Ph.D ANKITA JUIT, WAKNAGHAT 2018

COMPUTATIONAL STUDIES TO INVESTIGATE HUMAN DNA REPAIR ASSOCIATED MALIGNANCIES WITH SPECIAL RELEVANCE TO COLORECTAL CANCER

THESIS SUBMITTED IN FULFILLMENT
FOR THE REQUIREMENTS OF THE DEGREE OF

DOCTOR OF PHILOSOPHY

BY

ANKITA

Department of Biotechnology and Bioinformatics

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY

WAKNAGHAT, SOLAN-173234, HP, INDIA

DECEMBER, 2018

COMPUTATIONAL STUDIES TO INVESTIGATE HUMAN DNA REPAIR ASSOCIATED MALIGNANCIES WITH SPECIAL RELEVANCE TO COLORECTAL CANCER

THESIS SUBMITTED IN FULFILLMENT
FOR THE REQUIREMENTS OF THE DEGREE OF

DOCTOR OF PHILOSOPHY

BY

ANKITA

Department of Biotechnology and Bioinformatics

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY

WAKNAGHAT, SOLAN-173234, HP, INDIA

DECEMBER, 2018

Dedicated to My Family and Almighty God

IN THE LOVING MEMORY OF MY GRANDMOTHER

TABLE OF CONTENTS

DECLARATION

I certify that:

a. The work contained in this thesis is original and has been done by me under the guidance of my supervisor.
b. The work has not been submitted to any other organization for any degree or diploma.
c. Wherever, I have used materials (data, analysis, figures or text), I have given due credit by citing them in the text of the thesis.

Ankita **Date:**

(Enrollment No. 136502)

Department of Biotechnology and Bioinformatics

Jaypee University of Information Technology, Waknaghat, India

Email: shukla.ankita39@gmail.com

CERTIFICATE

This is to certify that the thesis entitled, **"Computational Studies to Investigate Human DNA Repair Associated Malignancies with Special Relevance to Colorectal Cancer"** which is being submitted by **Ankita (Enrollment No. 136502)** in fulfillment for the award of degree of **Doctor of Philosophy** in **Bioinformatics** at **Jaypee University of Information Technology, India** is the record of candidate's own work carried out by her under my supervision. This work has not been submitted partially or wholly to any other University or Institute for the award of this or any other degree or diploma.

Dr. Tiratha Raj Singh **Date:**

Associate Professor,

Department of BT and BI

Jaypee University of Information Technology, India

Email: tiratharaj@gmail.com

ACKNOWLEDGEMENT

"Every great person is always being helped by everybody; for their gift is to get good out of all things and all persons." - John Ruskin

The "Philosophy of Doctorate" has been a long and audacious journey for me. I am deeply indebted to lots of persons while completing this degree that stood by me through good and bad times and helped to conquer it.

*First and foremost, I would like to thank the almighty **"GOD"** to give me all the courage to complete this degree. I would like to thank the **Jaypee University of Information Technology (JUIT)** for providing teaching assistantship and all the required resources.*

*I owe my deepest sense of gratitude to my mentor **Dr. Tiratha Raj Singh** for the continuous support of my Ph.D study and related research, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Ph.D study. I would like to thank you for encouraging my research and for allowing me to grow as a researcher. Your advice on both research as well as on my career have been priceless. You haven't left any stones unturned for my research practices being it a review, research, general articles or, the book chapters. Thank you so much for each and every effort you have invested on me and believed in my abilities and accepted me as a potential candidate for such studies. Beyond a "**Guru**" you are a person with all valuable qualities along with research skills I have also learnt wonderful qualities from you only. Also **Dr. Ragini Raj Singh**, for their generous nature and words of wisdom whenever I felt disheartened. Both of you treated all group members like a family that overcame the feeling of away from home.*

*Also, I would like to thank **Dr. Ragothaman Yennamalli** for making me fit into the field of structure biology. His dedication and straightforward approach helped me to learn all the complex tasks very easily. I am indebted to all the necessary help and the suggestions you provided for my betterment. I would also like to thank my committee members **Dr. Pradeep Kumar Gupta**, **Dr. Rahul Srivastava**, and **Dr. Jayashree Ramana**, for their insightful comments and encouragement, and also for the queries that incented me to widen my research view from various perspectives. Special thanks to **Prof. Ghanshyam Singh** who provided me every possible guideline wherever and whenever possible.*

*I owe my deepest gratitude to **Prof. Vinod Kumar (Vice Chancellor, JUIT)**; **Maj.Gen. Rakesh Bassi (Registrar, JUIT)**; **Prof. S.D Gupta (Dean, Academic and Research, JUIT) Dr. Sudhir Kumar (HOD, BT & BI)**; for providing me with the opportunity to pursue my doctorate degree. I would like to sincerely thank*

Prof. R.S. Chauhan (former Dean and HOD, BT & BI) *for his encouragement and cooperation during past few years.*

My special thanks to my mentors during Masters study ***Dr. Suresh Sharma****,* ***Dr. Veena Puri****, and* ***Dr. Ashok Kumar*** *for their consistence guidance throughout my career.*

I would like to thank everyone in the BI Project Lab ***Ashwani, Kusum, Nupur, Imran, Smriti, Ankush*** *and* ***Nadia*** *for all the fun we have had throughout this journey that helped making it bit smoother. Also, I would like to thank* ***Mr. Ankur Choudhary*** *for his unconditional support during the most difficult phase of this journey; I could not have imagined getting through such situations without your help, thank you so much. I would especially like to thank* ***Manika mam****,* ***Priya mam****, and* ***Ira mam*** *for their guidance as seniors and sharing their work experiences. I would like to thank our senior lab assistant* ***Somlata mam*** *for providing technical assistance as and when required. I would also like to acknowledge* ***teaching****,* ***administration****,* ***server*** *and* ***library staff*** *for their help.*

*Last, but not the least I am deeply indebted to my parents (****Sh. Shyam Sunder Shukla*** *and* ***Smt. Rekha Shukla****) for all their love, support, patience and encouragement during this journey. I can't thank enough for all they have done for me to become what I am today. I highly believe that all my dedication and love for science comes from you and although for world it seems to be me but its "****you who did it****". During this journey we got the tag of research scholars and I truly believe that if there is any scholarity imbibed in me it's because of you. I would like to thank my dearest bhaiya and bhabhi* ***Gaurav Shukla*** *and* ***Poonam Shukla*** *for supporting me in every way for giving all the affection and making me feel happy whenever I felt sad.*

My special dedication to my loving nani maa ***Late Smt. Nirmala Awasthi*** *and my dear cousin* ***Late. Adarsh Awasthi*** *who wanted to see me getting doctorate degree, though you are not physically with me but your presence remained in my heart everytime. I missed you a lot through this journey.*

I appreciate the contribution of countless people I met along the way for making this journey an experience which I will always keep dearly in my heart.

Ankita

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

HGP	Human Genome Project
NIH	National Institutes of Health
DDR	DNA Damage Response
ROS	Reactive Oxygen Species
BER	Base Excision Repair
NER	Nucleotide Excision Repair
MMR	Mismatch Repair
DSBR	Double Strand Break Repair
UV	Ultraviolet
MGMT	O^6-methylguanine methyltransferase
AP	Apyrimidinic/Apurinic
HRR	Homologous Recombination Repair
NHEJ	Non-Homologous End Joining
SDSA	Synthesis Dependent Strand Annealing
BIR	Break Induced Replication
PKcs	Protein Kinases
Hh	Hedgehog
CSCs	Cancer Stem Cells
EMT	Epithelial-to-Mesenchymal Transition
TGFβ	Transforming Growth Factor–β
MK2	Protein Kinase 2
NF-kB	Nuclear Factor kappaB
HNPCC	Hereditary Nonpolyposis Colorectal Cancer
CRC	Colorectal Cancer
MSI	Microsatellite Instability
CIN	Chromosomal Instability
CIMP	CpG island methylator phenotype
LOH	Loss of Heterozygosity
IBD	Irritable Bowel Disease

FAP	Familial Adenomatous Polyposis
AFAP	Attenuated Familial Adenomatous Polyposis
CNV	Copy Number Variation
LS	Lynch Syndrome
EC	Endometrial Cancer
DREMECELS	DNA REpair MEchanism in Colorectal, Endometrium and Lynch Syndrome
GUI	Graphical User Interface
CDD	Conserved Domain Database
COSMIC	Catalog of Somatic Mutations in Cancer
Lgr5	Leucine-rich repeat-containing G protein-coupled Receptor 5
ISCs	Intestinal Stem Cells
FZD	Frizzled Receptors
DVL	Dishevelled
SIM	Single Input Module
MIM	Multiple Input Module
RMSD	Root Mean Square Deviation
RMSF	Root Mean Square Fluctuation
Rg	Radius of Gyration
SASA	Solubility Accessible Surface Area
MD	Molecular Dynamics

CHAPTER - 1

INTRODUCTION

~ Do not go where the path may lead, go instead where there is no path and leave a trail.

- Ralph Waldo Emerson

ABSTRACT

The DNA repair system sustains the genome integrity by upholding the poor responses through the diverse range of causal mechanisms. These responses are generally taken care by regulation of cell-cycle mechanism. During the checkpoint in the cell cycle, it is ensured that a DNA in the cell remains intact before undergoing the DNA replication process. If the checkpoints fail to mount their effect, the process can lead to the accumulation of damage that in long terms impose mutations. This can give rise to the variety of the DNA repair deficient disease mechanisms and cancer is one of them. These inflicted mutations could prompt chromosomal aberrations thus impact the functioning of key genes. Majorly, these genes are either having the role of an oncogene or, tumor suppressor ones. The oncogenes and tumor suppressors play a crucial role in maintaining the cellular state and check cells to undergo malignant transformation. This confirms that preserving genomic integrity is vital for maintaining cell homeostasis. The fidelity of the DNA repair processes promotes repair of the damages through extrinsic and intrinsic factors that help to manage cell survival. Maintaining genome integrity is, therefore, a key challenge for every living organism due to its central role in cell division, and regulation of the cell cycle.

There are varieties of cancers today that are contributed by the improper DNA repair mechanisms. These include skin cancer, breast cancer, lung cancer, prostate cancer, colorectal cancer and many more. My study has focused on the Lynch syndrome-associated cancers that include primarily colorectal cancer and endometrial cancer. In most of the study, the major work has been done implicated for DNA repair mechanisms and its deficient role in colorectal cancer. Various omics-based approaches have been implemented for determining biomarkers specifically for colorectal cancers, endometrial cancer and Lynch syndrome. The main aim of work is the identification and characterization of genes that play a role in genome stability in these cancer types. It is anticipated that these studies will lead to the identification of novel genetic variants associated with mutator phenotypes that potentially increase the possibility of cancer predisposition. Hopefully, this knowledge will facilitate in the future the diagnostic and treatment of cancer. The work complies the identification and characterization of genes that contribute with genome integrity. The current study investigated the mechanistic aspects of colorectal cancer and the novel genes that contribute to DNA replication fidelity and cancer.

1.1INTRODUCTION

The human genome is part of cells that are not visible through the naked eye; each of them encompasses DNA in their nucleus with more than 3-billion base pairs [1]. A genome comprises the genetic instructions that are compiled through a complete set of genes (~20,000-25,000) enclosed in the organism's DNA [2]. The genome contains all of the information required to build and maintain the organism that helps it to grow and develop. The information regarding human genome was obtained after the successful completion of the Human Genome Project (HGP). The Human Genome Project began in 1990 and was completed in 2003 by the National Institutes of Health (NIH) and the U.S. Department of Energy [3]. DNA sequencing technologies played a key role in this project [4]. These efforts aid in determining the sequence of human genome and identify the genes that it contains. As the human genome project gets completed, it helped to identify the cause of multiple genetic diseases, as some of the rare diseases are noticed to be caused by the change in a variant. Therefore examining the genome of a person affected by a rare disease can help to identify DNA variations that might be causing the problem.

Today world is fighting with diverse forms of complex diseases. The varieties of malignancies are there in the society that needs a quick measure to reduce the load of the patient's sufferings. In cancerous condition, major genomic alterations have been found in tumor cells in comparison to the healthy cells [5]. Such alterations have been detected through the comparison of tumor cells with the normal ones. This comparison helps to provide clues for the distinguished patterns that provide the ways to treat cancer. Therefore taking care of genome is one of the biggest challenges due to its vulnerability to the intrinsic and extrinsic damaging agents. The maintenance of genome integrity is essential for organism survival and for the inheritance of traits to offspring. Also, though organisms need genome stability, they must allow mutational changes to drive adaptation and evolution [6]. The damaging factors includes formation of oxidative species via metabolism, base loss, radiations, and chemicals exposure that severely compromise its integrity [7]. These factors cause the genomic instability due to DNA damage, aberrant DNA replication or uncoordinated cell division that in most cases leads to chromosomal aberrations and gene mutations [8]. The chromatin regulators are signaling coordinators that contour the epigenetic landscape by acting as potential gatekeepers [9].

Therefore to deal with such damages there is a requirement of the mechanism that takes care of all forms of the alterations. By default, the cells have an inbuilt mechanism called DNA

repair that take care of all occurring damages that happen to the DNA on daily basis [10]. Therefore, to overcome the damaging effect of the unusual mechanism and to preserve genomic integrity, three major and evolutionarily sealed cellular pathways have evolved [11]. The first one is a DNA damage response (also called DDR) that takes care of proficient repair for all type of damages. The second one is a chromosomal replication pathway that presides over accurate and unimpeded replication of DNA and the third one is a chromosome segregation pathway that preserves the actual number of chromosomes during cell division. These pathways work by a crosstalk method thus forming a network so that if one pathway gets disrupted the other ones come into the action to protect genome integrity thus help to maintain cell homeostasis. The signaling network of DNA damage contains key components like DNA damage sensors for sensing the damage, signal transducers for conveying the signal, mediators that help in cell-cell communication and effectors that show a response to a stimulus. Majorly among all sort of respondents, the DNA polymerases are crucial for maintenance of genome integrity in organisms [12]. Figure 1.1 shows the type of damages DNA faces; checkpoints to manage the damage and the consequences if the damage remains unrepaired.

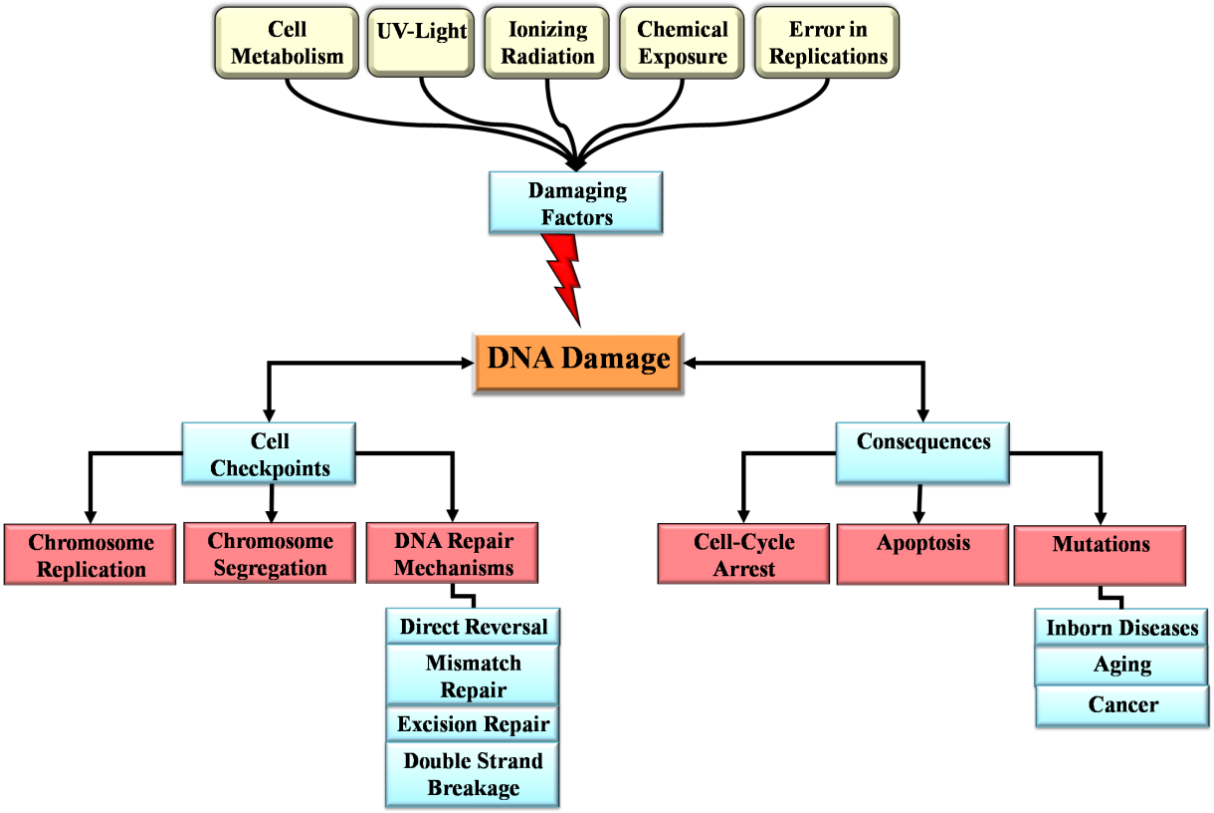


Figure 1.1 DNA Damage responses and its integrity in the maintenance of genome homeostasis.

1.2DNA REPAIR

The cellular DNA is most susceptible to the involuntary damage from a vast variety of sources like radiations, chemical exposure, reactive metabolic products, active oxygenic species, and replication errors [13]. Along with the above said damages the unprompted hydrolysis of nucleotide residues that occurs at 37°C is another unavoidable form of DNA damage that should be considered for instant repair [14]. The damage to the DNA is happening right from the beginning of life and for this, repair enzymes showed sturdy evolutionary conservation for almost all forms of the organism. The cells have evolved efficient repair mechanisms to handle the DNA damage for preventing the chromosomal aberrations and mutations that have a possibility to initiate cancer predisposition. Although repair mechanism effectively repairs damages, it has been found that sometimes improper activation of repair pathways causes tumor cells to become immortalized that make them treatment-resistant [15]. DNA serves as a permanent copy of the cell that carries genetic message and can undergo a variety of chemical reactions. As DNA is the basis of all, therefore, its structure is of utmost importance. The changes occurring in the DNA are brought by several types of mutations that affect its structure [16]. In addition, chemical alterations occur in DNA either impulsively or through chemicals or radiations [17]. This results in hindrance in replication or transcription processes and can lead to high rate of mutations that follows conditions intolerable during the cell cycle. The damages can be from endogenous or exogenous sources however endogenous sources were considered to be majorly contributing towards mutations that follow malignancy [18].

The damage to DNA through endogenous and exogenous sources intimidates genome and also the epigenome. This follows requirement of response pathways to operate on the damaging factors. The response pathways are used for sensing, processing and repairing the DNA damage. The endogenous damage occurs by reactive oxygen species (ROS) derived from metabolic products while the exogenous damage occurs by radiations, viruses, plant toxins, and hydrolysis [19]. While most DNA damage can be repaired but, these repair systems are not accurate to 100% and therefore cause accumulation of damages and cause disease progression [20]. The DNA-damaging agents though have possibility to cause human cancer but, satirically are among the primary means available to clinicians for treating cancerous malignancies.

1.2.1 Mechanisms of DNA Repair

Based on type of DNA damage, cell has evolved to gain a suitable method to repair the damages. These repair methods have been divided into two general classes that involve direct reversal method and damaged bases removal (Table 1.1) [21]. The damaged bases are further compensated by a different set of mechanisms depending on degree and types of DNA damage, which includes, base excision repair (BER), mismatch repair (MMR), nucleotide excision repair (NER), and double-strand break repair (DSBR) [21]. However, sometimes cells may proceed to follow apoptosis or senescence (biological aging) in case of intense damage. The rate of repair primarily depends on the cell type and age. The damage check is first and foremost duty of the DNA polymerases; that make a check while adding each base during the replication process, a process called proofreading [22]. Therefore, in case a wrong nucleotide has been detected by a polymerase it has to be replaced right away, before continuing DNA synthesis.

Table 1.1 DNA repair mechanisms and causal factors

Types		**Induced Factors**	**Type of Damages Corrected**	**Catalytic Agents**
Direct Reversal		N-methyl-N'-nitro-N-nitrosoguanidine (MNNG), N-methyl-N-nitrosourea (MNU), and methyl methanesulfonate (MMS)	O-alkylated and N-alkylated products	O6-methylguanine-DNA methyltransferases (MGMT) and ALKBH α-ketoglutarate Fe(II) dioxygenases (FeKGDs)
Excision Repair	Base Excision Repair	X-rays, Oxygen radicals, Alkylating agents, Spontaneous reactions	Uracil, Abasic site, 8-oxoguanine, Single-strand break	DNA glycosylase, AP endonuclease, DNA polymerase, Ligase
	Nucleotide Excision Repair	UV-light, Polycyclic, hydrocarbons	(6-4)PP, Bulky adduct, CPD	Excinuclease (UvrABC complex), helicase, DNA polymerase, Ligase
	Mismatch Repair	Replication Errors	A-G Mismatch, T-C Mismatch, Insertion, Deletion	MutH endonuclease, exonuclease, helicase, DNA polymerase and Ligase
Double Strand Breakage	Homologous Recombination	X-rays, Anti-tumor agents (cis-Pt, MMC)	Interstrand cross-link, Double-strand break	Exonuclease, endonuclease, polymerase, Ligase
	Non-Homologous End Joining	X-rays, Anti-tumor agents (*cis*-Pt, MMC)	Interstrand cross-link, Double-strand break	DNA Ligase, Ku-protein, Ligase

1.2.1.1 Direct Reversal of DNA Damage

The basic repair mechanism entails direct reversal of damage and is considered to be energy efficient method for damage restoration. The major type of damages like pyrimidine dimers and chemical adducts formed during reaction mechanisms are repaired in this way. The pyrimidine dimers are formed via exposure to ultraviolet (UV) light [23] while chemical adducts that include alkylated guanine residues are formed due to the addition of methyl or ethyl groups at the O^6-position in the purine ring [24]. These dimers deform double helical structure and follow transcription or replication blockage past the damaged site. In case of dimers, adjacent pyrimidines that lie on the same DNA strand are joined due to the formation of a cyclobutane ring. The cyclobutane ring is formed via saturation of double bonds between carbons at 5^{th} and 6^{th} positions. The mechanism to overcome the damage follows the photoreactivation process that makes use of the visible light causing a break in cyclobutane ring structure, thus restoring the original bases. If remains unrepaired, the damages become a major cause of almost all forms of skin cancer [25] (Figure 1.2 a).

Another form of damage occurs through the reaction between DNA and alkylating agents. The alkylating agents are reactive compounds formed by transfer of methyl or ethyl groups to a DNA base, therefore leading to the chemical modification. Methylation is a prominent form of DNA damage that occurs at the O^6-position of guanine that forms O^6-methylguanine product. This product forms complementary base pairs with thymine instead of cytosine. The O^6-methylguanine methyltransferase (MGMT) plays a major role in repairing the lesion by transfering methyl group from O^6-methylguanine to a cysteine residue that is bound to it [26]. This procedure helps in removing chemical modification and restoration of the original base pair (Figure 1.2 b).

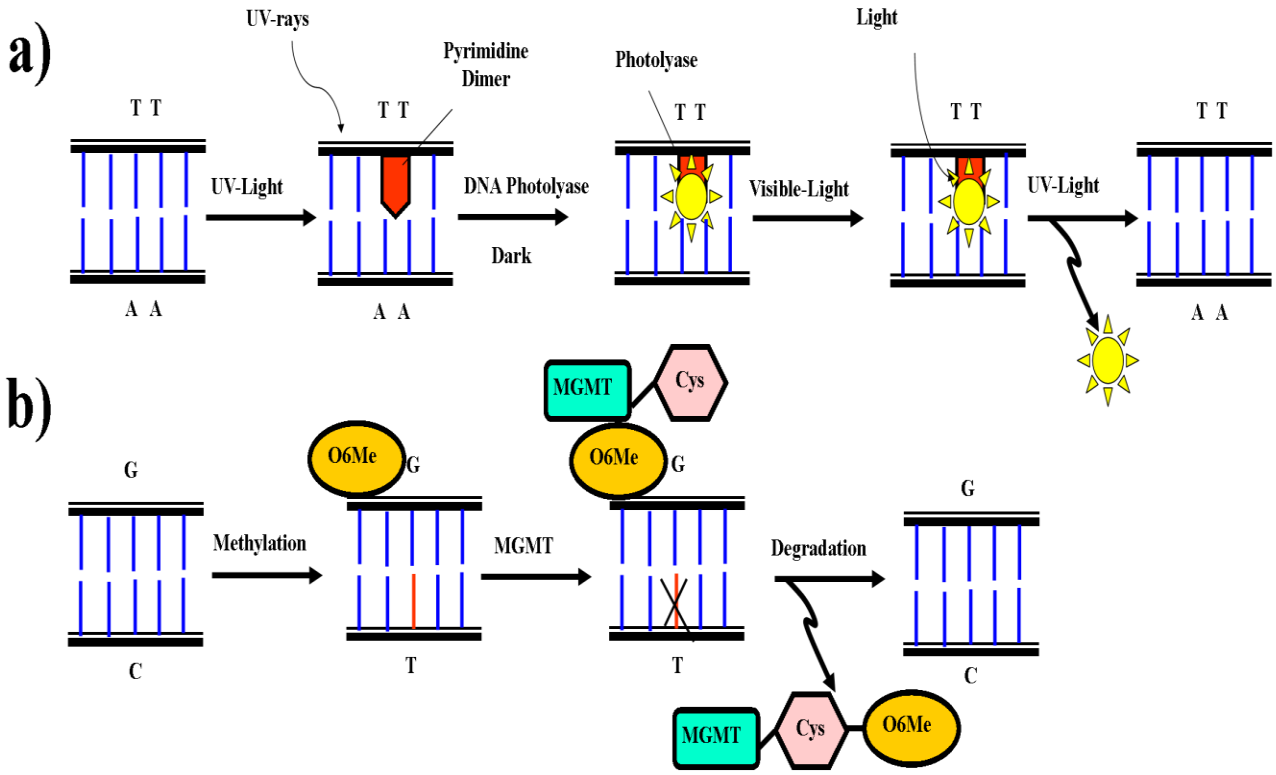


Figure 1.2 Direct DNA Damage Reversal a) The photoreactivation process to remove pyrimidine dimers b) The removal of the alkylating agent through MGMT implementation

1.2.2 Excision Repair

Generally, the reversal repair method takes care of DNA damages however, not all types of damages get repaired by this mechanism. The excision repair mechanism is a more general means of repairing a wide variety of chemical alterations that happen to the DNA. In this method, the damaged DNA is recognized and removed either in the form of free bases or as nucleotides. Therefore on the basis of specificity of the damage type, they are categorized as BER, NER, and the MMR methods. In BER**,** just the damaged base is removed, in NER, a patch of nucleotides is removed and in MMR, the non-complementary base pair is removed.

1.2.2.1 Base Excision Repair (BER)

As discussed above the BER mechanism is used to remove the damaged bases. It takes care of the damages that occur via processes such as oxidation, deamination and alkylation [27]. A group of enzymes called glycosylases play a key role in this repair mechanism [28]. These glycosylase notice and eradicate a specific kind of damaged base. Sometimes in a chemical reaction called deamination, there is a conversion of a Cytosine base into Uracil. During the

replication process, Uracil will pair with Adenine rather than Cytosine; such change can lead to a mutation. Therefore, to prevent such mutations glycosylase comes into the action for deaminated Cytosines removal. After the base gets removed, it leads to gap and forms abasic site. Then an apyrimidinic or, apurinic (AP) endonuclease or glycosylase or, lyase helps in cleaving the phosphodiester bond. Afterwards, DNA polymerase comes into the action and adds a complementary nucleotide. Finally, ligation of the DNA backbone restores the native structure and sequence [29]. Figure 1.3 illustrates the overall mechanism of the base excision repair, sometimes after the addition of the complementary base pair by DNA polymerase; the pathway undergoes two alternatives; one where the DNA ligase cause the binding of the backbone (short patch) and the other where DRPase along with ligase fills the backbone (long patch).

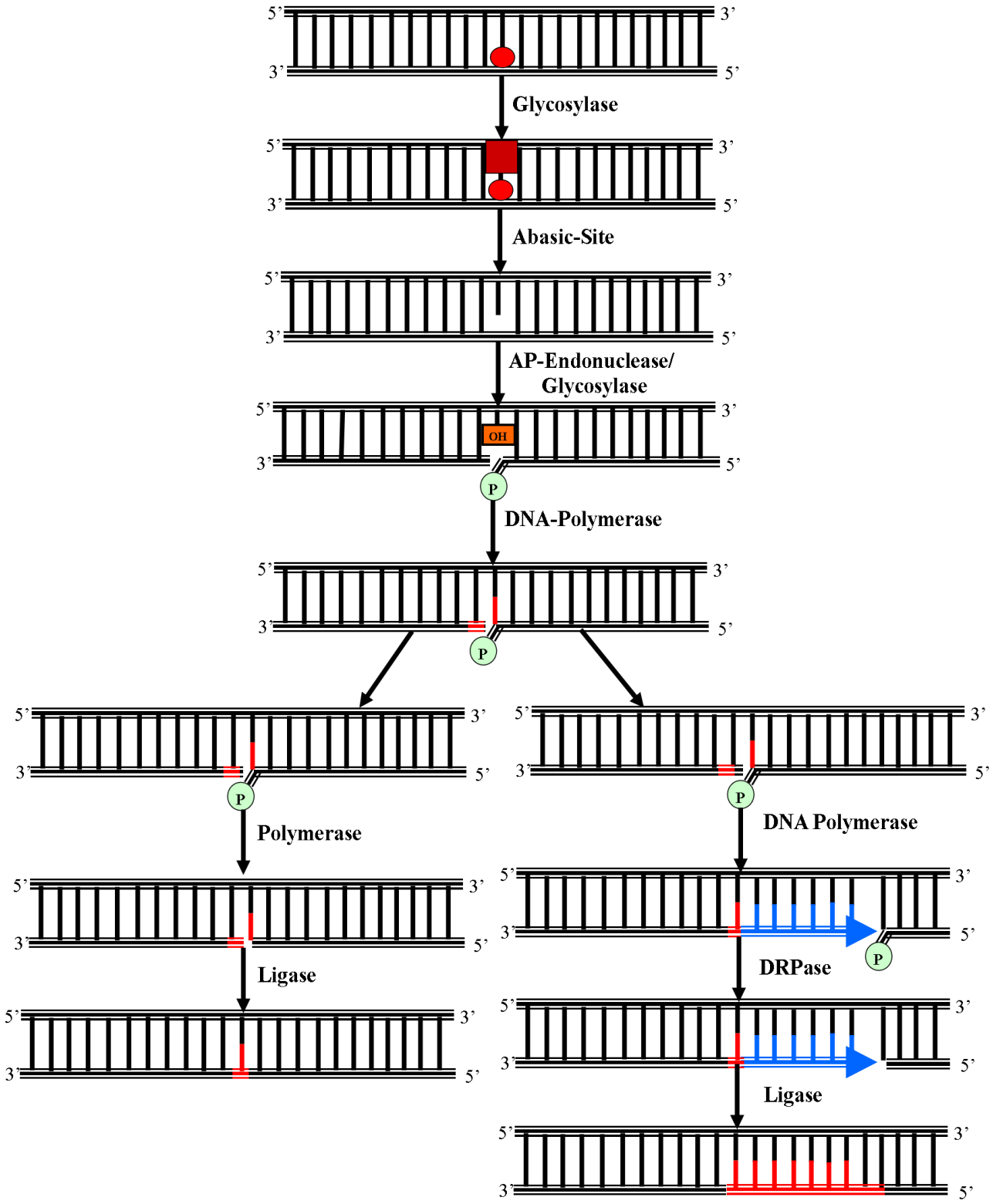


Figure 1.3 The Base-Excision Repair (BER) mechanism illustrating the repair of the short and long patch.

1.2.2.2 Nucleotide Excision Repair (NER)

The NER mechanism as its name suggests is a way to repair the damaged nucleotides. The method determines and corrects types of damage that distort the double-helical form of DNA. The pathway corrects bases modified through the bulky chemical groups formed by UV-light, environmental mutagens, and the chemotherapeutic agents [30]. The damage that occurs via the UV radiation can make cytosine and thymine bases react with neighboring similar bases, thus forming bonds that distort the double helix and induce replication errors. In NER the damaged nucleotide(s) are removed along with the surrounding DNA patch. The process involves a helicase that opens the DNA, the DNA-cutting enzymes then chop out the damaged part of the nucleotides. This form a gap in the DNA backbone that needs to be filled by the repairing enzyme. The DNA polymerase fills up the missing patch of DNA, and a ligase is implemented to seal the gap in the backbone of a strand [31]. Figure 1.4 illustrates the complete repair mechanism of the NER to replace the damaged nucleotide patch of the strand.

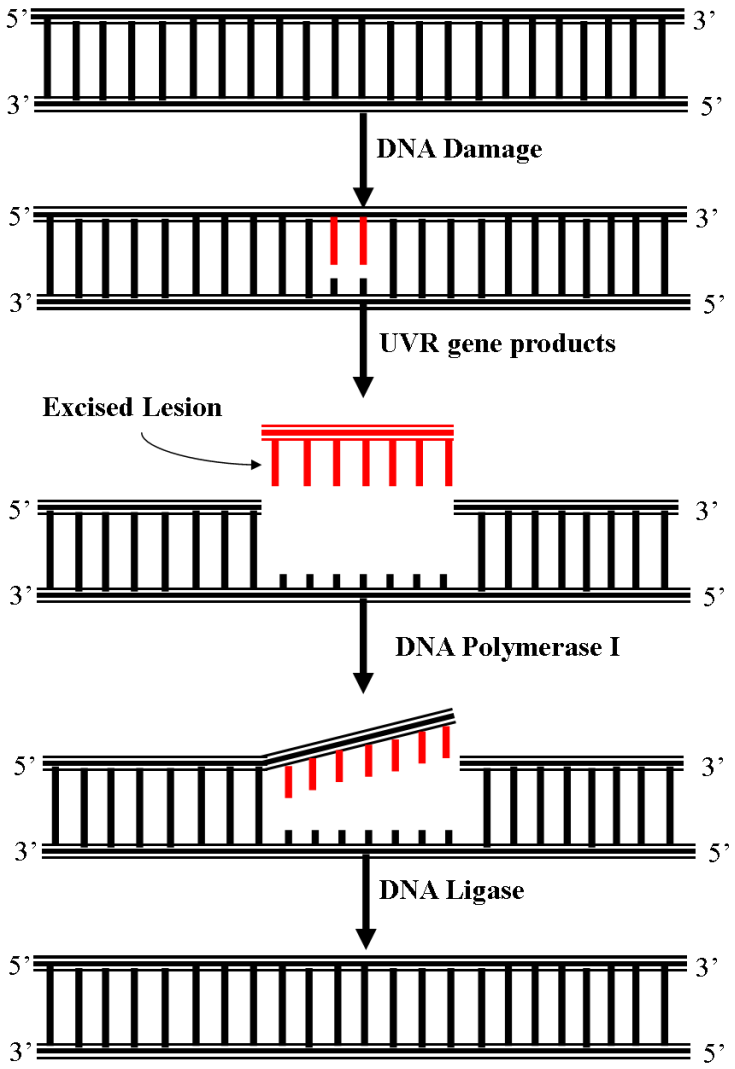


Figure 1.4 The Nucleotide Excision Repair (NER) mechanism repairing the damaged nucleotide patch.

1.2.2.3 Mismatch Repair (MMR)

The MMR repairs the bases that are wrongly placed as a complementary part [32]. Although many types of errors are corrected by the proofreading mechanism only, a few remains missed. The method comes into the action right after a new DNA has been made, and its job is to identify and replace the mispaired bases. The method has also a potential to sense and correct the small insertions and deletions. The insertion or the deletions happen in case the polymerases missed the checking of the base-pairs. In a method, first, a protein complex recognizes and attaches to the mispaired base. The human MMR pathway has two major components i.e. MutS and MutL that includes complex pairs. MutS has two basic forms; MutSα (consists of

heterodimer MSH2 and MSH6) and MutSβ (consists of heterodimer MSH2 and MSH3). Similarly, MutL heterodimer is also present in a number of forms, including the MutLα (consists of heterodimer MLH1 and PMS2), the MutLβ (consists of heterodimer MLH1 and PMS1), and MutLγ (consists of heterodimer MLH1 and MLH3 proteins). The role of MutSα heterodimer is to repair the base-substitutions and small mismatched loops, however, MutSβ repairs both small and large loop mismatches. The MutLα have a role in mismatch corrections, however, role of MutLβ and MutLγ is not clearly known [33]. After recognition of the mismatch by protein complexes the MutS and MutL bind to the site of damage with RFC and PCNA. In the next step, exonucleases remove the patch along with the mispaired base. This lead to the generation of a gap that needs constant repair by an enzyme. The DNA polymerase instantly activates to synthesize the damaged bases that are stitched by ligase enzyme to the strand [34], as shown in Figure 1.5.

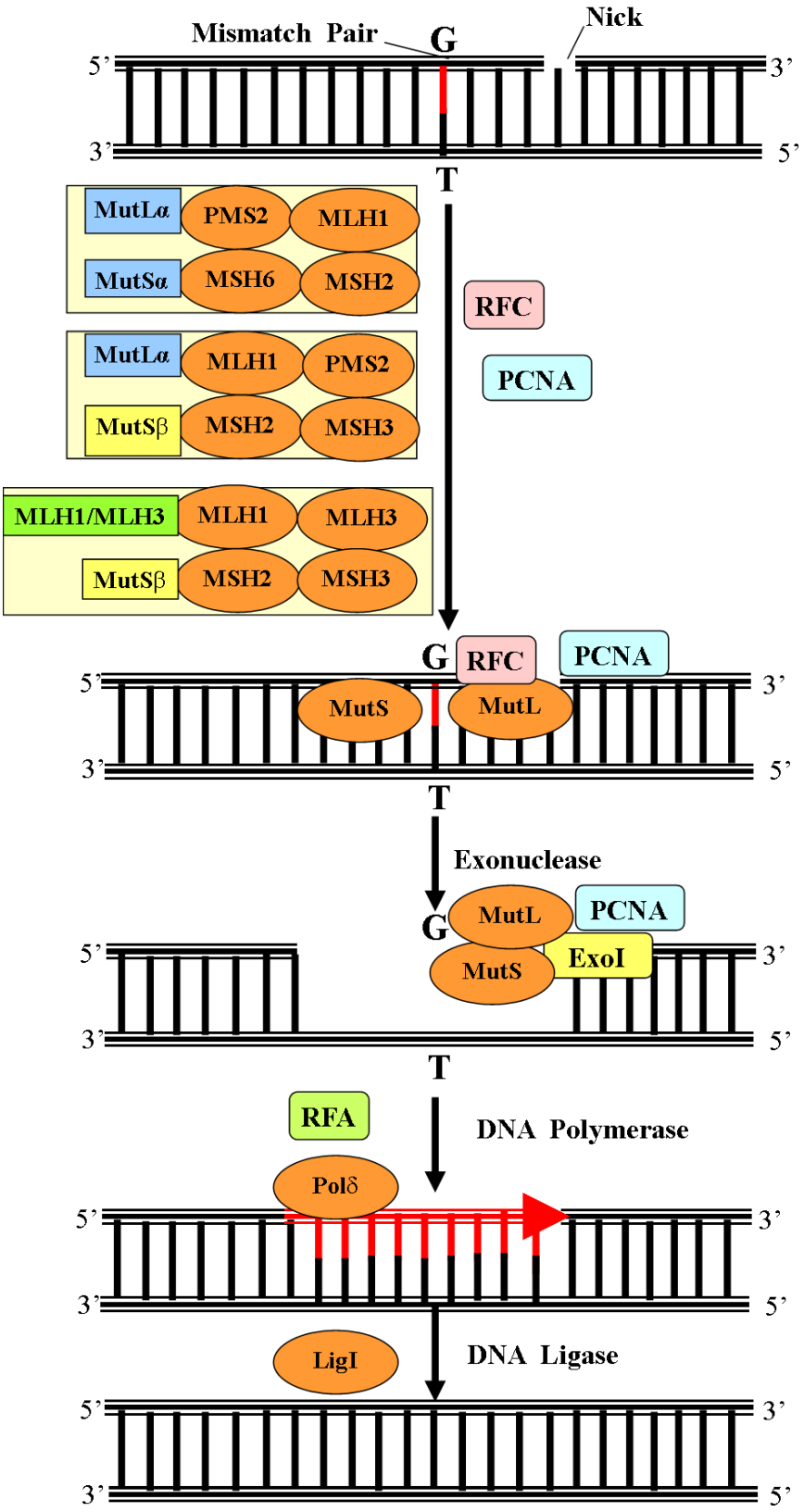


Figure 1.5 The Mismatch Repair (MMR) mechanism to repair the mispaired base.

1.2.3 Double Strand Breakage

The factors that introduce damages in the DNA sometimes also affect the double helical strand of the DNA along with single strand damage. There are many agents causing double strand breakage of the DNA, amongst all high-energy radiation are the major involving factors that split the chromosomes into two pieces. The double strand breakage is further categorized into two major pathways i.e. homologous recombination repair (HRR) and non-homologous end joining (NHEJ). These damages are dangerous for the genome integrity as hundreds of genes imbibed in large segments of chromosomes may be lost if the damage remains unrepaired.

