Evolution of AP Endonucleases: An *In-Silico* Genomic and Modelling study

THESIS

Submitted in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY

By

SWARNA KANCHAN
2007PHXF007P

Under the Supervision of
Prof. SHIBASISH CHOWDHURY

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE
PILANI (RAJASTHAN)
INDIA

2015

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE
PILANI (RAJASTHAN)

CERTIFICATE

This is to certify that the thesis entitled "**Evolution of AP Endonucleases: An *In-Silico* Genomic and Modelling study**" submitted by **Swarna Kanchan, ID No. 2007PHXF007P** for award of Ph. D. degree of the Institute embodies original work done by him under my supervision.

Signature in full of the supervisor :

Name in capital block letters : **SHIBASISH CHOWDHURY**

Designation : **Associate Professor**

Date:

ACKNOWLEDGEMENT

It's a matter of immense pleasure for me to acknowledge the efforts and blessings of one and all whose great endeavor culminated in producing this research work as a pious and fruitful attempt in investigating the intriguing facts of science. I devote my first and foremost thanking prayers to the Almighty God for indispensable blessings and grace to bestow me the opportunity to work in the renowned premises of Birla Institute of Technology and Science, Pilani founded by a great legend Ghanshyam Das Birla.

I feel blessed to grab the proud privilege of working under the exemplary guidance of my reverent supervisor Prof. Shibasish Chowdhury, Associate Professor, Department of Biological Sciences, Birla Institute of Technology and Science, Pilani, India. It would not be extravagant if I remain indebted to him all my life for his elusive guidance. His precious suggestions, consistent encouragement and moral support during the ups and downs of the path traveled and his inexhaustible efforts throughout the entire course of investigation remain intangible.

I am grateful to Prof. V. S. Rao, Vice-Chancellor (BITS, Pilani) and Dr. A. K. Sarkar, Director (Pilani campus), for allowing me to carry out my doctoral research work in the institute.

I extend my hearty thanks to Dr. Rajesh Mehrotra, Head of the Department of Biological Sciences, BITS, Pilani, and Prof. A. K. Das for kindly accepting to be my DAC member and helping me in all possible ways to provide all necessary facilities and guidance to fabricate the experimental works of my research. I dedicate my heartiest gratitude to departmental research committee convenor Prof. Jitendra Panwar for his consistent support and guidance.

I am extremely thankful to Prof. S. K. Verma, Dean, Academic Research Division, BITS, Pilani, for his co-operation and for providing me with all the necessary academic facilities and helping me relentlessly during my research work.

I also take this opportunity to express my sincere gratitude towards Prof. Dr. Rajeshwar P Sinha, Professor in Department of Botany, Banaras Hindu University, Varanasi, for his guidance, blessings and moral support.

I dedicate my heartfelt gratitude to Prof. Lalita Gupta, Prof. Sanjeev Kumar, Dr. Pankaj Kumar Sharma, Dr. Prabhat Nath Jha, Dr. Rajdeep Chowdhury, Dr. Sandhya Marathe, Dr. Sandhya Mehrotra, Dr. Santosh Kumar Padhi, Dr. Shilpi Garg, Dr. Sudeshna Mukherjee Chowdhury, Dr. Uma S Dubey, Dr. B.Vani, Dr. Vishal Saxena and Mr. Manoj Kannan. I also dedicate my sincere gratitude to Dr. Saumi Ray, Dr. Dilip Maithi, Dr. Chandra Shekhar, Mrs. Shikha Gupta, Dr. Satish Kumar Dubey, Mrs. Ranjana Dubey and Dr. Pramila Jha for their true blessings.

I am highly grateful to all the research scholars especially Aashish Runthala, Parva Sharma and Divya Niveditha and my friends Pallavi Sinha and Nidhi Gupta for their cooperation and moral support.

I express my thanks to the office staff members of Department of Biological Sciences for their help and cooperation.

I wish to dedicate my sincere gratitude and indebtness to my mother (Late) Mrs. Chandrakala Kesheri and father Dr. Ram Bilash Kesheri for their untiring efforts, moral support and immense patience without which the completion of my work in stipulated time could not be in its present concrete form. It would not be extravagant to thank my wife Dr. Minu Kesheri, for believing in me and being my moral support. I also count upon the support of my younger brother Dr. Saurabh Suman. I express my sincere thanks to my mother-in-law Mrs. Nayan Tara Kesheri, father-in-law Mr. Satya Narain Kesheri and sister-in-law Er. Deepti Kesheri for their affection and incessant blessings. Last but not the least I am thankful to my loving son Adarsh Kesheri for bearing with us and being a source of happiness during the entire work.

I deeply acknowledge the financial assistance provided by University Grants Commission in the form of Basic Science Research Fellowship. I am also grateful to Birla Institute of Technology and Science, Pilani for providing institute fellowship, required resources and support.

