, J., Ni, P., Fan, Q. (2010). Rab5A overexpression promoting ovarian cancer cell proliferation may be associated with APPL1- related epidermal growth factor signaling pathway. *Cancer Science*, 101(6), 1454-1462.

Zhu, G., Liu, J., Terzyan, S., Zhai, P., Li, G., & Zhang, X. C. (2003). High resolution crystal structures of human Rab5A and five mutants with substitutions in the

catalytically important phosphate-binding loop. *The Journal of Biological Chemistry*. 278(4), 2452-2460.

Zhong, H., Tran, L. M., Stang, J. L. (2009). Induced-fit docking studies of the active and inactive states of protein tyrosine kinases. Journal of Molecular Graphics and Modelling. 28(4), 336–346.

B. Books:

Costa, M.A., Z.Q. Zia, L.B Davin, and N.G. Lewis. (1999). "Chapter Four: Toward Engineering the Metabolic Pathways of Cancer-Preventing Lignans in Cereal Grains and Other Crops." In Recent Advances in Phytochemistry, vol. 33, Phytochemicals in Human Health Protection, Nutrition, and Plant Defense, ed. J.T. Romeo, 67-87. New York: Kluwer Academic/Plenum Publishers,

Mathai, K. (2000). Nutrition in the Adult Years. In Krause's Food, Nutrition, and Diet Therapy, 10th ed., ed. L.K. Mahan and S. Escott-Stump, 271, 274-275. Philadelphia: W.B. Saunders Co.

Meagher, E., & Thomson, C. (1999). Vitamin and mineral therapy. Medical Nutrition and Disease, 2nd Ed., G Morrison and L Hark, Malden, Massachusetts: Blackwell Science Inc,, 33-58.

Weinberg, R. (2013). The biology of cancer Garland Science, Chapter 2: The Nature of Cancer, second edition, Garland Science, Taylor & Francis Group, LLC, USA

Hejmadi, M. (2010). Introduction to cancer biology Chapter 7: Spread to other sites (metastasis), pp.40-43, Bookboon,

C. Thesis/Dissertation:

Tiikkainen, P. (2010). Study of ligand-based virtual screening tools in computer-aided drug design. [e- Thesis]- University of Turku, Finlannd.

Miller L.B. (2002). Phytochemicals as a motivational tool to change fruit, vegetable and whole grain consumption [e- Thesis]-Master Thesis, University of Wisconsin-Stout, Menomonie.