1.2.3.1 Homologous Recombination (HRR)

The HRR is a genetic recombination method, wherein the nucleotide sequences are exchanged between identical molecules of DNA [35]. The method follows usage of the homologous chromosome identical to the damaged one to repair the double-strand break. The two chromosomes that are homologous in relation come closer, and the non-damaged region of chromatid is used as a template for replacing the damaged portion. This method works in the S/G2 phases of the cell-cycle in presence of the intact sister chromatid [36]. During the repair of the double strand, HRR mechanism leads three possible alternative pathways. The first one involves double-strand break repair that performs initial invading using both strands. The confined second end performs annealing to the homologous template and initiates new DNA synthesis, resulting in a Holliday junction. The Holliday junction is then resolved by nucleases to form the crossover or non-crossover products. In the second one called synthesis dependent strand annealing (SDSA); the method undergoes series of annealing, synthesis, and ligation. In this method, the invading strand and a newly synthesized segment is unwound through the helicase activity that undergoes annealing with the resected end. The last one involves break-induced replication (BIR), wherein one end of the double strand break is lost and the remaining end overrun the homologous template performing DNA synthesis (Figure 1.6).

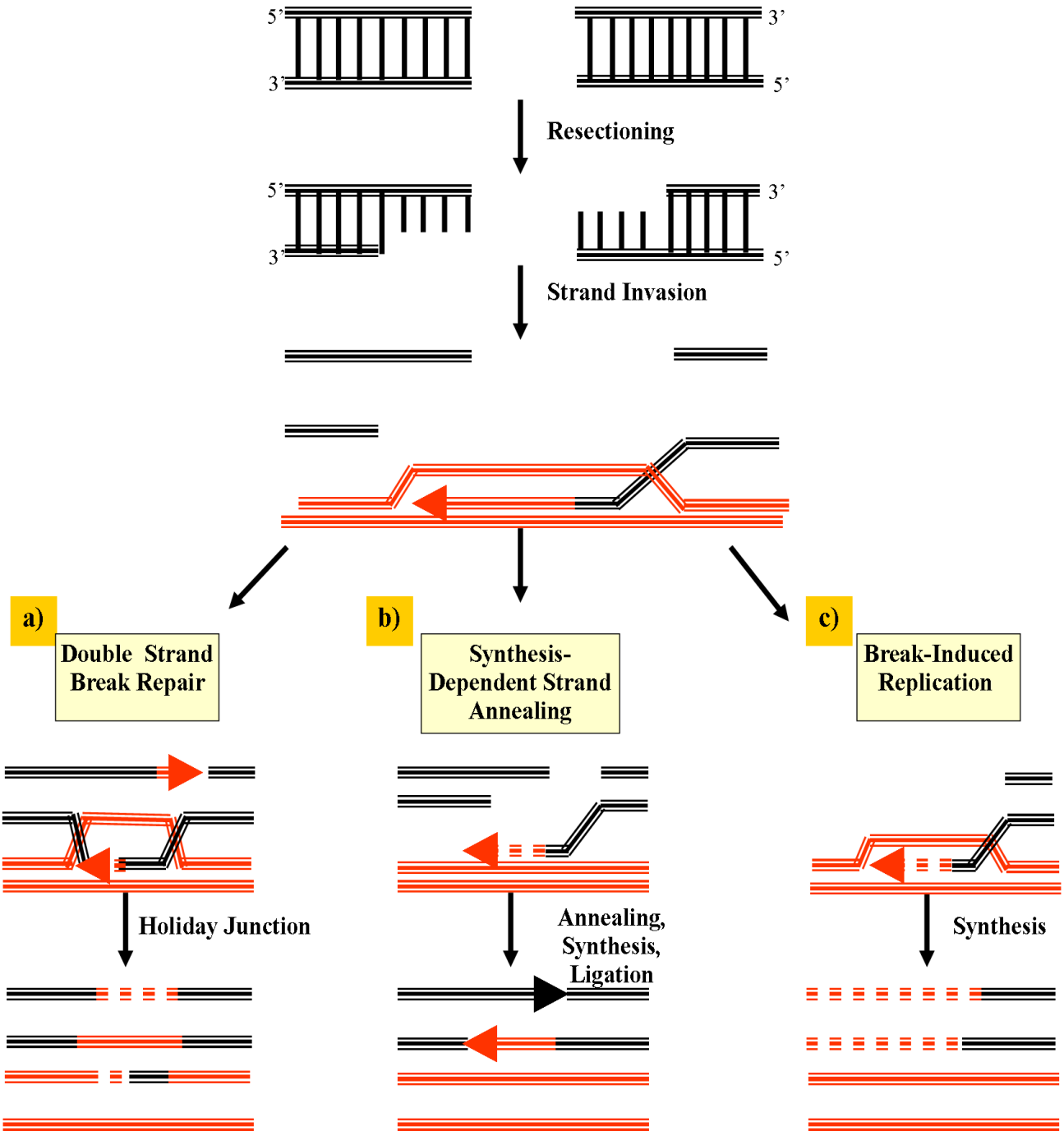


Figure 1.6 The Homologous recombination repair (HRR) mechanism. **a)** Double strand break repair **b)** Synthesis-dependent strand annealing **c)** and Break-induced replication.

1.2.3.2 Non-Homologous End Joining (NHEJ)

The non-homologous end joining (NHEJ) performs repair of the double strand break using a template that is not homologous to the one needs to be repaired [37]. In NHEJ, two broken ends of the DNA are repaired back with the aid of a set of repairing enzymes (Figure 1.7). The repair method usually involves addition or, loss of nucleotides at the damage site of double strand. It has been noticed that the mutations introduced in the HRR are not much damaging in comparison to the NHEJ [36].

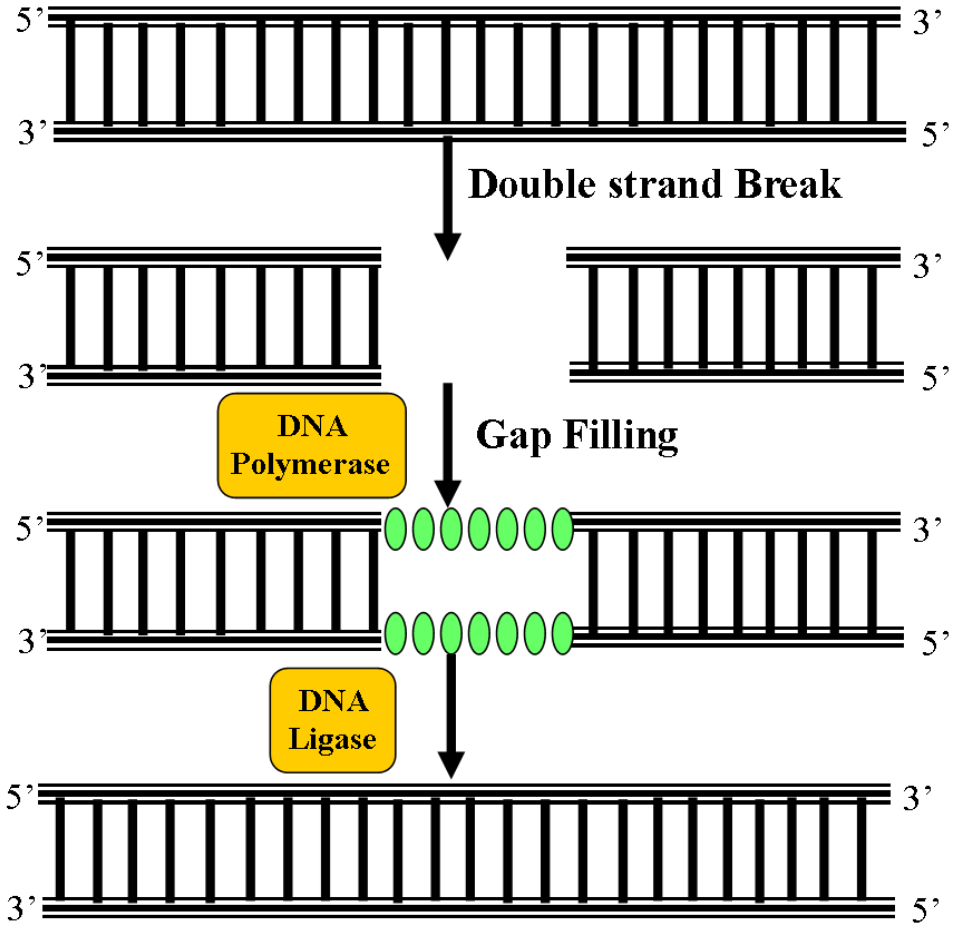


Figure 1.7 The Non-Homologous End Joining (NHEJ) repairing the double strand damage using non-homologous set of chromatid.

The method involves removal of the two broken ends of DNA that after a limited processing undergoes repair that sometimes include errors [38]. The NHEJ repair components involve Ku - complex that recognize the DNA-break, DNA-PKcs (protein kinase) that senses and signal presence of a break and perform the activation of the repair proteins at the breakage point,

Artemis (enzyme) that performs DNA-end processing and a ligase complex XRCC4–Ligase IV to seal the broken ends.

1.3 DNA REPAIR AND SIGNALING PATHWAYS

The DNA damage pathway and the checkpoint control signaling mechanisms irrespective of the stringent repair have a possibility to undergo mutational processes causing human tumors [39]. The pathways employ various oncogenes or, tumor suppressor genes that express in case of the adverse condition in the cell. The studies have shown, interaction of oncogenes or tumor suppressor genes with agents involved in damage response of the DNA has a significant effect in killing cancerous cells [40]. The damage in the DNA can be incorporated due to the various factors such as replication, metabolism, and exposure to radiations that forms chemical adducts [41]. In case damage remains unrepaired, they could result in the mutator phenotype. There are diverse pathways that a cell can employ to repair such damages. The DNA damage occurs inside chromatin and modifications of histones are vital for signaling the position of DNA damage and hence recruiting repair proteins at the damage site. The cross-talk between the repair pathways and signal transduction mechanism provides better understanding of the events that could help building strategies for treating cancer cells.

There are varieties of signaling pathways that have a role in combating the DNA damages which includes; Wnt, Hedgehog, TGFβ, MAPK, mTOR, JAK/STATs, COX, VEGF, NF-kB, and Notch (Figure 1.8). The Wnt signaling is a key pathway in gene regulation, cell polarity, adhesion and maintaining cell behavior [42]. It highly interacts with DNA damage response (DDR) at different levels and locations. The oxidative stress that leads to DNA lesions affects Wnt signaling in a variety of ways. However, understanding the cross-talk between damage processes and Wnt pathway could help in forming strategies for treatment of cells deficient in repair mechanisms and have a possibility of developing the malignancy. The Hedgehog (Hh) signaling is one of the important signaling pathways that play key roles in embryonic development, carcinogenesis, maintenance of cancer stem cells (CSCs), and acquisition of epithelial-to-mesenchymal transition (EMT) leading to metastasis [43]. The signaling process can repress almost all forms of DNA repair mechanisms (BER, MMR, NER, DSB repair including NHEJ and HRR) in case of any disparity. In a study Meng et.al suggested the inhibition of Hh/GLI as a major factor for reducing DNA repair activity in cancers [43].

The transforming growth factor–β (TGFβ) is a well-known signaling that regulates cell proliferation and helps to maintain the tissue homeostasis [44]. The TGFβ has shown to have a paradoxical behavior, due to its tumor suppressor activity in early stages of the carcinogenesis, and tumor promoters in established cancer forms. In a MAPK signaling, the protein kinase 2 (MK2) acts as a major mediator of damaged response as it suppresses progression of replication fork and equally enhances frequency of new replication origins in presence of replicative stress [45]. In mTOR signaling, normally it suppresses endogenous DNA damage and replication stress. However, downregulation of CHK1 by inhibitory activity of the mTOR kinase results defects in cell cycle during the S-phase that follows DNA damage [46]. The JAK/STAs signaling increase resistance to the damage by generation of a high level of heterochromatin. The decreased levels of activated JAK and increased levels of unphosphorylated STAT generate much higher levels of heterochromatin that suppresses formation of hematopoietic tumor-like masses [47]. The COX signaling is related to the DNA damage through COX-2 expression, that is involved in telomere malfunction and its induction in cells in absence of DNA damage that initiates cancer progression [48]. The VEGF increase the number of DNA damaged cells and lessened the induction of ERCC6 mRNAs in case of deficient repair [49]. The nuclear factor kappaB (*NF-kB*) signaling get activated as part of the DNA damage response that helps to coordinate a cell survival pathway. The process initiates altogether with the cell-cycle activation of checkpoints and DNA repair. The relevance of this pathway in protecting the DNA from damage can thus help accounting for chemotherapy resistance and thus providing effective cancer treatment [50]. The Notch receptor usually binds and inactivates ATM kinase and this is an evolutionarily conserved mechanism in *C. elegans*, *Xenopus laevis*, and *Homo sapiens*. The inactivation of ATM by Notch signaling contributes to the survival of tumor cells upon DNA damage [51]. Table 1.2 illustrates the diverse type of signaling pathways and their involvement in DNA repair mechanisms.

Table 1.2 Damaging effect on cellular signaling and its influence on DNA repair signaling pathways

Signaling Pathways	Effected Repair Mechanism	Key Genes	Factors	Role
Wnt signaling	Double strand break repair (DSBR) NHEJ, HRR	Wnt/β-catenin	oxidative stress	gene regulation, cell polarity, adhesion, maintaining cell-behavior
Hedgehog signaling	BER, MMR, NER, NHEJ, HRR	Hh/GLI	endogenous molecules, oxysterols	embryonic development, carcinogenesis, maintaining cancer stem cells (CSCs), mesenchymal transition (EMT)
TGFβ signaling	NHEJ, HRR	TGFβ, receptors I and II, SMADs	endogenous DNA damage, radiations	cell proliferation, tissue homeostasis
MAPK signaling	NHEJ, HRR	MK2, MAPKKK, MEKK	replicative stress, reactive oxygen species	progression of the replication fork
mTOR signaling	Direct reversal, NER	CHK1	endogenous DNA damage, replication stress	cell metabolism, growth, proliferation, survival
JAK/STATs signaling	NHEJ, HRR	JAK, STAT	cytostatic drugs, radiations	heterochromatin formation, development, homeostasis
COX signaling	NER	COX-2	UV-light	haematopoietic stem cell homeostasis, suppression of embryonic stem cell apoptosis
VEGF signaling	Direct reversal, excision repair mechanisms	VEGF-A	endogenous DNA damage	cell homeostasis
NF-kB signaling	NHEJ, HRR	NF-kB	metabolic and exogenous sources	DNA transcription, cytokine production, cell survival
Notch signaling	NHEJ, HRR	ATM	reactive oxygen species	cell-fate, homeostasis

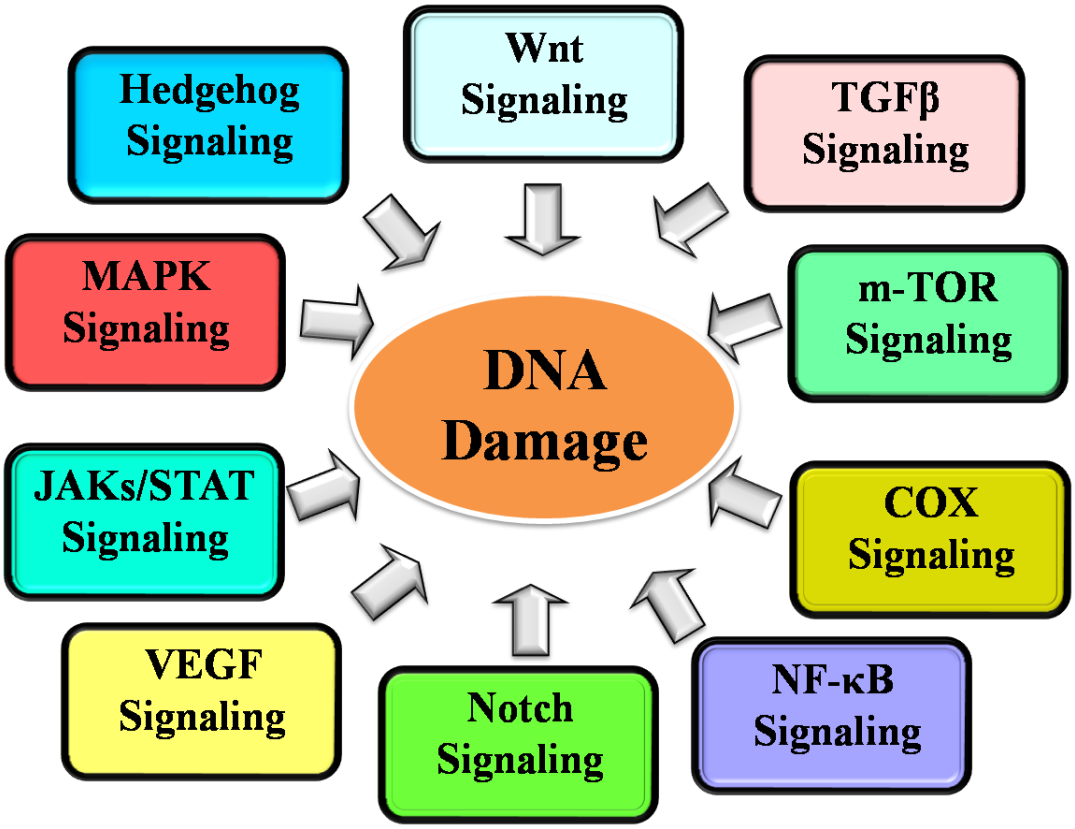


Figure 1.8 Common signaling pathways to conquer the DNA damage process.

1.4 DNA REPAIR AND CANCERS

Proof regarding the role of DNA repair mechanisms comes from a diverse form of genetic diseases and disorders. Many times mutations in proofreading genes and repair proteins are related to the hereditary forms of cancers [52]. The DNA every day faces massive destruction processes that are overpowered by the repair mechanisms. In most cases, the damage happens to a cell is destroyed and are replaced by new cells that are capable of performing same function as to that of the destroyed one. During this complex process of replacement of the cells, many errors get induced. Despite remarkably elegant systems to prevent the errors, the body still makes tens of thousands of mistakes on daily basis during the replacement of the damaged cells. Most of these mistakes are corrected by an advanced form of repair system else it leads to the destruction of the newly made cells. However, in extreme cases, a mistake is made and is not rectified. The multiple unrepaired mistakes have little effect on health, but in case that error allows the newly made cell to divide irrespective of the checkpoints that balance and control normal cell growth, it will lead to the division in an uncontrolled manner. The uncontrolled cell-

division can form an abundant mass of abnormal cells called a tumor. The tumor if not recognized at an early stage can start forming cancerous cells. There are a variety of cancers that are present to date that includes lung cancer, breast cancer, prostate cancer, colorectal cancer, skin cancer, endometrial cancer, bladder cancer, kidney cancer, thyroid cancer, uterine cancer, pancreatic cancer, oral cancer etc. [53]. In our study, we have focused on Lynch syndrome and associated cancer predominantly colorectal cancer.

1.4.1 Lynch Syndrome

Lynch syndrome that is often known as hereditary nonpolyposis colorectal cancer (HNPCC), is an autosomal inherited disorder that boosts the risk of multiple types of cancer, predominantly that of the colon and rectum, collectively known as colorectal cancer (CRC) [54]. The syndrome shows an autosomal dominant pattern that shows that even a single inherited copy of the altered gene is sufficient to amplify the cancer risk. This syndrome initiates due to faulty mismatch repair mechanism [55]. Because mismatched bases are not repaired the mutations accumulate at a very fast rate than in the cells of an unaffected individual. The other forms of the mismatch conditions are Muir-Torre and Turcot syndromes, these also put a person at high risk of cancers, particularly of skin lesions [56]. Although primary cancer related to Lynch syndrome is CRC yet, there are possibilities of other forms of cancer also such as of stomach, small intestine, liver, gallbladder, urinary tract, brain, and skin [57]. Women with this syndrome have a high possibility of initiating cancer of the uterus lining also known as endometrial cancer. People having Lynch syndrome sometimes occasionally have benign growths in the form of polyps (abnormal tissue growth) in the large intestinal parts [58]. However, polyps do not grow large in numbers for person having disorder. The person having Lynch syndrome inherits elevated risk of cancer but not the disease itself and not all who inherit mutations necessarily develop cancer. The Lynch syndrome is characterized by an enlarged risk for CRC, endometrium, stomach, ovary, small bowel, hepatobiliary tract, urinary tract, brain, and skin cancers [58]. The cancer cells often lack a method of DNA repair, and this deficit in most cases leads to tumorigenesis. The tumor cell has potential to break and reform chromosomes thus forming new oncogenic fusion genes, disruption of tumor suppressor genes, an increase of drug resistance genes, and progression towards a malignant state [59].

1.4.1.1 Genetic Alterations

The genetic variations such as *MLH1, MSH2, MSH6, PMS2,* and *EPCAM* increase the risk of developing Lynch syndrome [60]. The *MLH1, MSH2, MSH6,* and *PMS2* were found to overcome the inflicted errors that occur during the replication process of the DNA. Unable to correct the errors, it develops mutations in genes thus, interfere and prevent proper removal of the replication errors. This precedes the abnormal cell division and accumulation of errors initiating cancer pathways. The mutation in *EPCAM* gene has also significant effect, as it lies next to the *MSH2* gene. The mutational effect of the *EPCAM* gene reflects *MSH2* gene expression by turning it off. Therefore, interrupting the DNA repair mechanism and accumulation of errors thus predisposing individuals to cancer but, not all people who carry these mutations develop cancerous tumors.

1.4.1.2 Lynch Syndrome Screening

The screening process of Lynch syndrome follows the certain characteristics and is known as the *Amsterdam criteria*:

- In case three relatives have colorectal cancer (or another cancer linked with Lynch syndrome).
- One is a close relative (parent, brother or sister, or child) of the other two relatives.
- In case it involves two successive generations with Lynch syndrome associated cancers.
- In case anyone relative had cancer history before the age of 50 years.

In case all of the above criteria apply to a family, then they need to have regular checkups. However, it's not necessary to have this syndrome even if a family history covers any of the Amsterdam criteria. Also, there are cases where families with Lynch syndrome do not satisfy the Amsterdam criteria [61]. There is another set of criteria, called the Bethesda guidelines that is used to identify that a person with colorectal cancer should test for genetic changes (such as microsatellite instability (MSI)) seen in case of Lynch syndrome. The criteria includes following ones:

- For the person younger than age of 50 years.
- If a person has or had a second CRC or any another cancer linked to Lynch syndrome.
- For a person younger than 60 years, and the cancer shows certain characteristics of Lynch syndrome.

- In case the person has a close relative (parent, sibling, or child) younger than 50 years who was diagnosed with CRC or any another cancer linked to Lynch syndrome.
- In case a person has two or more relatives (such as aunts, uncles, nieces, nephews, or grandparents) who had CRC history or any another Lynch syndrome-related cancer.

In case a person having CRC shows anyone of the Bethesda criteria, testing for MSI should be done. If MSI is found, patient is further tested for Lynch syndrome-associated gene mutations.

1.4.2 Colorectal Cancer (CRC)

CRC also known as bowel cancer, colon cancer, or rectal cancer affects the colon and the rectum. According to the American Cancer Society, about 1 in 21 men and 1 in 23 women in the United States (US) will develop CRC during their lifetime [62]. The CRC is relatively uncommon in India in comparison to the western world [63]. In India the age-standardized rates of CRC have been estimated to be 4.2/100,000 for males and 3.2/100,000 for females, compared to 35.3 and 25.7, respectively, in the USA. Comparing the statistics of India and the USA, the incidence, mortality, and prevalence rates are all consistently higher in the USA although the incidence is higher in males in both countries [64]. Although the incidence of CRC in Indian older age group is currently very low when compared to the Western older population, yet, it seems to be increasing in the younger generation. It is the fourth most common cancer in the world with 1.3 million new cases each year and a 5-year prevalence rate of 3.2 million [64]. In USA, CRC falls behind prostate, breast, and lung cancer however in India, it is the fifth most common cancer following breast, cervix, oral, and lung cancer. It is the second leading cause of cancer-related death in women, and the third for men. With advances in screening techniques and improvised treatments, the death rate for CRC has been falling from past few years.

The studies demonstrated that patients in India are found with same symptoms as to that western world; however, the cases are frequent for younger generation presented at a later stage of the cancer [64]. By 2035, the incidence rate has been predicted to increase by ~80%, with 114,986 incidences and 87,502 mortalities [64]. The younger CRC patient individuals have been found to have aggressive forms of cancer like mucinous adenocarcinoma and synchronous liver metastasis [65]. Today China and India are amongst the most heavily populated countries with low incidence rates of CRC; but as their economies have developed the incidences for CRC has

shown to be increased [65]. The development of CRC involves multiple sequential steps, initiating from normal colon epithelium to aberrant crypt foci, followed by the formation of early and advanced polyps and then subsequent progression towards cancer. The classical approach involves development of tubular adenomas that progress towards adenocarcinomas. However, an alternate pathway involves formation of serrated polyps and their progression towards cancer.

1.4.2.1 Mechanisms of CRC Carcinogenesis

The CRC carcinogenesis follows three major mechanisms i.e. chromosomal instability (CIN), MSI, and CpG island methylator phenotype (CIMP) [66]. The study have shown that the CIN pathway begins with mutations in the *APC*, followed by the mutational activation of *KRAS*, and the inactivation of the tumor suppressor gene, *TP53* [67]. These mutations leads to the condition called aneuploidy, and loss of heterozygosity (LOH) in CIN tumors. It constitutes sporadic tumors (~85% cases) and also involves hereditary condition called familial adenomatous polyposis (FAP). FAP is generally associated with germline mutations in the *APC* gene [68]. In MSI pathway there is inactivation of genetic alterations in short repeated sequences. Mutations in DNA mismatch repair genes are the major factor in developing MSI condition, due to lack in correcting replication errors. This condition is a hallmark for the familial Lynch syndrome (LS). Generally it appears in ~15% of the sporadic CRC cases. The MSI pathway has been found to be associated with the CIMP pathway [69]. MSI tumors have better prognosis although they are associated with proximal colon and have poor differentiation [70]. The CIMP pathway is characterized by promoter hypermethylation of various tumor suppressor genes, most importantly *MGMT* and *MLH1*. The hypermethylation in the promoter region cause transcriptional inactivation of genes. These genes have tumor suppressive roles or, are involved in the cell cycle. This hypermethylation is often associated with BRAF mutation and MSI [71]. As illustrated in Figure 1.10 these mechanisms initiate tumorigenesis by activating various signaling pathways that in presence of tumor suppressors and oncogenes progress towards cancer. The key mutation genes, their genetic consequences, hereditary mechanisms, and their possibility to become cancerous are given in Table 1.3.

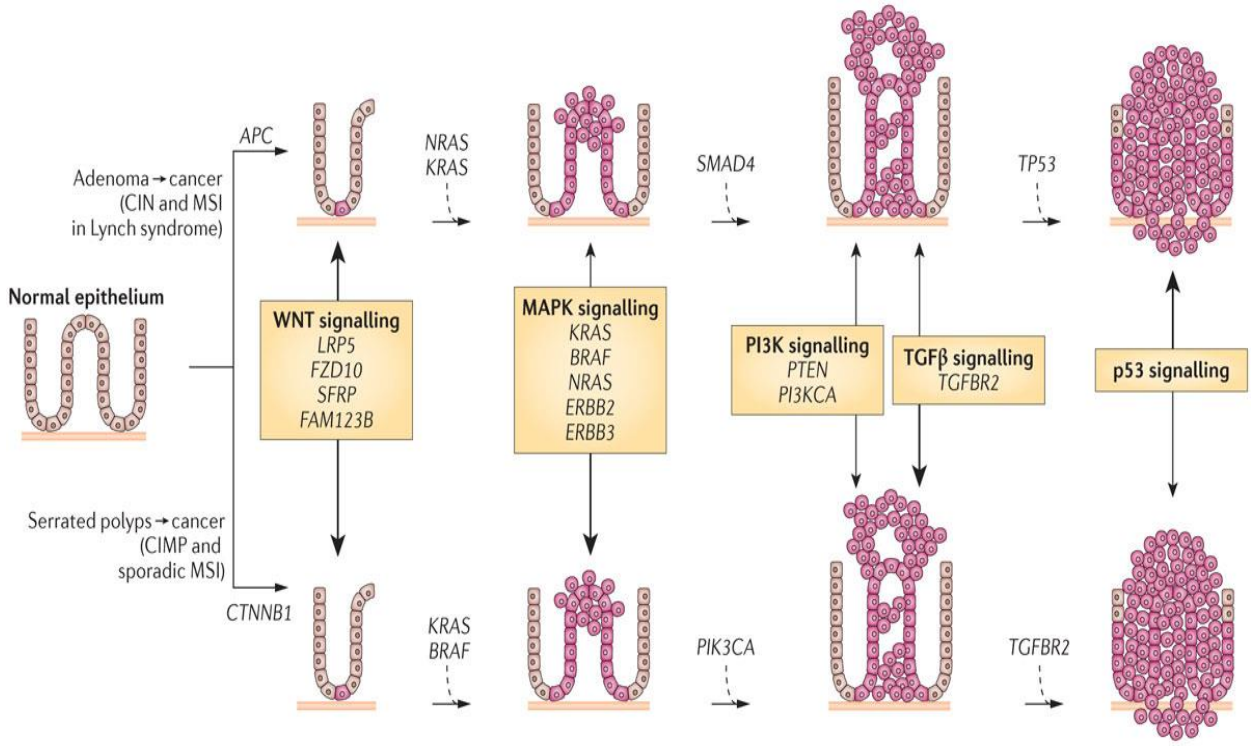


Figure 1.9 Formation of Cancer (Retrieved from Davies, R. J., Miller, R. & Coleman, N. Colorectal cancer screening: prospects for molecular stool analysis. Nat. Rev. Cancer 5, 199–209 (2005), Nature Publishing Group)[72]

Table 1.3 Mechanisms and other important details for colorectal cancer (CRC) carcinogenesis

CRC Mechanisms	Mutated Genes	Genetic Consequences	Hereditary Mechanisms	Involvement in CRC
Chromosomal instability (CIN)	Adenomatous polyposis coli (APC), Kirsten rat sarcoma virus (KRAS), Tumor protein p53 (TP53)	Aneuploidy, Loss of heterozygosity (LOH)	Familial Adenomatous Polyposis (FAP)	~ 85%
Microsatellite instability (MSI)	MutL Homolog 1 (MLH1), MutL Homolog 3 (MLH3), MutS homolog 2 (MSH2), MutS homolog 3 (MSH3), MutS homolog 6 (MSH6), Postmeiotic segregation increased 1 (PMS1) and postmeiotic segregation increased 2 (PMS2)	genetic hypermutability (predisposition to mutation)	Hereditary nonpolyposis colorectal cancer (HNPCC)	~15%
CpG island methylator phenotype (CIMP)	O-6-methylguanine-DNA methyltransferase (MGMT) and MutL Homolog 1 (MLH1), B-Raf Proto-Oncogene (BRAF)	Hyper and Hypomethylation in the satellite regions of chromosome	No	~40%

1.4.2.2 Risk Factors

There are vast varieties of risk factors that cause CRC which includes older age, red meat intake, saturated fats, high calories, a diet low in fiber, alcohol consumption, having history of breast, ovary, or uterine cancers, a family history of CRC, ulcerative colitis, Crohn's disease, irritable bowel disease (IBD), obesity, smoking, physical inactiveness, the presence of polyps in the colon or rectum. Most cancer starts with the formation of polyps in the colon or the rectum region and is generally called adenoma carcinoma. The rate of CRC is however equally prevalent in men and women yet, men tend to develop cancerous state at a younger age [73]. Figure 1.11 illustrates all the possible condition that hikes the risk of CRC.

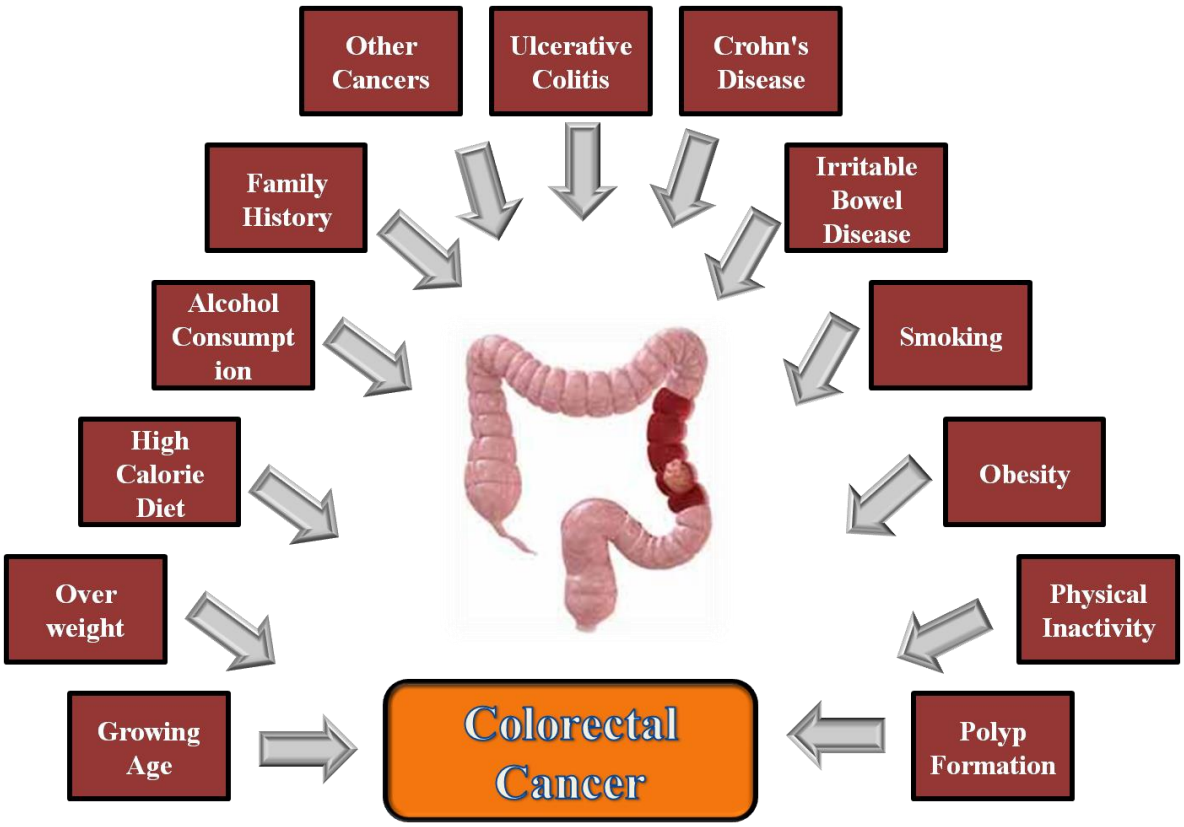


Figure 1.10: Various causal factors involved in CRC progression. The factors may be extrinsic or intrinsic in nature such as: Aging, overweight, Diet high in calories, alcohol intake, smoking, family history, other cancerous condition, ulcerative colitis, crohn's disease etc.

1.4.2.3 Stages of Cancer

There are majorly four stages of the cancer that determine how far it has been progressed (Figure 1.12). The complete details regarding the individual state are given in Table 1.4.

Table 1.4 Stages of the colorectal cancer (CRC)

Stages	Progression	Rate of Survival
Stage 0	This is the initial stage, when the cancer develops in the mucosa, or inner layer, of the colon or rectum. It is also called carcinoma in situ.	~99%
Stage 1	The cancer starts growing through the inner layer of the colon or rectum but has not yet spread beyond.	~90%
Stage 2	Cancer has spread through the colon wall to nearby organs, but not metastasized to nearby lymph nodes.	~70%
Stage 3	Cancer has spread to nearby lymph nodes but not invade other normal tissues.	~50%
Stage 4	Cancer has spread to distant organs like the liver, lungs, myolemma, and ovary.	~0-40%

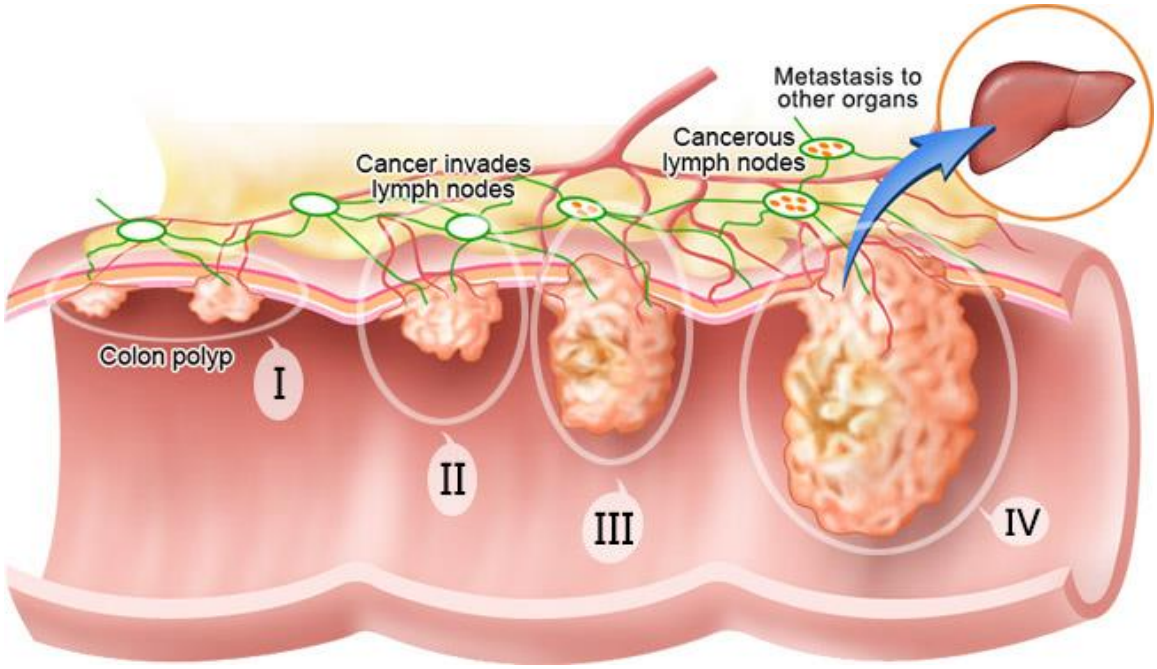


Figure 1.11 Stages of the Colorectal Cancer (**Retrieved From:** http://www.asiancancer.com/cancer-topics/new-intestinal/)

1.4.2.4 Hereditary Mechanisms

To find the future possibility of a person to have CRC or not can be determined through his/her family history. However, a family history in CRC plays important role in calculating the risk factor, but the research indicated that only some case upto 20% are associated with the family

history. Yet ~80% of the CRC cases found to happen sporadically without any family history. In sporadic form of the cancer the chromosome damages occur during later stages but, in hereditary form they are inherited at birth only. Person with the disease-inherited chromosomes have high risk of developing polyps at young ages and thus develop cancer at an early stage of life [74]. There are number of forms of the hereditary mechanisms of cancer and it involves the following ones:

- **Familial adenomatous polyposis (FAP):** In FAP, the affected individual will develop hundreds or thousands of colon polyps. The condition arises due to mutation of the APC gene [75]. The person having this inherited form will surely develop cancer during his/her lifetime.

- **Attenuated familial adenomatous polyposis (AFAP):** This is a milder version of FAP that also develop through APC mutations. Similarly, in this case also numerous polyps will occur in the colonic region. Such individuals are at very high risk of developing cancer at a young age and in addition the risk for gastric and duodenal polyps [76].
- **Hereditary nonpolyposis colon cancer (HNPCC/Lynch Syndrome):** This syndrome occurs due to impaired mismatch repair genes. Patients with HNPCC are at risk of developing uterine cancer, stomach cancer, ovarian cancer, and cancers of the ureters and the bile ducts [77].

- **MYH polyposis syndrome:** This syndrome is usually found in the individual with around 40 years of age having inactivated form of MYH gene. The affected person usually develops multiple polyps and is at high risk for cancer condition [78].

1.4.2.5 Diagnosis

There are diverse methods to diagnose the CRC, the following are the most common screening and diagnostic procedures for CRC [79].

- **Blood stool test:** In this test diagnosis is done through the stool test, if blood is found in stools then further screening is recommended for the confirmation. However, test may give false results as blood shedding also happens in case of the hemorrhoids.

- **Stool DNA test:** This test involve analysis as several of the DNA markers of the polyp cells shed into the stool. However, a problem with this test is that it cannot detect all DNA mutations.
- **Flexible sigmoidoscopy:** The method is used to examine the patient's rectum and sigmoid colon. In case polyps have been detected in this case further colonoscopy is recommended as a precautionary initiative.
- **Barium enema X-ray:** The test follows the use of Barium dye that is given in an enema form in the bowel of the patient. In case of any abnormality colonoscopy is recommended.
- **Colonoscopy:** This is one of the advanced techniques used currently and comprises of a long, flexible, slender tube, attached to a video camera and monitor. This help to see the whole colonic and rectum region of the intestine. The polyps discovered during this procedure are removed during the procedure itself or, are sometime taken for the biopsies.
- **CT colonography:** This is a machine that takes images of the colon. In this case also any unusual appearance of tissues makes conventional colonoscopy necessary.
- **Imaging scans:** It involves Ultrasound or MRI scans, these scans help to see if a cancer has spread to another part of the body.

1.4.2.6 Treatment

Treatment of the CRC depends on several factors such as the size, location, and stage of the cancer, the following techniques help in the treatment of CRC [80].