Swarna Kanchan

ABSTRACT

DNA repair processes restore the normal nucleotide sequence and DNA structure after damage. In spite of the critical need for DNA repair, ‘‘evolvability’’ i.e., the ability to generate a certain level of mutations also seems to be selected in the course of evolution. Organisms with an optimal level of evolvability have the best chance to survive due to the virtue of having variation in its genome, which provides the raw material for natural selection. The complex interplay between the two opposing forces, namely the need for fidelity of transmission of genetic information and the need for evolvability, seems to define the organization of the repair system. Thus, the study of evolution of DNA repair system can throw light on the role of environment on evolvability. Among the different repair mechanisms, BER is the prominent pathway for repair of small DNA lesions in which AP endonuclease protein family plays a crucial role. AP endonuclease protein family mainly includes endonuclease III, endonuclease IV and exonuclease III proteins. In the present study, the evolution mechanisms of endonuclease III, endonuclease IV and exonuclease III proteins in all the three domains of life are investigated. In addition to that we have analyzed the sequences of different DNA repair proteins from four different DNA repair pathways belonging to six model organisms to understand the unified evolution mechanism of different repair systems.

In each case, we have identified evolutionary conserved residues, domains and motifs and crucial events during evolution process. We observe that horizontal gene transfer is one of the most common event within repair proteins. However, synonymous nucleotide substitution of the GC content at third codon position as well as the overall GC content of gene makes gene sequence-based phylogenetic tree construction noisy. On the basis of these analysis we also propose models for evolutionary history of these proteins.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

5'dRP	5'Deoxy Ribose Phosphate
6-4PP	6-4 photoproducts
8-OH-G	8-hydroxyguanine
AGT	Alkyl Guanine Transferase
AP	Apurinic/Apyrimidinic
Ape1	Apurinic/Apyrimidinic Endonuclease 1
Ape2	Apurinic/Apyrimidinic Endonuclease 2
BER	Base Excision Repair
BLASTP	Basic Local Alignment Search Tool Protein
CD	Conserved Domain
COG	Clusters of Orthologous Groups
CPD	Cyclobutane Pyrimidine Dimers
CS	Cockayne's Syndrome
DPD	Divergence Prior to Duplication
DR	Direct Repair
E value	Expect value
EMBOSS	European Molecular Biology Open Software Suite
FAD	Flavin Adenine Dinucleotide
FCL	Iron–sulfur Cluster Loop
GGR	Global Genomic Repair
HGT	Horizontal Gene Transfer
HhH	Helix Hairpin Helix
HNPCC	Hereditary Non Polyposis Colon Cancer

INDEL	INsertion or DELetion
JTT	Jones Taylor Thornton
LCA	Last common ancestor
LP-BER	Long-Patch - Base Excision Repair
MGMT	Methyl Guanine Methyl Transferase
ML	Maximum Likelihood
MMR	MisMatch Repair
MMS	Methyl Methane Sulfonate
mRNA	Messenger RNA
MSA	Multiple Sequence Alignment
NCBI	National Center for Biotechnology Information
NER	Nucleotide Excision Repair
NJ	Neighbor Joining
NR	Non Redundant
O6-meG	O6-methylguanine
PARP-1	Poly ADP Ribose Polymerase-1
PDB	Protein Data Bank
PFAM	Protein FAMily
PIR	Protein Information Resource
Pol ß	Polymerase ß
PROCHECK	PROtein Structure CHECK
PSI-BLAST	Position Specific Iterative - Basic Local Alignment Search Tool
PSSM	Position Specific Score Matrix
RMSD	Root Mean Square Deviation
ROS	Reactive Oxygen Species

SCOP	Structural Classification of Protein
SMART	Simple Modular Architecture Research Tool
SP-BER	Short Patch - Base Excision Repair
SSB	Single Strand Breaks
TCR	Transcription Coupled Repair
TTD	Tricho Thio Dystrophy
UDGs	Uracil DNA glycosylases
UPGMA	Unweighted Group Method with Arithmatic Mean
UV	Ultraviolet Rays
UVSS	Ultra Violet Sensitive Syndrome
XP	Xeroderma Pigmentosum
XPD	Xeroderma Pigmentosum group-D

Chapter I

Introduction

Chapter I

1.1 Evolution

Evolution can be defined as descent with modification in which genetic material of a population changes. These modifications take place over long period of time. During the process of evolution simpler forms of life usually changes into more complex form. In most cases evolution leads to adaptation of organisms to their changing environment. Biological evolution may be of two types: small-scale or micro evolution and large-scale or macro evolution. Microevolution refers to changes in DNA sequences and gene frequencies which are mainly due to the mutation and large-scale evolution refers to the descent of species from a common ancestor which is mainly due to the accumulation of many micro-scale evolutions over a long period of time. Along with mutation and migration, natural selection and genetic drift are two major elements that drive evolution. Natural selection is the processes by which individual organisms are adapted to the environment with favorable traits Theses traits are passed from one generation to another and hence beneficial heritable traits become more common in offsprings in successive generations (Kondrashov 1988). During this process variation within genetic makeup is created through mutation and genetic recombination. On the other hand, genetic drift refers to gene frequency change due to random sampling of organisms. However, both natural selection and genetic drift can lead to evolutionary changes if genetic variation exists within population.