- **Surgery:** In this method surgery is done that help to remove the affected malignant tumors and nearby lymph nodes.
- **Chemotherapy:** The chemotherapy includes the chemical in the form of drugs that destroys the cancerous growth. Drugs such as bevacizumab (Avastin) and ramucirumab (Cyramza) are majorly given in chemotherapy.
- **Radiation therapy:** This therapy involves exposure of the high energy radiation beams to destroy the cancer cells and prevent them from dividing.
- **Ablation:** This method uses a radiofrequency, ethanol, or cryosurgery. These are delivered using a probe that is guided by the scanning technology (Ultrasound/CT-scan).

1.4.2.7 Prevention

The primary preventive measures [81] of the CRC are the following ones:

- **Regular screenings:** The person who have a family history, or age over 50, or have a Crohn's disease should have regular screenings.
- **Nutrition:** A person should follow a high fiber diet, use good quality fats instead of saturated ones.
- **Exercise:** Person should make a routine exercise schedule to lower the risk of CRC.
- **Bodyweight:** The body weigh should be proportionate to the height of the person.

Knowledge Gap

Although vast majority of advancements have been done to treat various forms of the cancers but till day it is the most challenging disease and enlarging its feet a way more. There are variety of cancers in the society that are currently being studied, CRC is the third most among the prevalent cancers still not sufficient knowledge have been gained regarding this cancer. The knowledge gap for this cancer persists in following forms:

- ✓ Limited international, national and state surveillance to monitor urgent and emerging state of the cancer.
- ✓ Mismanagement of cancer diagnosed cases to take effective measures as most of them go unnoticed.
- ✓ Lack of awareness amongst all level of people.
- ✓ Less effort invested in development of new drug and diagnostic tests.

An attempt has been made to contribute an effective study by taking care of the addressed gaps, such that effective measures can be taken to get the maximum genetic and genomic level knowledge of the disease so that therapeutic intervention could be developed.

Research Problem Statement

The methods applied various bioinformatics and systems biology approaches along with several statistical approaches to determine the structural and functional impact of the study towards CRC. Some new gene regulators have been identified for the genetic disorders and the cancer conditions that provide clues about putative predicted biomarkers for these diseases. The bottom-up and top-down approaches was implemented for annotation of the complex biological networks. The annotation and functional enrichment was made through various pathways and

network motifs to generate biological knowledge. Our lab work is on DNA repair mechanisms and my target of study is DNA repair associated disease condition specifically cancer. Till date our lab implemented various methods for diverse type of repair mechanisms be it in the form of an archive [82], methylation studies [83], network studies [84], and the differential gene expression studies [85].

My overall study is focused on the role of DNA repair mechanisms in the Lynch syndrome associated cancers. Although there are varieties of cancers that are involved with this syndrome but the colorectal cancer (CRC) is the predominant of these all. The second most cancer allied to this syndrome is the endometrial cancer that has been involved in the first objective of the study. Therefore, based on the worldwide limitations in research area for CRC and the Lynch syndrome associated cancer (specifically CRC, and cancer of endometrium), the study has been performed with following three objectives:

- Development of the database, DREMECELS: A Curated Database for Base Excision and Mismatch Repair Mechanisms Associated Human Malignancies.
- Network-Based Approach to Study Dynamics of Wnt Pathway Regulatory Elements in Colorectal Cancer (CRC).
- Network and Structure Based study of Functional Single Nucleotide Polymorphisms of TGFβ1 Gene and its Role in CRC.

The study focused on the two different signaling mechanisms i.e. Wnt and TGFβ pathways; Wnt is the first signaling mechanism for the CRC triggering and TGFβ is in the middle of the carcinogenesis and many CRC cases were detected at that stage; therefore these two pathways were targeted in the study so that therapeutic work can be conducted at an initial level of diagnosis. The very first objective followed an archive development to store variety of information for genes involved in DNA repair mechanisms specifically base excision and mismatch repair mechanisms in colorectal cancer, endometrial cancer and the Lynch syndrome. The information is orchestrated in the form of genes, proteins, annotations, references, miRNAs, transcription factors, conserved domains, gene-interactions, pathways, available drugs, somatic mutations, and copy number variations. The data not only provides the possible biomarkers for these diseases but help to retrieve the direct links for further details for their original sources of information. This would provide an aid for the computational biologist as well as the experimental scientists to have direct access to all available data known for the disease type. This

will not only save time but help to put efforts in right directions to find the possible therapeutic interventions. In the second objective, network based study has been implied for Wnt signaling pathways to find the putative biomarkers for the colorectal cancer. The study entails behavioral analysis to determine effect of individual component of the Wnt signaling pathway and network motif detection to identify the ones working in close association.

The outcome of the proposed research would certainly assist to further carry out the research work with additional known factors and the regulatory mechanisms. Eventually, the exhaustive mechanistic perspective of the DNA repair mechanisms and its impact on these cancer types will help in finding the advanced knowledge of the cancers and thus to apply the correct strategies to overcome them. Therefore, the work performed will provide innovative paradigms for genetic susceptibility, prevention, diagnosis at an earliest condition.

REFERENCES

[1] F. H. Ruddle, "Mapping and sequencing of the human genome," *Jpn J Cancer Res,* vol. 89, p. inside front cover, Dec 1998.

[2] L. D. Stein, "Human genome: end of the beginning," *Nature,* vol. 431, p. 915, 2004.

[3] G. J. van Ommen, "The Human Genome Project and the role of genetics in health care," *Clin Chem Lab Med,* vol. 36, pp. 515-7, Aug 1998.

[4] H. Chial, "DNA sequencing technologies key to the Human Genome Project," *Nature Education,* vol. 1, p. 219, 2008.

[5] L. Chin, W. C. Hahn, G. Getz, and M. Meyerson, "Making sense of cancer genomic data," *Genes Dev,* vol. 25, pp. 534-55, Mar 15 2011.

[6] A. Livnat, "Interaction-based evolution: how natural selection and nonrandom mutation work together," *Biology direct,* vol. 8, p. 24, 2013.

[7] E. C. Friedberg, G. C. Walker, W. Siede, and R. D. Wood, *DNA repair and mutagenesis*: American Society for Microbiology Press, 2005.

[8] M. Papamichos-Chronakis and C. L. Peterson, "Chromatin and the genome integrity network," *Nat Rev Genet,* vol. 14, pp. 62-75, Jan 2013.

[9] C. Dinant, A. B. Houtsmuller, and W. Vermeulen, "Chromatin structure and DNA damage repair," *Epigenetics Chromatin,* vol. 1, p. 9, Nov 12 2008.

[10] H. Lodish, A. Berk, S. L. Zipursky, P. Matsudaira, D. Baltimore, and J. Darnell, "DNA damage and repair and their role in carcinogenesis," 2000.

[11] S. F. Bakhoum, L. Kabeche, J. P. Murnane, B. I. Zaki, and D. A. Compton, "DNA-damage response during mitosis induces whole-chromosome missegregation," *Cancer discovery,* vol. 4, pp. 1281-1289, 2014.

[12] R. Barnes and K. Eckert, "Maintenance of genome integrity: how mammalian cells orchestrate genome duplication by coordinating replicative and specialized DNA polymerases," *Genes,* vol. 8, p. 19, 2017.

[13] R. Hakem, "DNA-damage repair; the good, the bad, and the ugly," *The EMBO journal,* vol. 27, pp. 589-605, 2008.

[14] T. Lindahl and R. D. Wood, "Quality control by DNA repair," *Science,* vol. 286, pp. 1897-1905, 1999.

[15] A. L. Harris, "DNA repair and resistance to chemotherapy," *Cancer Surv,* vol. 4, pp. 601-24, 1985.

[16] H. Lodish, A. Berk, S. L. Zipursky, P. Matsudaira, D. Baltimore, and J. Darnell, "Mutations: types and causes," *Molecular Cell Biology,* vol. 4, 2000.

[17] J. Ward, C. Limoli, P. Calabro-Jones, and J. Evans, "Radiation vs chemical damage to DNA," in *Anticarcinogenesis and radiation protection*, ed: Springer, 1987, pp. 321-327.

[18] A. Tubbs and A. Nussenzweig, "Endogenous DNA damage as a source of genomic instability in cancer," *Cell,* vol. 168, pp. 644-656, 2017.

[19] E. C. Friedberg, L. D. McDaniel, and R. A. Schultz, "The role of endogenous and exogenous DNA damage and mutagenesis," *Curr Opin Genet Dev,* vol. 14, pp. 5-10, Feb 2004.

[20] K. Shimada, T. R. Crother, and M. Arditi, "DNA Damage Responses in Atherosclerosis," in *Biological DNA Sensor*, ed: Elsevier, 2014, pp. 231-253.

[21] G. M. Cooper and R. E. Hausman, *The cell: Molecular approach*: Medicinska naklada, 2004.

[22] V. Khare and K. A. Eckert, "The proofreading 3′→ 5′ exonuclease activity of DNA polymerases: a kinetic barrier to translesion DNA synthesis," *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis,* vol. 510, pp. 45-54, 2002.

[23] L. Grossman, P. R. Caron, S. J. Mazur, and E. Y. Oh, "Repair of DNA-containing pyrimidine dimers," *FASEB J,* vol. 2, pp. 2696-701, Aug 1988.

[24] N. Kondo, A. Takahashi, K. Ono, and T. Ohnishi, "DNA damage induced by alkylating agents and repair pathways," *J Nucleic Acids,* vol. 2010, p. 543531, Nov 21 2010.

[25] G. P. Pfeifer and A. Besaratinia, "UV wavelength-dependent DNA damage and human non-melanoma and melanoma skin cancer," *Photochem Photobiol Sci,* vol. 11, pp. 90-7, Jan 2012.

[26] C. Fan, W. Liu, H. Cao, C. Wen, L. Chen, and G. Jiang, "O 6-methylguanine DNA methyltransferase as a promising target for the treatment of temozolomide-resistant gliomas," *Cell death & disease,* vol. 4, p. e876, 2013.

[27] I. D. Hickson, *Base excision repair of DNA damage*: Springer Science & Business Media, 1999.

[28] V. Sidorenko and D. Zharkov, "Role of base excision repair DNA glycosylases in hereditary and infectious human diseases," *Molecular biology,* vol. 42, pp. 794-805, 2008.

[29] H. E. Krokan and M. Bjørås, "Base excision repair," *Cold Spring Harbor perspectives in biology,* vol. 5, p. a012583, 2013.

[30] A. N. Evdokimov, O. I. Lavrik, and I. O. Petruseva, "Model DNA for investigation of mechanism of nucleotide excision repair," *Biopolymers and Cell,* vol. 30, pp. 167-183, 2014.

[31] O. D. Schärer, "Nucleotide excision repair in eukaryotes," *Cold Spring Harbor perspectives in biology,* vol. 5, p. a012609, 2013.

[32] D. MacPhee, "Mismatch repair as an important source of new mutations in non-dividing cells," *Experientia,* vol. 52, pp. 357-363, 1996.

[33] K. Fukui, "DNA mismatch repair in eukaryotes and bacteria," *J Nucleic Acids,* vol. 2010, Jul 27 2010.

[34] T. A. Kunkel and D. A. Erie, "DNA mismatch repair," *Annu. Rev. Biochem.,* vol. 74, pp. 681-710, 2005.

[35] E. C. Greene, "DNA sequence alignment during homologous recombination," *Journal of Biological Chemistry,* vol. 291, pp. 11572-11580, 2016.

[36] L. H. Thompson and D. Schild, "Recombinational DNA repair and human disease," *Mutat Res,* vol. 509, pp. 49-78, Nov 30 2002.

[37] A. J. Davis and D. J. Chen, "DNA double strand break repair via non-homologous end-joining," *Transl Cancer Res,* vol. 2, pp. 130-143, Jun 2013.

[38] M. R. Lieber, Y. Ma, U. Pannicke, and K. Schwarz, "The mechanism of vertebrate nonhomologous DNA end joining and its role in V(D)J recombination," *DNA Repair (Amst),* vol. 3, pp. 817-26, Aug-Sep 2004.

[39] L. Li and L. Zou, "Sensing, signaling, and responding to DNA damage: organization of the checkpoint pathways in mammalian cells," *J Cell Biochem,* vol. 94, pp. 298-306, Feb 1 2005.

[40] C. G. Broustas and H. B. Lieberman, "DNA damage response genes and the development of cancer metastasis," *Radiat Res,* vol. 181, pp. 111-30, Feb 2014.

[41] E. A. Williamson, J. W. Wray, P. Bansal, and R. Hromas, "Overview for the histone codes for DNA repair," *Prog Mol Biol Transl Sci,* vol. 110, pp. 207-27, 2012.

[42] A. Karimaian, M. Majidinia, H. B. Baghi, and B. Yousefi, "The crosstalk between Wnt/β-catenin signaling pathway with DNA damage response and oxidative stress: Implications in cancer therapy," *DNA repair,* vol. 51, pp. 14-19, 2017.

[43] E. Meng, A. Hanna, R. S. Samant, and L. A. Shevde, "The Impact of Hedgehog Signaling Pathway on DNA Repair Mechanisms in Human Cancer," *Cancers (Basel),* vol. 7, pp. 1333-48, Jul 21 2015.

[44] S. Du, S. Bouquet, C. H. Lo, I. Pellicciotta, S. Bolourchi, R. Parry, *et al.*, "Attenuation of the DNA damage response by transforming growth factor-beta inhibitors enhances radiation sensitivity of non-small-cell lung cancer cells in vitro and in vivo," *Int J Radiat Oncol Biol Phys,* vol. 91, pp. 91-9, Jan 1 2015.

[45] F. Kopper, C. Bierwirth, M. Schon, M. Kunze, I. Elvers, D. Kranz, *et al.*, "Damage-induced DNA replication stalling relies on MAPK-activated protein kinase 2 activity," *Proc Natl Acad Sci U S A,* vol. 110, pp. 16856-61, Oct 15 2013.

[46] X. Zhou, W. Liu, X. Hu, A. Dorrance, R. Garzon, P. J. Houghton, *et al.*, "Regulation of CHK1 by mTOR contributes to the evasion of DNA damage barrier of cancer cells," *Scientific reports,* vol. 7, p. 1535, 2017.

[47] L. Silver-Morse and W. X. Li, "JAK-STAT in heterochromatin and genome stability," *JAKSTAT,* vol. 2, p. e26090, Jul 1 2013.

[48] C. Fordyce, T. Fessenden, C. Pickering, J. Jung, V. Singla, H. Berman, *et al.*, "DNA damage drives an activin a–dependent induction of cyclooxygenase-2 in premalignant cells and lesions," *Cancer Prevention Research,* vol. 3, pp. 190-201, 2010.

[49] Z. J. Yang, W. L. Bao, M. H. Qiu, L. M. Zhang, S. D. Lu, Y. L. Huang, *et al.*, "Role of vascular endothelial growth factor in neuronal DNA damage and repair in rat brain following a transient cerebral ischemia," *J Neurosci Res,* vol. 70, pp. 140-9, Oct 15 2002.

[50] S. Janssens and J. Tschopp, "Signals from within: the DNA-damage-induced NF-κB response," *Cell death and differentiation,* vol. 13, p. 773, 2006.

[51] J. Vermezovic, M. Adamowicz, L. Santarpia, A. Rustighi, M. Forcato, C. Lucano, *et al.*, "Notch is a direct negative regulator of the DNA-damage response," *Nature Structural and Molecular Biology,* vol. 22, p. 417, 2015.

[52] G. Goyal, T. Fan, and P. T. Silberstein, "Hereditary cancer syndromes: utilizing DNA repair deficiency as therapeutic target," *Familial cancer,* vol. 15, pp. 359-366, 2016.

[53] M. C. Poirier, "Chemical-induced DNA damage and human cancer risk," *Discov Med,* vol. 14, pp. 283-8, Oct 2012.

[54] G. Poulogiannis, I. M. Frayling, and M. J. Arends, "DNA mismatch repair deficiency in sporadic colorectal cancer and Lynch syndrome," *Histopathology,* vol. 56, pp. 167-79, Jan 2010.

[55] R. M. Byrne and V. L. Tsikitis, "Colorectal polyposis and inherited colorectal cancer syndromes," *Ann Gastroenterol,* vol. 31, pp. 24-34, Jan-Feb 2018.

[56] R. Kleinerman, J. Marino, and E. Loucas, "Muir-Torre Syndrome / Turcot Syndrome overlap? A patient with sebaceous carcinoma, colon cancer, and a malignant astrocytoma," *Dermatol Online J,* vol. 18, p. 3, May 15 2012.

[57] A. K. Win, N. M. Lindor, J. P. Young, F. A. Macrae, G. P. Young, E. Williamson, *et al.*, "Risks of primary extracolonic cancers following colorectal cancer in lynch syndrome," *J Natl Cancer Inst,* vol. 104, pp. 1363-72, Sep 19 2012.

[58] R. Gryfe, "Inherited colorectal cancer syndromes," *Clin Colon Rectal Surg,* vol. 22, pp. 198-208, Nov 2009.
[59] I. R. Kirsch, *The causes and consequences of chromosomal aberrations*: CRC Press, 1992.
[60] Q. Wang, "Cancer predisposition genes: molecular mechanisms and clinical impact on personalized cancer care: examples of Lynch and HBOC syndromes," *Acta Pharmacol Sin,* vol. 37, pp. 143-9, Feb 2016.
[61] H. Hampel, W. L. Frankel, E. Martin, M. Arnold, K. Khanduja, P. Kuebler, *et al.*, "Screening for the Lynch syndrome (hereditary nonpolyposis colorectal cancer)," *N Engl J Med,* vol. 352, pp. 1851-60, May 5 2005.
[62] R. L. Siegel, K. D. Miller, S. A. Fedewa, D. J. Ahnen, R. G. S. Meester, A. Barzi, *et al.*, "Colorectal cancer statistics, 2017," *CA Cancer J Clin,* vol. 67, pp. 177-193, May 6 2017.
[63] P. S. Patil, A. Saklani, P. Gambhire, S. Mehta, R. Engineer, A. De'Souza, *et al.*, "Colorectal Cancer in India: An Audit from a Tertiary Center in a Low Prevalence Area," *Indian J Surg Oncol,* vol. 8, pp. 484-490, Dec 2017.
[64] J. Ferlay, I. Soerjomataram, M. Ervik, R. Dikshit, S. Eser, C. Mathers, *et al.*, "GLOBOCAN 2012 v1. 0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11 [Internet]. 2013; Lyon, France: International Agency for Research on Cancer," ed, 2014.
[65] T. Patra, S. Mandal, N. Alam, and N. Murmu, "Clinicopathological trends of colorectal carcinoma patients in a tertiary cancer centre in Eastern India," *Clinical Epidemiology and Global Health,* 2017.
[66] K. Tariq and K. Ghias, "Colorectal cancer carcinogenesis: a review of mechanisms," *Cancer Biol Med,* vol. 13, pp. 120-35, Mar 2016.
[67] E. R. Fearon and B. Vogelstein, "A genetic model for colorectal tumorigenesis," *Cell,* vol. 61, pp. 759-767, 1990.
[68] G. Smith, F. A. Carey, J. Beattie, M. J. Wilkie, T. J. Lightfoot, J. Coxhead, *et al.*, "Mutations in APC, Kirsten-ras, and p53—alternative genetic pathways to colorectal cancer," *Proceedings of the National Academy of Sciences,* vol. 99, pp. 9433-9438, 2002.
[69] J. E. East, B. P. Saunders, and J. R. Jass, "Sporadic and syndromic hyperplastic polyps and serrated adenomas of the colon: classification, molecular genetics, natural history, and clinical management," *Gastroenterology Clinics,* vol. 37, pp. 25-46, 2008.
[70] A. S. Sameer, S. Nissar, and K. Fatima, "Mismatch repair pathway: molecules, functions, and role in colorectal carcinogenesis," *Eur J Cancer Prev,* vol. 23, pp. 246-57, Jul 2014.
[71] D. J. Weisenberger, K. D. Siegmund, M. Campan, J. Young, T. I. Long, M. A. Faasse, *et al.*, "CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with BRAF mutation in colorectal cancer," *Nat Genet,* vol. 38, pp. 787-93, Jul 2006.
[72] R. J. Davies, R. Miller, and N. Coleman, "Colorectal cancer screening: prospects for molecular stool analysis," *Nature Reviews Cancer,* vol. 5, p. 199, 2005.
[73] F. A. Haggar and R. P. Boushey, "Colorectal cancer epidemiology: incidence, mortality, survival, and risk factors," *Clin Colon Rectal Surg,* vol. 22, pp. 191-7, Nov 2009.
[74] K. W. Jasperson, T. M. Tuohy, D. W. Neklason, and R. W. Burt, "Hereditary and familial colon cancer," *Gastroenterology,* vol. 138, pp. 2044-2058, 2010.
[75] R. Fodde, "The APC gene in colorectal cancer," *European journal of cancer,* vol. 38, pp. 867-871, 2002.

[76] A. L. Knudsen, M. L. Bisgaard, and S. Bülow, "Attenuated familial adenomatous polyposis (AFAP): a review of the literature," *Familial cancer,* vol. 2, pp. 43-55, 2003.

[77] H. T. Lynch, S. Lanspa, T. Smyrk, B. Boman, P. Watson, and J. Lynch, "Hereditary nonpolyposis colorectal cancer (Lynch syndromes I & II): genetics, pathology, natural history, and cancer control, Part I," *Cancer genetics and cytogenetics,* vol. 53, pp. 143-160, 1991.

[78] A. Bolocan, D. Ion, R. Stoian, and M. Serban, "Map syndrome (MYH Associated Polyposis) colorectal cancer, etiopathological connections," *Journal of medicine and life,* vol. 4, p. 109, 2011.

[79] P. Vega, F. Valentín, and J. Cubiella, "Colorectal cancer diagnosis: Pitfalls and opportunities. World J Gastrointest Oncol.[Internet]. 2015 [Access Dec 19, 2016]; 7 (12): 422-33," ed.

[80] B. M. Wolpin and R. J. Mayer, "Systemic treatment of colorectal cancer," *Gastroenterology,* vol. 134, pp. 1296-1310. e1, 2008.

[81] P. J. T. López, J. S. Albero, and J. A. Rodríguez-Montes, "Primary and secondary prevention of colorectal cancer," *CliniCal MediCine insights: gastroenterology,* vol. 7, p. CGast. S14039, 2014.

[82] M. Sehgal and T. R. Singh, "DR-GAS: a database of functional genetic variants and their phosphorylation states in human DNA repair systems," *DNA repair,* vol. 16, pp. 97-103, 2014.

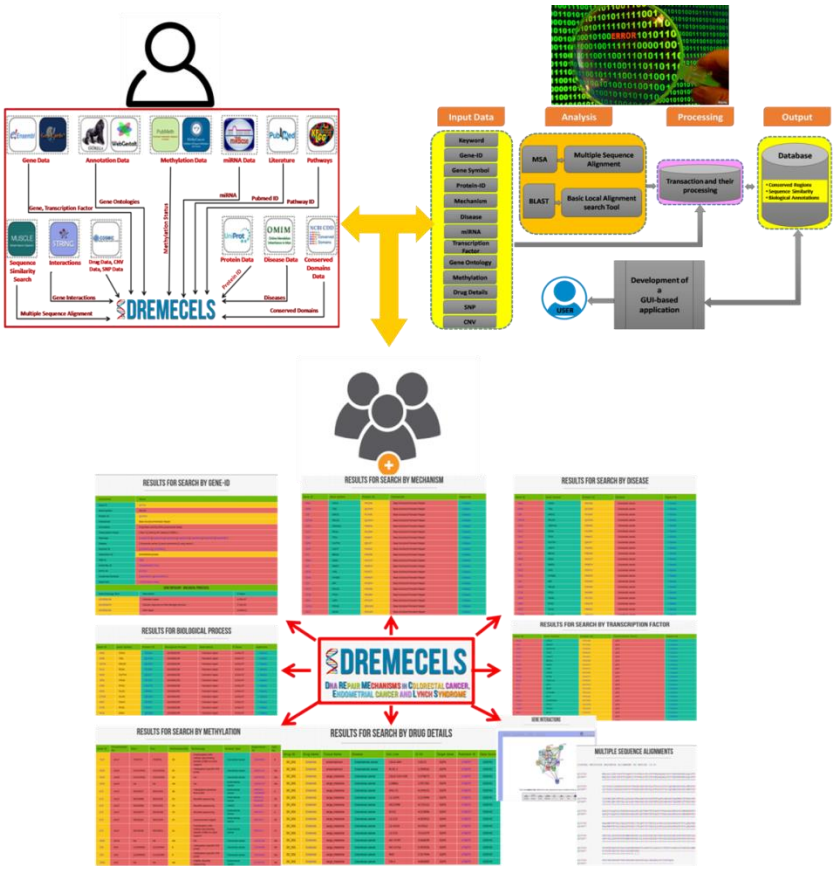
[83] A. Shukla, M. Sehgal, and T. R. Singh, "Hydroxymethylation and its potential implication in DNA repair system: A review and future perspectives," *Gene,* vol. 564, pp. 109-18, Jun 15 2015.

[84] A. Shukla and T. R. Singh, "Computational Network Approaches and Their Applications for Complex Diseases," in *Translational Bioinformatics and Its Application*, ed: Springer, 2017, pp. 337-352.

[85] M. Sehgal, R. Gupta, A. Moussa, and T. R. Singh, "An integrative approach for mapping differentially expressed genes and network components using novel parameters to elucidate key regulatory genes in colorectal cancer," *PloS one,* vol. 10, p. e0133901, 2015.

CHAPTER - 2

DREMECELS: A Curated Database for Base Excision and Mismatch Repair Mechanisms Associated Human Malignancies



~ Strive not to be a success, but rather to be of value.
- Albert Einstein

ABSTRACT

DNA repair is a key regulatory mechanism to overcome faults happening to the DNA on daily basis through endogenous or exogenous agents. It acts as a rebellion in opposition to the diverse form of damaging processes that have a tendency to develop a critical form of the disease. There are varieties of diseases that occur through the inefficient repair mechanisms and cancers are the most common of them. Although several cancer databases have been built so far, no one focused primarily on the colorectal or endometrial cancer. Bearing in mind, a database named DREMECELS has been developed for the colorectal and endometrial cancers, in association with Lynch syndrome. Lynch syndrome (also known as hereditary nonpolyposis colorectal cancer) set- off due to frequent alterations in DNA repair pathways particularly base excision repair and mismatch repair. The Lynch syndrome has high risk in both cancer types and accounts to 40-60% cases; considering this all three diseases were included in the archive. The DREMECELS is imbibed with the data of 156 genes focusing on base excision and mismatch repair mechanisms as they are the major contributor towards these diseases. The database is featured with the parameters that include a variety of regulatory processes having a role in progression of a disease. The database also offers information regarding somatic mutations, copy number variation (CNV), miRNAs, methylation status, and about drug sensitivity that makes it a complete package of fully featured genetic content embraced at one spot. The aim of database is to provide integrated information of disease types to serve the scientific community, thus supporting the diagnostic and therapeutic processes development. The repository will not only serve the researchers working in this field but also serve as an exceptional auxiliary for biomedical professionals thus facilitate understanding of the critical diseases. The database is free and easily accessible for public availability at http://www.bioinfoindia.org/dremecels.

2.1 INTRODUCTION

Genomes have constantly been vulnerable to damage from internal and external sources; including replication errors (leading to mismatches, insertion, and deletion); anti-tumor agents (cis-Pt, and MMC forming interstrand cross-links or double strand breaks); polycyclic aromatic hydrocarbons (forming bulky adducts); oxygen radicals, alkylating agents, and spontaneous reactions (creating abasic site) [1]. Organisms have undergone evolutionary changes since decades allowing the mutational alterations to stabilize their genome integrity. The maintenance of the genome integrity is taken care by inbuilt DNA repair mechanisms of the cells that cope up with the regulatory processes and help to identify and correct damages imposed to the DNA. There are various mechanisms of the DNA repair such as direct reversal of the damage, excision repair (base excision repair (BER), mismatch repair (MMR), and nucleotide excision repair), strand repair mechanism (single strand breaks and double strand breaks (via direct Joining and homologous recombination methods)), that help repairing the lost information [2]. Although if the repair mechanisms become non-functional it is an emergency signal to the system and can lead to the massive errors in the DNA impending genetic or epigenetic mutations. If the system fails to repair such alterations it can lead to cellular adversity and such condition cause a high risk of a disease outbreak. Studies revealed the prime role of MMR and BER mechanisms in colorectal cancer (CRC), endometrial cancer (EC), and Lynch syndrome (LS) allied CRC and EC [3-5].

Both the cancers CRC and EC are sturdily linked to the autosomal dominant syndrome LS. LS is also known as hereditary nonpolyposis colorectal cancer (HNPCC), occurs primarily through inept MMR mechanism [6]. The major mechanism follows the germline mutation coupled with tumor leading to the abnormal chromosomal condition called microsatellite instability (MSI) [7]. It is a prime reason for multiple cancers but predominantly it cause CRC, and up to less extent comparatively found to develop EC [8, 9]. Quantitatively it results in CRC up to 50%–80% (mean age 40–60 years) cases and EC up to 20%–60% (mean age 45–60 years) cases [9]. Several studies demonstrated the effect of genes that undergo alterations in CRC, EC, and LS through various *in-vitro* (genome hybridization arrays, miRNA arrays, methylation arrays, ChIp-on-chip, proteomic and functional genomics, genome-wide association) and *in-silico* approaches (bioinformatics, biostatistical) [10-14]. These approaches to date assist in

providing bulk data for genes related to cancer and facilitate scientists deciphering the underlying mechanisms for diseases.

Despite the presence of huge databases there was lack of the database that considers CRC, EC, and also both cancers in association with the LS. An extensive vision is required to understand the genetic mechanism of these constantly emerging cancers. An exponential growth of the data makes it necessary to have a resource where all information can be preserved at one place to make the manual cumbersome process easy. It not only provides the easy access to the data but also keep updated with what new has been come so far and hence decline the probability to miss out the important things for research consideration. To understand these cancers there is a dire need for considerable advances in the research area of biomarkers for risk assessment, monitoring disease progression, and causal factors recognition. This facilitates a need for a platform to easily access the information that is scattered in public domains and literature and to integrate into organized form for all the concerned parameters of the diseases under consideration. The DREMECELS is a (DNA REpair MEchanism in Colorectal, Endometrium and Lynch Syndrome) is first of its type with comprehensive information for the CRC, EC, and LS. It is a manually curated database that comprises genes concerned with the repair mechanisms in association with the three disease forms. The database is diversified with various genetic parameters, for instance, gene markers, protein markers, gene annotation, PubMed links, respective regulatory miRNAs, transcription factors, conserved domains, gene-interactions, and pathways. The mutation information for the cancers such as somatic mutation at the single base position and copy number variation (CNV), a form of structural variation is also incorporated. The epigenetic details that are essential to determine the cancerous conditions are being incorporated in the form of methylation patterns. Also, the respective drug details have been provided for the individual disease type.

DREMECELS is an inclusive catalog for DNA repair genes in association with the disease pathway. The database is incorporated into a web based Graphical User Interface (GUI) and developed for 156 base excision and mismatch repair genes/proteins. The information presented in a database could be of high relevance to the researchers involved in the study of these diseases. It provides the biomarkers information for BER and MMR mechanisms allied CRC, EC, and LS that could provide a feasible way for disease diagnosis and hence determining suitable drug for same.

2.2 MATERIALS AND METHODS

An extensive literature search has been done to collect the data manually as well as a variety of standard resources have been used to gather cancer-related genes. Events like gene interactions, enrichment, drug sensitivity, and methylation have added substantial credence to the database. The overall architecture of the database is given below (Figure 2.1); it represents four steps; the input data it comprises, the analysis tools used, processing phase of the gathered information, output generation phase and finally the development of a web based resource enema.

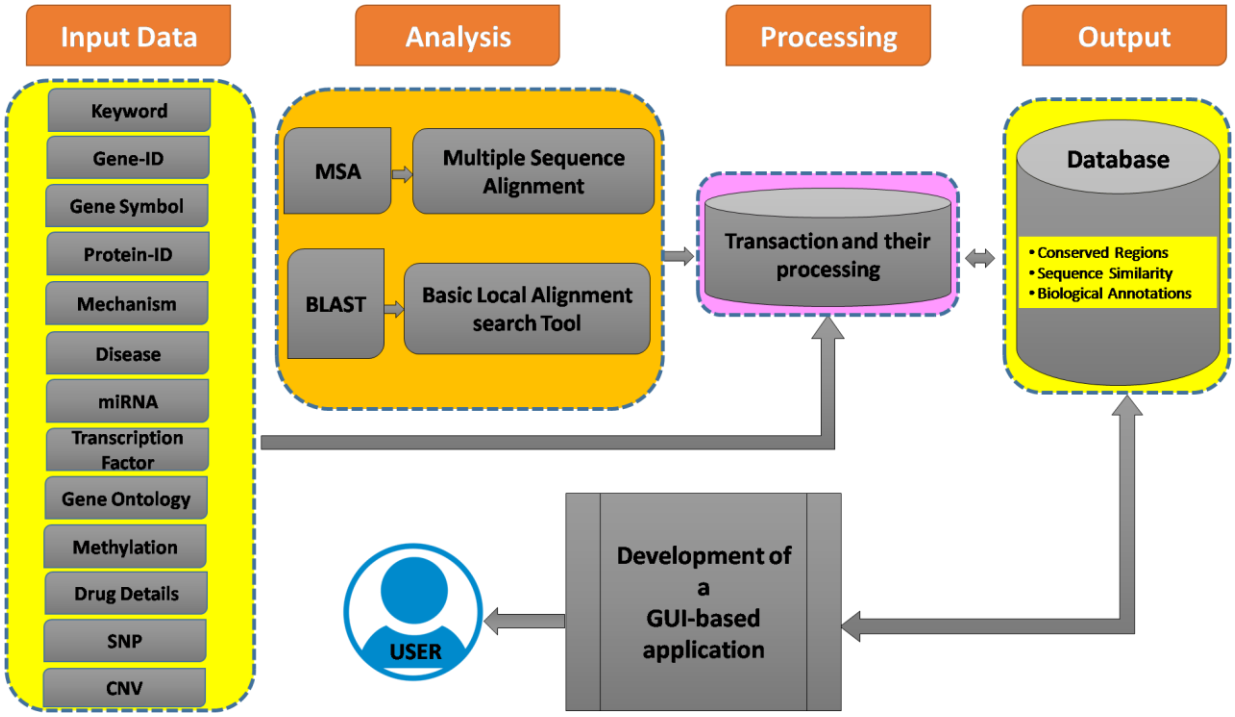


Figure 2.1 The comprehensive architecture of the DREMECELS.

2.2.1 Data Collection

Various standard databases have been used to collect the data; it includes UniProt (proteins archive) for proteins and their sequences with respect to the disease-causing genes [15]. A program called MUSCLE has been used to determine the evolutionary relationship among close homologs via MSA through conserved regions of the alignment [16]. Conserved domain database (CDD) was used to determine the domain regions (a region that function autonomously) for a protein sequence [17]. Gorilla [18] and WebGestalt [19] tools were used for gene annotation studies that provide the functional enrichment to the genes through statistical analysis. As post-transcription regulatory mechanisms provide hidden details for the futuristic processes, therefore, inclusion of miRNAs data was done, which was collected from miRBase [20]. The

pathways data has also been incorporated by providing an external link to the KEGG database for every related set of genes as it entails information for gene regulatory elements [21]. PubMed was extensively searched for the literature and therefore for every set of gene, the hyperlinked references have been added. GeneCards have been used to determine the transcription factors for the related gene; the information is of high significance as it reveals the factors regulating transcription rate of a gene [22].

Methylation data has been added from literature survey and by taking disease associated methylation patterns from the PubMeth database [23, 24]. It is most widely studied epigenetic mechanism for cancers and therefore can prove to be a crucial biomarker for cancer detection. It targets the gene regulatory mechanism and therefore can fortify therapeutic intervention process. The methylation patterns for the diseases have been added in terms of hypo- or hyper-methylation forms to the database (Table 2.1). The catalog of somatic mutations in cancer (COSMIC) is a database for almost all forms of human cancer; the drug sensitivity data of 140 drugs and that of somatic mutations and CNV were taken for CRC, EC, and LS [25]. Duplication or deletion via germline or somatic mutation events can give rise to the condition called CNV's [26, 27] however in somatic mutations it comprises of single base-pair alteration in a genome relative to the other members of a species [28]. The overall data collection process for diverse parameters and features are shown in Figure 2.2.

Table 2.1 Genes involved at different methylation level

Methylation status	Disease	No. of genes	Name of genes
Methylated	Colorectal cancer	32	TP53, MGMT, MPG, APC, MLH1, PMS1, EXO1, ATM, PTGS2, BCL2, KRAS, CDX2, TERT, FHIT, RB1, HNF1A, WRN, SMAD2, DCC, DKK1, IGF2, STK11, BAX, WNT5A, RASSF2, TWIST1, AXIN1, ERBB2, MYO1A, MEIS1, SFRP4, HLTF
	Endometrial cancer	9	MGMT, BRCA1, MLH1, MSH6, PMS1, RB1, CTNNB1, ABCB1, RPS6KA6,
	Lynch syndrome (HNPCC)	3	BRCA1, MLH1, PMS1
Hypomethylated	Colorectal cancer	2	IGF2, MUC5AC
Hypermethylated	Colorectal cancer	11	MSH2, TDG, OGG1, PTEN, ERCC1, XPC, WRN, IGF2, ARID1A, IDO1, SDC1
	Endometrial cancer	2	PTEN, IGF2
	Lynch syndrome (HNPCC)	1	MSH2

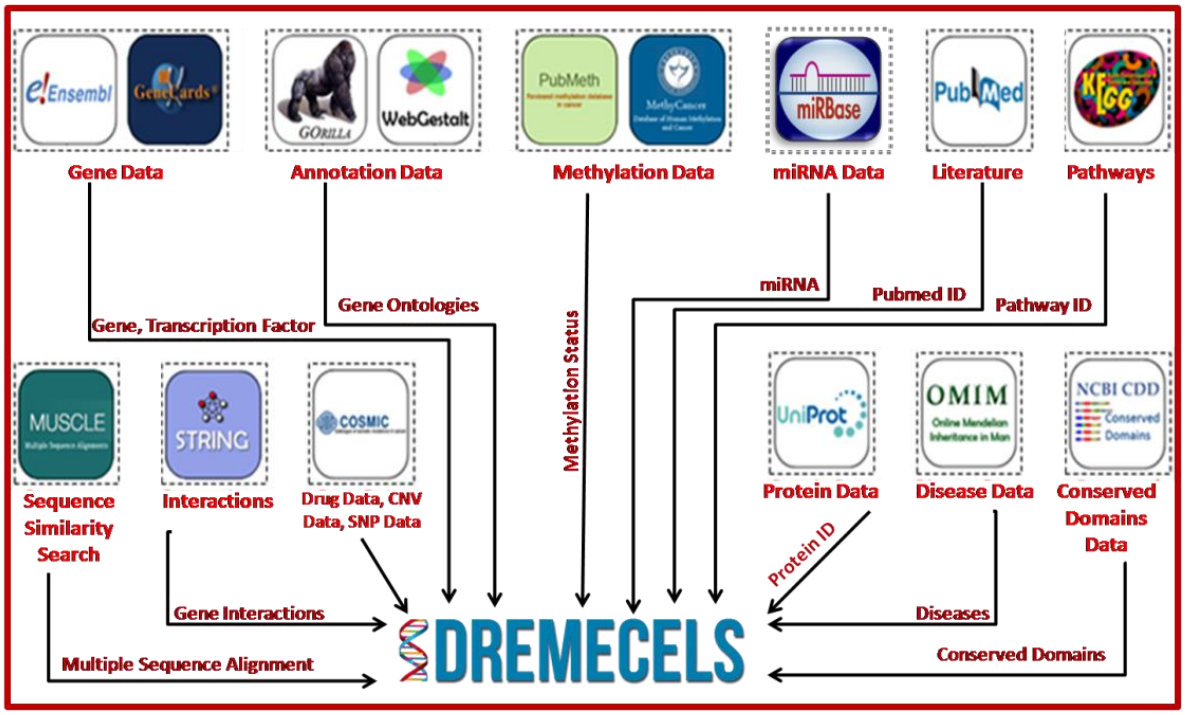


Figure 2.2 Data Collection and compilation for DREMECELS from various standardized resources.

2.2.2 Database Configuration

The database follows relational database system as it integrates data through MySQL implementation. The programming languages like HTML, CSS, JavaScript, and PHP were used for creating a GUI. The connection between the interface and the back-end was established using phpMyAdmin being hosted through Apache server. It is expected to be critical for both researchers and clinicians in understanding and determining the molecular mechanisms underlying these diseases.

2.2.3 Back-End Design

In our database back-end comprises of tables devoid of redundancy; there are a set of 12 tables that are interconnected through primary key. Since the database follows relational database system, therefore, the data in all the tables are linked.

- **Gene Table**: It comprises of gene data containing information of all non-redundant set of genes with gene ID as a primary key.
- **Disease Table**: It comprises information of contributing diseases relative to the individual gene.
- **Conserved Domains Table**: It comprises information on the functional units of the proteins that has potential to work independently.
- **GO Table**: It comprises information on gene annotation for biological processes, molecular function, and cellular component providing GO-ID, its description, and the level of significance (p-value) with respect to the protein.
- **miRNA Table**: It comprises miRNA information given in terms of standardized HUGO gene nomenclature.
- **Pathways Table**: It comprises information on the pathways associated with the disease type.
- **PubMed Table**: It comprises literature references to the respective gene and protein.
- **Transcription Factor Table**: It comprises information on gene regulatory factors.
- **Methylation Table**: It comprises information on epigenetic mechanism data (the hypo and hypermethylation).
- **Drug Sensitivity Table**: It comprises information on all the available drugs along with their target.

- **CNV Table**: It comprises information on copy number alterations along with expression level corresponding to the disease.
- **Somatic Mutations Table**: It comprises information on various types of polymorphisms, its histology and the corresponding reference related to the gene.

2.2.4 Biological Enrichment Analysis

For all 156 set of genes present in this database, classification has been performed for the cancer causing genes depending upon their molecular function and the biological processes they are involved. It was found that for the molecular functions they were involved in the transcription regulation, tumor suppression, binding of DNA, enzymatic activity, and control of cell cycle. This shows significance of their gene products in cancerous condition. The gene products for biological processes are found to be involved in developmental, metabolic, and multicellular organismal processes, stimulating a response, and in cell signaling. This suggests the role of biological processes for understanding the cell behavior studies and its transition into diseased condition. Likewise in the following chapters the cell signaling processes have been discussed; specifically the Wnt and the TGFβ pathways and their regulatory effects on the CRC.

2.3 RESULTS AND DISCUSSION

DREMECELS comprises of 156 DNA repair (MMR and BER) gene biomarkers associated with the CRC, EC, and LS. The database offers an easy to use and efficient method to search and retrieve data of associated genes along with other significant genetic parameters. The percentage of genes falling under the category of BER is 23%, for MMR it is 48% and the genes falling under both categories cover 29% of total genes in the database (Figure 2.3).

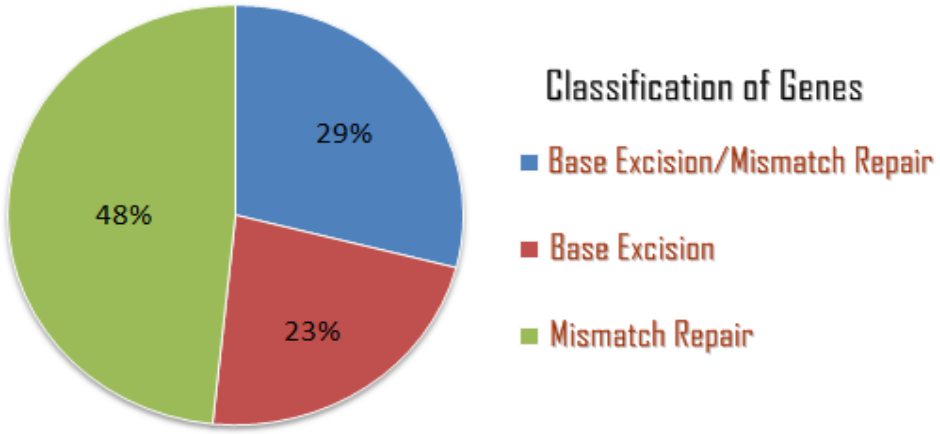


Figure 2.3 DNA repair genes and associated mechanisms; the percentage (%) shows the number of genes present in each mechanism. Mismatch repair mechanism was found more dominant.

2.3.1 The Graphical User Interface (GUI)

The database is integrated with the web interface which offers quick accessibility of data to the users for browsing the repository. There are thirteen different types of search options provided for accessing the data through a keyword search, gene id, gene symbol, protein id, mechanism, disease, miRNA, transcription factor, gene ontology, methylation, drug details, CNV and SNP search (Figure 2.4).

By Keyword: It takes the keyword e.g. a disease (colorectal, endometrial) or repair mechanism (mismatch, base excision) or a gene symbol (MLH1) etc. in the search box and return the results accordingly.

By Gene ID: The gene ID is a stable identifier which is unique integer generated by Entrez [29], provide the query search through the respective ID of a gene. The result page shows data with respect to the given geneID as a search query.

By Gene Symbol: Gene symbol is a standardized symbol approved by HGNC, it's a short abbreviation given to the gene name comprising of alphanumeric characters (e.g. APEX1). Whenever a search is made using gene symbol the corresponding gene-details were provided correspondingly.

By Protein ID: The database used for accessing protein ids was UniProt. It is a freely accessible resource of protein sequence and functional annotation. The ID represents length of 6

alphanumeric characters. The query search using the protein ID will retrieve the comprehensive information for protein of interest.

By Mechanism: It represents the search through the DNA repair mechanisms i.e. base excision repair, mismatch repair, and both. The user search query will retrieve the results for the repair mechanism of interest.

By Disease: This search option includes the search through disease type, i.e. colorectal cancer, endometrial cancer, Lynch syndrome and other associated diseases. The query search results into the page displaying information for the mentioned disease in search box.

By miRNA: The search through the standard miRNA symbol will display complete information about its feature with respect to the gene and protein it belongs.

By Transcription Factor: The search through the transcription factors provides an easy retrieval of the genes regulated by the mentioned transcription factors. It can be searched using HGNC-approved symbol.

By Gene Ontology: The gene ontology search is mainly focused on three types of the annotations that include biological process, molecular function and the cellular component. It can be searched using standard GO Id.

By Methylation: Methylation search is focused on the pattern of methylation like hyper and the hypo form of the methylation. The search can be performed through the in-built options: methylated, hypomethylated, and hypermethylated.

By Drug Details: The drug details search option provides all the drugs available for the disease type and its relative information. It can be searched using drug ids. The results will display detailed information for the respective drug searched.

By CNV: The search through the CNV can be searched by either disease name or by gene name.

By SNP: To determine the single nucleotide mutation related to gene, SNP search option is provided that will search for the single point mutation for a gene. It can be searched using six mutations types i.e. substitution-missense, substitution-intronic, substitution-nonsense, substitution-coding silent, deletion-frameshift, and deletion-in frame.

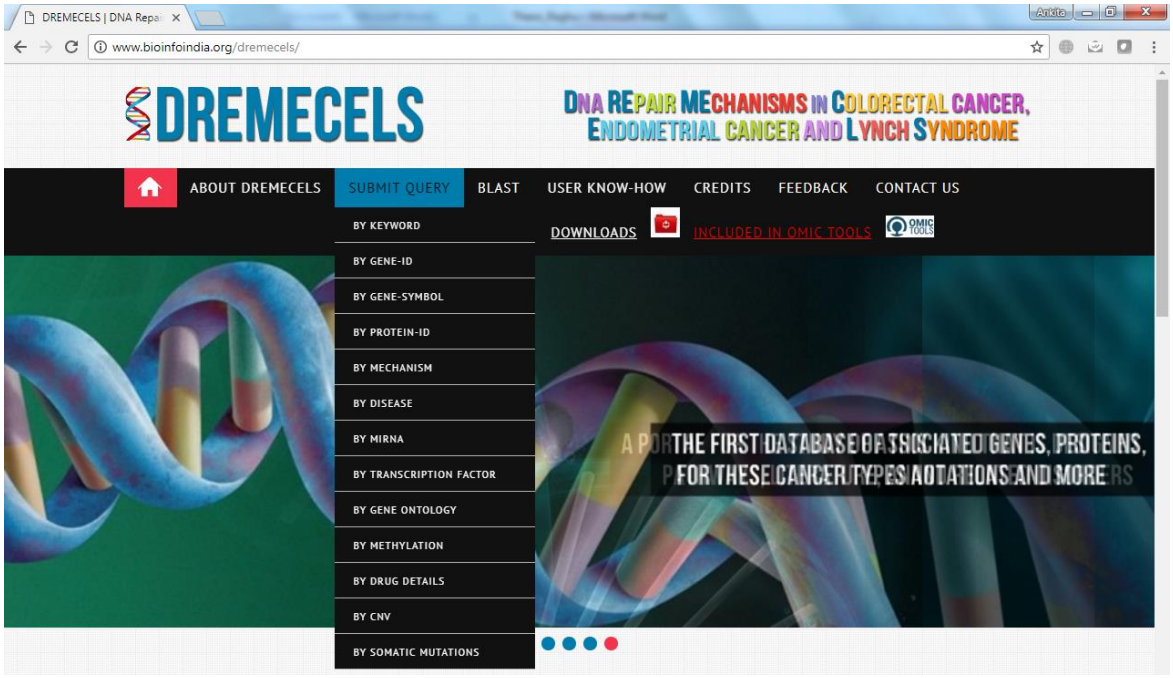


Figure 2.4 Screenshot of the homepage of a DNA Repair Mechanisms in colorectal cancer, endometrial cancer, and Lynch syndrome (DREMECELS).

After the query get submitted, if the query search matches the entry in the database will redirect to the results page, otherwise return an error message *"No such record found"*. The results page will display following outputs:

- The comprehensive information in a tabular format covering gene ID, gene symbol, protein ID, protein name, mechanism, annotation, transcription factor, pathways, disease, PubMed ID, miRNA, interaction ID, PDB ID, ENSEMBL ID, OMIM ID, conserved domains that are directly linked to the gene card.
- The pictographic illustration of the interacting genes through embedded STRING page [30].
- The sequence conservation has been done through MUSCLE, thus represented in the form of aligned sequences signifying level of similarity.
- The methylation data include gene ID, chromosome number, start and end positions, gene methylation (%age), technology, disease type, experiment ID, and CpG number.
- The drug details include manually created drug ID, drug name, tissue name, disease, cell line, IC-50, target gene, PubChem ID, and data source.
- The CNV details include sample, gene, expression, CN_Type, copy number, copy number segment position and disease.

- ❖ The SNP detail include gene, transcript, sample name, amino acid mutation, somatic mutations for substitution-missense, substitution-intronic, substitution-nonsense, substitution-coding silent, deletion-frameshift and deletion-in frame and disease.

Also, BLAST has been integrated [31] in our database for performing similarity search for identifying close homolog. The method employs user-defined protein in a fasta format as a query search against the sequences available in the protein archive of this database. The main usage of deploying BLAST tool in this portal is to characterize hypothetical sequences related to the BER, MMR and the disease implicated. It results in the homologous sequences from the database on the basis of sequence similarity scores which is counted through E-value. Standard parameters have been implicated for protein BLAST to make the search smooth and effective. An illustration of the results obtained through the database has been given in Figure 2.5.

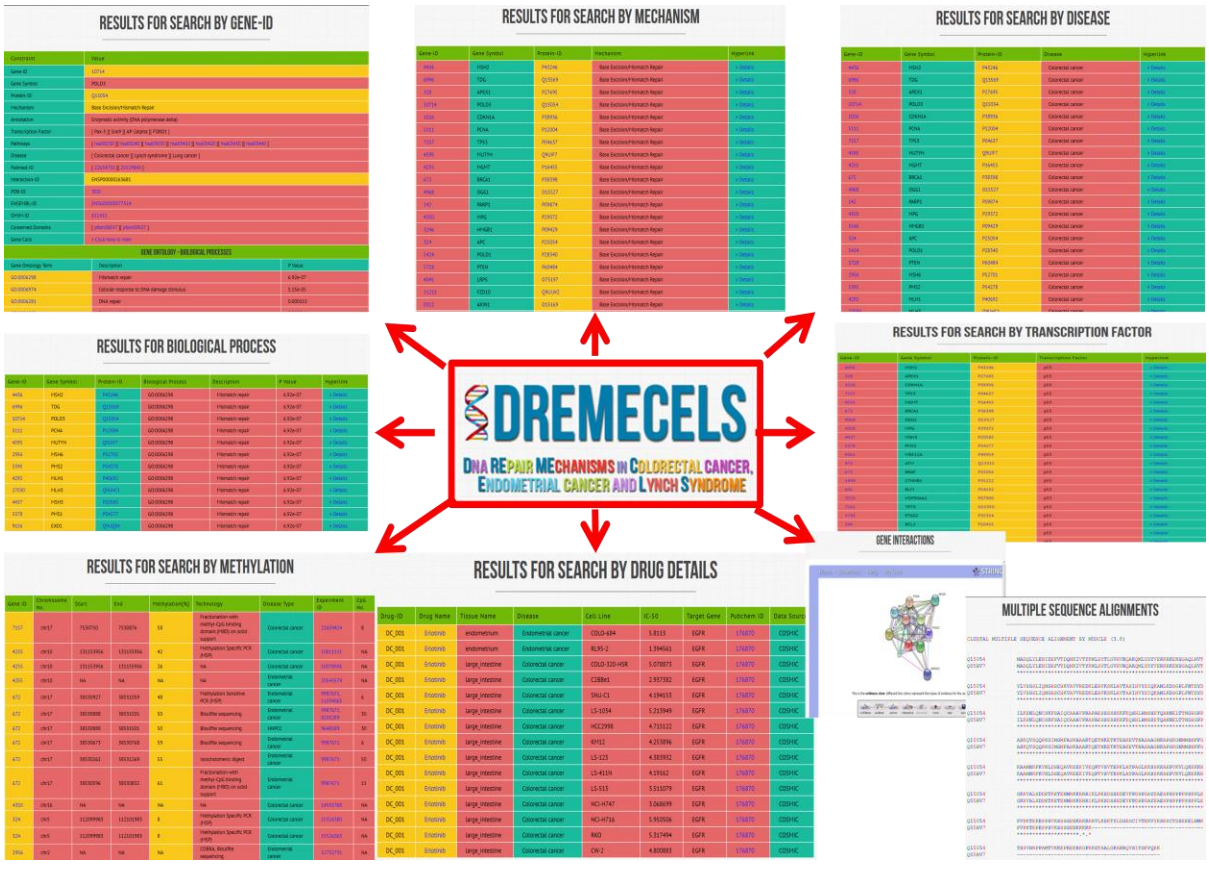


Figure 2.5 Manifestation and implementation of DREMECELS with various available searches. The exemplified output from the archive is represented in an integrated results form for all search options.

2.3.2 The Statistics

It reports the number of genes present in a respective disease and the functional annotation of the genes by category. According to the disease statistics of this portal majority of the genes were found to be involved in CRC (142), then in Lynch syndrome (114), afterwards in endometrial cancer (113) and in a small number for the rest of the disease like multiple myeloma, gastric cancer, and breast cancer etc (Figure 2.6). However, in annotation statistics that is calculated depending upon the function, the genes were found to have a role in DNA damage repair, then in signal transduction and afterward in enzymatic activity (Figure 2.7). This suggests role of the genes and their involvement in the multiple disease conditions. This will entail not only the diagnosis for the targeted DNA repair disease but also for the other prevalent disease forms.

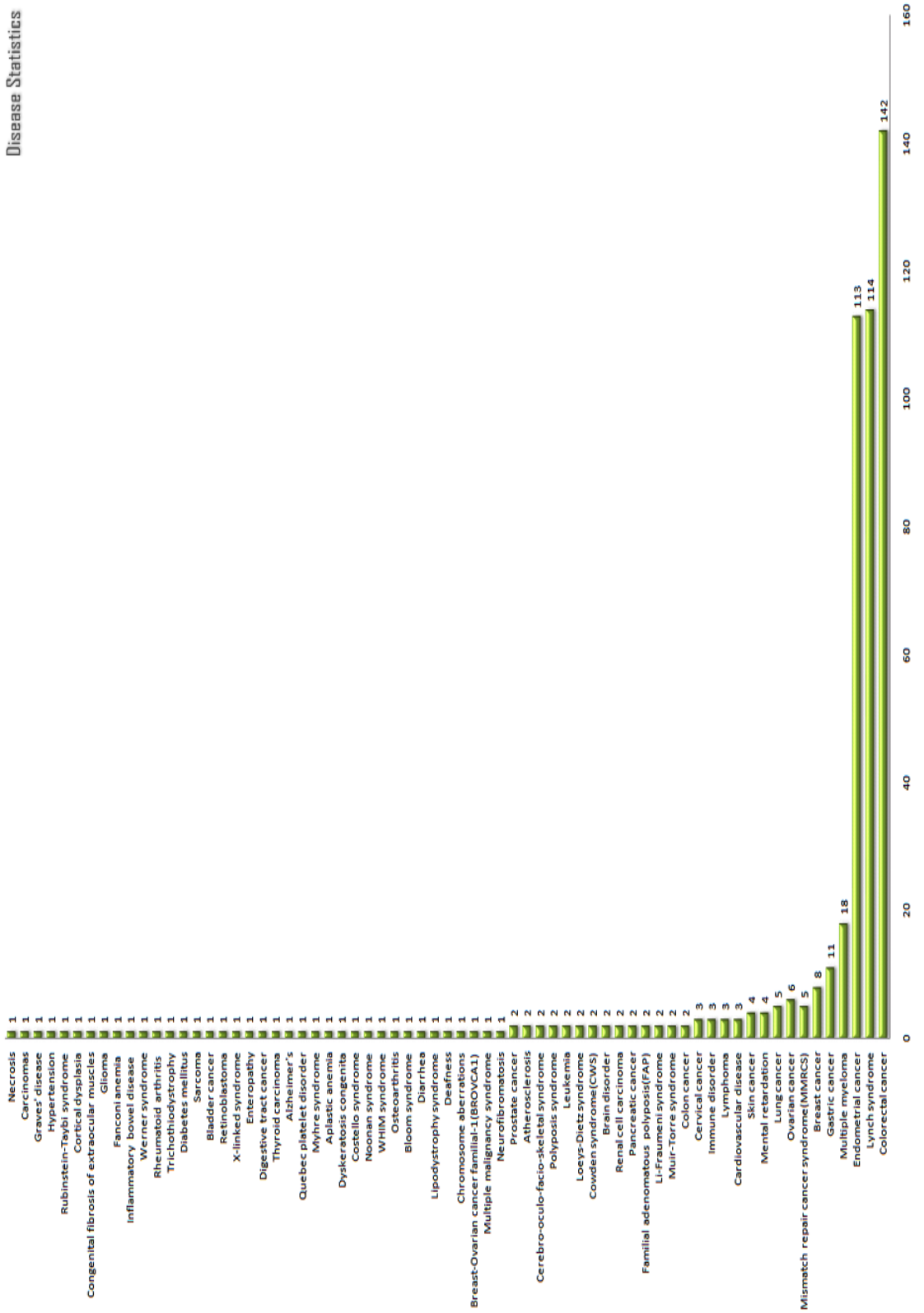


Figure 2.6 Graph displaying the number of genes present in each disease type.

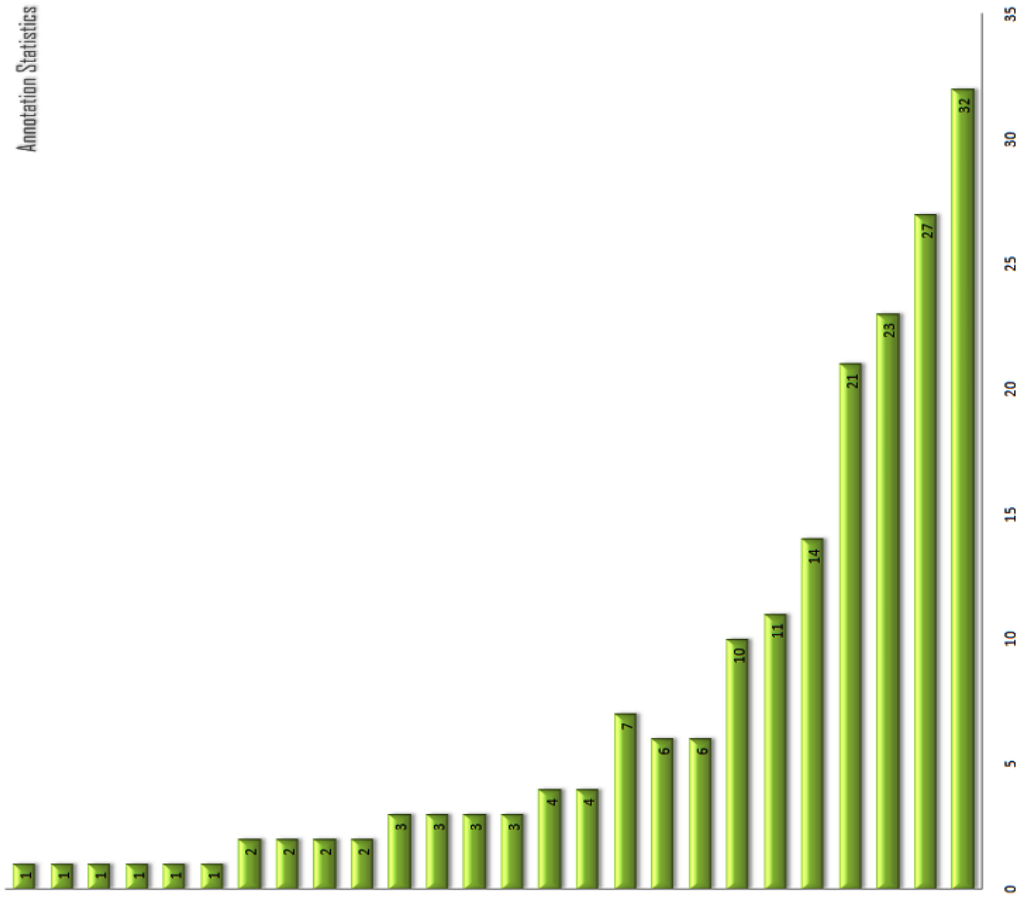


Figure 2.7 The annotation statistics of the genes involved in the colorectal cancer, endometrial cancer and Lynch syndrome.

2.4 CONCLUSION

DREMECELS is a first-ever resource predominantly for colorectal, endometrial cancers and both cancers allied to Lynch syndrome. It serves as an extensive compilation for base excision, and mismatch repair genes comprising of numerous genetic parameters such as comprehensive gene-details, methylation, pathways, diseases, mutations etc. This database will assist researchers to study the gene markers in depth and will provide useful insight for future analysis and studies. This repository will also help for easy understanding and investigation of many other related disease and disorders and provide useful genetic information. The database will prove to be useful to the scientists aiming new therapeutic targets not only for these three forms of disease but also to the other complex forms like multiple myelomas, gastric cancers, breast cancers etc. via genetic factor's information, which is basis for the disease diagnosis. The information integrated into the database will not only assist molecular biologists but also therapeutic developers to encounter biologically meaningful information. All data has been collected manually supported by literature references for these malignancies that provide a knowledge-based resource and thus allow researchers and clinicians to have a biological overview of the genes implicated in disease. The database would save time and efforts of researchers involved in the field through easy accessibility to data, and thus will facilitate in biological discoveries. The database will be updated on a regular basis to keep updated information to the academicians and researchers. The database is organized to provide clarity, ease of access, download and fast browsing capability. Further work based upon the information compiled in DREMECELS has been presented in other remaining objectives of this thesis.

REFERENCES

[1] E. C. Friedberg, "A history of the DNA repair and mutagenesis field: The discovery of base excision repair," *DNA Repair (Amst)*, vol. 37, pp. A35-9, Jan 2016.

[2] J. M. Furgason and E. M. Bahassi, "Targeting DNA repair mechanisms in cancer," *Pharmacology & therapeutics*, vol. 137, pp. 298-308, 2013.

[3] T. J. Kinsella, "Coordination of DNA mismatch repair and base excision repair processing of chemotherapy and radiation damage for targeting resistant cancers," *Clin Cancer Res*, vol. 15, pp. 1853-9, Mar 15 2009.

[4] E. Chow, C. Thirlwell, F. Macrae, and L. Lipton, "Colorectal cancer and inherited mutations in base-excision repair," *Lancet Oncol*, vol. 5, pp. 600-6, Oct 2004.

[5] V. Shilpa and K. Lakshmi, "Molecular Mechanisms of Mismatch Repair Genes in Cancer–A Brief Review," *Journal of Proteomics and Genomics*, vol. 1, p. 1, 2014.

[6] J. V. Martín-López and R. Fishel, "The mechanism of mismatch repair and the functional analysis of mismatch repair defects in Lynch syndrome," *Familial cancer*, vol. 12, pp. 159-168, 2013.

[7] J. C. Strafford, "Genetic testing for lynch syndrome, an inherited cancer of the bowel, endometrium, and ovary," *Reviews in Obstetrics and Gynecology*, vol. 5, p. 42, 2012.

[8] E. Stoffel, B. Mukherjee, V. M. Raymond, N. Tayob, F. Kastrinos, J. Sparr, *et al.*, "Calculation of risk of colorectal and endometrial cancer among patients with Lynch syndrome," *Gastroenterology*, vol. 137, pp. 1621-1627, 2009.

[9] R. A. Pagon, M. P. Adam, H. H. Ardinger, S. E. Wallace, A. Amemiya, L. J. Bean, *et al.*, "GeneReviews (®)," 1993.

[10] K. Yano, K. Imai, A. Shimizu, and T. Hanashita, "A new method for gene discovery in large-scale microarray data," *Nucleic acids research*, vol. 34, pp. 1532-1539, 2006.

[11] C.-G. Liu, G. A. Calin, S. Volinia, and C. M. Croce, "MicroRNA expression profiling using microarrays," *Nature protocols*, vol. 3, pp. 563-578, 2008.

[12] B. A. Bejjani and L. G. Shaffer, "Application of array-based comparative genomic hybridization to clinical diagnostics," *J Mol Diagn*, vol. 8, pp. 528-33, Nov 2006.

[13] D. J. Huebert, M. Kamal, A. O'Donovan, and B. E. Bernstein, "Genome-wide analysis of histone modifications by ChIP-on-chip," *Methods*, vol. 40, pp. 365-369, 2006.

[14] X. Chen, E. Jorgenson, and S. T. Cheung, "New tools for functional genomic analysis," *Drug Discov Today*, vol. 14, pp. 754-60, Aug 2009.

[15] M. Magrane and C. UniProt, "UniProt Knowledgebase: a hub of integrated protein data," *Database (Oxford)*, vol. 2011, p. bar009, 2011.

[16] R. C. Edgar, "MUSCLE: a multiple sequence alignment method with reduced time and space complexity," *BMC Bioinformatics*, vol. 5, p. 113, Aug 19 2004.

[17] A. Marchler-Bauer, J. B. Anderson, P. F. Cherukuri, C. DeWeese-Scott, L. Y. Geer, M. Gwadz, *et al.*, "CDD: a Conserved Domain Database for protein classification," *Nucleic acids research*, vol. 33, pp. D192-D196, 2005.

[18] E. Eden, R. Navon, I. Steinfeld, D. Lipson, and Z. Yakhini, "GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists," *BMC Bioinformatics*, vol. 10, p. 48, Feb 03 2009.

[19] J. Wang, D. Duncan, Z. Shi, and B. Zhang, "WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013," *Nucleic Acids Res*, vol. 41, pp. W77-83, Jul 2013.

[20] S. Griffiths-Jones, R. J. Grocock, S. van Dongen, A. Bateman, and A. J. Enright, "miRBase: microRNA sequences, targets and gene nomenclature," *Nucleic Acids Res,* vol. 34, pp. D140-4, Jan 1 2006.

[21] E. Altermann and T. R. Klaenhammer, "PathwayVoyager: pathway mapping using the Kyoto Encyclopedia of Genes and Genomes (KEGG) database," *BMC Genomics,* vol. 6, p. 60, May 3 2005.

[22] M. Safran, I. Dalah, J. Alexander, N. Rosen, T. Iny Stein, M. Shmoish, *et al.*, "GeneCards Version 3: the human gene integrator," *Database,* vol. 2010, p. baq020, 2010.

[23] X. He, S. Chang, J. Zhang, Q. Zhao, H. Xiang, K. Kusonmano, *et al.*, "MethyCancer: the database of human DNA methylation and cancer," *Nucleic acids research,* vol. 36, pp. D836-D841, 2007.

[24] M. Ongenaert, L. Van Neste, T. De Meyer, G. Menschaert, S. Bekaert, and W. Van Criekinge, "PubMeth: a cancer methylation database combining text-mining and expert annotation," *Nucleic acids research,* vol. 36, pp. D842-D846, 2007.

[25] S. A. Forbes, N. Bindal, S. Bamford, C. Cole, C. Y. Kok, D. Beare, *et al.*, "COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer," *Nucleic acids research,* vol. 39, pp. D945-D950, 2011.

[26] P. J. Hastings, J. R. Lupski, S. M. Rosenberg, and G. Ira, "Mechanisms of change in gene copy number," *Nat Rev Genet,* vol. 10, pp. 551-64, Aug 2009.

[27] F. Speleman, C. Kumps, K. Buysse, B. Poppe, B. Menten, and K. De Preter, "Copy number alterations and copy number variation in cancer: close encounters of the bad kind," *Cytogenet Genome Res,* vol. 123, pp. 176-82, 2008.

[28] B. N. Ford, C. C. Ruttan, V. L. Kyle, M. E. Brackley, and B. W. Glickman, "Identification of single nucleotide polymorphisms in human DNA repair genes," *Carcinogenesis,* vol. 21, pp. 1977-1981, 2000.

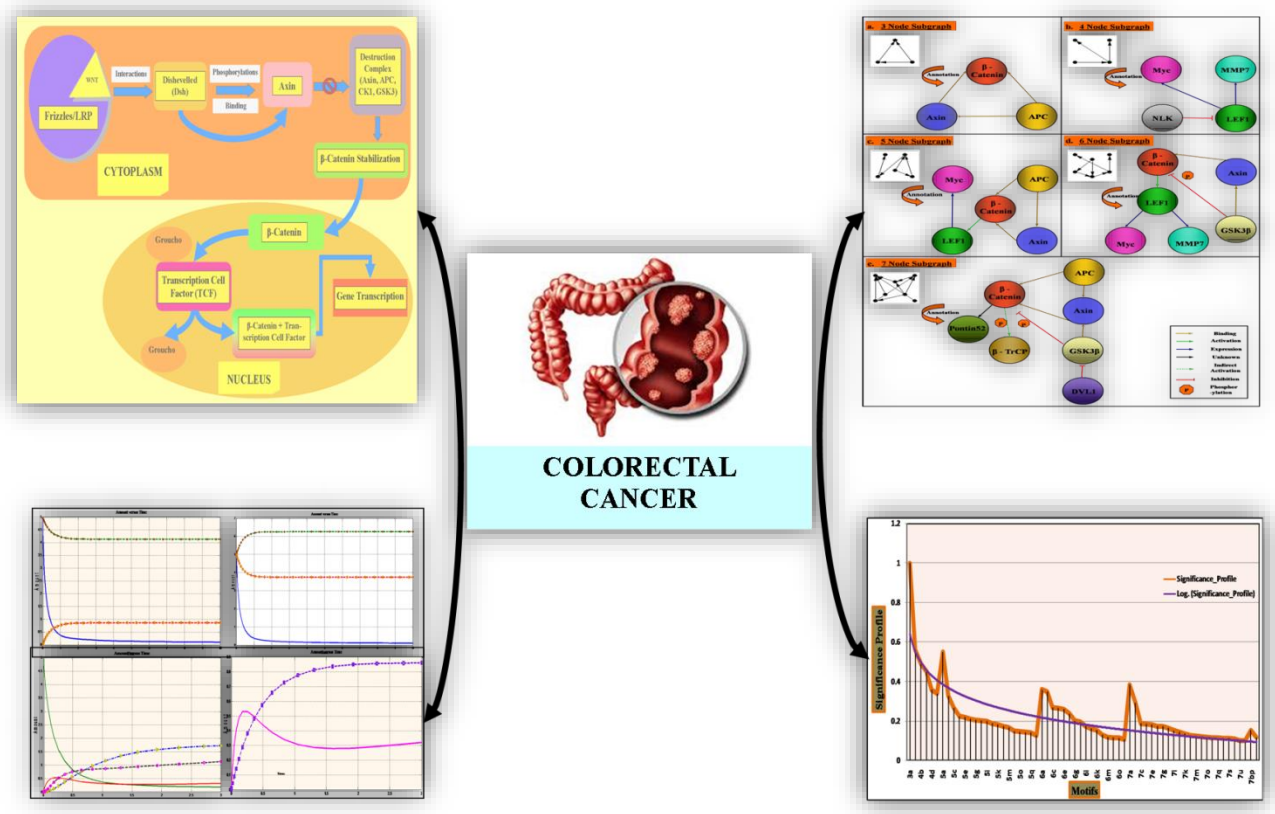
[29] D. Maglott, J. Ostell, K. D. Pruitt, and T. Tatusova, "Entrez Gene: gene-centered information at NCBI," *Nucleic acids research,* vol. 33, pp. D54-D58, 2005.

[30] C. v. Mering, M. Huynen, D. Jaeggi, S. Schmidt, P. Bork, and B. Snel, "STRING: a database of predicted functional associations between proteins," *Nucleic acids research,* vol. 31, pp. 258-261, 2003.

[31] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of molecular biology,* vol. 215, pp. 403-410, 1990.

CHAPTER - 3

Network-Based Approach to Study Dynamics of Wnt Pathway Regulatory Elements in Colorectal Cancer



~ Perfection is not attainable, but if we chase perfection we can catch excellence.
-Vince Lombardi

ABSTRACT

The systems biology facilitates understanding complex biological interactions at the genome, proteome, and organelle level. Involved biological processes include regulatory and metabolic mechanisms. To understand the complex form of disease like cancer identification and characterization, individual molecule is not sufficient, there is need to study them whole instead of considering specific cells or, tissues. Therefore, it is necessary to attain a thorough understanding of the interaction among molecules and pathways. With advancement in technologies, huge data has been coming across researchers; therefore, it's essential to correctly interpret them and to understand the system at the molecular level. Computational models following mathematical procedures allow researchers to investigate complex regulatory processes connection, and how processes disruptions contribute towards disease development. Computational modeling assists investigators to systematically analyze systems perturbations, thus help to develop hypotheses, and guide the design of new experimental tests, and ultimately assess the suitability of specific molecules as novel therapeutic targets. In this study, Wnt pathway analysis has been performed to determine the role of potential biomarkers in colorectal cancer. The simulations were carried out to identify the dynamics of an individual component that helps to attain their behavioral role in colorectal carcinogenesis. Also, network motifs were determined to decipher the significant transcription factors or regulators. The standard statistical parameters such as z-score, p-value, and significance profile were used to find the candidate genes in the pathway. Five key genes were found to be statistically significant i.e. *AXIN*, *APC*, *β-catenin*, *LEF1*, and *Myc*. It is hoped that these putative biomarkers could be efficient in disease diagnosis process and help to solve the mystery for the abnormal regulation of Wnt signaling in colorectal cancer.

3.1 INTRODUCTION

Colorectal cancer (CRC) also known as bowel cancer that affects colon and rectum part of the large intestine. CRC is one of the major causes of morbidity and mortality, representing the third majorly spotted malignancy and the fourth leading cause of cancer-allied deaths worldwide that is expected to increase the burden 60% by 2030 [1]. It is the second leading cause of cancer death in women, and the third for men [2, 3]. If considered worldwide, incidence appears to be highly variable among different countries and the trend is unexpectedly increasing in countries once considered being at lower risk. The disease is most common in the developed countries, however; CRC incidence and mortality rates are rising rapidly in low and middle income countries also [1]. Also, if considered geographically, the variations in cancer incidence and mortality rate are more evident in racial and ethnic minority populations in comparison with the white population [4]. On daily basis, the gastro-intestinal tract undergoes the key process for food digestion and nutrient absorption and intrudes damage to the intestinal epithelium [5]. The replenishment of lost cells is constantly supported by the leucine-rich repeat-containing G protein-coupled receptor 5 (Lgr5) [6]. Colorectal cancer initiates as a growth called a polyp often, and some polyps become cancer over time [7]. Most colorectal cancers are adenocarcinomas (cancers that begin in cells that make and release mucus and other fluids). Multiple factors are linked to the pathogenesis of CRC; some are related to the dietary effects (smoking, alcohol habit, a diet rich in fat, physical inactivity, and obesity) [8, 9], lifestyle, and others are linked with the genetic predisposition [10-12]. The trends in CRC incidence and mortality correlate and reflect the adoption of more western lifestyles. The pathogenesis of CRC varies according to the genetic or epigenetic changes that lead to alterations within the sequence and contribute to the transformation of healthy epithelium towards disease progression and in adverse condition leads to malignancy [13].

The proliferation and differentiation of intestinal epithelium are regulated by diverse signaling pathways including Wnt, BMP, EGF, and Notch pathways [14]. Among these, the Wnt pathway contributes as a primary driving force for intestinal cell proliferation and maintenance [15]. The over-activation of the pathway is a hallmark of CRC. The Wnt pathway is a key regulator of both early and the later stages of CRC progression. In the normal epithelium, it controls homeostasis of intestinal stem cells (ISCs) [16]. Recent advances in high-throughput

sequencing reveal many novel recurrent Wnt pathway mutations in addition to the well-characterized *APC* and *β-catenin* mutations in CRC.

DNA repair mechanisms have a pivotal role in repairing damages occurring to the intestinal epithelium and also in the regulation of disease progression [12,17,18]. Among all other repair mechanisms, mismatch repair (MMR) is found to be the main player in CRC development [19]. Diverse forms of signaling pathways regulate different interrelated cellular processes, but these pathways do not work alone but are interconnected in some ways regulating complex cell networks. The sets of processes that happen within the cell vary for every cellular milieu and produce distinct responses accordingly. The inherent complexity of cellular signaling pathways and their importance for a wide range of cellular functions necessitates its understanding in cancer cells that could be cognitive to the scientific community.

In this study, Wnt signaling pathway has been targeted for the quantitative simulation and motif identification study for determining putative biomarkers for CRC. Many studies have shown its role in maintaining the stability of the intestinal epithelium [20-23]. There are two forms of *Wnt* signaling pathways i.e. canonical and non-canonical pathways. It is well established that the canonical *Wnt* signaling plays key roles in physiology (by maintaining intestinal crypts) and in pathology (via mutation causing cancer). In the signaling process if *Wnt* is absent, in such case *β-catenin* is phosphorylated by *CK1* and *GSK3* in a complex followed by recruitment of the *TrCP* to the complex for ubiquitination and proteasomal degradation. However when *Wnt* is there, it first binds to the Frizzled receptors (*FZD*) and *LRP5/6*, destruction complex is then recruited to the membrane. Dishevelled (*DVL*) play critical role in this mechanism and transduces extracellular Wnt signals from receptors to downstream effectors initiating accumulation of nuclear *β-catenin*. *DVL* has the ability to bind to *FZD* receptor and *AXIN* protein respectively thus inhibiting destruction complex as soon as the *DVL-AXIN* interacts. Therefore, the complex bound to *β-catenin* is no longer degraded. This results in the accumulation of free *β-catenin* in the cytoplasm and its subsequent nuclear translocation. In the nucleus, *β-catenin* displaces the repressor Groucho from T-cell factor (*TCF*). This leads to formation of an active transcriptional complex (*β-catenin/TCF*), along with other co-activators, that leads to the expression of *Wnt* target genes (Figure 3.1). Subsequent cancer progression requires stepwise accumulation of other mutations, such as in *KRAS*, *PI3K*, *TGFβs*, *p53*, and *SMAD4*.

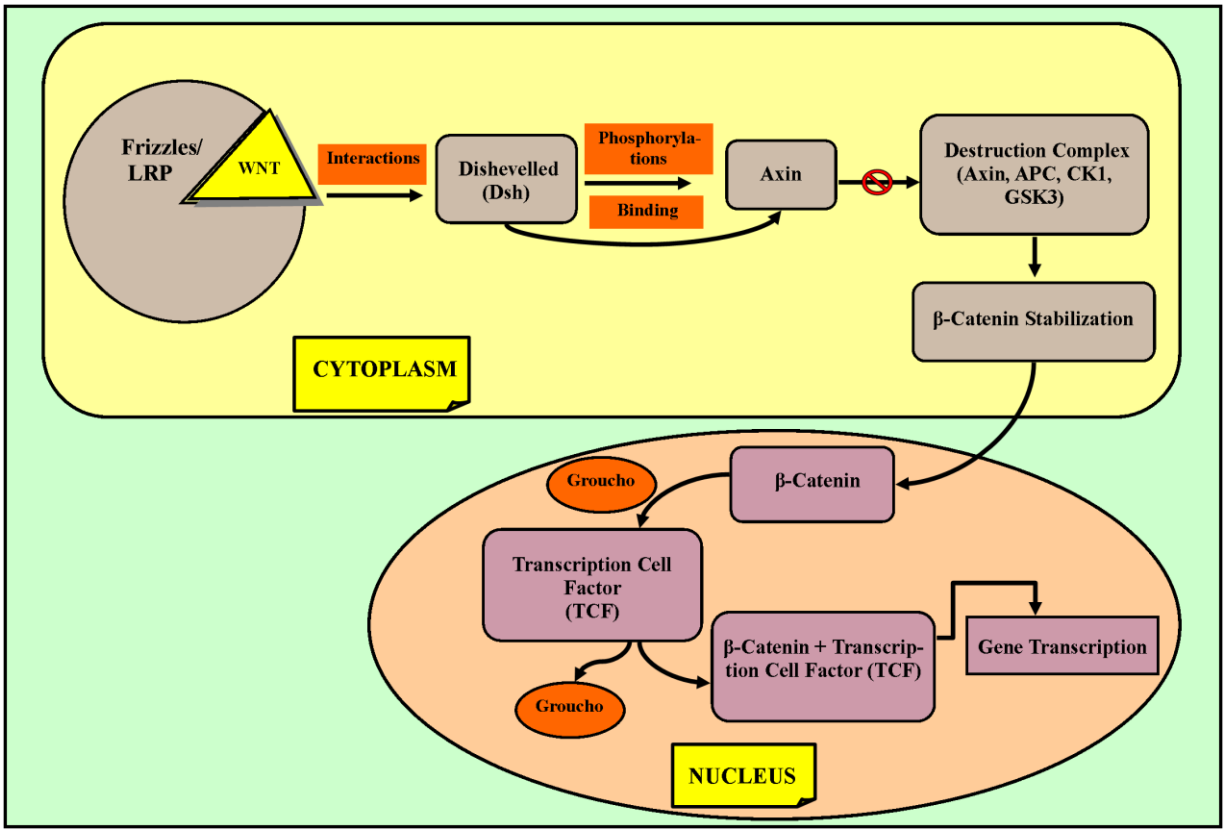


Figure 3.1 The pathway representing canonical Wnt signaling mechanism.