1 .2 Genetic variation

During the process of genetic variation, evolution of an entirely new gene without any functional progenitor has been regarded as an exceedingly rare and improbable event (Andersson et al. 2015). Gene flow (migration), sexual reproduction and mutation within DNA sequence are major

source of genetic variation. While during migration, individual or genetic material that one individual carries move from one population to another, a large scale genetic shuffling occurs during sexual reproduction. Although single mutation in DNA sequence can cause huge effect, generally many mutations within DNA sequence produces evolutionary change. Organisms acquire new genes by *de novo* origin from noncoding sequences, duplication/fusion, horizontal gene transfer, endosymbiotic transfer as well as mutational event and thereafter divergence from one copy of the existing genes (Alder et al. 2014) provides the genetic variation. Thus, large scale DNA recombination (through duplication, horizontal gene transfer and endo symbiotic transfer) and many mutational events primarily provide genetic variation within population which is a key to the evolution process.

1.2.1 Gene duplication and gene fusion

The genetic variation, thus evolution through gene duplication process played an important role in the diversity of living individual (Hughes 2005; Megadum et al. 2013). Gene duplication generates functional redundancy, whereby one copy of the gene accumulates mutations and finally becomes a pseudogene, which is either deleted or becomes much diverged from the parental genes (Sankoff 2001). Only the young pseudogenes can be identified because of sequence similarity. The most important contribution of gene duplication is to provide new genetic material for mutation, drift and natural selection to act upon which finally emerge as gene with new functions. Duplication is an effective method to increase complexity during the course of evolution (Vision et al. 2000). Duplication also makes organism fit to survive in different habitat due to the increased fitness of the involved organisms by doubling gene dosage

or neo-functionalization (Pinheiro et al. 2004). Although gene duplication within organism may increase its fitness yet in many cases, it may also result in a simple division of ancestral functions into daughter genes, which need not promote adaptation (Qian & Zhang 2014). Three basic models of gene duplication event are known till date. These are Ohno's Model, sub-functionalization model and divergence prior to duplication (DPD) model (Soskine & Tawfik 2010). In Ohno's model (Ohno 1973), one copy of a duplicated gene is considered as completely redundant copy of the gene and is kept free from functional responsibility. This redundant copy can accumulate mutations at random. Occasionally, by sheer chance, these randomly accumulated mutations will give rise to a protein with a new useful role. Sub-functionalization model (Rastogi & Liberles 2005) postulates that duplicated genes are retained largely with the help of neofunctionalization. In both Ohno's model and the sub-functionalization model, duplication is considered as a neutral event which is not selected. By contrast, the 'divergence prior to duplication' (DPD) model considers that duplication is positively selected and delivers immediate advantage. Gene duplication has become a common event in bacteria, archaea and eukaryotes in which most of genes were generated by gene duplication. Gene duplication may take place by unequal crossing over, retro-position, or chromosomal (or genome) duplication. Unequal crossing over is responsible for tandem gene duplication in which duplicated genes are linked in a chromosome. Retro-position is a common mechanism in which messenger RNA (mRNA) is retro transcribed to complementary DNA first then it is inserted into the genome. Chromosomal/ genome duplication takes place by lack of disjunction among daughter chromosomes after DNA replication. Chromosomal/genome duplication is frequent in plants (Li 1997; Hughes 1994). Duplication takes place in an individual which may be either heritable or lost in the population. If a new allele which contains duplicate genes is selected, it only has a

small probability of 1/2N of being fixed in a diploid population in which N is the effective population which suggests that many duplicated genes become lost in the population. If the new allele becomes fixed, it takes too much time and nearly 4N generations for a neutral allele to become fixed in the population (Zhang 2003).

1.2.2 Horizontal Gene Transfer (HGT)

Horizontal gene transfer is the transfer of gene among different species i.e. from bacteria to archaea or eukaryotes and vice versa. Genomic sequencing of prokaryotes suggested horizontal gene transfer as one of the important evolutionary mechanism among prokaryotes. Horizontal gene transfer from prokaryotes to eukaryotes as well as transfer of genetic information/genes from mitochondria and chloroplasts to the nuclear genome (Zhang 2003) is also very common event. It has been shown that eukaryotic nuclear genome mostly contains prokaryotic archaeal and bacterial sequences (Hall et al. 2005). Among several methods, phylogenetic tree and local pairwise sequence analysis are two popular tools to detect horizontally transferred genes. One important method Basic local alignment search tool **(**BLAST) identified HGT events among bacteria, archaea and eukaryotes (Virel & Backman 2007; Takahashi et al. 2008, Santos et al. 2008; Altschul et al. 1997). HGT also plays an important role in emergence and spread of virulence as well as resistance to antibiotics.

1.2.3 Endosymbiotic transfer of genes

Many genes in eukaryotes are found to transfer from organellar genomes to the nucleus. Most of these genes have become functional genes in nucleus which were found essential for biogenesis of mitochondria/chloroplasts, while other genes were found to control other cellular processes. Sequence comparisons and many other evidences showed that cyanobacteria and proteobacteria are the ancestors of chloroplasts and mitochondria, respectively (Gray & Doolittle 1982). Many proteins which are encoded by the nuclear genome were found essential to chloroplasts and mitochondria (Bogorad 1975; Ellis 1981), suggesting that these genes were transferred from the genome of the organelles to the nucleus genome during evolution (Weeden 1981). The closest relatives of organelles like mitochondrial and chloroplast genomes are α-proteobacteria and cyanobacteria respectively (Gray et al. 1999; Martin et al. 2002).