Here in behavior analysis has been conducted by performing quantitative simulations for a set of genes. To look at specific network level components, top down approach was used. The network motifs are a small set of recurring patterns that serve as basic building blocks of transcription networks [24]. These regulatory sub-network patterns play a key role in revealing explicit patterns in biological networks and thus provide significant insights for understanding complex biological processes [25].

The systems biology approaches developed in recent years provides novel and significant information relevant to research and biological applications. A novel integrative approach has been applied; wherein reductionist approach has been implemented for the elucidation of components involved. It is anticipated that the combined outcome of this study would provide biologically meaningful results and will be of utmost use to the researchers and biomedical scientists. This extensive *in silico* analysis is assumed to improve the diagnosis, treatment and other therapeutics for CRC.

3.2 MATERIALS AND METODS

The Wnt model was developed by considering the Sivakumar et al. as a reference model to understand the biological signaling processes [26]. The pathway analysis is performed using the MATLAB toolbox, SimBiology [27]. The program helps to model, simulate, and analyze the dynamic systems focusing principally on biological systems. The overall step-wise description of the method followed is given in Figure 3.2. The method comprises of two forms of analysis; one is based on pathway quantitative simulations and the other is for network motif determination to find the putative candidates.

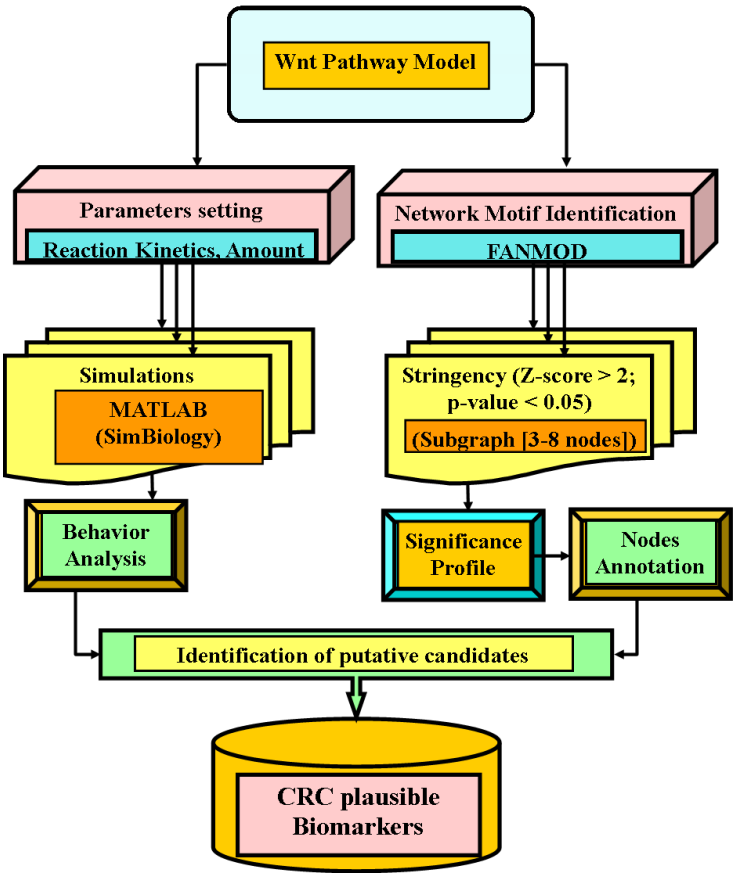


Figure 3.2 The flowchart of the methodology followed for pathway analysis.

3.2.1 Pathway-Based Quantitative Simulations

Here the pathway analysis has been performed to identify the dynamic behavior of the species (biological entities in systems biology terms) involved. The method followed the quantitative simulations study for the species by setting them at varied concentrations with respect to the time. This study focused on the impact of canonical *Wnt* signaling towards CRC, by considering individual component involved in the signaling process [28]. The simulation included reaction

kinetics for the type of reaction followed to form the product. The reaction kinetics states that the reaction mechanism takes place at a rate proportional to the product of its reactants. The in-built Ordinary Differential Equations (ODE) solver was used to perform the simulations for all these reactions and their respective parameters. The reaction mechanisms are involved in the form of ODEs and their mathematical representation are given as follows:

$$Dsh_i' = -k1Dsh_i + k2Dsh_a \quad (0.1)$$

$$Dsh_a' = k1Dsh_i - k2Dsh_a \quad (0.2)$$

$$Dest\ Complex_a' = k3Dest\ Complex_i - k4Dest\ Complex_a - k10\beta - catenin\ .\ Dest\ Complex_a + k11 (Complex)\beta - catenin\ .\ Dest\ Complex + k13(Complex)\beta - catenin\ .\ Dest\ Complex(proteasomal\ degradation) \quad (0.3)$$

$$Dest\ Complex_i' = k6GSK3.(Complex)Axin\ .\ APC - k5Dsh_a\ .Dest\ Complexi - k3Dest\ Complex_i + k4 Dest\ Complex_a - k7Dest - Complex_i \quad (0.4)$$

$$GSK3' = k5Dsh_a.\ Dest\ Complex_i - k6GSK3.(Complex)Axin\ .\ APC + k7Dest\ Complex_i \quad (0.5)$$

$$(Complex)Axin\ .\ APC' = k5Dsha.\ Dest\ Complex_i - k6GSK3.(Complex)Axin\ .\ APC + k7Dest\ Complex_i + k8Axin\ .\ APC - k9(Complex)Axin\ .\ APC \quad (0.6)$$

$$APC' = -k8Axin\ .\ APC + k9(Complex)Axin\ .\ APC - k21\beta\ catenin\ .\ APC + k22(Complex)\beta - catenin\ .\ APC \quad (0.7)$$

$$(Complex)\beta\ catenin\ .\ Dest\ Complex' = k10\beta\ catenin\ .\ Dest\ Complex_a - k11(Complex)\beta - catenin\ .\ Dest\ Complex - k12(Complex)\beta - catenin\ .\ Dest\ Complex \quad (0.8)$$

$$(Complex)\beta\ catenin\ .\ Dest\ Complex(proteasomal\ degradation)' = k12(Complex)\beta\ catenin\ .\ Dest\ Complex - k13(Complex)\beta\ catenin\ .\ Dest\ Complex(proteasomal\ degradation) \quad (0.9)$$

$$\beta\ catenin(proteasomal\ degradation)' = k13(Complex)\beta\ catenin\ .\ Dest\ Complex\ (proteasomal\ degradation) - k14\beta\ catenin(proteasomal\ degradation) \quad (0.10)$$

$$\beta\ catenin' = -k10\beta\ catenin\ .\ Dest\ Complex_a + k11(Complex)\beta\ catenin\ .\ Dest\ Complex + k15 - k16\beta\ catenin - k19\beta\ catenin\ .\ TCF + k20(Complex)\beta\ catenin\ .\ TCF - k21\beta\ catenin\ .\ APC + k22(Complex)\beta\ catenin\ .\ APC \quad (0.11)$$

$$Axin' = -k8Axin\ .\ APC + k9(Complex)Axin\ .\ APC + k17 - k18Axin \quad (0.12)$$

$$TCF' = -k19\beta\ catenin\ .\ TCF + k20\left(Complex\right)\beta\ catenin\ .\ TCF \quad (0.13)$$

$$(Complex)\beta\ catenin\ .\ TCF' = k19\beta\ catenin\ .\ TCF - k20(Complex)\beta\ catenin\ .\ TCF \quad (0.14)$$

$$\left(Complex\right)\beta\ catenin\ .\ APC' = k21\beta\ catenin\ .\ APC - k22\left(Complex\right)\beta\ catenin\ .\ APC \quad (0.15)$$

In the reaction kinetics shown above primes (') denote the differentiation with respect to time and k_n (n = 1, 2,....., 22) signifies the rate constants. The entities specifying the reaction kinetics thus

help to study the dynamic behavior of the biochemical pathway and provide inference for the crucial components need to be considered.

3.2.2 Network Motifs

The network motifs were determined through FANMOD, the statistical parameters with z-value > 2 and p-value < 0.05 were considered to generate sub-graphs [29]. This result in the sub-graphs of 3-8 nodes size (Appendix: Table 2). Also, significance profile (SP) has been computed to assess the statistical inference of the motifs generated through the tool. SP generates normalized *z*-score values for individual network motif. SP in general given as:

$$SP\left(m_i\right) = \frac{Z\left(m_i\right)}{\sqrt{\sum_{i=1}^{n} Z\left(m_i\right)^2}}$$

Where m_i represents network motif and *Z(mi)* the *Z-score* for individual network motif. This method thus provides network motifs having statistical significance (Appendix: Table 2). The logarithmic conversion has been applied for the network motifs and those with frequent occurrence have been annotated for their specific role.

3.3 RESULTS AND DISCUSSION

3.3.1 Pathway Analysis

The pathway analysis provides information regarding the crucial components involved in the biological networks. The signaling processes are the key regulatory mechanisms that maintain the cell integrity. In this study, simulations were performed for all the entities by changing the parameters for every set of experiment. Ode45 (Dormand-Prince) solver is used to run the simulations; the solver provides an explicit method to solve set of ODEs with minimum error rate. It is therefore considered highly suitable for the high-order reactions and integration process [30]. Different parameters were implicated depending on various standards provided in the literature (Appendix: Table 1).

For initial simulation, end-time was set as 3 milliseconds with an absolute tolerance 1.0E-6 and relative tolerance is set to 0.001. The entire set of components in the model was considered for initial set. The concentrations of the model were obtained via the SBMLsqueezer; a

CellDesigner plugin that generates reaction kinetics for the biochemical network [31]. In each reaction mechanism, the kinetic equation is derived from its stoichiometry, regulatory mechanisms and the involved species (such as simple molecules, proteins etc.). The rate laws are derived by considering individual reaction for every set of participating substrates, products, and the regulators.

Various studies to date have been conducted for the Wnt signaling mechanism. In a study conducted by Lee et al., the role of *AXIN* and *APC* in the formation of degradation complexes has been envisaged along with fidelity of *AXIN* degradation on *APC* [32]. The Kruger et al. considered the vigorousness of the signaling pathway with respect to the parameter fluctuations [33]. Cho et al. explained the preference of *APC* mutations on the CRC study along with the effect of *APC, AXIN, and β-catenin* mutations [34]. In a study conducted by Van Leeuwen et al., the results showed that amount of *APC* is a deciding factor towards the fate of Wnt signaling for carrying normal or diseased phenotype [35]. In a Goldbeter et al. study, Wnt pathway has exerted the negative feedback mechanism on axin through the formation of the destruction complex due to the *β-catenin* degradation that leads to the oscillatory effects [36]. Mirams et al. represented crucial details of the pathway at distinct timescales; with the fastest timestamp representing action of the destruction complex for the *β-catenin*; the intermediate timescale with the impressions to regulate the level of destruction complex with axin removal via influence of the *Wnt* and *Dishevelled*; and the slowest timescale with alterations in the *β-catenin* level [37].

The above-mentioned studies analyzed effect of the individual component, though the study is being planned and performed to capture the effect of the key components to the various complexes in the pathway by considering them at varying concentrations in different environments (nucleus or, cytoplasm). The analysis conducted thus provides a basis for understanding the behavioral dynamics of genes and proteins while performing the quantitative perturbation. The quantitative approach used, could be fruitful in understanding behavioral dynamics and thus proved to be useful for the researchers working in this field. The simulations preformed via considering sets of entities and rate kinetics is integrated into Appendix: Figure 1. Behavior of all set of entities was taken care and only some of the entities were selected based upon their dynamical behavior concerning overall pathway. Therefore, separate simulations were run for some specific set of entities depending upon their individual or cumulative contributions. The study illustrates the activation and suppression phenomenon of the proteins captured at

different time-frames thus inferring their role in carcinogenesis. The work presented a basis for understanding the dynamics involved for the key regulatory genes (specifically tumor suppressor, and oncogene), thus will be helpful to develop pharmacological strategies to regulate the pathway through therapeutic interventions.

Table 3.1 Quantitative parameters for all the analyzed species in simulation studies with varied concentrations

Type	Name	Location	Initial Concentrations (μM)	Solver
Species	β-catenin	Plasma Membrane	5	Ode45 (Dormand-Prince)
Complex	APC, β-catenin, GSK3β, Axin, PP2A, Diversin, CK1	Plasma Membrane	5	
Complex	APC, Axin, PP2A, Diversin, CK1, β-catenin, β- TrCP, GSK3β	Plasma Membrane	0	
Complex	APC, Axin, PP2A, Diversin, CK1, β-catenin, β- TrCP, β- TrCP, GSK3β	Plasma Membrane	5	
Complex	Complex_br_(Wnt/Frizzled)	Plasma Membrane	0	Ode45 (Dormand-Prince)
Species	β-catenin	Plasma-Membrane	5	
Species	β-catenin	Nucleus	0	
Complex	TCF, Smad4, β-catenin	Nucleus	0	
Species	β-catenin	Nucleus	0.53	
Complex	APC, Axin, PP2A, Diversin, CK1, β-catenin, β- TrCP, GSK3β	Nucleus	0	
Species	β-catenin	Nucleus	5	
Complex	APC, Axin, , PP2A, Diversin, CK1, β-catenin, β- TrCP, GSK3β	Nucleus	5	
Species	β-catenin	Nucleus	6	

For better understanding, the signaling processes interrelated components were considered so as to capture the effect of one over another. The overall signaling process revolves around the *β-catenin* level and the destruction complex. The check is made regarding the overproduction of the *β-catenin* that in case not controlled, initiate the transcription of the targeted genes hence lead its way towards cancer progression. The simulations have been performed at different time period by considering different entities; that uncovers some interesting behavior that was not targeted in the studies performed earlier. The simulation was run by taking into account the *β-catenin* (plasma membrane) with concentration of 5 μM, the complex (*APC, β-catenin, GSK3β,*

Axin, PP2A, Diversin, CK1) with 5 μM concentration and the complex (*APC, Axin, PP2A, Diversin, CK1, β-catenin, β-TrCP, GSK3β*) considering to be at 0 μM initially (Table 3.1).

The results have shown that even though there is high amount of *β-catenin* (a subunit of the cadherin protein complex and has a role in the regulation and coordination of cell-cell adhesion and gene transcription) but as long as it is captured by the destruction complex, *β-catenin* slows down its activity and its concentrations diminishes from its initial amount of 5 μM to 0.2 μM. This illustrates the fact even with the high concentration *β-catenin* will not work until the destruction complex is there and it will regulate the level of *β-catenin* irrespective of its own higher or lower concentrations as shown in the graph at a time it is having a concentration of 5 μM and 0 μM (Figure 3.3). Complex-1 and *β-catenin* have same behavior while complex-2 is showing the opposite behavior (Figure 3.3a). However, all three reaching at steady state at the almost same time and this reflects that initial few milliseconds (i.e. 3-4 milliseconds) are crucial for their activity to finally reach and follow the steady state.

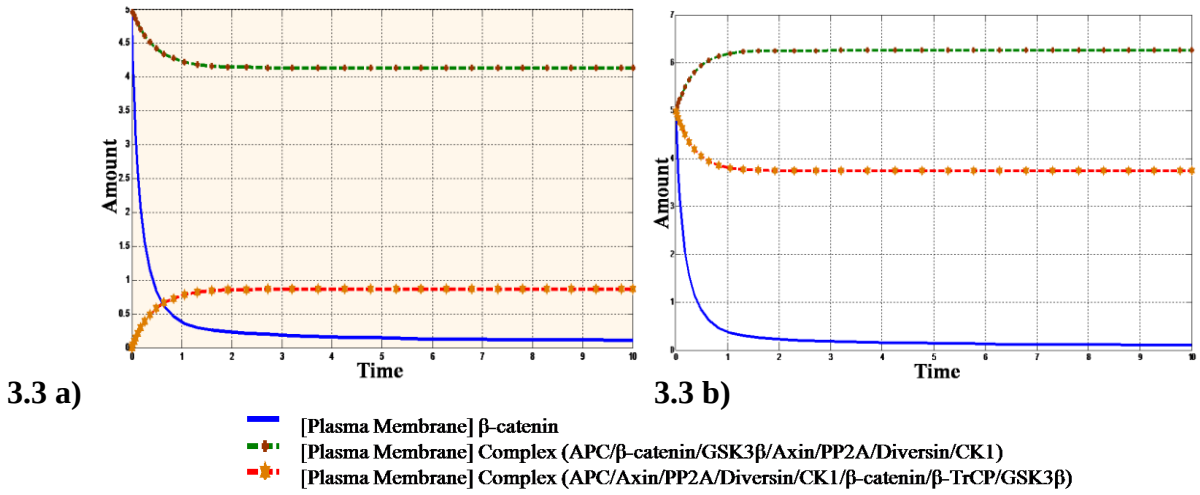
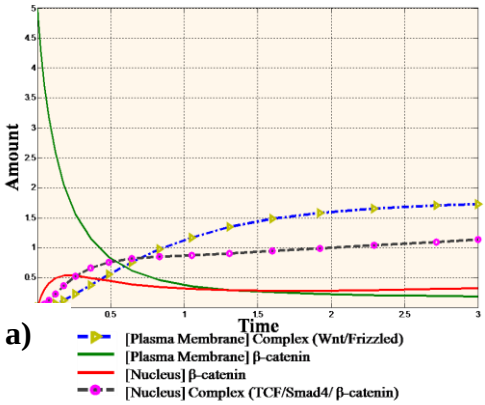


Figure 3.3 Dynamic behavior of components in different compartments. **a.)** Behavior of the β-catenin (plasma membrane), Complex (*APC, β-catenin, GSK3β, Axin, PP2A, Diversin, CK1*), and Complex (*APC, Axin, PP2A, Diversin, CK1, β-catenin, β-TrCP, GSK3β*) at varied amount. **b.)** Behavior of the *β-catenin* (plasma membrane), Complex (*APC, β-catenin, GSK3β, Axin, PP2A, Diversin, CK1*), and Complex (*APC, Axin, PP2A, Diversin, CK1, β-catenin, β-TrCP, GSK3β*) at equal amount. Concentration (Amount) in μM, Time in milliseconds.

The next iteration was performed by considering the same set of species but with equal concentrations (Figure 3.3b); for this an interesting behavior has been noticed as both the

complexes are overruling and thus not allowing the *β-catenin* to overcome the signaling process, this states that the amount of these two complexes are crucial for controlling the *β-catenin* level. The above two simulations were indicative of the initial activity at 3-4 milliseconds, therefore to look for this effect next iterations were performed for 3 milliseconds by considering the species (Complex (*Wnt/Frizzled*), *β-catenin*, and Complex (*TCF, Smad4, β-catenin*)) (Table 3.1). In the simulation (Figure 3.4a), it's been found that once the *Wnt* bind to the *FZD* receptor the signaling initiate and the *β-catenin* level in the plasma membrane decrease and then rise up to certain extent (0.2 milliseconds) in the nucleus. It leads to the transcription when binds to the *TCF* via tumor genes activation that leads a step towards the cancer progression. The next iteration was performed by considering the species (*β-catenin*, and *Complex (APC, Axin, PP2A, Diversin, CK1, β-catenin, β-TrCP, GSK3β*)). Through the simulation, it was noticed that the complex has strong efficiency to suppress the level of *β-catenin* in the nucleus and thus overcome the negative stimulation of *β-catenin* (Figure 3.4b).

3.4 a)



3.4 b)

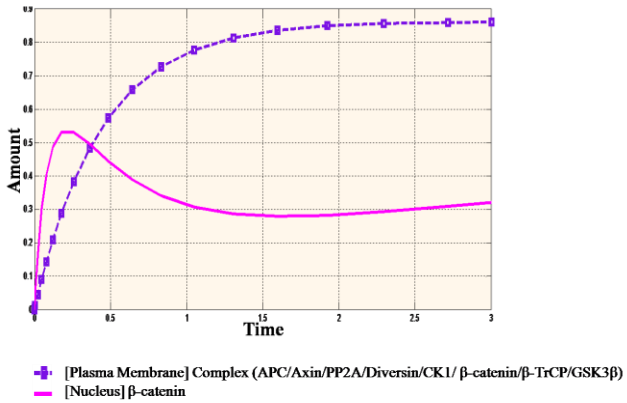


Figure 3.4 Dynamic behavior of components in different compartments. **a.)** Behavior of the Complex (*Wnt/Frizzled*), *β-catenin*, and Complex (*TCF, Smad4, β-catenin*) and **b.)** Behavior of the complex (i.e. *β-catenin*, and Complex (*APC, Axin, PP2A, Diversin, CK 1, β-catenin, β-TrCP, GSK3β*)) in plasma membrane vs. the *β-catenin* in nucleus. Concentration (Amount) in μM, Time in milliseconds.

The next iteration was performed by considering the species (*β-catenin*, and Complex (*APC, Axin, PP2A, Diversin, CK1, β-catenin, β-TrCP, GSK3β*)) but at varying concentrations. Two different simulations were run; Figure 3.5a is showing the overruling behavior of the complex over the *β-catenin*., however when the concentration of the *β-catenin* is set high by one-unit, there is high possibility towards cancer progression (Figure 3.5b). Along with this many other simulations were performed with varying concentrations and behavior of these regulatory

elements were found to be similar in almost all cases (results not shown). The detailed information regarding the parameters for pathway entities might further improve the analysis and investigations towards the pathway dynamics. With the informed parameter guesses, simulation studies could have improved to understand the system's behavior as getting knowledge related to the dynamics of the individual pathway entities is itself a cumbersome process. The powerful analytic tools highlight the successful study of the biological processes that occurs *in-vivo*. However, such success is possible through the inexorable endeavors and in-depth understanding of both computational methods and the biological problems of interest.

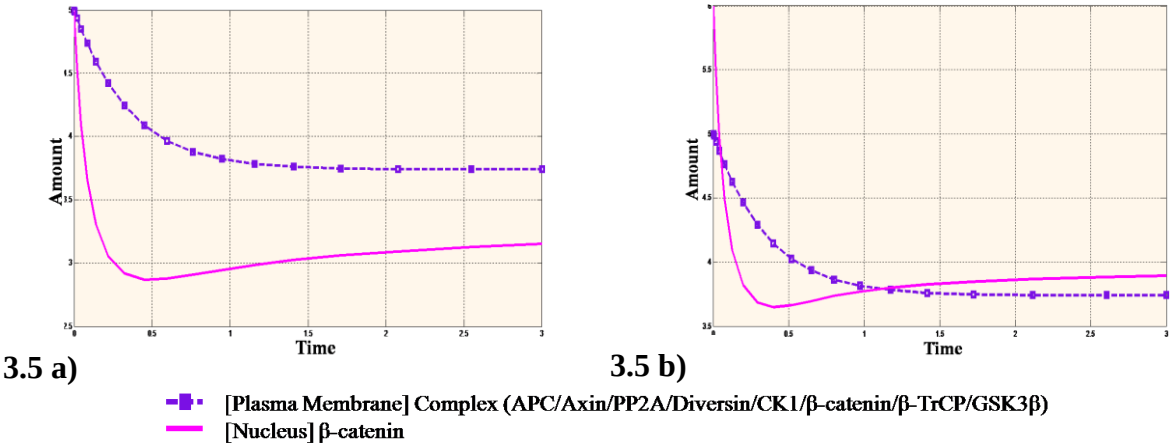


Figure 3.5 Dynamic behavior of components in different compartments. **a.)** Behavior of the Complex (*APC, Axin, PP2A, Diversin, CK1, β-catenin, β-TrCP, GSK3β*) and *β-catenin*) at equal amount. **b.)** Behavior of the Complex (*APC, Axin, PP2A, Diversin, CK1, β-catenin, β-TrCP, GSK3β)* and *β-catenin*) at alternative amount. Concentration (Amount) in μM, Time in milliseconds.

3.3.2 Network Motif Detection

The network motifs are sub-graphs that occur frequently in a specific network or even among various networks. These sub-graphs have specific interaction pattern that reflects a framework to achieve particular functional efficiently. For subgraph network detection, the network motifs were generated using statistical standards (with $Z > 2$; p-value < 0.05) that produced in total 595 sub-graphs having 3-8 nodes (Appendix: Table 2). SP has been calculated for normalizing the Z-score to filter out the sub-graphs of high statistical significance. This procedure reduced the sub-graphs from 595 to 64 only (Appendix: Table 2). The network motifs generated had 4-chain motifs, single input module (SIM), multiple input module (MIM), bifan motifs etc. supported by

significant *Z*-scores and *p*-values. Other regular 4-node motifs confirmed the presence of diamond, biparallel and bifan motifs (often built by two regulatory and two regulated genes). The graph has been plotted on a logarithmic scale for the respective significance profile generated corresponding to the individual sub-graph type (Figure 3.6). From the significance profile, five subgraphs (3a, 4a, 5a, 6a, and 7a) were found to be overrepresented. The motif significance profiles illustrate that there is high difficulty in recognizing the genes once the network start intensifying. Therefore, the overrepresented sub-graphs were annotated to determine the interacting gene partners. However there were multiple instances of the particular sub-graph type so the ones with utmost frequency were selected, this signifies the importance of three genes in the pathway (Figure 3.7). Overall five key genes were found i.e. *Axin*, *APC*, *β-catenin*, *LEF1*, and *Myc* with high statistical significance. Some additional significant components that could be important for *in vitro* and *in vivo* studies are *MMP7*, *NLK* and *DVL1*. The interaction of *MMP7* with *LEF1* as represented by multiple motifs indicates a close association of these entities. *NLK* and *DVL1* could be crucial for regulating the biological processes in the Wnt pathway due to their role as potent inhibitors through *LEF1* and *GSK3β* that negatively phosphorylate *β-catenin* (Appendix: Table 3).

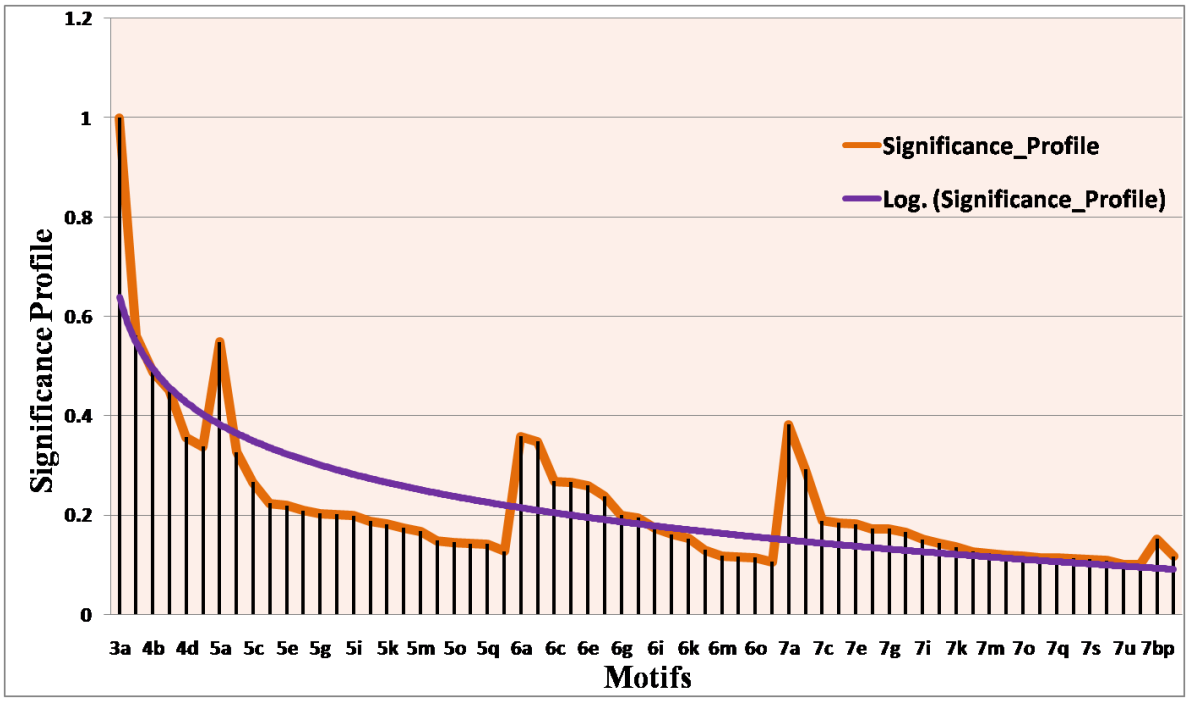


Figure 3.6 The Significance profile (SP) of 3-7 nodes sub-graphs.

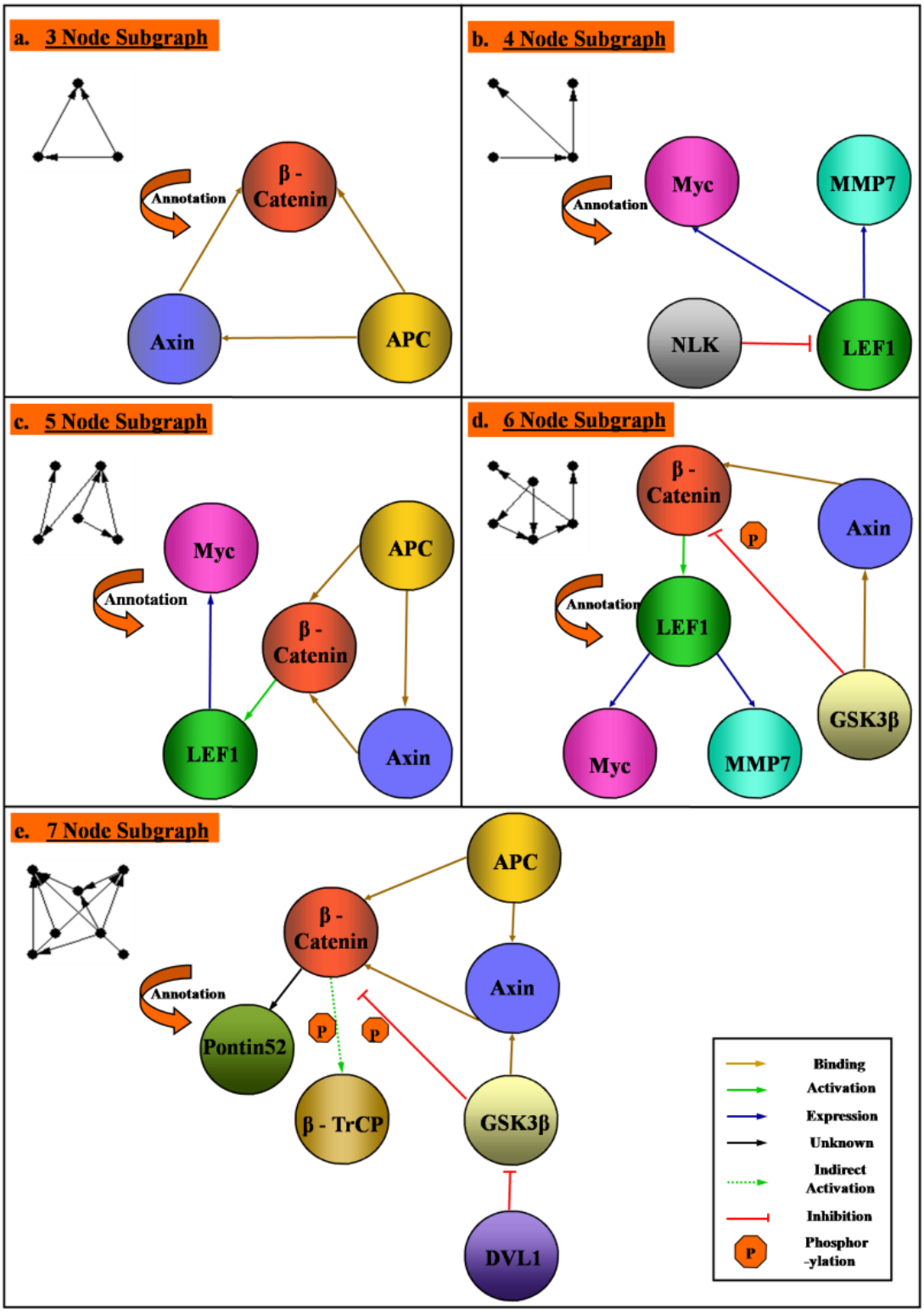


Figure 3.7 The Sub-graph annotation of overrepresented ones with vital interactions.

The study followed a combination of top-down and bottom-up approaches to uncover the key interacting partners that were not targeted yet for their abrupt regulation via Wnt signaling in CRC. The network motifs provide a way to understand the complex signaling as it divides the large complex network into the smaller sub-networks that can be evaluated through statistical means to determine its significance in biological processes. The behavior analysis entails dynamic activities of the individual components that provide plausible biomarkers through the bottom-up approach. The exemplified *in silico* approach could be applied to any other diseases at the pathway level for identifying biomarkers that will save the time of experimental biologists and will help them focus on the crucial components of the pathway for an early detection of the disease condition.

3.4 CONCLUSION

The systematic exploration and modeling of the biological networks serve both fundamental research as well as the industrial application. Technologies have revolutionized the ways via generation of huge amount of data the *"BIG DATA"* that need to be visualized with extreme care as even the best data become useless if not presented clearly. Numerous computational models have been developed to address different categories of biological processes, such as metabolic processes or signaling and regulatory pathways. Today, modeling approaches are vital for biologists working in the field of systems biology, enabling them to analyze complex physiological processes, and also for the pharmaceutical industry as a means for supporting drug discovery and development. In this work, pathway simulations have been done along with network motif determination by considering all possible targets of the *Wnt* pathway for the disease progression study. The study illustrates the effect of the *β-catenin* and its regulation via different complexes that include destruction complex which captures *β-catenin* in the cytoplasm and prevent its transcription, and another one in the nucleus where it is bound to the transcription cell factor (*TrCP*) and made its progress towards CRC. The five key genes i.e. *Axin, APC, β-catenin, LEF1,* and *Myc* were detected to be putative regulatory elements of the pathway. Also, *MMP7, NLK,* and *DVL1*, are found to be the essential component of the study that can be possible putative elements for the disease progression study, thus should be considered for the experimental validation. In this chapter, an attempt has been made to evaluate the role of *Wnt* signaling pathway involved in colorectal cancer progression via target genes. The

overrepresented patterns play a significant role in the biological processes; the subgraph study helped to identify closely interacting partners although their role in CRC is still unknown. These interacting patterns could be helpful to unravel complex threads of the disease condition. Despite the advent strategies in drug development for *Wnt* pathway, the major hurdles in therapeutic intervention of the pathway still persist. With the aid of computerized modeling efforts, researchers can make sense of all the different bits and pieces of information that are accumulating at an ever-increasing pace. This will help the experimental biologists to understand the workings of living organisms with comprehensive models that enable them to connect often disparate pieces of information and to look for all possible applications for the betterment of living systems.

REFERENCES

[1] M. Arnold, M. S. Sierra, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global patterns and trends in colorectal cancer incidence and mortality," *Gut*, pp. gutjnl-2015-310912, 2016.

[2] K. Tariq and K. Ghias, "Colorectal cancer carcinogenesis: a review of mechanisms," *Cancer Biol Med*, vol. 13, pp. 120-35, Mar 2016.

[3] I. M. Hisamuddin and V. W. Yang, "Molecular genetics of colorectal cancer: an overview," *Current colorectal cancer reports*, vol. 2, pp. 53-59, 2006.

[4] C. S. Jackson, M. Oman, A. M. Patel, and K. J. Vega, "Health disparities in colorectal cancer among racial and ethnic minorities in the United States," *Journal of gastrointestinal oncology*, vol. 7, p. S32, 2016.

[5] C. D. Davis and J. A. Milner, "Gastrointestinal microflora, food components and colon cancer prevention," *The Journal of nutritional biochemistry*, vol. 20, pp. 743-752, 2009.

[6] L. Novellasdemunt, P. Antas, and V. S. Li, "Targeting Wnt signaling in colorectal cancer. A Review in the Theme: Cell Signaling: Proteins, Pathways and Mechanisms," *Am J Physiol Cell Physiol*, vol. 309, pp. C511-21, Oct 15 2015.

[7] W. M. Grady and S. D. Markowitz, "The molecular pathogenesis of colorectal cancer and its potential application to colorectal cancer screening," *Digestive diseases and sciences*, vol. 60, pp. 762-772, 2015.

[8] F. A. Haggar and R. P. Boushey, "Colorectal cancer epidemiology: incidence, mortality, survival, and risk factors," *Clin Colon Rectal Surg*, vol. 22, pp. 191-7, Nov 2009.

[9] W. M. Grady and S. D. Markowitz, "Genetic and epigenetic alterations in colon cancer," *Annual review of genomics and human genetics*, vol. 3, pp. 101-128, 2002.

[10] G. S. Martin, "Cell signaling and cancer," *Cancer cell*, vol. 4, pp. 167-174, 2003.

[11] S. K. Li and A. Martin, "Mismatch Repair and Colon Cancer: Mechanisms and Therapies Explored," *Trends in molecular medicine*, vol. 22, pp. 274-289, 2016.

[12] S. D. Markowitz and M. M. Bertagnolli, "Molecular origins of cancer: Molecular basis of colorectal cancer," *N Engl J Med*, vol. 361, pp. 2449-60, Dec 17 2009.

[13] D. Colussi, G. Brandi, F. Bazzoli, and L. Ricciardiello, "Molecular pathways involved in colorectal cancer: implications for disease behavior and prevention," *Int J Mol Sci*, vol. 14, pp. 16365-85, Aug 7 2013.

[14] S. J. Cohen, R. B. Cohen, and N. J. Meropol, "Targeting signal transduction pathways in colorectal cancer—more than skin deep," *Journal of Clinical Oncology*, vol. 23, pp. 5374-5385, 2005.

[15] T. Zhan, N. Rindtorff, and M. Boutros, "Wnt signaling in cancer," *Oncogene*, 2016.

[16] K. Nalapareddy, K. J. Nattamai, R. S. Kumar, R. Karns, K. A. Wikenheiser-Brokamp, L. L. Sampson, *et al.*, "Canonical Wnt Signaling Ameliorates Aging of Intestinal Stem Cells," *Cell Reports*, vol. 18, pp. 2608-2621, 2017.

[17] N. S. Gavande, P. S. VanderVere-Carozza, H. D. Hinshaw, S. I. Jalal, C. R. Sears, K. S. Pawelczak, *et al.*, "DNA repair targeted therapy: The past or future of cancer treatment?," *Pharmacol Ther*, vol. 160, pp. 65-83, Apr 2016.

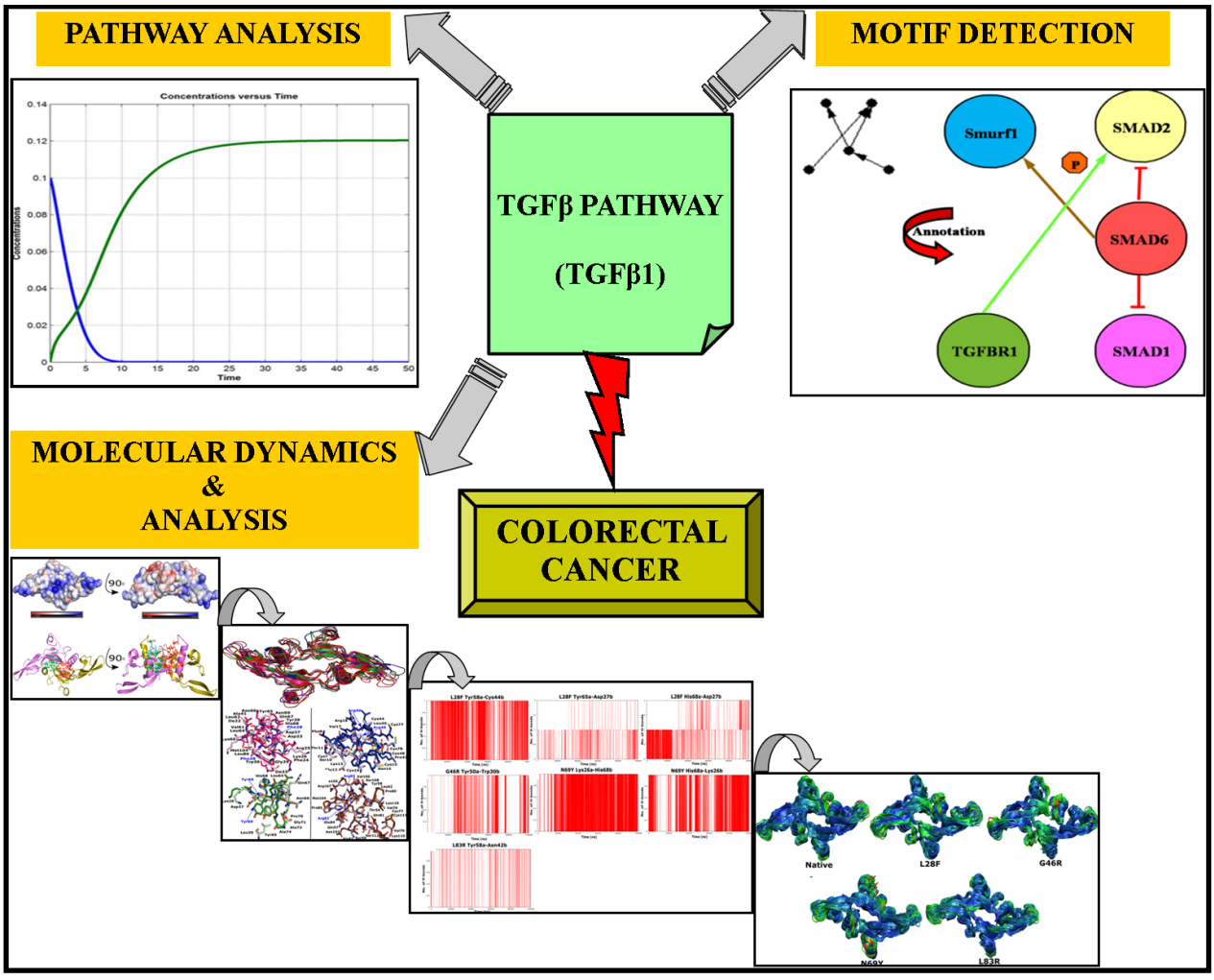
[18] F. Dietlein, L. Thelen, and H. C. Reinhardt, "Cancer-specific defects in DNA repair pathways as targets for personalized therapeutic approaches," *Trends in Genetics*, vol. 30, pp. 326-339, 2014.