1.2.4 Mutations

Mutation is considered as change in DNA sequence which is considered as hereditary material of life. However, all mutations are not considered as a part of evolution as a particular mutation can happen within non-reproductive cells which are not passed to next generation. Even mutation in germ line or reproductive cell may not produce any or small phenotypic change. Depending on nature and position of mutation, it can be beneficial, neutral or harmful. Among many different reasons, faulty DNA replication and environmental influence are two major reasons for mutation.

Most of the mutations that are related to evolution have occurred naturally during DNA replication process of cell division. During the cell division a copy of its DNA is made which sometime is not perfect and the small difference between the original DNA strand and copied strand are not detected by DNA repair system. External influence is other major cause of

mutation. DNA from genomes of living organisms are always exposed to exogenous and endogenous agents which attack and destroy DNA, thus integrity of our genome is always on threat. In general, the most of DNA modifications in humans are endogenous in origin (De Bont & Van Larebeke 2004). Endogenous DNA damage is caused by spontaneous hydrolysis as well as chemical modifications by reactive molecules. Theses reactive molecules are created during normal cellular metabolism called reactive oxygen species which include O^{2-}, H_2O_2, and •OH (Apel & Hirt 2004; Kesheri et al. 2014). These reactive species damage or modify DNA molecules, and sometime even break DNA molecule. Endogenous genomic damage could be also due to mismatches or insertion/deletion of bases by DNA polymerases during DNA replication reactions (McCulloch & Kunkel, 2008). Besides the numerous endogenous sources of DNA damage, cellular DNA damage is also due to attack from exogenous/environmental agents. These include ultraviolet light (UV) which damages cell by affecting DNA and protein molecules. UV rays can affect cell indirectly through the production of reactive oxygen species (ROS) (Vincent & Neale, 2000). Ionizing radiation is another source of DNA damaging agent, which can originate from both natural (e.g., cosmic and gamma radiation) and artificial sources (e.g., medical treatments, such as X-rays and radiotherapy). It causes various types of DNA lesions, among which double-strand breaks are most dangerous (Dexheimer 2013). It should be reliably repaired to maintain genomic integrity/cellular homeostasis (Panier & Boulton 2014). Numerous other exogenous chemical agents have the ability to induce stable and bulky DNA adducts. Alkylating agents like methyl methanesulfonate and temozolomidec induce alkylation of the DNA bases (Irigaray & Belpomme 2010). Therefore, alkylation, oxidation, deamination, depurination/depyrimidation of DNA create large number of DNA base lesions per day (Lindahl 1993).

When the cellular repair system repairs damaged DNA, it might not do a perfect job of the repair. So the cell carries on with the mutated DNA molecules. In spite of critical need for DNA repair, "evolvalibilty" i.e. the ability to generate a certain level of mutations also seems to be selected during evolution. Organisms with optimum level of evolvalibilty have the best chance to survive due to virtue of having variation in the genome. The complex interplay between two opposing forces, namely the need for transmission of genetic information to the offsprings and need for evolvability define the organization of the DNA repair system.

1.3 DNA repair system

The molecular mechanisms of DNA repair can be classified into four main categories: direct r epair (DR), nucleotide excision repair (NER), mismatch repair (MMR) and base excision repair (BER) (Dmitry& Kosuke 1997). In the direct repair pathway, the unusual chemical bonds of bases, nucleotides or substituents are directly modified with the help of DNA repair enzymes. UV induced DNA damages are repaired mainly by this pathway. DNA photolyase, Spore photoproduct lyase, Photoproduct photolyase and O^{6-}methylguanine DNA methyl transferases are major enzymes that catalyzes direct DNA repair (Aziz 1995; Lindahl 1982; Richard 2001). In NER, phosphodiester bonds of both sides of damaged base are cleaved and finally damaged base containing nucleotide is removed by NER pathway enzymes like exonucleases and excision nucleases (Aziz 1996). In MMR pathway, repair of mismatch during DNA replication and recombination is done by MutS, MutH and MutL family of proteins (Thomas & Dorothy 2005). Among the different repair mechanisms, BER is the prominent pathway for repair of small DNA

lesions resulting from exposure to either environmental agents or cellular metabolic processes that produces alkylating agents, reactive oxygen species or reactive metabolites.

1.3.1 Direct repair pathway

In the direct repair pathway, the abnormal chemical bonds which are caused by UV induced DNA damage between bases or between a nucleotide are cleaved and thus repaired. UV when absorbed by DNA may cause cell death, mutations and cancer. DNA photolyases are highly efficient light dependent DNA repair enzymes which are responsible for reverting the DNA damage by ultraviolet (UV) radiation (Essen & Klar 2006). Photolyase can repair UV-induced DNA damage by blue light called as DNA photoreactivation which depends on a cofactor, flavin adenine dinucleotide (FAD) (Weber2005). Photolyases enzyme removes the major UV-induced DNA damage e.g. cyclobutane pyrimidine dimers (CPDs) and 6-4 photoproducts (6-4PPs) (Garinis et al. 2006). These enzymes are present in almost all living organisms except placental mammals like humans and mice. Consequently, for these lesions, which are less efficient, humans and mice are dependent on nucleotide excision repair (NER) pathway (Garinis et al. 2006).