[19] P. Peltomaki, "Deficient DNA mismatch repair: a common etiologic factor for colon cancer," *Hum Mol Genet,* vol. 10, pp. 735-40, Apr 2001.
[20] J. Schneikert and J. Behrens, "The canonical Wnt signalling pathway and its APC partner in colon cancer development," *Gut,* vol. 56, pp. 417-25, Mar 2007.
[21] A. T. Mah, K. S. Yan, and C. J. Kuo, "Wnt pathway regulation of intestinal stem cells," *J Physiol,* vol. 594, pp. 4837-47, Sep 1 2016.
[22] W. Y. Zou, S. E. Blutt, X. L. Zeng, M. S. Chen, Y. H. Lo, D. Castillo-Azofeifa*, et al.*, "Epithelial WNT Ligands Are Essential Drivers of Intestinal Stem Cell Activation," *Cell Rep,* vol. 22, pp. 1003-1015, Jan 23 2018.
[23] H. Du, Q. Nie, and W. R. Holmes, "The interplay between Wnt mediated expansion and negative regulation of growth promotes robust intestinal crypt structure and homeostasis," *PLoS computational biology,* vol. 11, p. e1004285, 2015.
[24] U. Alon, "Network motifs: theory and experimental approaches," *Nat Rev Genet,* vol. 8, pp. 450-61, Jun 2007.
[25] A.-L. Barabási, N. Gulbahce, and J. Loscalzo, "Network medicine: a network-based approach to human disease," *Nature Reviews Genetics,* vol. 12, pp. 56-68, 2011.
[26] K. C. Sivakumar, S. B. Dhanesh, S. Shobana, J. James, and S. Mundayoor, "A systems biology approach to model neural stem cell regulation by notch, shh, wnt, and EGF signaling pathways," *Omics: a journal of integrative biology,* vol. 15, pp. 729-737, 2011.
[27] T. MathWorks, "MATLAB and Statistics Toolbox Release 2012a,", *The MathWorks, Inc., Natick, Massachusetts, United States.*, 2012.
[28] A. L. MacLean, H. A. Harrington, M. P. Stumpf, and H. M. Byrne, "Mathematical and statistical techniques for systems medicine: The Wnt signaling pathway as a case study," *Systems Medicine,* pp. 405-439, 2016.
[29] S. Wernicke and F. Rasche, "FANMOD: a tool for fast network motif detection," *Bioinformatics,* vol. 22, pp. 1152-1153, 2006.
[30] J. R. Dormand and P. J. Prince, "A family of embedded Runge-Kutta formulae," *Journal of computational and applied mathematics,* vol. 6, pp. 19-26, 1980.
[31] A. Drager, N. Hassis, J. Supper, A. Schroder, and A. Zell, "SBMLsqueezer: a CellDesigner plug-in to generate kinetic rate equations for biochemical networks," *BMC Syst Biol,* vol. 2, p. 39, Apr 30 2008.
[32] E. Lee, A. Salic, R. Kruger, R. Heinrich, and M. W. Kirschner, "The roles of APC and Axin derived from experimental and theoretical analysis of the Wnt pathway," *PLoS Biol,* vol. 1, p. E10, Oct 2003.
[33] R. Kruger and R. Heinrich, "Model reduction and analysis of robustness for the Wnt/beta-catenin signal transduction pathway," *Genome Inform,* vol. 15, pp. 138-48, 2004.
[34] K. H. Cho, S. Baek, and M. H. Sung, "Wnt pathway mutations selected by optimal beta-catenin signaling for tumorigenesis," *FEBS Lett,* vol. 580, pp. 3665-70, Jun 26 2006.
[35] I. M. van Leeuwen, H. M. Byrne, O. E. Jensen, and J. R. King, "Elucidating the interactions between the adhesive and transcriptional functions of β-catenin in normal and cancerous cells," *Journal of theoretical biology,* vol. 247, pp. 77-102, 2007.
[36] A. Goldbeter and O. Pourquié, "Modeling the segmentation clock as a network of coupled oscillations in the Notch, Wnt and FGF signaling pathways," *Journal of theoretical biology,* vol. 252, pp. 574-585, 2008.

[37] G. R. Mirams, H. M. Byrne, and J. R. King, "A multiple timescale analysis of a mathematical model of the Wnt/β-catenin signalling pathway," *Journal of mathematical biology,* vol. 60, pp. 131-160, 2010.

CHAPTER - 4

Network and Structure Based Study of Functional Single Nucleotide Polymorphisms of TGFβ1 Gene and its Role in CRC



~ The only way to do great work is to love what you do.
- Steve Jobs

ABSTRACT

Cell signaling administers vital events and coordinates multiple actions like development, repair, immunity-check, and homeostasis. The fault in signaling leads to inaccurate cellular processing that increases the possibility of disease progressions such as autoimmunity disorders, diabetes, and cancer. The methods established for large-scale quantitative analysis of genome provides systems-level insights for the signal transduction processes such as rate of signal transmission, propagation, regulation, and signaling specificity for cellular differentiation. There are many cellular signaling pathways such as *Wnt/β-catenin* (as discussed in previous objective), *TGF-β/Smad, MAPK, JAKs/STAT3*, *VEGF*, *Notch*, *NF-κB, COX,* and *p53* that leads to colorectal cancer (CRC) due to disproportionate signaling. Out of these signaling cascades *Wnt* and *TGFβ* have been shown to have a cross regulatory role and that is why the main emphasis has been given to these two pathways only. High-throughput technologies that have been devised so far for genome and proteome analysis though have guided a lot but not provided an inclusive concern to infer the disease diagnosis. This has lead to the requirement of an approach that offers a complete solution by considering the whole system. The system-level understanding furnishes an inclusive aspect of the interactions among the system components thereby inferring the key outcome of the activity. These functional processes include gene expression, protein-protein interactions, transcriptional regulations, post-translational modifications etc. The *TGFβ*-signaling is thought to be one of the important signal transduction mechanisms in the CRC. Although many studies have focused on *TGFβ* pathway, most of them targeted *TGFβ* receptors and very few have taken care of *TGFβ1* ligand. In this chapter, network as well as the structure-based analysis has been done to decipher the mechanism of *TGFβ1* gene in CRC conditions. The network-based study involved the pathway level and network motif level analysis to determine the key regulatory genes in the disease pathway. The structure-based study involved the analysis of the four ns-SNPs rs199946261, rs768250306, rs763943753, and rs541829714 that were detected to be highly damaged through the computational tools. These ns-SNPs were found to highly affect the structure at functional level thus hindering the activity of the *TGFβ1* protein.

4.1 INTRODUCTION

Cancer has a major vulnerability among the other types of diseases that are affecting the people in almost all parts of the globe. Cancer has a tendency to affect almost all parts of the human body that is why all types of cancers are prevalent these days depending on the patient's exposure to diverse factors. Amongst all forms of cancer the breast cancer, lung cancer, prostate cancer and the colorectal cancer (CRC) are the topmost malignancies these days [1]. Although the life expectancy has increased over the time for the cancer patients throughout the world still the malignancy is not under control. It is one of the most prevalent and deadly forms of cancer states. If counted it is third most detected malignancy that causes the death of the people worldwide. By 2030 the risk has been expected to increase by 2.2 million leading to the death of half of the estimated population [2]. Though men and women equally likely to die from this cancer but, CRC has been found to be more prevalent in men than in women [3]. The developed countries have been found to be more exposed to the condition in comparison to the ones that are underdeveloped [4, 5]. Many factors are involved with the CRC carcinogenesis, one of them is an adaptation to bad food habits and other cause can be an inherited genetic makeup for CRC [3, 4, 6]. Majority of CRC build up sporadically, and the remaining cases develop in the form of hereditary syndromes, mainly familial adenomatous polyposis coli (FAP) and hereditary nonpolyposis colon cancer (HNPCC or Lynch syndrome) [7].

The primary role of DNA damage has become evident in cancer development due to the genetic defects in DNA repair systems. The damage to the DNA leads to cancer development via erroneous repair mechanism that causes mutations at the chromosomal level either by oncogenes activation or tumor suppressor genes inactivation [8]. DNA replication fidelity majorly relies on the nucleotide selection, polymerase activity, proofreading, and mismatch repair (MMR) mechanism [9]. The replication errors that evade the proofreading functions are corrected by MMR [10]. In case of mutations in the MMR genes, it results in the mutator phenotypes that lead to early onset of cancer.

The transforming growth factor-beta (*TGFβ*) signaling regulates adhesion, proliferation, differentiation, migration, and other cell functions. A large number of colorectal tumors hold mutations that interrupt *TGFβ* member signaling. The signaling initiates by binding of the ligand such as *TGFβ1/TGFβ2/TGFβ3* to the surface of the receptor *TGFβR2*. This induces the formation of the complex between serine/threonine kinase receptors; *TGFβR2* and *TGFβR1*.

After binding the *TGFβR2* phosphorylates *TGFβR1* receptor that leads to its activation. The transcription factors (*SMAD's*) thus forward the signal downstream to the *TGFβ* receptors. The activated *TGFβR1* then phosphorylates a *SMAD/3* that dimerizes with the *SMAD4* and form a stable complex. After then the complex translocates into the nucleus and through the DNA binding partners execute genes transcription (Figure 4.1).

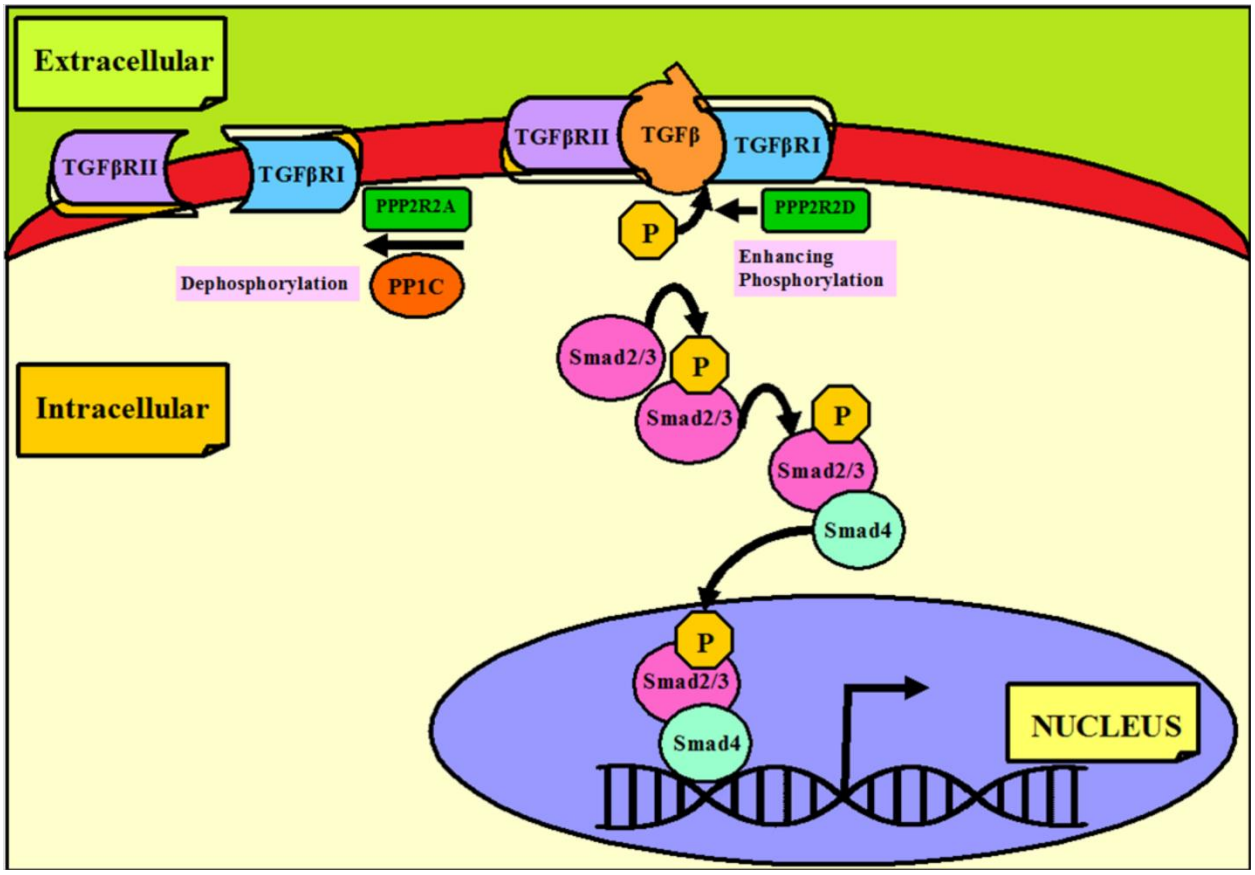


Figure 4.1 The signaling mechanism of TGFβ pathway.

The Transforming growth factor-beta 1 (*TGFβ1*) is the most abundant amongst its isoforms (*TGFβ2/TGFβ3*). It is a multifunctional cytokine that plays a crucial role in tumor progression and cancer initiation by regulating cell-motility and invasive capability. *TGFβ1* shows a pleiotropic effect as it possesses dual role; one having tumor inhibition effect in precancerous lesions and early stage of cancer and other as a tumor promoter in the later stage that causes malignancy [11, 12]. The switch from a tumor inhibitor function to a tumor promoter might be due to a variety of alterations in *TGFβ* signaling pathway, like mutations or expression loss in *TGFβ* receptors and *SMAD* proteins. The CD105 is a receptor for the *TGFβ1* and has a crucial

role in angiogenesis, and vascular development in CRC [13]. The *TGFβ1* promotes the DNA instability via down-regulating *Rad51*-mediated expression [14]. Generally, *TGFβ1* persuade and regulates apoptosis through the *SMAD*-dependent pathway [15]. However in case of the *SMAD*-independent pathway that includes *Ras/Raf* mediated mitogen-activated protein kinases (*MAPK*) pathway [16, 17] drives the human colon cancer cells proliferation [18, 19]. Hence targeting *TGFβ* pathway would be useful in determining the putative biomarkers for CRC progression. Through the studies it has been found that the mediator of the *Wnt* (i.e. *β-catenin*) and the *TGFβ* (i.e. *SMADs*) pathways bind to the common transcription factor i.e. Lymphoid enhanced binding factor/T-cell factors (i.e. *LEF/TCFs*) to induce expression of genes that control cell-fate [20]. Wnt negative regulator "*AXIN*" (a core component of *β-catenin* destruction component) has been also reported to be associated with various *SMADs* to modulate *TGFβ* signaling [21]. Also, both the pathways have been found to be commonly involved in the embryogenesis or in the tumorigenesis.

In this chapter, network and structure-based method has been implicated to decipher the candidate genes of the *TGFβ* pathway along with their mutational effect on *TGFβ1* and its progression towards CRC. The study conducted uses the systems biology and structural bioinformatics approaches to understand the dynamics of a biological system entailing the functional impact on the structure and dynamics of the system. Thus, the system level study helps in the discovery of new biomarkers or therapeutic targets for complex diseases, providing a key step in the development of personalized or population based medicine.

4.2 MATERIALS AND METHODS

The methodology for the network analysis followed the quantitative simulations method and determination of the network motifs that help to study the individual behavior of the pathway components to identify the putative markers. The structural variation study follows the effect of the damaging SNPs on the activity of the *TGFβ1* protein and its progression towards CRC. The methodology for performing these analyses is given in figure 4.2.

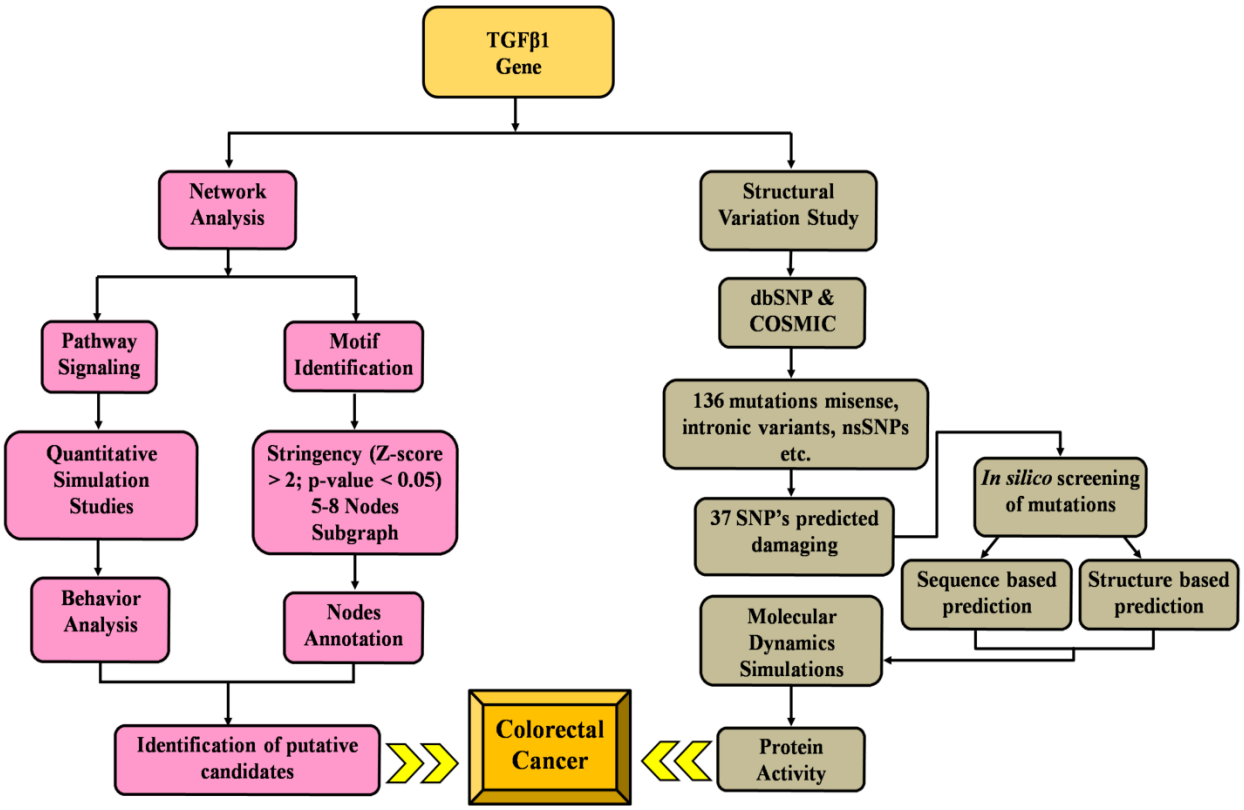


Figure 4.2 The flowchart of the procedure followed for analysis.

4.2.1 Pathway Simulation

The simulation process was initiated by considering the model given by Vizan et al. [22]. To analyze biological processes the MATLAB toolbox SimBiology [23], was used to analyze the simulations. It provides an application based programmatic tools for modeling, simulations, and analysis of the dynamic systems. The reaction kinetics of the *TGFβ* pathway with their respective (Ordinary Differential Equations) ODEs are given as follows:

$$\frac{1}{k_d} R' = 1 - r_c^0 - (\alpha + 1)R \tag{1}$$

$$\frac{1}{k_d} R_{com}' = \alpha R - R_{com}^I - k_T' TGF.R_{com}^s \tag{2}$$

$$\frac{1}{k_d} R_T' = k_T' TGF.R_{com}^s - (k_{act}' + D)R_T \tag{3}$$

$$\frac{1}{k_d} R_{act}' = k_{act}' R_T - D.R_{act} \tag{4}$$

$$\frac{1}{k_d} TGF' = -(k_T' R_{com}^S + k_{cc}')TGF \tag{5}$$

$$S2_c' = k_{ex} S2_n - S2_c(k_{in} + k_p R_{act}') \quad (6)$$

$$S4_c' = k_{ex4} S4_n^f - k_{in}[S4_c^f + CIF.S24_c] \quad (7)$$

$$pS2_{tot}' = \frac{1}{\alpha+1}[\alpha k_p R_{act}' S2_c - k_{dp} pS2_n^f] \quad (8)$$

$$pS2_c' = k_p R_{act}' S2_c + k_{ex} pS2_n^f - k_{in}[pS2_c^f + CIF(S24_c + 2S22_c)] \quad (9)$$

$$S24' = \frac{1}{\alpha+1}[k_{on}(\alpha S4_c^f\, pS2_c^f + S4_n^f\, pS2_n^f) - k_{off}(\alpha S24_c + S24_n) \quad (10)$$

$$S24_c' = k_{on} S4_c^f\, pS2_c^f - (k_{off} + k_{in} CIF) S24_c \quad (11)$$

$$S22' = \frac{1}{\alpha+1}[k_{on}(\alpha\, pS2_c^{f^2} + pS2_n^{f^2}) - k_{off}(\alpha S22_c + S22_n)] \quad (12)$$

$$S22_c' = k_{on}\, pS2_c^{f^2} - (k_{off} + k_{in} CIF) S22_c \quad (13)$$

4.2.2 The Network-Motifs

The network motifs for the *TGFβ* pathway were obtained through FANMOD [24] tool. The FANMOD is a tool that performs fast detection of the network motif through RAND-ESU algorithm for calculating sample sub-graphs. The network motifs comprise small recurring patterns in a pathway under study and have a potential to work independently. The motifs thus generated using the tool were obtained to have 5-8 nodes; these were in total four that are forming the active sub-networks. The statistical parameters like *Z*-score and *p*-value were used to determine the significant set of sub-graphs. It included the stringency of greater than value 2 (>2) for *Z*-score and less than 0.05 (<0.05) for the level of significance. Comparing motif occurrence in active sub-network and random network decipher the statistical significance of identified motif.

4.2.3 SNP Collection and Damage Predictions

The data collection of 136 SNP's (i.e. non-synonymous, missense and intronic variants) for *TGFβ1* gene were obtained through the dbSNP [25] and COSMIC databases [26]. Out of 136 SNP's final 37 were selected only with non-synonymous types. Various damage prediction tools (Table 4.1) i.e. sequence (PolyPhen [27], I-Mutant Suite [28], PROVEAN [29], MutPred [30], SNP & GO [31], PredictSNP [32], MAPP [33], SNAP [34], SIFT [35], Mutation Accessor [36]) or structure-based (SNP Effect 4.0 [37], Eris [38], SNP&GO^{3D} [39]) were used for the study

(Table 4.2). Through the prediction only four (rs199946261, rs768250306, rs763943753, and rs541829714) SNPs were found to be damaged by all of the damage prediction tools and these are considered for the molecular dynamics (MD) simulations.

Table 4.1 Damaging nsSNPs Prediction through sequence prediction tools

		PolyPhen	I-Mutant Suite	Provean	MutPred		SNP&GO		PredictSNP	MAPP	SNAP	SIFT	MutationAccessor
SNP_ID	**AA Change**	**Prediction**	**Prediction**	**Prediction**	**Score**	**PhD-SNP**	**PANTHER**	**SNPs&GO**	**Prediction**	**Prediction**	**Prediction**	**Prediction**	**Functional_Impact**
rs201700967	R107H	PD	D	N	0.573	N	U	N	D	N	N	D	L
rs747563417	V106L	PD	N	N	0.855	D	D	N	D	D	N	D	M
rs781566009	N103T	PD	N	D	0.891	D	D	D	D	D	D	D	H
rs770505137	R94C	PD	D	D	0.448	D	D	D	D	D	D	D	M
rs758966134	P85R	PD	D	D	0.913	D	D	N	D	D	D	D	H
rs541829714	L83R	PD	D	D	0.918	D	D	D	D	D	D	D	M
rs199849225	V79G	PD	N	D	0.83	D	D	D	D	D	D	D	M
rs775550741	A74V	PD	N	D	0.685	N	D	N	N	D	N	D	L
rs763943753	N69Y	PD	D	D	0.776	D	D	D	D	D	D	D	M
rs199699574	S59N	PD	N	N	0.798	D	N	N	D	D	N	D	L
rs199699574	S59T	PD	N	N	0.747	N	N	N	N	D	N	N	N
rs753287325	T56M	PD	D	D	0.612	D	D	D	D	D	D	D	M
rs200230083	D55N	PD	D	D	0.572	N	D	D	N	D	N	D	L
rs199713772	S53N	B	N	N	0.597	N	D	D	N	D	N	N	N
rs199713772	S53T	B	N	N	0.571	N	D	N	N	D	N	D	L
rs745482429	S53G	PD	N	D	0.58	N	D	N	N	D	D	D	L
rs56361919	P49T	PD	D	D	0.862	D	D	D	N	D	N	N	M
rs200209614	P47A	B	D	N	0.486	N	N	N	N	N	N	N	N
rs768250306	G46R	PD	D	D	0.962	D	D	D	D	D	D	D	H
rs748135261	L45F	PD	D	N	0.429	N	N	N	N	N	N	N	N
rs769434404	F43L	PD	D	D	0.861	D	D	D	D	D	D	D	L
rs763101036	H40R	PD	D	D	0.561	N	N	N	N	N	D	D	L
rs766572720	K37N	PD	D	D	0.577	N	D	D	D	N	D	D	M
rs199516461	K37R	PD	D	N	0.441	D	D	D	N	N	N	D	L
rs200763912	E35D	PD	N	N	0.749	D	D	D	D	D	D	D	M
rs767685429	H34Q	PD	D	D	0.595	D	D	D	D	D	D	D	M
rs369182751	K31R	PD	D	N	0.621	N	N	N	N	N	N	N	L
rs761279576	W30R	PD	D	D	0.823	D	D	D	D	D	D	D	H
rs199946261	L28F	PD	D	D	0.757	N	D	N	D	D	N	D	M
rs775861573	R25H	PD	D	D	0.729	D	U	N	D	D	D	D	M
rs200527282	R25C	PD	D	D	0.673	D	U	D	D	D	D	D	M
rs201635147	I22T	PD	D	D	0.773	D	N	D	D	D	D	D	M
rs200164212	R18Q	PD	D	D	0.75	N	D	D	D	D	N	D	L
rs747968669	K13T	PD	D	N	0.523	N	U	N	N	N	N	D	N
rs770754343	T11M	PD	D	N	0.405	N	D	N	N	D	N	N	M
rs749275172	Y6C	PD	D	D	0.432	N	D	N	N	D	N	D	M
rs778711968	L2P	PD	D	D	0.677	D	D	D	D	D	N	D	L

***PD** – Probably Damaging, **B** – Benign, **D** – Disease, **N** – Neutral, **U** – Unclassified, **L** – Low, **M** – Medium, **H** – High

Table 4.2 Damaging nsSNPs Prediction through structure prediction tools

	SNP_Effect_4.0(A-chain)					SNP_Effect_4.0(B-chain)					ERIS	SNP&GO$^{3D\ (A\text{-}CHAIN)}$				SNP&GO$^{3D\ (B\text{-}CHAIN)}$			
SNP_ID	**FOLDX (kcal/mol)**	**dLIMBO**	**WALTZ**	**TANGO**	**Prediction**	**FOLDX (kcal/mol)**	**dLIMBO**	**WALTZ**	**TANGO**	**Prediction**	**Prediction (ΔΔG (kcal/mol))**	**S3D-PROF**	**PANTHER**	**SNPs&GO**	**S3Ds&GO**	**S3D-PROF**	**PANTHER**	**SNPs&GO**	**S3Ds&GO**
rs201700967	2.14	0	0.03	-3.89	RS	1.52	0	0.03	-3.89	RS	0.34	D	U	N	N	D	U	N	N
rs747563417	3.84	0	-0.07	-2.09	RS	3.12	0	-0.07	-2.09	RS	1.86	D	D	N	D	D	D	N	D
rs781566009	2.28	0	-0.29	57.05	RS	1.57	0	-0.29	57.05	RS	1.27	D	D	N	D	D	D	N	D
rs770505137	0.24	-6.17	-119.79	56.03	NE	0.41	-6.17	-119.79	56.03	NE	-1.59	D	D	D	D	D	D	D	D
rs758966134	1.63	0.01	1.13	-1.42	RS	1.75	0.01	1.13	-1.42	RS	-3.85	D	D	D	D	D	D	D	D
rs541829714	2.41	0	0	0	RS	1.61	0	0	0	RS	2.32	D	D	D	D	D	D	D	D
rs199849225	3.09	0	0	0	RS	3.16	0	0	0	RS	5.28	D	D	D	D	D	D	D	D
rs775550741	2.24	0	0	0	RS	1.96	0	0	0	RS	-1.63	D	D	N	D	D	D	N	D
rs763943753	1.83	0	0.04	0.08	RS	1.68	0	0.04	0.08	RS	0.28	D	D	D	D	D	D	D	D
rs199699574	6.05	-2.01	0.52	-4.45	SRS	2.52	-2.01	0.52	-4.45	RS	2.83	D	N	N	D	D	N	N	D
rs199699574	1.99	-1.15	-2.56	29.64	RS	1.84	-1.15	-2.56	29.64	RS	4.34	N	N	N	N	N	N	N	N
rs753287325	-1.09	0	0.37	9.8	NE	0.3	0	0.37	9.8	NE	0.83	N	D	D	D	N	D	D	D
rs200230083	0.28	0	3.73	32.43	NE	0.28	0	3.73	32.43	NE	-0.18	N	D	D	N	N	D	D	N
rs199713772	-0.74	0	11.7	-19.45	NE	-0.7	0	11.7	-19.45	RS	-3.31	N	D	N	D	N	D	N	D
rs199713772	0.41	0	-0.96	94.03	NE	0.21	0	-0.96	94.03	NE	-1.02	D	D	N	D	D	D	N	D
rs745482429	-0.03	0	0.1	-8.38	NE	-0.1	0	0.1	-8.38	NE	0.95	N	D	D	D	N	D	D	D
rs56361919	2.19	0	0.5	47.34	RS	2.3	0	958.82	47.34	RS	-1.51	D	D	D	D	D	D	D	D
rs200209614	1.26	0	0.05	-0.01	RS	0.74	0	958.82	-0.01	RS	-1.57	N	N	N	N	N	N	N	N
rs768250306	17.12	0	0	0.01	SRS	14.95	0	0	0.01	SRS	>10	D	D	D	D	D	D	D	D
rs748135261	0.22	0	0.22	-0.02	NE	0.15	0	0.22	-0.02	NE	2.51	N	N	N	N	N	N	N	N
rs769434404	3.05	0	-0.7	0	RS	3.41	0	-0.7	0	RS	-0.5	D	D	D	D	D	D	D	D
rs763101036	0.04	0	-0.65	0	NE	0.1	0	-0.65	0	NE	-0.26	N	N	N	N	N	N	N	N
rs766572720	-0.14	0	-0.01	0	NE	0.36	0	-0.01	0	NE	-3.12	D	D	D	D	D	D	D	D
rs199516461	0.03	0	0	0	NE	-0.02	0	0	0	NE	-1.29	N	D	D	N	N	D	D	D
rs200763912	0.64	0.04	-0.01	0	SRS	1.1	0.04	-0.01	0	RS	0.98	D	D	D	D	D	D	D	D
rs767685429	-0.58	0.06	0.25	0	ES	-0.21	0.06	0.25	0	NE	0.26	D	D	D	D	D	D	D	D
rs369182751	0.12	0	-0.08	0	NE	0.28	0	-0.08	0	NE	-1.22	N	N	N	N	N	N	N	N
rs761279576	3.61	0.04	0.1	0	RS	3.45	0.04	0.1	0	RS	-0.96	D	D	D	D	D	D	D	D
rs199946261	2.89	0	0.27	0	RS	4.5	0	0.27	0	RS	1.9	N	D	N	D	N	D	N	D
rs775861573	0.22	-1.12	0.67	0	NE	1.93	-1.12	0.67	0	RS	0	D	U	N	N	D	U	N	N
rs200527282	0.16	-1.09	-0.11	6.62	NE	2.01	-1.09	-0.11	6.62	RS	-1.31	D	U	D	N	D	U	D	N
rs201635147	2.9	-0.13	-128.78	0	RS	3.17	-0.13	-128.78	0	RS	4.24	D	N	D	D	D	N	D	D
rs200164212	1.46	0	421.2	1.3	RS	1.46	0	421.2	1.3	RS	1.28	D	D	D	D	D	D	D	D
rs747968669	0.79	0	-0.14	0	SRS	0.67	0	-0.14	0	RS	-1.17	N	U	N	N	N	U	N	N
rs770754343	-0.08	0	-0.01	0	NE	-0.1	0	-0.01	0	NE	-2.44	N	D	N	N	N	D	N	N
rs749275172	3.02	0	-2.33	0.11	RS	2.92	0	-2.33	0.11	RS	3.77	D	D	N	D	D	D	N	D
rs778711968	2.2	0	-2.19	0	RS	2.26	0	-2.19	0	RS	>10	D	D	D	D	D	D	D	D

***RS –** Reduces Stability, **NE –** No Effect, **SRS –** Severely Reduces Stability, **ES –** Enhances Stability, **D** – Disease, **U** – Unclassified, **N** – Neutral

4.2.4 Selection of Structure

The structural dynamics study was performed to determine the effect of the mutations on structural stability. For molecular dynamics it is important to have the structure of the *TGFβ1* protein to be analyzed. Therefore, PDB database has been searched [38] for the corresponding structures, and this resulted in five structures (1KLA, 1KLC, 1KLD, 3KFD, and 4KV5) from *Homo sapiens*. The next challenge was to select a single structure among the five for studying the dynamics. Theseus-3D tool was used to identify the median structure based on the maximum likelihood method [41, 42]. The median structure is considered to be most similar to the average structure and therefore considered to be the most typical structure in the ensemble. Theseus-3D aligns multiple structures by extracting the corresponding aligned sequences and performs the superposition using that alignment. Therefore, the process starts with identifying the single model among the NMR structures i.e. 1KLA, 1KLC, and 1KLD. For 1KLA, the 17^{th} model out of 17 different conformations was identified as the median structure; similarly for 1KLD, the 2nd model was the median structure out of the 16 different models. However, for 1KLC there was only a single conformer deposited in PDB. Theseus-3D was now run for the three median NMR structures and the two X-ray structures i.e. 3KFD and 4KV5. This resulted in identifying 4KV5 as the median structure. Since, 4KV5 is reported to be functionally active as a heterodimer. Both chains (A and B) were considered. The maximum likelihood for both chains was calculated to be 0.36628. The mutagenesis tool in PyMOL was used to introduce the mutations in the structure [43], and for each mutant structure, the rotamer of the mutated residue with least steric clashes was selected.

4.2.5 Molecular Dynamics (MD)

MD simulation calculates the time-dependent behavior of the physical movements of atoms and molecules. It provides understanding for the physical basis of the structure and function of biological macromolecules. For MD simulation the GROMACS v5.1.2 package [44] with OPLS all-atom force field parameters [45] was used. The simulations were performed for a time period of 100ns each for the set of four variants i.e. L28F, G46R, N69Y, L83R, and including the native form of a protein as a control. The first step for the simulation was to perform system solvation using TIP3P water model [46]. The genion tool was used to add the counter ions to neutralize the system, i.e., to bring the net charge to zero. The monoatomic ions (i.e. Cl^-) were added to each structure i.e. 4 Cl^- for L28F, 10 Cl^- for G46R, 4 Cl^- for N69Y and 10 Cl^- for L83R and 4 Cl^- for

the wild type. The particle-based method was used to calculate the potential of all atoms. This was followed by energy minimization (50000 steps) to reduce the total energy of the stable conformers. After this, solvent equilibration step was performed followed by a system equilibration for time period of 100ps. At 1 atmospheric pressure, the production run was performed for 100ns by integrating the equation of motion in the NPT at 100ps time. For analysis, the coordinates of MD simulations were saved at 100ps time-period for all the structures (wild and native). At the end of the production run, the trajectory was analyzed for global parameters of stability and hydrogen bonding analysis was performed to identify structural changes.

4.3 RESULTS AND DISCUSSION

This chapter aims at gaining insight into the key components of the *TGFβ* pathway and the mutational effects occurring to the structure of the *TGFβ1* protein. It is anticipated that this network and structure-based computational analysis not only provide the potential candidates for the study but, also covers the extensive view of the *TGFβ* allied studies towards disease condition. This will provide new direction to the work related to the CRC that can help the biological researchers to develop the effective methods for the therapeutic intervention.

4.3.1 The Pathway Analysis

The *TGFβ* pathway simulations were performed through the Ode45 (Dormand-Prince) solver for all set of genes with varying parameters [47]. The method is effective for a higher-order solution and for performing the simulations with minimal error-rate. The parameters were selected based upon the standards given in the literature and absolute tolerance is set to 1.0E-6 with a relative tolerance of 0.001 (Table 4.3). Depending on the function of the entities, separate simulations were run each time and at different time-period. Therefore, more than a hundred simulations were run to analyze the behavior of the entities and only those with consistent behavior were represented.

Table 4.3 The concentration values for the respective entities.

Species	Type	Description	Initial Concentrations (μM)
S22	COMPLEX	Total cellular homomeric S22 complexes	0
S24	COMPLEX	Total cellular heteromeric S24 complexes	0
pS2tot	PHOSPHOPROTEIN	Total cellular pS2	0
TGF	PROTEIN	TGFβ1	4
R	PROTEIN	Nascent Receptors	0.9
S2C	PROTEIN	Cytoplasmic, unphosphorylated Smad2	1.2
Rcom	PROTEIN	TGFβ bound receptors	0.12
pS2c	PHOSPHOPROTEIN	Total cytoplasmic pS2	0
RcomS	PROTEIN	Mature, competent receptors	0.05
S2n	PROTEIN	Nuclear unphosphorylated Smad2	0.6
S22n	COMPLEX	Nuclear homomeric S22 complexes	0
S4n	COMPLEX	Total nuclear Smad4	1
S22c	COMPLEX	Cytoplasmic homomeric S22 complexes	0
pS2n	PHOSPHOPROTEIN	Total nuclear pS2	0
pS2fn	PHOSPHOPROTEIN	Monomeric nuclear pS2	0
S24n	COMPLEX	Nuclear heteromeric S24 complexes	0
S24c	COMPLEX	Cytoplasmic heteromeric S24 complexes	0
S4fc	PROTEIN	Monomeric cytoplamic Smad4	1
S4c	PROTEIN	Total cytoplamic Smad4	1
pS2fc	PHOSPHOPROTEIN	Monomeric cytoplasmic pS2	0
S4fn	PROTEIN	Monomeric nuclear Smad4	1
Rtot	GENE	Total receptors	1
RT	COMPLEX	Active receptors	0
RcomI	PROTEIN	-	0.07
Ract	PROTEIN	-	0

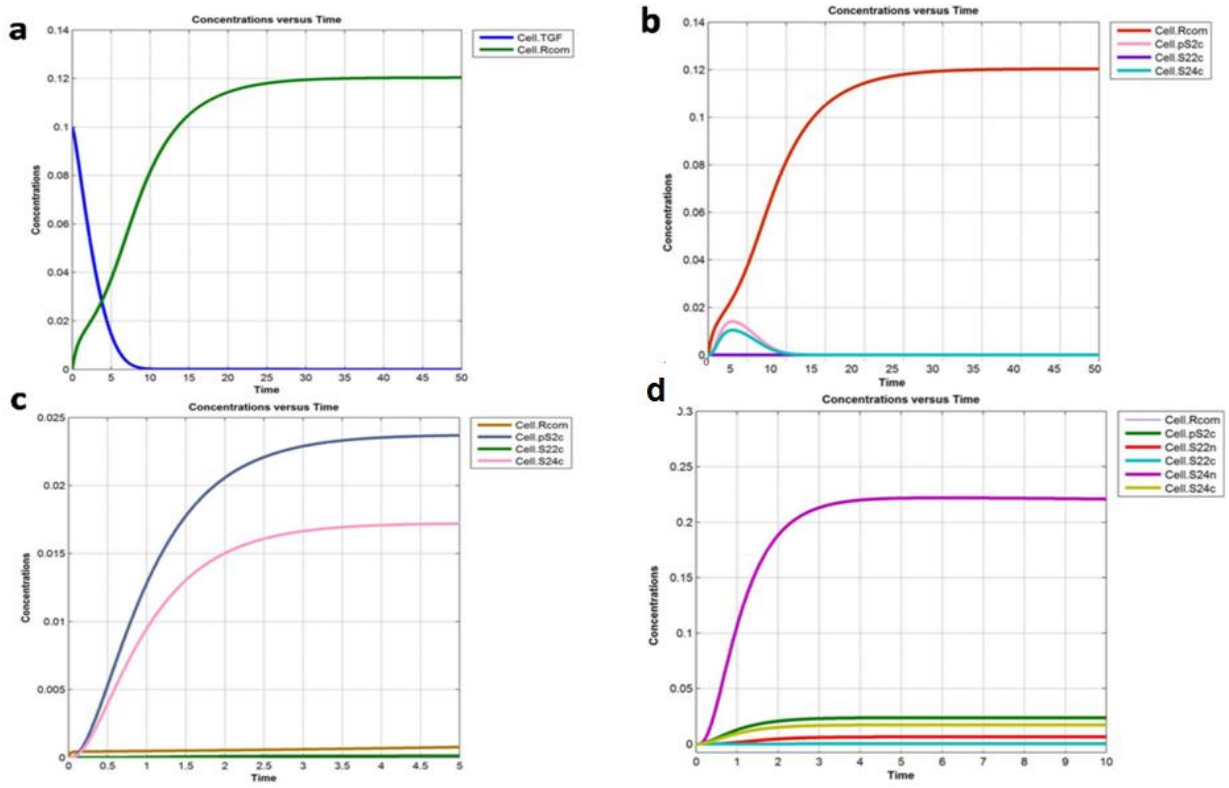


Figure 4.3 Quantitative simulations performed at different time-scale. a) Initiation of cell signaling via *TGFβ1* b) Activation of the receptors and the *SMAD* complexes c) Phosphorylation of the *SMADs* d) Enhanced activity of the *SMADs*; Concentration in μM, Time in milliseconds.