O6-methylguanine DNA Methyl Transferase is other important enzyme in direct repair pathway which is present in all species which transfers the methyl group from O6-methylguanine as well other alkyl groups to a cysteine residue within the enzyme (Lindahl et al. 1998). This is an irreversible process resulting into the inactivation of the protein (Fig. 1.1). O6-methylguanine (O6-meG) is formed when cellular DNA interacts with endogenous alkylating agent S-adenosyl-t-methionine, as well as exogenous electrophiles. O6-Alkylguanine, mainly O6-

methylguanine (O6 MeG) and O6-ethylguanine, are the main source of GC→AT transition mutations.

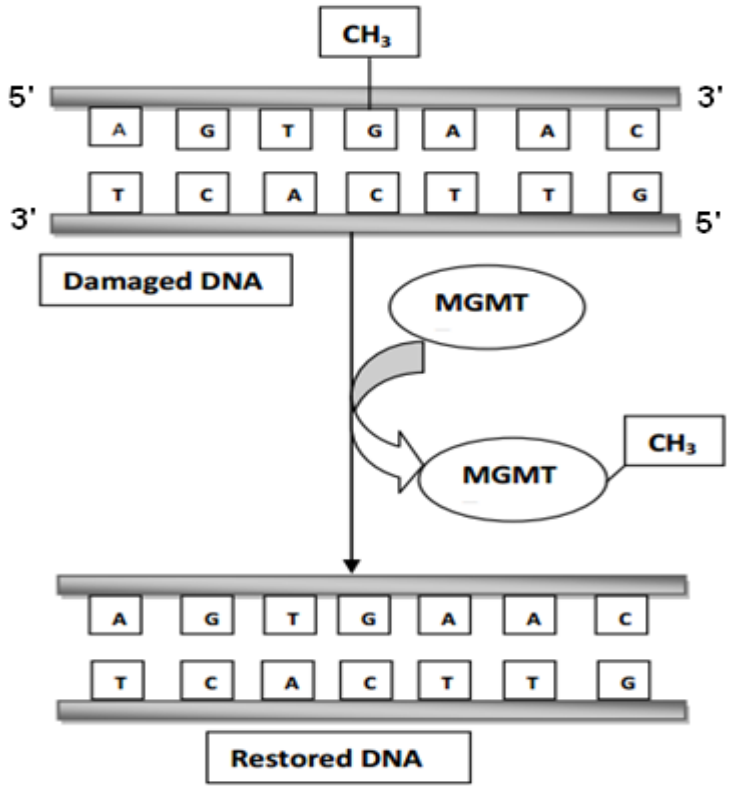


Fig. 1.1 Mechanism of action of 6-methylguanine DNA Methyl Transferase (MGMT) from direct repair pathway (adapted from Lindahl et al. 1998). MGMT acts on 6-metylgunanine and transfers the methyl group from guanine to its own enzyme molecule and thus restores the metylguanine into normal guanine molecule.

1.3.2 Nucleotide excision repair pathway

Bulky DNA adducts such as UV-light-induced photo-lesions [(6-4) photoproducts (6-4PPs) and cyclobutane pyrimidine dimers (CPDs)] are large chemical adducts which are generated due to exposure to aflatoxine, benzopyrene. Other genotoxic agents and intrastrand cross-links are repaired through nucleotide excision repair (NER) pathway in *Homo sapiens* (Friedberg 2001).

UV-hypersensitive disorders such as xeroderma pigmentosum (XP), trichothiodystrophy (TTD), Cockayne's syndrome (CS) and UV-sensitive syndrome (UVSS) (Vermeulen et al.1997) are the diseases caused due to defective NER. NER consists of two distinct pathways which are termed as global genomic repair (GGR) and transcription-coupled repair (TCR) (Fig. 1.2). GGR removes lesions mainly from the non-transcribed portions from the genome while TCR removes different RNA-polymerase-blocking lesions from the transcribed portions from the genome which belong to active genes (Bohr et al. 1985; Mellon et al. 1987).

During global genomic repair (GGR), recognition of the DNA damage is made by XPC–HR23B which recognizes UV-induced DNA lesion 6-4PPs (Hey et al. 2002), while RPA–XPA recognizes 6-4 photoproducts (6-4PPs). RPA–XPA also recognizes DNA treated with cisplatin (Burns et al.1996) or damaged DNA binding protein DDB1–DDB2and also stimulate the excision of cyclobutane pyrimidine dimmers (CPDs) *in vitro* with high efficiency (Tang & Chu 2002). DNA unwinding is made by the transcription factor TFIIH which is composed of seven different proteins (MNAT1, XPB, XPD, CDK7, GTF2H1, GTF2H2, GTF2H3, GTF2H4, and CCNH). After that, XPA and RPA not only stabilize the opening but also help in positioning other factors. Then, XPA and RPA bind to the damaged and undamaged DNA strand respectively. Possibly, RPA plays an important role in opening the complex. XPG (Habraken et al.1994) and XPF–ERCC1 (Sijbers et al. 1996) perform excision at 3'and 5' end respectively. Finally, resynthesis of DNA is performed by Polδ or Polε. Thereafter that ligation of DNA is made by DNA ligase I (Aboussekhra et al. 1995; Araujo et al. 2000; Mu et al. 1995).