Figure 4.3 shows entities with significant behavior; the low concentration of the *TGFβ1* while binding to the receptor (Figure 4.3 a) activates the signaling nearby 4-units time-period at 0.03 μM concentration. This states that even small amount of *TGFβ1* protein is enough to initiate the pathway signaling. Once the receptor gets activated there is phosphorylation event of the *SMAD* complexes that lead to the complex stabilization (Figure 4.3b). Thus, inferring that the cytoplasm signaling occurs at a fast pace. By changing the time from 50 μM to 5 μM, there is activation of the *SMAD* complexes for first 5-units time period as the signal is consumed by the receptors (Figure 4.3 b and 4.3 c). This could be the crucial step, because if the complexes remain active for a longer period of time, then there is a chance that they will violate the gene regulation event and will overpower the process. When the concentration of the *SMAD4* is increased by 4-units the event is thought to be crucial as it's over expression has a tendency to perturb the cell cycle and initiate the disease progression (Figure 4.3 d).

The regulatory patterns were determined in the *TGFβ* pathway to find the regulatory path of the transcription factors involved. The retrieved sub-graphs were those containing 5-8 nodes and representative of them is given in Figure 4.4 a-d. In a 5-8 nodes sub-graphs, three variants of 3-node motifs and six variants of 4-node motifs were found to be overrepresented (Figure 4.4 a-i). These variants are TFs regulating the target gene, and the reprehensive SIMs (single input modules). The disintegration of large sub-graphs into small sub-graphs provided us some motifs matching with the motif-id's in the standard network motif dictionary provided by Alon et al [48]. The sub-graphs were annotated to determine the type of nodes that are significant to the pathway. The processes like binding, activation, expression, inhibition, and phosphorylation have been seen amongst the nodes of the sub-graphs. A large number of events have been captured through the sub-graphs, such as the role of *SMAD6* and *INBHA* in dual inhibition. The nodes like *SMAD4* and *TNFα* in rebuilding the network as evident from Figures 4.4 a-d. The *TGFβR1, SMAD1, SMAD2, SMAD4, SMAD6* were found to be frequently occurring in almost all node of the sub-graphs, thus representing their key role in the pathway.

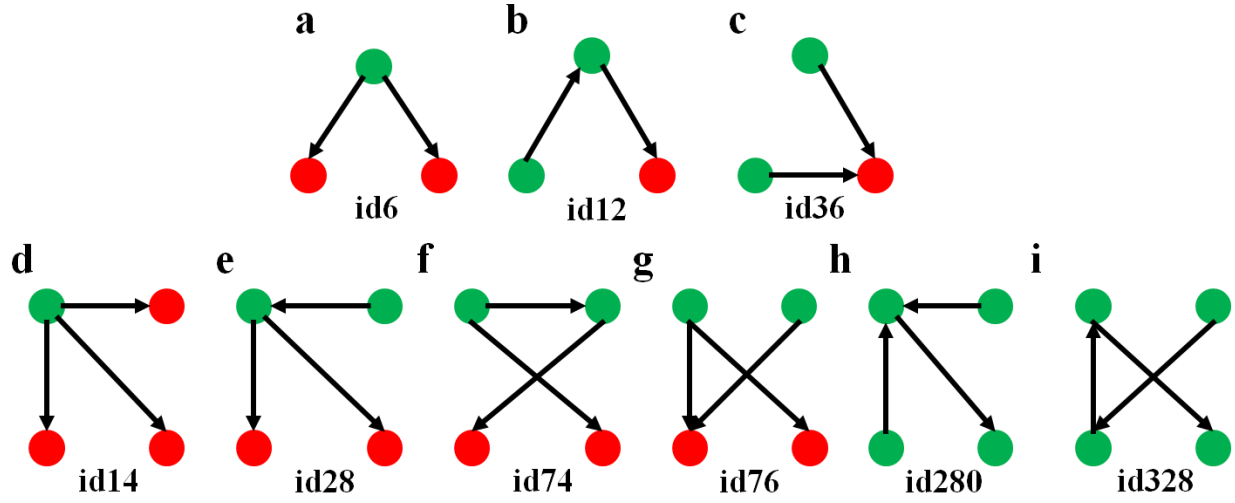


Figure 4.4: Network motifs Network motifs obtained from the 5-8 nodes subgraph.

In general, *SMAD6* activation by *TGFβR1* regulates *SMAD2* expression; it also has a tendency to bind to *Smurf1* that in case of over expression leads to CRC progression. *Smurf1* can also inhibit *TGFβR1* but, there is nothing to make a check on *Smurf1*. The analysis reveals that these motifs are involved in regulating expression of a large number of genes by turning them on or off. Delineation of inhibition and activation processes has been defined through the network motifs, which reflects their role in normal and malignant conditions. Besides six key genes i.e. *Smurf1*,

SMAD2, TGFβR1, SMAD1, SMAD6, and *SMAD4* other new ones were determined such as *TNFα, INHBA, LTBP1, TGFβ1, NODAL, PPP2CA,* and *ROCK1*. The results postulate the computational prediction of the active sub-network that needs to be confirmed experimentally to determine their regulatory role.

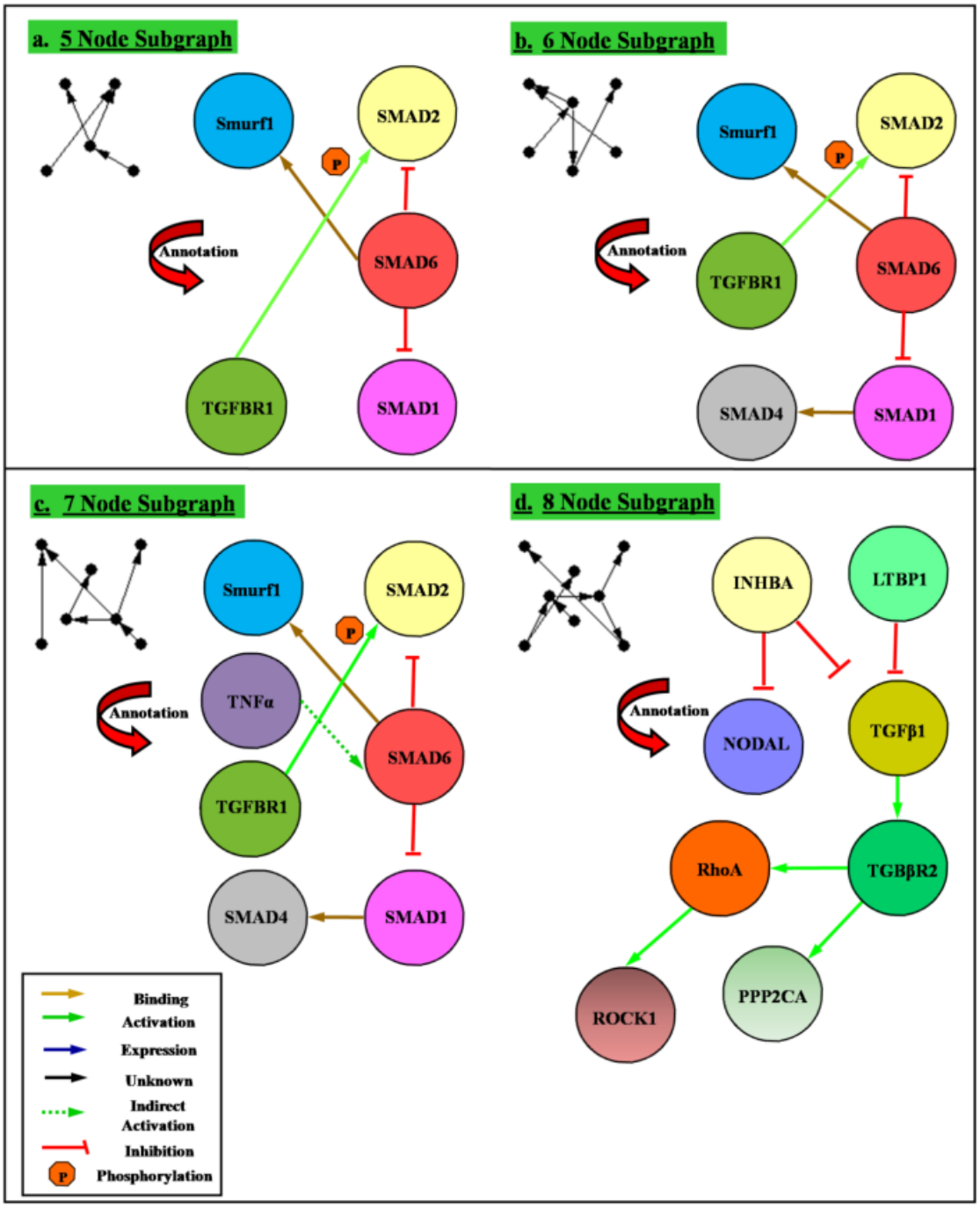


Figure 4.5 The over-represented sub-graphs annotated portray the crucial interactions.

4.3.2 Identification of Binding Pocket

The CASTp program was used to determine the binding pocket [49]. The binding pocket of the protein is an area with concave regions, where the interacting partner makes interaction with the protein. The active site is presumed to play important roles while binding on the receptor site. CASTp characterizes the binding site by measuring the void area using weighted Delaunay triangulation and the alpha complex [49]. CASTp measures the area and volume of each pocket and cavity through the solvent accessible surface model (Richard's surface) and molecular surface model (Connolly's surface). This helps in identifying the active site residues that lie on the surface of a protein. Using this tool the binding site prediction was conducted as illustrated in Figure 4.6 for the *TGFβ1* (4KV5) protein, after performing Theseus 3D step. The electrostatic potential elucidate the presence of mainly positive charges at the active site as blue color illustrates the positive charges, red color for negative charges and white for the neutral ones (Figure 4.6a). The results were confirmed through the CASTp as two pockets were selected (i.e. pocket 30 and pocket 29) based on the spatial position and relatively larger surface area, where the surface area for the pocked id 30 was 149.93$Å^2$ and for the pocket id 29 was 127.20$Å^2$. The pocket id 30 is comprised of 13 residues from both chain A (Tyr58, Leu62, Ser73, Ala75, Cys77, and Ser112) and chain B (Phe43, Cys44, Leu45, Cys77, Cys78, Val79, and Pro80) (shown as orange sticks in Figure 4.6b). The residues that make up the pocket were principally all hydrophilic. In the case of the pocket id 29 the binding site is comprised of 12 residues from both chain A (Phe43, Cys44, Leu45, Cys78, Val79, and Pro80) and chain B (Tyr58, Leu62, Ser73, Ala75, Cys77, and Ser112) (shown as cyan sticks in Figure 4.6b). The binding of *TGFβ1* to its receptor at the predicted cavities site was validated by performing structural analysis and simulation studies.

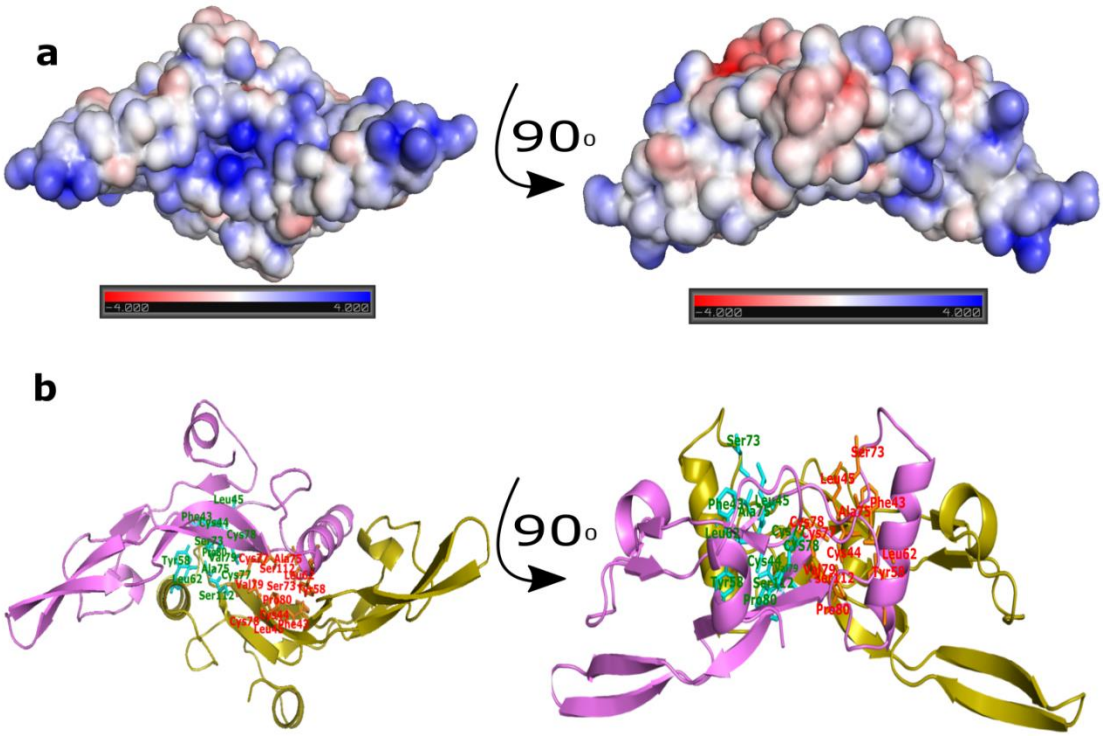


Figure 4.6 a) Surface representation of the 4KV5 representing the electrostatic potential measured through the APBS **b)**The binding pocket of protein 4KV5 obtained through CASTp results.

4.3.3 MD (The Global Parameter Study)

The MD simulation was performed for all five structures (i.e. four mutants and one native) for 100ns time-period, each. The root mean square deviation (RMSD) (Figure 4.7a), root mean square fluctuation (RMSF) (Figure 4.7b), radius of gyration (Rg) (Figure 4.7c), and solubility accessible surface area (SASA) (Figure 4.7d), was performed to analyze the impact of mutations on the global parameters using gnuplot [50]. The significant changes, if any, were observed for these set of global parameters. Through RMSD analysis, the mutants L83R (2.4Å), N69Y (3.0Å), and L28F (3.1Å) were found to be similar to the native (2.9Å) trajectory up to 50ns. However, after 50ns timestamp, variations were observed in comparison to the native's trajectory, thus reflecting the structural shift. In comparison to the other mutants L28F and G46R, have shown major variations (Figure 4.7a). Through RMSF analysis, significant changes were found in comparison to the native. The residues Arg94 and Tyr50 in the mutants L28F, N69Y, and L83R were shown to be highly fluctuating (Figure 4.7b).

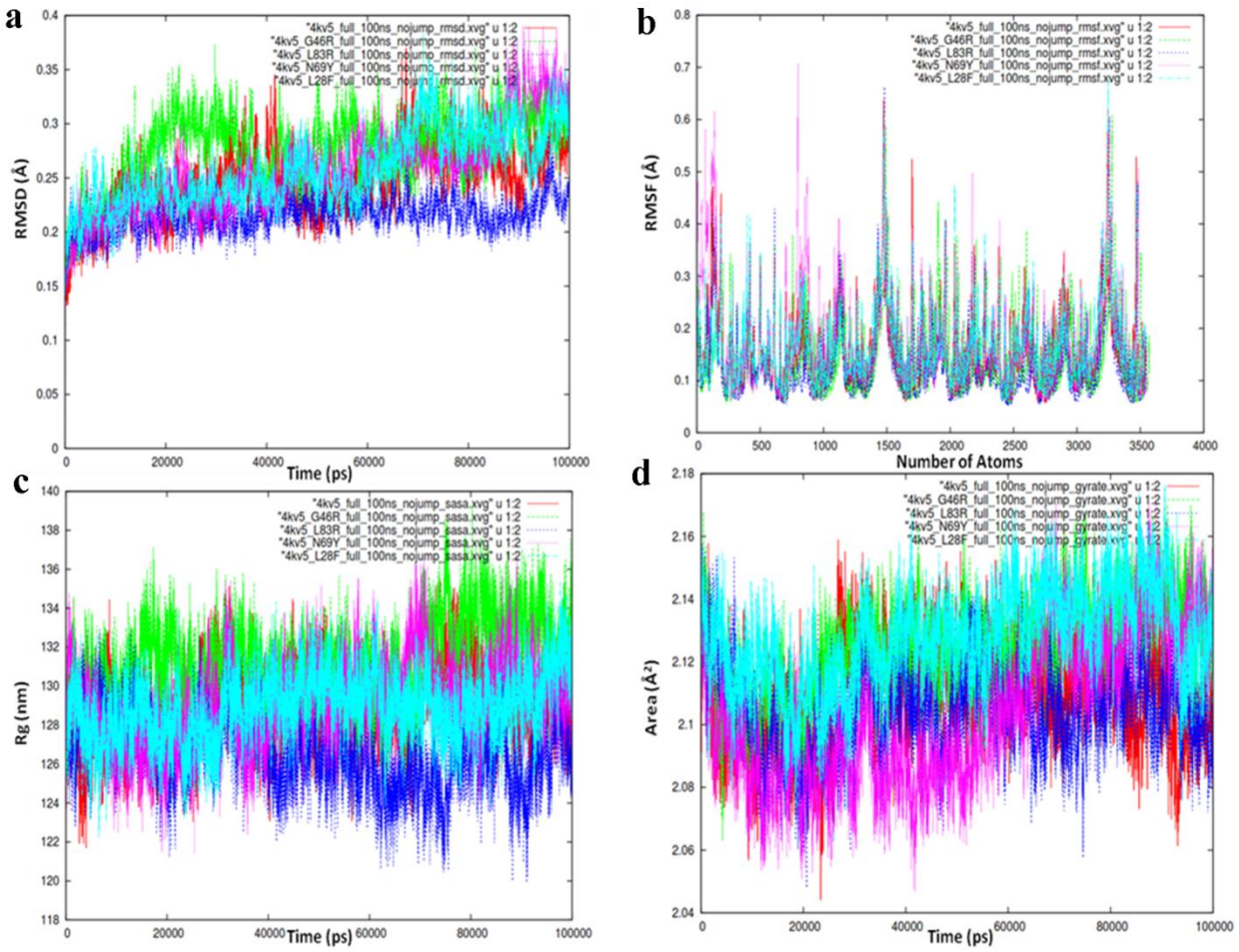


Figure 4.7: Molecular Dynamics (MD) analysis **a)** Graph displaying the root mean square deviation (RMSD) **b)** Graph displaying the root mean square fluctuation (RMSF) **c)** Graph displaying the radius of gyration (Rg) **d)** Graph displaying the solvent accessible surface area (SASA).

The Rg measures the stability index of the protein structure that undergoes mutation or denaturation. In other words, whether the protein unfolds due to a mutation or not can be analyzed by measuring its Rg, where lower the value indicates globularity. The Rg was measured to be 2.12nm and 2.14nm, respectively for the mutants (Figure 4.7c). This states that the structures remain intact and globular irrespective of the mutational event thus stabilizing the MD trajectories and not disturbing the overall structure.

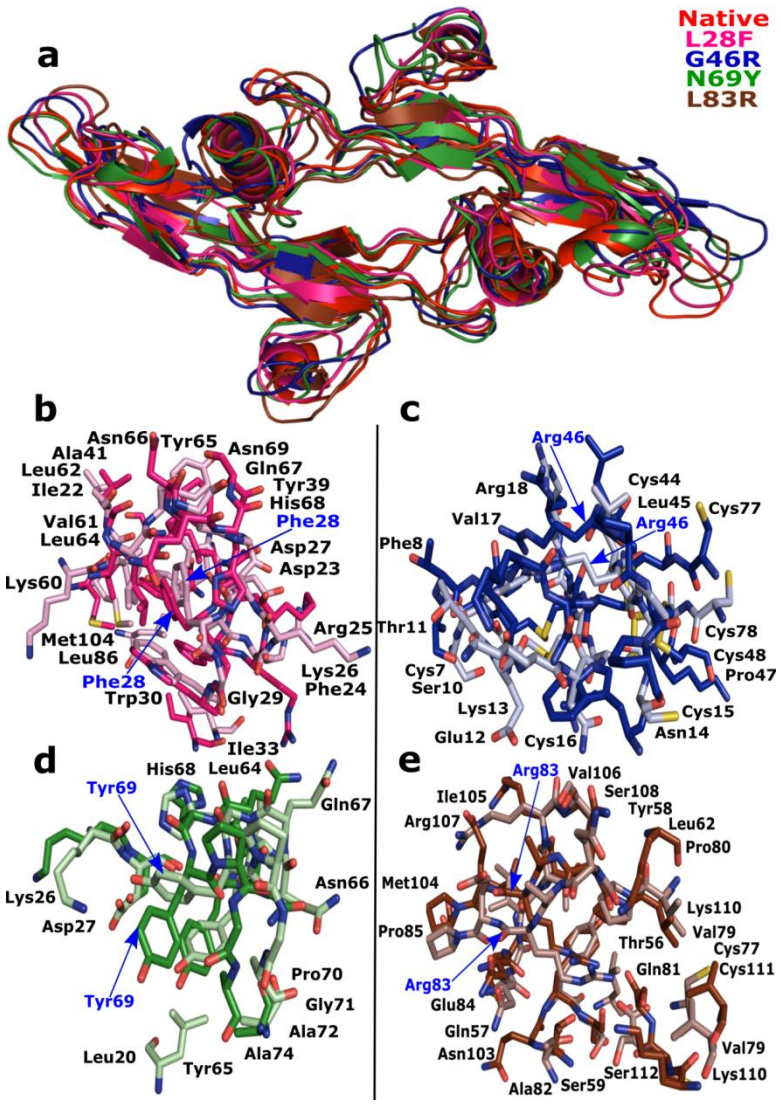


Figure 4.8 The comparisons of the first and last frame of the native and mutant structures obtained through MD. **a)** Superposing native and the mutant structures **b)** Superposing 0ns and the 100ns frames for the F28 mutant **c)** Superposing 0ns and the 100ns frames for the R46 mutant **d)** Superposing 0ns and the 100ns frames for the Y69 mutant **e)** Superposing 0ns and the 100ns frames for the R83 mutant.

The SASA is one of the major contributing factors for protein stability as it predicts the conformational changes to the proteins upon binding to the substrate [51]. Through SASA analysis, the G46R is found to be distinctly different in comparison to other mutants (Figure 4.7d). Specifically, G46R residue was found to be exposed after a time period of 10ns. However, for L83R residue it gets exposed after 30ns and not much exposure was found for the N69Y and L28F variants.

Also, the superposition of frames, i.e. the first timeframe (0ns) to the last timeframe (100ns) reflected the conformational changes in terms of secondary structure specifically at the loop region, helix region and also at the beta-sheet region (Figure 4.8a). Again mapping was performed by considering the residues around the 5Å regions of the residues Phe28, Arg46, Tyr69, and Arg83. The structural mapping states the significant structural differences in the residues within a 5Å radius of mutation for the first and last time frames (Figure 4.8 b-e).

4.3.4 Polar and non-Polar Bonds (The Local level Analysis)

The longevity of the hydrogen bonds in the structure was determined using LIGPLOT [52], as the polar interactions help to stabilize the conformation in the event of a mutation. LIGPLOT is a program that takes standard PDB file as input and generates schematic 2D-representations of protein-ligand complexes. It provides an information about various intermolecular interactions like hydrogen bonds (H-bonds), hydrophobic interactions etc. Hydrogen bonds provide information for directing the protein's structure, folding, and molecular recognition; hence new polar interactions that are formed in the mutants were analyzed using the final snapshot of the trajectory (at the end of 100ns) (Table 4.4). The polar interactions determined by LIGPLOT were cross-validated using the H-bond command of the GROMACS. Simultaneously, hydrogen bonds have been determined throughout the trajectory (Table 4.5 and Figure 4.9). After cross-checking, it has been found that certain bonds were formed at the beginning of the simulation and maintained throughout the 100ns trajectory, while other were formed at the end of the simulation only. This concludes that the former bonds were formed via structural adjustments at internal level thus compensating the entropic cost of a mutating residue by stabilizing the dimer structure. The bonds found only in the last timeframe of the trajectory (at 100ns) are transient bonds that are formed due to the minor fluctuations and not contributing to the mutant structure stability (Figure 4.7). The interactions for the residues around the 5Å region were also identified at both 0ns and 100ns time frames using PyMOL (Table 4.4). Also for these pair of interactions, cross-check was done by the hydrogen bond analysis in GROMACS (Table 4.6). Many new polar interactions were created during the 100ns simulations, and many of them vanished at the end of the simulation, clearly depicting the new interactions occurring among the residues of the *TGFβ1* mutants. This illustrates the high possibility that the protein alters its function via structural modification thus, altering the normal activity and have potential to lead its way towards CRC progression.

Table 4.4. Polar interactions obtained through the Ligplot

4kv5_Native_t0		**4kv5_Native_t100**		**4kv5_G46R_t0**		**4kv5_G46R_t100**		**4kv5_L83R_t0**		**4kv5_L83R_t100**	
ChainA-ChainB	**H-Bond**	**ChainA-ChainB**	**H-Bond**	**ChainA-ChainB**	**H-Bond**	**ChainA-ChainB**	**H-Bond**	**ChainA-ChainB**	**H-Bond**	**ChainA-ChainB**	**H-Bond**
Gln57-Ser102	2.98	Asp27-Asn69	2.74	Asn69-Asp27	2.86	Tyr58-Asn103	3.1	Asn42-Tyr58	2.69	Asn42-Tyr58	2.67
Tyr58-Asn103	3.01	Asn103-Tyr58	2.87	Asp27-Asn69	2.84	Tyr58-Asn42	2.81	Asn103-Tyr58	3.2	Asn103-Tyr58	2.85
Asn69-Asp27	2.83	Asn42-Tyr58	2.74	Asn103-Tyr58	2.98	Asn103-Tyr58	2.94	Tyr58-Asn42	3.03	Tyr58-Asn42	2.83
Asp27-Asn69	3.08	Gln57-Ser102	3.31	Asn42-Tyr58	2.66	Asn42-Tyr58	2.69	Gln57-Ser102	2.73		
Asn103-Tyr58	2.82	Gln57-Asn103	2.99	Tyr58-Asn103	3.03	Tyr50-Trp30	3.29				
Asn42-Tyr58	3.03	Tyr58-Asn103	2.89	Gln57-Ser102	2.9	Lys26-His68	3.24				
		Tyr58-Asn42	2.84								
4kv5_N69Y_t0		**4kv5_N69Y_t100**		**4kv5_L28F_t0**		**4kv5_L28F_t100**					
ChainA-ChainB	**H-Bond**	**ChainA-ChainB**	**H-Bond**	**ChainA-ChainB**	**H-Bond**	**ChainA-ChainB**	**H-Bond**				
Tyr58-Asn103	3.18	Asp27-Tyr69	2.59	Asp27-Asn69	3.19	His68-Asp27	2.6				
Gln57-Ser102	2.92	Lys26-His68	2.96	Asn103-Tyr58	3.13	Tyr65-Asp27	2.98				
Asn42-Tyr58	2.97	Asn42-Tyr58	2.86			Tyr58-Cys44	3.13				
Asn103-Tyr58	3.11	Tyr58-Asn103	3.07			Tyr58-Asn103	2.78				
		Gln57-Ser102	2.68			Asn103-Tyr58	2.98				
		Tyr69-Asp27	2.51								
		His68-Lys26	3.01								

Table 4.5 The polar contacts recognized for the structural variants

Chains (0-100ns Time Frame)	**Polar Contact Residues (ChainA-ChainB)**
L28F	Tyr58- Cys44
	Tyr65- Asp27
	His68- Asp27
G46R	Tyr50- Trp30
N69Y	Lys26- His68
	His68- Lys26
L83R	Tyr58- Asn42

Table 4.6 The polar interactions identified at diverse time-frames

Chain Type	Polar Contact Residues	Time Frames	Presence
Chain A	**4kv5 native G46 around 5Å**		
	Cys15-Pro47	remains at 0ns	TP
	4kv5 mutant R46 around 5Å at 0ns		
	Glu12-Lys13	vanishes 100ns	FP
	Glu12-Arg46	vanishes 100ns	TP
	4kv5 mutant R46 around 5Å at 100ns		
	Cys15-Asn14	forms at 100ns	TP
Chain A	**4kv5 native L28 around 5Å**		
	Tyr39-Met104	remains at 0ns	TP
	Asp23-Lys26	remains at 0ns	TP
	Asp23-Asp27	remains at 0ns	TP
	4kv5 mutant F28 around 5Å at 0ns		
	Tyr39-Ala41	vanishes 100ns	TP
Chain B	Val61-Leu64	vanishes 100ns	TP
	4kv5 mutant F28 around 5Å at 100ns		
Chain A	Asp23-Phe24	forms at 100ns	TP
	Ile22-Phe24	forms at 100ns	FP
Chain B	Leu64-His68	forms at 100ns	TP
Chain A - Chain B	Asp27-His68B	forms at 100ns	TP
	4kv5 native N69 around 5Å		
Chain A	Leu64-His68	remains at 0ns	TP
	Leu64-Gln67	remains at 0ns	FP
	4kv5 mutant Y69 around 5Å at 0ns		
Chain A	Asn66-Tyr69	vanishes 100ns	TP
	Tyr65-His68	forms at 0ns&100ns	TP
	Tyr65-Tyr69	forms at 0ns&100ns	FP
	4kv5 mutant Y69 around 5Å at 100ns		
Chain B	Lys26-Asp27	forms at 100ns	TP
ChainA - ChainB	Tyr69-Asp27	forms at 100ns	FP
	Tyr69-Lys26	forms at 100ns	FP
	His68-Lys26	forms at 100ns	TP
	4kv5 native L83 around 5Å		
Chain A	Gln81-Ser108	remains at 0ns	FP
	Ala82-Ser108	remains at 0ns	TP
	Ala82-Arg107	remains at 0ns	FP
	Glu84-Arg107	remains at 0ns	FP
	4kv5 mutant R83 around 5Å at 0ns		
Chain B	Ser59-Ser112	forms at 0ns_vanishes_at100ns	FP
ChainA - ChainB	Arg83-Ser112	forms at 0ns_remains100ns	FP
	4kv5 mutant R83 around 5Å at 100ns		
Chain B	Cys77-Ser112	forms at 100ns	FP

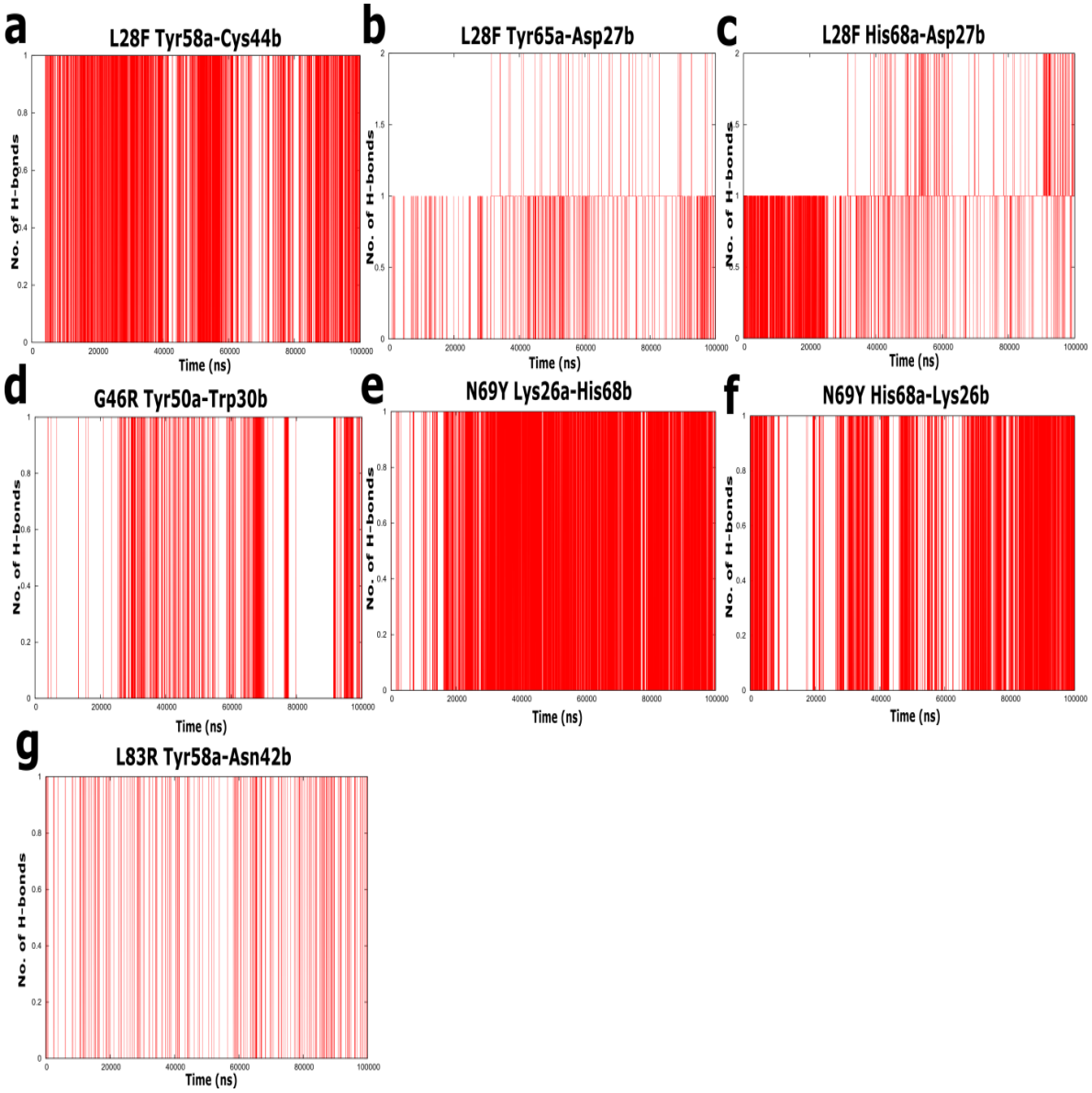


Figure 4.9 Polar-interactions in respective time-frame. **a)** One new H-bond was identified for the L28F mutant at Tyr58 of chain A and Cys44 of chain B **b)** Two new H-bonds were identified for the mutant L28F at Tyr65 of chain A and Asp27 of chain B **c)** Two new H-bonds were identified for the mutant L28F at His68 in chain A and Asp27 of chain B **d)** One new H-bond was identified for the G46R mutant at Tyr50 of chain A and Trp30 of chain B **e)** One new H-bond was identified for the N69Y mutant at Lys26 of chain A and His68 of chain B **f)** One new H-bond was identified for the N69Y mutant at His68 of chain A and Lys26 of chain B **g)** One new H-bond was identified for the L83R mutant at Tyr58 of chain A and Asn42 of chain B.

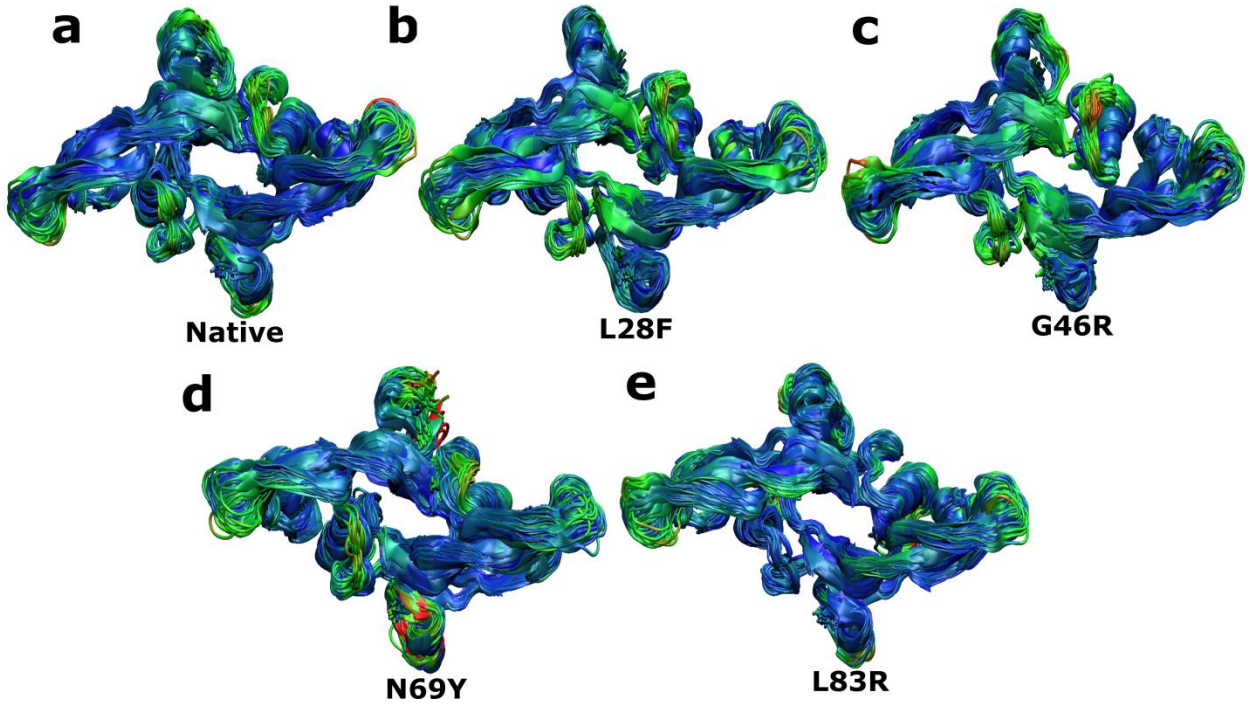


Figure 4.10 The superposed frames snapshots of 100ns simulation of *TGFβ1* dimer site mutants. **a)** Comparison of the structure deviation via superposition of 100 frames for the native structure **b)** Comparison of the structure deviation via superposition of 100 frames for the L28F mutant structure **c)** Comparison of the structure deviation via superposition of 100 frames for the G46R mutant structure **d)** Comparison of the structure deviation via superposition of 100 frames for the N69Y mutant structure **e)** Comparison of the structure deviation via superposition of 100 frames for the L83R mutant structure.

Further, the VMD tool was used to visualize the trajectories for all 10000 frames at 100ns for the native and four mutants [53]. This shows variations of the structure at different time frames through localized RMSD differences, as the red regions were displaying the regions with high deviation, green regions with intermediate deviations, and blue ones with slightest deviations (Figure 4.10). This shows that although no major changes are happening to the overall protein structure, the deviation has been found at the secondary structure regions (i.e. loop region, helix region) that have chances of structure conformational change thus altering functionality. Correlating the CASTp and MD results, it was observed that polar interaction was formed between the Tyr58 and Cys44 residue of the L28F mutant at 100ns time period as represented in Figure 4.6a. Since the hydrogen bond contributes towards the stability of protein structure, so

there is a possibility that the bonds newly created will lead to the structure stabilization. The MD simulation can observe and predict the structural effects of mutations; it provided possible insights into CRC's carcinogenesis through structure variation. The results presented here will be helpful for an experimental biologist to test the effect of SNPs and their subsequent phenotypes.

4.4 Conclusion

The complexity of the cancer signaling network presents a huge challenge to understand the interactions among the pathway modules. The pathway signaling processes regulate multiple distinct features which characterizes cancer. The components of the pathways, or upstream receptors, are so commonly mutated in pathway causing a variety of cancers. This confirms the need of the approaches targeting these modules as a whole with more cost-effectiveness and less time-taking alternatives in comparison to the traditional ones. The systems biology approaches have great potential for understanding the complex behavior of the signaling processes thus enhancing the drug discovery and development processes. In this study, comprehensive view of the pathway simulation (*TGFβ*) has been given along with the mutational event on *TGFβ1* to determine their mechanisms in CRC. It is anticipated that this pathway and structure level study will give a new insight into the computational and experimental biologists for designing therapeutics for the CRC. The network analysis identified six key genes i.e. *Smurf1*, *SMAD2, TGFβR1*, *SMAD1*, *SMAD6*, and *SMAD4* regulating the TGFβ pathway along with the new regulatory elements such as *TNFα, INHBA, LTBP1, TGFβ1*, *NODAL, PPP2CA*, and *ROCK1* that were found to be involved in the major processes in a pathway and thus should be considered experimentally to validate their role in CRC. Thus, computer simulations can quickly investigate different experimental conditions for the biological system of interest, and through the analysis, only the most relevant cases can be assessed easily thus saving time and money. From the structural study, the SNP's rs199946261, rs768250306, rs763943753, and rs541829714 were found to be the damaging and have a potential to hinder the functionality of the protein. The study provides a broad view of the disease to the biological researchers in terms of the pathway and the structural analysis. It is anticipated that this would help to develop effective methods for the disease eradication at the earlier stage.

REFERENCES

[1] A. C. Society, "Cancer Facts and Figures 2017.," *Atlanta, Ga: American Cancer Society, 2017,* January 18, 2017.

[2] M. Arnold, M. S. Sierra, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global patterns and trends in colorectal cancer incidence and mortality," *Gut,* pp. gutjnl-2015-310912, 2016.

[3] F. A. Haggar and R. P. Boushey, "Colorectal cancer epidemiology: incidence, mortality, survival, and risk factors," *Clin Colon Rectal Surg,* vol. 22, pp. 191-7, Nov 2009.