During transcription-coupled repair (TCR) DNA damage causes the blockage of RNAPII (Selby et al.1997) which promotes the assembly of CSA, CSB and/or TFIIS at the site of the damage.

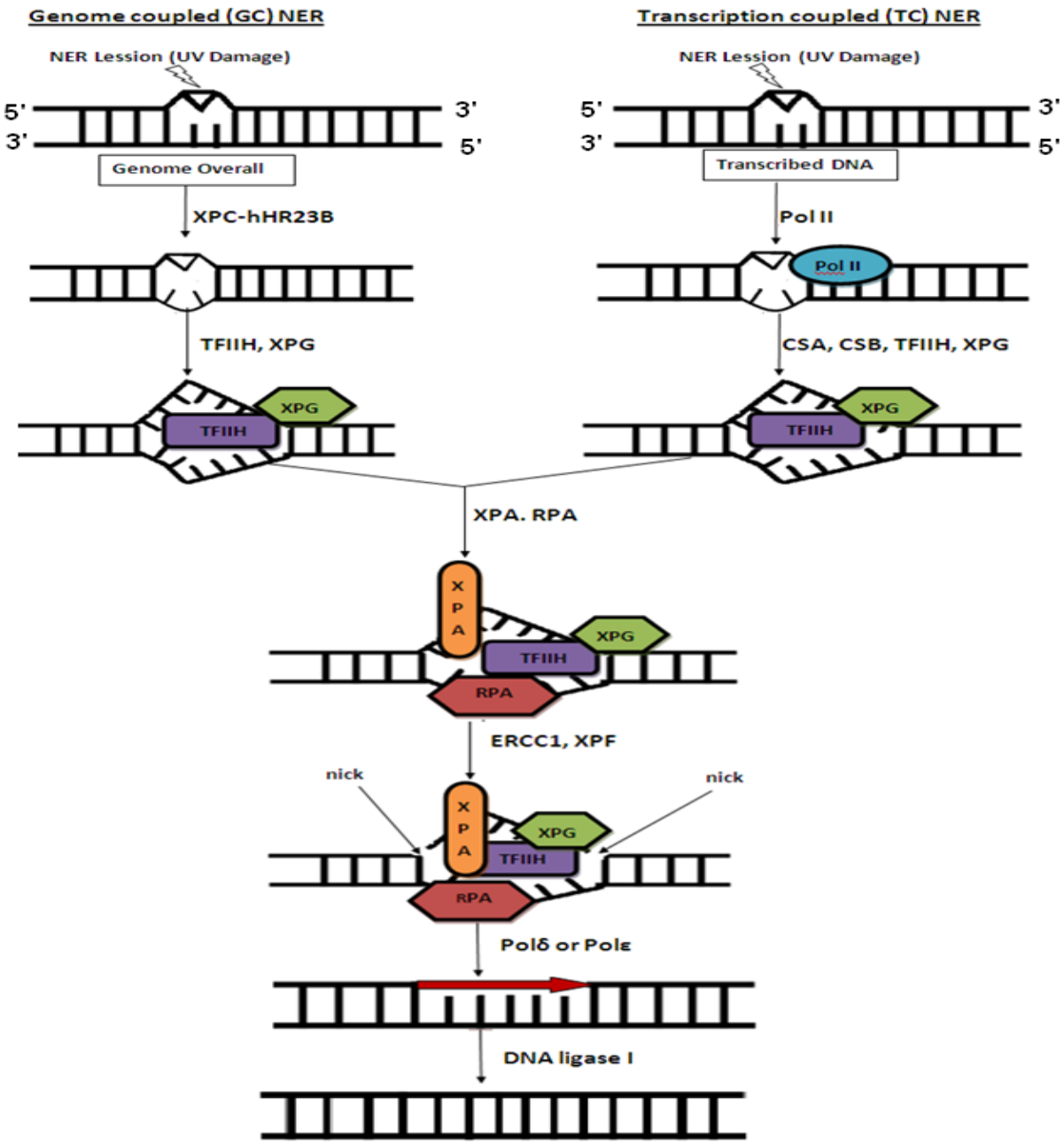


Fig. 1.2 Nucleotide excision repair pathway is shown schematically (adapted from De Laat et al. 1999)**.** XPC–hHR23B recognizes lesions in GG–NER, while in TC–NER lesions are recognized by elongating RNA Pol II. In GG–NER, XPC–hHR23B performs assembly of TFIIH and XPG. TFIIH creates a 10- to 20-nucleotide open DNA complex around the lesion due to its helicase activity. In TC–NER, CSA, CSB, TFIIH and XPG displace and remove the stalled Pol II from the lesion. Now in both GG–NER and TC–NER, open DNA complex becomes accessible for XPA and RPA and causes full opening and stabilization of the complex. XPG, TFIIH and RPA performs the incision at 3' end, while ERCC1–XPF, RPA and XPA performs the incision at 5'end of the lesion which is followed by gap-filling DNA synthesis and ligation.

After that, RNAPII is removed from the DNA strand and finally displaced from the site of damage. Similarly, in GC coupled NER, XPA and RPA stabilize the opening as well as position other factors at the site of damage. XPA binds to the damaged nucleotides while RPA to the opposite and undamaged DNA strand and finally full open complex is formed. Now DNA lesion becomes accessible to the exonucleases XPF–Ercc1 and XPG which are finally cleaved. Finally, resynthesis of DNA again occurs by Polδ or Polε which are ligated by DNA ligase I .

1.3.3 Mismatch repair pathway

The mismatch repair (MMR) pathway is mainly responsible for removal of base mismatches which are caused by spontaneous as well as induced base deamination, and replication errors (Modrich & Lahue 1996).

A mismatch can be defined as any non-complementary base pair present in double stranded DNA. *In vivo,* mismatching among base pairs are resulted from: (a) errors in DNA replication, (b) hybrid DNA formation during recombination and (c) chemical modification of bases in double stranded DNA (*e.g.* deamination 5-methylcytosine to form thymine). Mismatch repair pathway is mainly responsible for recognition and correction of incorrectly paired nucleotides in DNA which is finally excised and new DNA is synthesized. In *E.coli,* mismatch repair pathway is called as MutHLS system which is composed of MutS, MutL,MutH proteins along with DNA polymerase, single-stranded binding proteins, and DNA ligase. Homologs of *E.coli* MutS are called MSH genes in eukaryotes which are designated as MSH1, MSH2, MSH3, MSH4, MSH5 and MSH6 (Fishel & Wilson 1997). The stepwise procedure of MMR (Fig. 1.3) may be summarized as follows.

A. Recognition of DNA lesions: The recognition of base mismatches is performed by MutS alpha complex (MSH2+MSH6). This complex binds primarily to base–base mismatches and insertion/deletion mismatches (Umar et al.1994). MSH2 also form a complex with MSH3 which is termed as MutS beta complex that binds to insertion/deletion mismatches only (Genschel et al.1998).

B. Strand discrimination: According to the molecular switch model, binding of MutS alpha–ADP performs the recognition and binding to the mismatch (Fishel, 1998), which triggers ADP→ATP transition by stimulating intrinsic ATPase activity. In the hydrolysis-driven translocation model, ATP hydrolysis leads to release of the energy which is utilized by MutS alpha to translocate actively along the DNA from the site of mismatch recognition to a site responsible for signaling the strand specificity.

C. **Excision and repair synthesis:** After binding to the mismatch, MutS alpha interacts and binds with MutL alpha complex (MLH1 and PMS2) (Li & Modrich 1995). Excision of the DNA strand containing the mismatch base is performed by exonuclease I (Genschel et al. 2002) and new DNA is synthesis by Pol δ (Longley et al.1997).

1.3.4 Base excision repair pathway

DNA damage due to mutagenic agent as well as cytotoxic effects due to spontaneous hydrolytic and non-enzymatic alkylation of DNA is taken care by BER pathway.