[4] K. Tariq and K. Ghias, "Colorectal cancer carcinogenesis: a review of mechanisms," *Cancer Biol Med,* vol. 13, pp. 120-35, Mar 2016.

[5] I. M. Hisamuddin and V. W. Yang, "Molecular genetics of colorectal cancer: an overview," *Current colorectal cancer reports,* vol. 2, pp. 53-59, 2006.

[6] S. D. Markowitz and M. M. Bertagnolli, "Molecular origins of cancer: Molecular basis of colorectal cancer," *N Engl J Med,* vol. 361, pp. 2449-60, Dec 17 2009.

[7] E. R. Fearon, "Molecular genetics of colorectal cancer," *Annu Rev Pathol,* vol. 6, pp. 479-507, 2011.

[8] A. Torgovnick and B. Schumacher, "DNA repair mechanisms in cancer development and therapy," *Front Genet,* vol. 6, p. 157, 2015.

[9] R. A. Ganai and E. Johansson, "DNA replication—a matter of fidelity," *Molecular cell,* vol. 62, pp. 745-755, 2016.

[10] T. A. Kunkel and D. A. Erie, "Eukaryotic Mismatch Repair in Relation to DNA Replication," *Annu Rev Genet,* vol. 49, pp. 291-313, 2015.

[11] A. Villanueva, C. Garcia, A. B. Paules, M. Vicente, M. Megias, G. Reyes, *et al.*, "Disruption of the antiproliferative TGF-β signaling pathways in human pancreatic cancer cells," *Oncogene,* vol. 17, pp. 1969-1978, 1998.

[12] W. M. Grady, L. L. Myeroff, S. E. Swinler, A. Rajput, S. Thiagalingam, J. D. Lutterbaugh, *et al.*, "Mutational inactivation of transforming growth factor β receptor type II in microsatellite stable colon cancers," *Cancer research,* vol. 59, pp. 320-324, 1999.

[13] J. Wang, B. Zhang, H. Wu, J. Cai, X. Sui, Y. Wang, *et al.*, "CD51 correlates with the TGF-beta pathway and is a functional marker for colorectal cancer stem cells," *Oncogene,* vol. 36, pp. 1351-1363, 2017.

[14] T. Kanamoto, U. Hellman, C. H. Heldin, and S. Souchelnytskyi, "Functional proteomics of transforming growth factor- β1- stimulated Mv1Lu epithelial cells: Rad51 as a target of TGFβ1- dependent regulation of DNA repair," *The EMBO journal,* vol. 21, pp. 1219-1230, 2002.

[15] Y. Yamamura, X. Hua, S. Bergelson, and H. F. Lodish, "Critical role of Smads and AP-1 complex in transforming growth factor-β-dependent apoptosis," *Journal of Biological Chemistry,* vol. 275, pp. 36295-36302, 2000.

[16] S. P. Fink, S. E. Swinler, J. D. Lutterbaugh, J. Massagué, S. Thiagalingam, K. W. Kinzler, *et al.*, "Transforming growth factor-β-induced growth inhibition in a Smad4 mutant colon adenoma cell line," *Cancer research,* vol. 61, pp. 256-260, 2001.

[17] H. Hanafusa, J. Ninomiya-Tsuji, N. Masuyama, M. Nishita, J.-i. Fujisawa, H. Shibuya, *et al.*, "Involvement of the p38 mitogen-activated protein kinase pathway in transforming

growth factor-β-induced gene expression," *Journal of Biological Chemistry,* vol. 274, pp. 27161-27167, 1999.

[18] B. J. Park, J. I. Park, D. S. Byun, J. H. Park, and S. G. Chi, "Mitogenic conversion of transforming growth factor-beta1 effect by oncogenic Ha-Ras-induced activation of the mitogen-activated protein kinase signaling pathway in human prostate cancer," *Cancer Res,* vol. 60, pp. 3031-8, Jun 01 2000.

[19] Z. Yan, X. Deng, and E. Friedman, "Oncogenic Ki-ras confers a more aggressive colon cancer phenotype through modification of transforming growth factor-β receptor III," *Journal of Biological Chemistry,* vol. 276, pp. 1555-1563, 2001.

[20] S. Edlund, S. Y. Lee, S. Grimsby, S. Zhang, P. Aspenström, C.-H. Heldin, *et al.*, "Interaction between Smad7 and β-catenin: importance for transforming growth factor β-induced apoptosis," *Molecular and cellular biology,* vol. 25, pp. 1475-1488, 2005.

[21] X. Guo, A. Ramirez, D. S. Waddell, Z. Li, X. Liu, and X.-F. Wang, "Axin and GSK3-β control Smad3 protein stability and modulate TGF-β signaling," *Genes & development,* vol. 22, pp. 106-120, 2008.

[22] P. Vizán, D. S. Miller, I. Gori, D. Das, B. Schmierer, and C. S. Hill, "Controlling long-term signaling: receptor dynamics determine attenuation and refractory behavior of the TGF-β pathway," *Sci. Signal.,* vol. 6, pp. ra106-ra106, 2013.

[23] T. MathWorks, "MATLAB and Statistics Toolbox Release 2012a,", *The MathWorks, Inc., Natick, Massachusetts, United States.,* 2012.

[24] S. Wernicke and F. Rasche, "FANMOD: a tool for fast network motif detection," *Bioinformatics,* vol. 22, pp. 1152-1153, 2006.

[25] S. T. Sherry, M.-H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, *et al.*, "dbSNP: the NCBI database of genetic variation," *Nucleic acids research,* vol. 29, pp. 308-311, 2001.

[26] S. A. Forbes, N. Bindal, S. Bamford, C. Cole, C. Y. Kok, D. Beare, *et al.*, "COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer," *Nucleic acids research,* vol. 39, pp. D945-D950, 2011.

[27] I. Adzhubei, D. M. Jordan, and S. R. Sunyaev, "Predicting functional effect of human missense mutations using PolyPhen-2," *Curr Protoc Hum Genet,* vol. Chapter 7, p. Unit7 20, Jan 2013.

[28] E. Capriotti, P. Fariselli, and R. Casadio, "I-Mutant2. 0: predicting stability changes upon mutation from the protein sequence or structure," *Nucleic acids research,* vol. 33, pp. W306-W310, 2005.

[29] Y. Choi and A. P. Chan, "PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels," *Bioinformatics,* p. btv195, 2015.

[30] B. Li, V. G. Krishnan, M. E. Mort, F. Xin, K. K. Kamati, D. N. Cooper, *et al.*, "Automated inference of molecular mechanisms of disease from amino acid substitutions," *Bioinformatics,* vol. 25, pp. 2744-2750, 2009.

[31] R. Calabrese, E. Capriotti, P. Fariselli, P. L. Martelli, and R. Casadio, "Functional annotations improve the predictive score of human disease-related mutations in proteins," *Hum Mutat,* vol. 30, pp. 1237-1244, 2009.

[32] J. Bendl, J. Stourac, O. Salanda, A. Pavelka, E. D. Wieben, J. Zendulka, *et al.*, "PredictSNP: robust and accurate consensus classifier for prediction of disease-related mutations," *PLoS computational biology,* vol. 10, p. e1003440, 2014.

[33] E. A. Stone and A. Sidow, "Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity," *Genome research,* vol. 15, pp. 978-986, 2005.

[34] Y. Bromberg and B. Rost, "SNAP: predict effect of non-synonymous polymorphisms on function," *Nucleic acids research,* vol. 35, pp. 3823-3835, 2007.

[35] N.-L. Sim, P. Kumar, J. Hu, S. Henikoff, G. Schneider, and P. C. Ng, "SIFT web server: predicting effects of amino acid substitutions on proteins," *Nucleic acids research,* vol. 40, pp. W452-W457, 2012.

[36] B. Reva, Y. Antipin, and C. Sander, "Predicting the functional impact of protein mutations: application to cancer genomics," *Nucleic acids research,* p. gkr407, 2011.

[37] G. De Baets, J. Van Durme, J. Reumers, S. Maurer-Stroh, P. Vanhee, J. Dopazo, *et al.*, "SNPeffect 4.0: on-line prediction of molecular and structural effects of protein-coding variants," *Nucleic acids research,* vol. 40, pp. D935-D939, 2012.

[38] S. Yin, F. Ding, and N. V. Dokholyan, "Eris: an automated estimator of protein stability," *Nature methods,* vol. 4, pp. 466-467, 2007.

[39] E. Capriotti, R. Calabrese, P. Fariselli, P. L. Martelli, R. B. Altman, and R. Casadio, "WS-SNPs&GO: a web server for predicting the deleterious effect of human protein variants using functional annotation," *BMC genomics,* vol. 14, p. S6, 2013.

[40] J. L. Sussman, D. Lin, J. Jiang, N. O. Manning, J. Prilusky, O. Ritter, *et al.*, "Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules," *Acta Crystallographica Section D: Biological Crystallography,* vol. 54, pp. 1078-1084, 1998.

[41] D. L. Theobald and P. A. Steindel, "Optimal simultaneous superpositioning of multiple structures with missing data," *Bioinformatics,* vol. 28, pp. 1972-1979, 2012.

[42] D. L. Theobald and D. S. Wuttke, "Accurate structural correlations from maximum likelihood superpositions," *PLoS Comput Biol,* vol. 4, p. e43, Feb 2008.

[43] W. L. DeLano, "The PyMOL molecular graphics system," 2002.

[44] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, *et al.*, "GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers," *SoftwareX,* vol. 1, pp. 19-25, 2015.

[45] S. W. Siu, K. Pluhackova, and R. A. Böckmann, "Optimization of the OPLS-AA force field for long hydrocarbons," *Journal of chemical theory and computation,* vol. 8, pp. 1459-1470, 2012.

[46] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, "Comparison of simple potential functions for simulating liquid water," *The Journal of chemical physics,* vol. 79, pp. 926-935, 1983.

[47] J. R. Dormand and P. J. Prince, "A family of embedded Runge-Kutta formulae," *Journal of computational and applied mathematics,* vol. 6, pp. 19-26, 1980.

[48] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, "Network motifs: simple building blocks of complex networks," *Science,* vol. 298, pp. 824-827, 2002.

[49] J. Liang, H. Edelsbrunner, and C. Woodward, "Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design," *Protein Sci,* vol. 7, pp. 1884-97, Sep 1998.

[50] T. Williams, C. Kelley, H. Bröker, J. Campbell, R. Cunningham, D. Denholm, *et al.*, "Gnuplot 4.6: An interactive plotting program, 2012," *URL http://www. gnuplot. info*.

[51] J. A. Marsh and S. A. Teichmann, "Relative solvent accessible surface area predicts protein conformational changes upon binding," *Structure*, vol. 19, pp. 859-867, 2011.

[52] A. C. Wallace, R. A. Laskowski, and J. M. Thornton, "LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions," *Protein Eng*, vol. 8, pp. 127-34, Feb 1995.

[53] W. Humphrey, A. Dalke, and K. Schulten, "VMD: visual molecular dynamics," *J Mol Graph*, vol. 14, pp. 33-8, 27-8, Feb 1996.

CHAPTER - 5

Overall Conclusions
&
Future Prospects

~Every new beginning comes from some other beginning's end.

- Seneca

5.1 CONCLUSIONS

In overall study, analysis has been done for the DNA repair associated malignancies (primarily colorectal cancer and also for endometrial cancer and Lynch syndrome) at an extensive level. The overall goal of this research was to decipher the key biomolecules such as genes and proteins involved in these malignancies at cellular and molecular level and to determine their structural and functional impact. The study also helped in finding the genetic association with disorder for better understanding of the complex molecular mechanism involved in these cancers. The overall research problems that have been implemented for our research work is given in Figure 5.1 along with the implied methodology and the processed outcome.

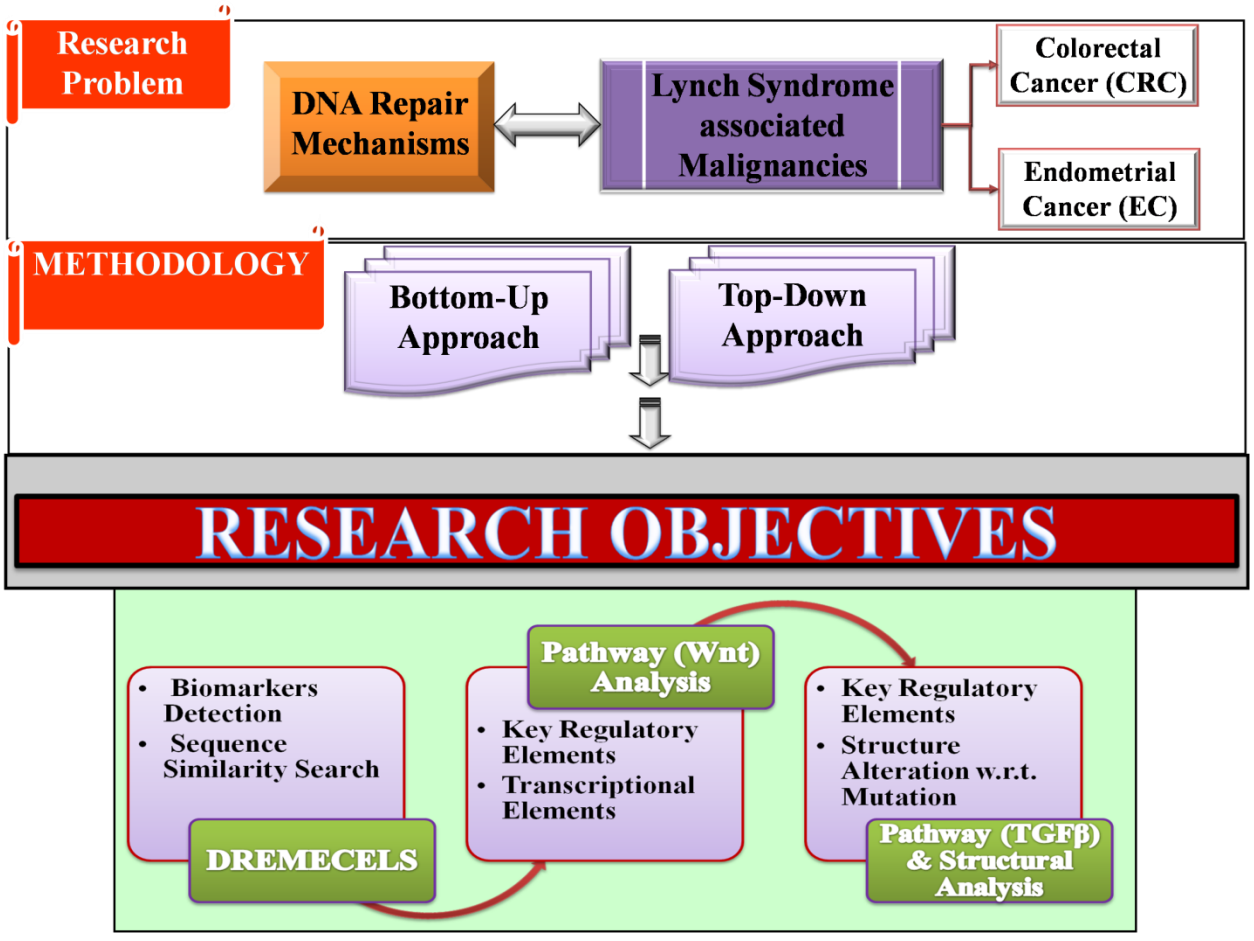


Figure 5.1 Representation of the overall applied approaches in the fulfillment of proposed objectives and their outcomes.

The study has vastly covered the colorectal cancer (CRC) in most of the studies, however in depth information is provided for the Lynch syndrome and the endometrial cancers in the first objective that covers the database. The critical role of DNA repair mechanisms in Lynch syndrome associated malignancies is deliberated through computational means. An extensive analysis has been done for the DNA repair genes/proteins, genetic variations, PPIs, intricate networks and pathways concerned with human DNA repair system. Taken as a whole, the study aimed at widespread assessment of these malignancies and immersed DNA repair mechanisms as manifested from the created database. Both top-down and bottom-up approaches have been utilized for deciphering candidate markers for CRC and Lynch syndrome allied cancer forms. Various biomolecules such as genes, TFs, proteins, vital regulatory elements and interactions among them have been elucidated at systems level for better understanding of the bioprocesses concerned with DNA repair.

Important findings of this research work are summarized as follows:

- The DREMECELS is imbibed with the data of 156 genes focusing on base excision and mismatch repair mechanisms as they are the major contributor towards disease. The database is featured with the parameters (such as genes, proteins, diseases, conserved domains, gene ontology, pathways, literature link, and transcription factors) that include a variety of regulatory processes having a role in the progression of a disease. The database also offers information regarding somatic mutations, copy number variation (CNV), miRNAs, methylation status, and about drug sensitivity that makes it a complete package of fully featured genomic descriptors embraced at one spot. This repository is unique and first of its kind as there is no such archive that comprises such detailed information specifically on the covered malignancies. The aim of the database is to provide integrated information of disease types to serve the scientific community, thus supporting the diagnostic and therapeutic processes development. The repository will not only serve the researchers working in this field but also serve as an exceptional auxiliary for biomedical professionals thus facilitate understanding of the critical diseases.

- The systems level approach for the Wnt signaling pathway help in understanding its complex behavior in CRC study. The study provided the detailed view to determine the role of potential biomarkers in CRC. The simulations were carried out to identify the dynamics of an individual component that helps to attain their behavioral role in colorectal carcinogenesis. Also, network motifs were determined to decipher the significant transcription factors or relevant regulatory elements. The standard statistical parameters such as *Z*-score, p-value, and significance profile were used to find the candidate genes in the pathway. Five key genes were found to be statistically significant i.e. *AXIN, APC, β-catenin, LEF1, and MYC*. It is estimated that these putative biomarkers could be efficient in disease diagnosis process and help to solve the mystery for the abnormal regulation of Wnt signaling in colorectal cancer.

- In our last objective, the comprehensive view of the pathway simulation (TGFβ) has been provided along with the mutational events for *TGFβ1* to determine their mechanisms in CRC. It is anticipated that this pathway and structure level study will give a new insight into the computational and experimental biologists for designing therapeutics for the CRC. The network analysis identified six key genes i.e. *Smurf1, SMAD2, TGFβR1, SMAD1*, *SMAD6*, and *SMAD4* regulating the TGFβ pathway along with few newly identified regulatory elements such as *TNFα, INHBA, LTBP1, TGFβ1, NODAL*, *PPP2CA*, and *ROCK1* that were found to involved in the major processes in a pathway and thus should be considered experimentally to validate their role in CRC. Thus computer simulations can quickly investigate different experimental conditions for the biological system of interest, and through the analysis, only the most relevant cases can be assessed easily thus saving time and money. From the structural study, the SNP's rs199946261, rs768250306, rs763943753, and rs541829714 were found to be the damaging and have a potential to hinder the functionality of the involved protein. It is anticipated that this computational analysis will provide a broad view of the disease to the biological researchers so that effective methods can be developed to eradicate this disease.

5.2 FUTURE PROSPECTS

- It is projected that this web based comprehensive resource would serve as a valuable accompaniment for retrieving crucial information for the CRC, endometrial cancer, and Lynch syndrome. The database is organized to provide clarity, ease of access and download, and fast browsing capability. Thus, it would save time and efforts of researchers involved in the field through easy accessibility to data, and will facilitate in biological discoveries. The database will assist researchers to study the gene markers in depth and will provide useful insight for future analysis and studies. This repository will also help for easy understanding and investigation of many other related disease and disorders and provide useful genetic information. The database will prove to be useful to the scientists aiming new therapeutic targets not only for these three forms of disease but also to the other complex forms like multiple myelomas, gastric cancers, breast cancers etc. The database will be updated on a regular basis to keep rationalized information to the academicians and researchers to perform the research in the right direction.

- The proposed five key genes (*AXIN, APC, β-catenin, LEF1,* and *MYC*) in the *Wnt* signaling pathway can be experimentally verified and the consequences of their impact on CRC could be observed. The analysis of these putative candidates could provide significant insights in the progression of CRC. Identification of these key genes and the manner they behave during cancer progression can make them plausible biomarkers for the disease conditions at various stages of the disease.

- The projected key genes of the *TGFβ* pathway and the four potential damaging SNP candidates (rs199946261, rs768250306, rs763943753, and rs541829714) may be authenticated experimentally thus saving time and resources; contributing to timely systematic understanding of CRC. The novel parameters planned for identifying key candidates in the pathway could be applied on other diseases also having known biological pathways. Thus, the prior information of the damaging SNPs not only aid in fast drug discovery process but help the biologist to take necessary actions to target the appropriate region of a protein instead of applying strategies randomly. Structural level mutational

studies will assist in computer aided drug design for several populations based study upon the selection of relevant SNP for a particular population involved.

APPENDIX

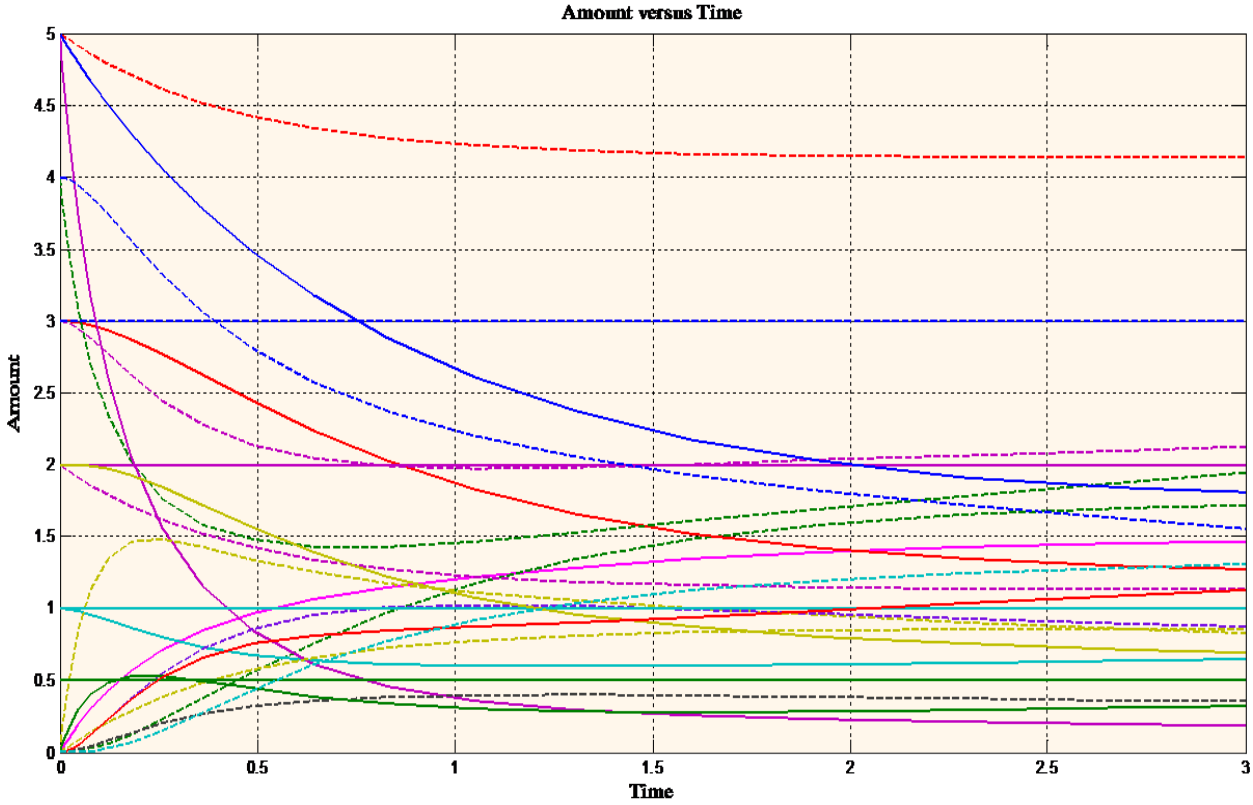


Figure 1. Simulations preformed via considering sets of entities.

Table 1. Entities, their types and associated concentration values

Entity	Type	Concentrations
Wnt	PROTEIN	0
Complex (Wnt/Frizzled)	COMPLEX	0
Frizzled	PROTEIN	3
Complex (Frizzled/Wnt/LRP5/6)	COMPLEX	0
LRP5/6	PROTEIN	3
Casein Kinase 1	PROTEIN	1
Casein Kinase 2	PROTEIN	3
ATP	SIMPLE_MOLECULE	1
ADP	SIMPLE_MOLECULE	0
Glycogen Synthase Kinase-3Beta (GSK3β)	PROTEIN	0
Diversin	PROTEIN	1
Complex (Ebi/Siah-1)	COMPLEX	3
Complex (Siah-1/Ebi)	COMPLEX	0
FRAT	PROTEIN	3
Complex (Dishevelled/Beta-Arrestin/Frodo)	COMPLEX	3
Complex (Axin/PP2A/APC)	COMPLEX	4
Complex (APC/Axin/Diversin/Casein Kinase 1/Glycogen Synthase Kinase-3Beta/PP2A)	COMPLEX	0
Complex (APC/Axin/PP2A)	COMPLEX	0

Pygo	PROTEIN	2
CBP	PROTEIN	2
SWI/SNF	PROTEIN	2
Bcl9	PROTEIN	2
Wnt Target Genes	GENE	0
Complex (TCF/Smad4)	COMPLEX	4
Wnt	PROTEIN	5
β-catenin	PROTEIN	5
Complex (APC/Axin/ β-catenin /PP2A)	COMPLEX	0
Complex (APC /Axin/Diversin/ β-catenin /PP2A)	COMPLEX	0
Complex (APC /Axin/PP2A/β-catenin /Siah-1/Ebi)	COMPLEX	0
β-catenin	PROTEIN	0
Complex(TCF/Smad4/β-catenin)	COMPLEX	0
Complex(TCF/β-catenin/Smad4/Bcl9)	COMPLEX	0
Complex(Bcl9/ β-catenin /TCF/Smad4/Pygo)	COMPLEX	0
Complex(β-catenin /TCF/Smad4/Bcl9/Pygo/SWI/SNF)	COMPLEX	0
Complex(APC/_Beta_-Catenin/Axin/PP2A/Diversin/Casein Kinase 1)	COMPLEX	0
Complex (APC/β-catenin/Glycogen Synthase Kinase-3 Beta/Axin/PP2A/Diversin/Casein Kinase 1)	COMPLEX	5
Ubiquitin	PROTEIN	1
β-catenin	PROTEIN	0
β-catenin	PROTEIN	0
Complex (APC/ β-catenin /Siah-1/Ebi/Axin/PP2A)	COMPLEX	0
Complex (APC/ β-catenin /Axin/PP2A)	COMPLEX	0
Complex (Dishevelled/Beta-Arrestin/Frodo)	COMPLEX	0
Complex (Dishevelled/Beta-Arrestin/Frodo/Casein Kinase 2)	COMPLEX	0
Complex (Dishevelled/Casein Kinase 2/Beta-Arrestin/Frodo/FRAT)	COMPLEX	0
beta-TrCP	PROTEIN	2
Complex (APC/Axin/PP2A/Diversin/Casein Kinase 1/β-catenin/beta-TrCP/Glycogen Synthase Kinase-3 Beta)	COMPLEX	0
Complex (APC/Axin/PP2A/Diversin/Casein Kinase 1/β-catenin/beta-TrCP/Glycogen Synthase Kinase-3 Beta)	COMPLEX	0
Complex (beta_TrCP/β-catenin)	COMPLEX	0
Complex (Bcl9/Pygo/../Smad4)	COMPLEX	0

Table 2. Network motif generated through FANMOD

Subgraph_Node	Subgraph_ID	Adj-Matrix	Frequency	Mean-Freq	Standard-Dev	Z-Score	p-Value	Network_ID	Significance_Profile
			[Original]	[Random]	[Random]				
3	**38**	**000100110**	**4.86%**	**2.18%**	**0.0083081**	**3.2253**	**0.004**	**3a**	**1**
4	28	0000000000011100	11.53%	7.88%	0.0080639	4.5234	0	4a	0.562009898
4	**2116**	**0000100001000100**	**21.78%**	**16.86%**	**0.012492**	**3.9323**	**0**	**4b**	**0.48856867**
4	2188	0000100010001100	5.67%	2.22%	0.0094928	3.6363	0.001	4c	0.451792146
4	**2118**	**0000100001000110**	**1.56%**	**0.61%**	**0.0032806**	**2.8671**	**0.012**	**4d**	**0.356222881**
4	2184	0000100010001000	19.67%	16.87%	0.010282	2.7227	0.005	4e	0.338281901
5	**74252**	**0000000010010001000001100**	**0.99%**	**0.09%**	**0.00065832**	**13.759**	**0**	**5a**	**0.548901545**
5	533260	0000010000010001100001100	0.19%	0.02%	0.00021739	8.1963	0	5b	0.32698314
5	**1080**	**0000000000000010000111000**	**7.88%**	**3.65%**	**0.0063326**	**6.6786**	**0**	**5c**	**0.266436026**
5	60	0000000000000000000111100	3.43%	1.39%	0.0036484	5.6019	0	5d	0.223482163
5	**533004**	**0000010000010001000001100**	**1.04%**	**0.20%**	**0.001534**	**5.4924**	**0**	**5e**	**0.219113771**
5	541208	0000010000100001000011000	3.82%	1.16%	0.0050724	5.2441	0	5f	0.209208092
5	**1176**	**0000000000000010010011000**	**4.26%**	**1.19%**	**0.0060048**	**5.099**	**0.002**	**5g**	**0.203419474**
5	532748	0000010000010000100001100	1.98%	0.56%	0.0028009	5.0617	0	5h	0.201931428
5	**147724**	**0000000100100000100001100**	**1.77%**	**0.41%**	**0.0027244**	**4.9867**	**0.004**	**5i**	**0.198939378**
5	147740	0000000100100000100111100	0.22%	0.02%	0.00041141	4.7113	0.004	5j	0.187952572
5	**533020**	**0000010000010001000011100**	**0.31%**	**0.04%**	**0.00060412**	**4.5542**	**0.003**	**5k**	**0.181685226**
5	17178	0000000000100001100011010	0.29%	0.03%	0.00060467	4.3527	0.006	5l	0.173646586
5	**541460**	**0000010000100001100010100**	**0.12%**	**0.02%**	**0.00025109**	**4.1998**	**0.007**	**5m**	**0.167546795**
5	541212	0000010000100001000011100	0.29%	0.04%	0.00067738	3.7121	0.013	5n	0.148090494
5	**67632**	**0000000010000100000110000**	**6.29%**	**2.93%**	**0.0092336**	**3.6371**	**0.004**	**5o**	**0.145098444**
5	532744	0000010000010000100001000	8.34%	5.44%	0.0081163	3.5779	0.001	5p	0.14273672
5	**532520**	**0000010000010000000101000**	**9.94%**	**5.69%**	**0.011985**	**3.5433**	**0.001**	**5q**	**0.141356388**
5	30	0000000000000000000011110	0.36%	0.16%	0.00064787	3.1969	0	5r	0.127537108
6	**560140**	**000000000000000010001000110**	**0.58%**	**0.01%**	**0.00016705**	**34.08**	**0**	**6a**	**0.358767602**
6	142673428	000000001000100000010000011	0.24%	0.00%	7.11E-05	33.028	0	6b	0.347692986
6	**135397656**	**000000001000000100100000000**	**1.26%**	**0.05%**	**0.0004749**	**25.468**	**0**	**6c**	**0.268107211**
6	37816348	000000000010010000010000100	0.08%	0.00%	2.95E-05	25.257	0	6d	0.265885968
6	**142740538**	**000000001000100000100000110**	**0.03%**	**9.89E-07**	**1.27E-05**	**24.607**	**0**	**6e**	**0.259043276**
6	142675002	000000001000100000010000110	0.03%	8.96E-07	1.39E-05	22.597	0.001	6f	0.237883566
6	**135397912**	**000000001000000100100000001**	**1.38%**	**0.07%**	**0.00068507**	**19.018**	**0**	**6g**	**0.200206649**
6	142673170	000000001000100000010000010	0.09%	0.00%	4.65E-05	18.488	0	6h	0.194627223
6	**134809628**	**000000001000000010010000100**	**0.04%**	**0.00%**	**2.62E-05**	**16.456**	**0**	**6i**	**0.173235914**
6	135398936	000000001000000100100000011	0.08%	0.00%	4.79E-05	15.343	0	6j	0.16151912
6	**1116714**	**000000000000000100010000101**	**0.04%**	**0.00%**	**2.55E-05**	**14.485**	**0**	**6k**	**0.152486766**
6	37882418	000000000010010000100000101	0.04%	0.00%	3.03E-05	12.29	0	6l	0.12937952
6	**142641676**	**000000001000100000001000101**	**0.25%**	**0.01%**	**0.00020669**	**11.166**	**0**	**6m**	**0.117546926**
6	541199384	000000100000010000100000110	0.23%	0.01%	0.00019945	11.003	0	6n	0.11583099
6	**545326354**	**000000100000010000001000010**	**0.10%**	**0.00%**	**8.93E-05**	**10.822**	**0**	**6o**	**0.113925563**
6	541132828	000000100000010000010000100	0.06%	0.00%	6.07E-05	10.058	0.001	6p	0.105882768
7	**5.67071E+11**	**000000000100001000000100000**	**0.02%**	**3.18E-08**	**8.00E-07**	**208.63**	**0**	**7a**	**0.382715884**
7	5.67004E+11	000000000100001000000010000	0.04%	1.66E-07	2.62E-06	159.33	0	7b	0.292278799
7	**570699804**	**000000000000000000010001000**	**0.05%**	**7.93E-07**	**4.87E-06**	**102.66**	**0**	**7c**	**0.188321983**
7	5.67004E+11	000000000100001000000010000	0.01%	2.62E-08	8.28E-07	100.77	0	7d	0.184854921
7	**2.20775E+12**	**000000010000000100000100000**	**0.01%**	**5.13E-08**	**1.01E-06**	**99.194**	**0**	**7e**	**0.181963868**
7	302287968	000000000000000000001001000	0.05%	0.00%	5.49E-06	94.046	0	7f	0.172520253
7	**1.46097E+11**	**000000000001000100000010000**	**0.01%**	**2.81E-08**	**8.89E-07**	**93.785**	**0**	**7g**	**0.172041468**
7	1.46164E+11	000000000001000100000100000	0.01%	3.52E-08	1.11E-06	90.032	0	7h	0.165156874
7	**1.46097E+11**	**000000000001000100000010000**	**0.03%**	**7.10E-07**	**3.64E-06**	**82.439**	**0**	**7i**	**0.151228092**
7	2282760314	000000000000000001000100000	0.01%	6.19E-08	1.07E-06	78.014	0	7j	0.143110765
7	**1.46097E+11**	**000000000001000100000010000**	**0.02%**	**3.01E-07**	**2.71E-06**	**73.826**	**0**	**7k**	**0.135428197**
7	2215906388	000000000000000001000010000	0.01%	1.69E-07	1.46E-06	68.628	0	7l	0.12589286
7	**302271584**	**000000000000000000001001000**	**0.15%**	**0.00%**	**2.25E-05**	**67.246**	**0**	**7m**	**0.123357686**
7	17314882134	000000000000001000000100000	0.01%	6.53E-08	1.28E-06	65.022	0	7n	0.119277927
7	**2.20124E+12**	**000000010000000010000100000**	**0.01%**	**5.47E-08**	**1.31E-06**	**63.686**	**0**	**7o**	**0.116827136**
7	302255200	000000000000000000001001000	0.90%	0.01%	0.00014428	61.983	0	7p	0.113703112
7	**71001705010**	**000000000001000010001000000**	**0.01%**	**1.61E-07**	**1.61E-06**	**61.915**	**0**	**7q**	**0.113578371**
7	2229872	000000000000000000000000000	0.20%	0.00%	3.25E-05	61.309	0	7r	0.11246671
7	**2.77035E+11**	**000000000010000001000000010**	**0.10%**	**0.00%**	**1.64E-05**	**60.734**	**0**	**7s**	**0.111411916**
7	2215905492	000000000000000001000010000	0.01%	2.03E-07	1.69E-06	58.97	0	7t	0.108175992
7	**2.83603E+11**	**000000000010000100000100000**	**0.02%**	**7.45E-07**	**3.64E-06**	**54.869**	**0**	**7u**	**0.100653018**

Table 3. Statistically significant entities with their biological annotation and literature references

Gene	Description	Molecular Function	Reference
Axin	Axis inhibitor, Axin1	signal transducer activity, GTPase activator activity, protein binding, beta-catenin binding, enzyme binding	9601641, 10644691, 19759537
β – catenin/CTNNB1	Catenin Beta 1	RNA polymerase II transcription factor binding, DNA binding, chromatin binding	18936100, 19443654
APC	Adenomatous Polyposis Coli	protein binding, beta-catenin binding, microtubule binding, protein kinase binding	8638126, 7890674, 11166179, 11972058
LEF1	Lymphoid Enhancer Binding Factor 1	transcriptional activator activity, RNA polymerase II transcription regulatory region sequence-specific binding	23001182
Myc	Frequently Rearranged In Advanced T-Cell Lymphomas 1	transcriptional activator activity, RNA polymerase II core promoter proximal region sequence-specific binding, DNA binding	10723141, 10597290, 9924025
MMP7	Matrix Metallopeptidase 7	metalloendopeptidase activity, serine-type endopeptidase activity, heparin binding, zinc ion binding	21207220, 25677090
GSK3β	Glycogen Synthase Kinase 3 Beta	RNA polymerase II transcription factor binding, p53 binding, protein kinase activity	20864106, 14744935, 22988876
NLK	Nemo Like Kinase	magnesium ion binding, protein serine/threonine kinase activity, protein binding	10863097, 15764709
Jun	Jun Proto-Oncogene, AP-1 Transcription Factor Subunit	RNA polymerase II distal enhancer sequence-specific DNA binding, RNA polymerase II transcription factor activity, sequence-specific DNA binding, transcription factor activity, RNA polymerase II core promoter proximal region sequence-specific binding	19861239, 21113145, 19962668
Pontin52/ RUVBL1	RuvB Like AAA ATPase 1	DNA helicase activity, protein binding, ATPase activity	9843967, 10966108
DVL1	Dishevelled Segment Polarity Protein 1	frizzled binding, protein binding, enzyme binding	19388021, 10330181, 15454084
β – TrCP/FBXW11	F-Box And WD Repeat Domain Containing 11	ubiquitin-protein transferase activity, protein binding	14532120

PUBLICATIONS & PRESENTATIONS

"Life doesn't require that we be the best, only that we try our best." —*H. Jackson Brown Jr.*

Research Publications

- **Ankita Shukla**, Manika Sehgal, Tiratha Raj Singh. Hydroxymethylation and its potential implication in DNA repair system: A review and future perspectives. ***Gene. 2015 Jun 15; 564(2):109-118.***

- **Ankita Shukla**, Ahmed Moussa, and Tiratha Raj Singh. DREMECELS: A Curated Database for Base Excision and Mismatch Repair Mechanisms Associated Human Malignancies. ***PLoS One. 2016 Jun 8; 11(6):e0157031.***

- **Ankita Shukla**, Ragothaman Yennamalli, and Tiratha Raj Singh. Study of Network and Structure Based Inference of Functional Single Nucleotide Polymorphisms of TGFβ1 Gene and its Impact on Colorectal Cancer (CRC). ***GeneReports. 2018 Jun ,11; 131-142.***

- **Ankita Shukla**, and Tiratha Raj Singh. Network-Based Approach to Study Dynamics of Wnt Pathway Regulatory Elements in Colorectal Cancer. ***Network Modeling Analysis in Health Informatics and Bioinformatics. 2018 Sep;7(1):14.***

Conference Publications

- **Ankita Shukla**, Manika Sehgal, Tiratha Raj Singh (2014). Hydroxymethylation and its Role in DNA Repair Associated Disease. ***Proceedings of the International Conference on Life Sciences, Informatics, Food and Environment [Noida, India: 29-30, August 2014].***

- **Ankita Shukla**, Ashwani Kumar, Manika Sehgal, Tiratha Raj Singh (2015). Modeling and Simulation of Colorectal Cancer Pathways: An Insight into Regulatory Elements and Biological Processes. ***Proceedings of the Annual International Conference on Advances in Biotechnology [5th: Kanpur, India: 13-15 March 2015], pp.85-89.***

- **Ankita Shukla**, Ragothaman Yennamalli, Tiratha Raj Singh (2017). Structure Based Inference of Functional Single Nucleotide Polymorphism "L28F" and to Determine

its Role in TGFβ1 Allied Colorectal Cancer (CRC). ***Proceedings of the International Journal of Bioinformatics Research and Applications [Inbix'17].***

- **Ankita Shukla,** Tiratha Raj Singh (2018). Role of Androgen Receptor (AR) pathway in Prostate Cancer through Gene Expression Studies. ***Proceedings of the Biotechnology International (BTI) [Chandigarh, India: 4-6, April 2018].***

Book Chapters

- **Ankita Shukla** and Tiratha Raj Singh. "Computational Network Approaches and Their Applications for Complex Diseases." Translational Bioinformatics and Its Application. ***Springer Netherlands, 2017. 337-352.***

- Tiratha Raj Singh, **Ankita Shukla**, Bensellak Taoufik, Ahmed Moussa and Brigitte Vannier. "Metabolome Analysis: A Disease Biomarker Discovery". ***Encyclopedia of Bioinformatics and Computational Biology. vol. 3, pp. 476–488. Oxford: Elsevier.***

General Publications

- Tiratha Raj Singh, **Ankita Shukla** (2017). Bioinformatics to Systems Biology: A Journey of Knowledge Discovery. ***CSI Communications, 41 (4), 10-13***.