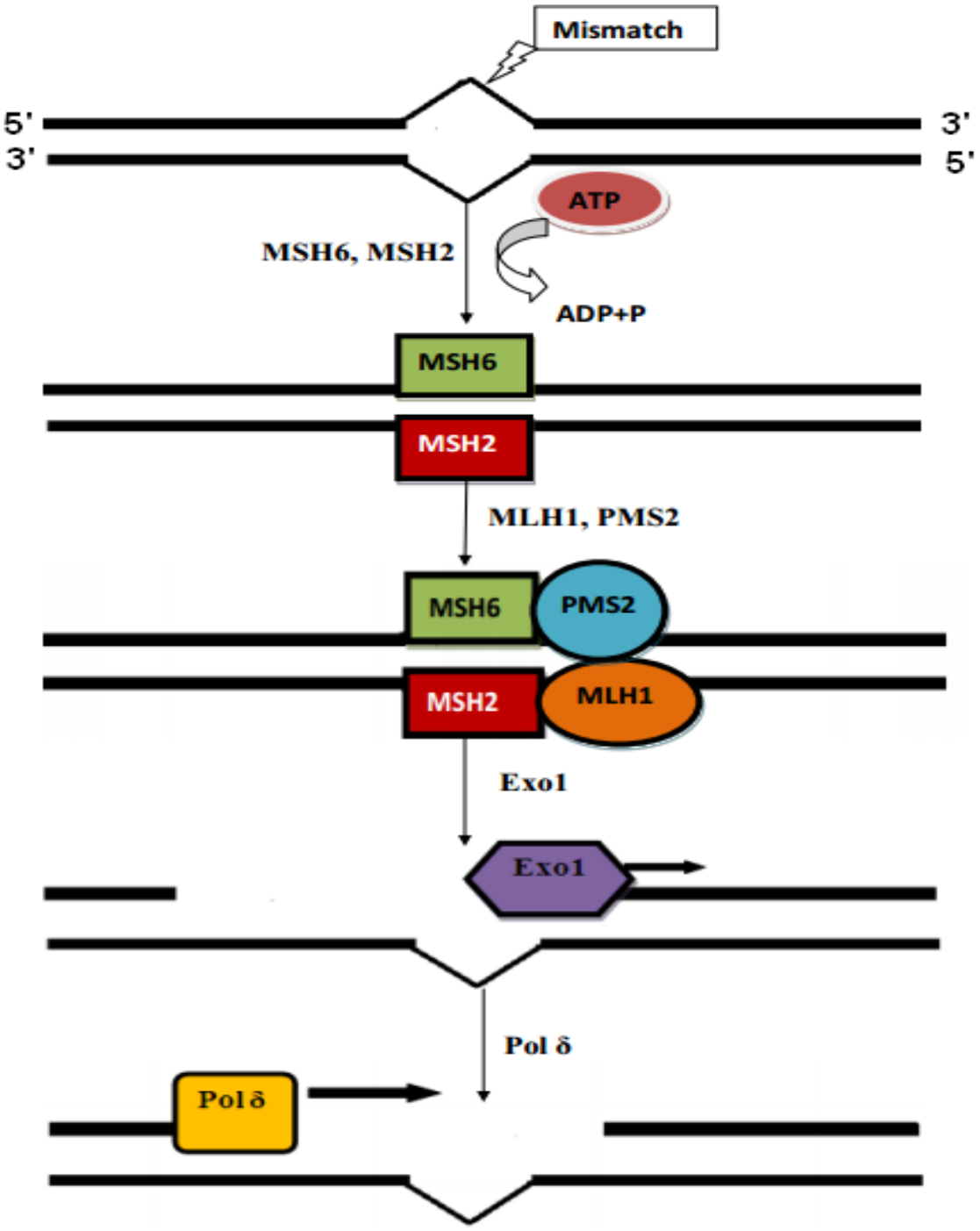


Fig. 1.3 A schematic diagram depicting mismatch repair pathway (adapted from Kunkel &Erie 2005). The recognition of mismatches is performed by MutS alpha complex (MSH2 and MSH6) first recognizes mismatches through binding which requires ATP→ADP transition. This triggers intrinsic ATPase activity. After binding to the mismatch, MutS alpha complex bindsto MutL alpha (MLH1 and PMS2) complex which is followed by excision of the mismatch containing DNA strand by exonuclease I. Finally new DNA strand is synthesized by Pol lambda.

BER is also responsible for repairing oxidized DNA bases which arise within the cell during inflammatory responses, or damages due to ionising radiation and long-wave UV light. Base excision pathway repairs 8-hydroxyguanine (8-OH-dG), 3-methyladenine,single-strand breaks (SSBs) and apurinic/apyrimidinic (AP) sites(Wilson & Bohr 2007). To avert the deleterious consequences of DNA damage base excision repair (BER) pathway needs to recognize, excise, and replace specific forms of genetic modifications accurately. The overall BER process (Fig. 1.4) consists of the following steps:

(i) Recognition and excision of an inappropriate base

(ii) Nucleotide insertion

(iii) Separation of short- and long-patch BER pathways

(iv) Strand displacement and DNA-repair synthesis by long-patch BER

(v) Sealing of the nick (Ligation).

The first step of BER is to recognize and remove damaged and incorrect base. To initiate the first step in base excision repair, cell is equipped with specific protein called DNA glycosylases. DNA glycosylases recognize limited types of modified bases and then catalyze hydrolysis of the N-glycosylic bond (Scharer & Jiricny 2001). These DNA glycosylases are categorized into two types: type I and type II glycosylases. Type I glycosylases are responsible for removing modified bases and creates an AP site in DNA (e.g. MPG). Type II glycosylases first remove the base and then cleave the AP site due to endogenous 3' endonuclease activity and create single-strand break (e.g. OGG1). For type I glycosylases, phosphodiester bond of the AP site is broken by AP endonuclease (APE1 alias APEX, Ref-1 or HAP1) leaving 5'-deoxyribose-5-phosphate (5'dRP)

and 3'-OH (Wilson & Barsky 2001) at two ends. AP endonucleases are a group of enzymes which are responsible for breaking the phosphodiester bond at AP site. Based on the location of incision of the phosphodiester cleavage,AP endonucleases are categorized into two types which are class I and class II AP endonucleases. Class I AP endonucleases cleave DNA at the 3' end to AP site while class II AP endonuclease enzymes cleave at the 5' end leaving 5' -phosphate and a 3'-OH groupsin both the cases (Shida et al. 1999). Endonuclease III (*E.coli*) is the most studied enzyme from class I, and exonuclease III and endonuclease IV in *E.coli*from class II are most studied enzymes and their three dimensional structures are also available with protein data bank.

In the second step of BER, displacement of 5'dRP and insertion of a single nucleotide are made by DNA polymerase β (Pol β), during short-patch (Dianov et al. 1992) and long-patch BER (Dianov et al. 1999).

In third step of BER, separation of short-patch BER (SP-BER) and long-patch BER (LP-BER) repair takes place after the removal of 5'dRP and insertion of the first nucleotide. Patch length depends on the nature of damage of nucleotides which is shown in many examples

In fourth stage both SP-BER and LP-BER repair start with DNA polymerase which add one dNMP into DNA, after adding one nucleotide 3'-end is processed accurately (Podlutsky et al. 2001). *In vitro*, SP-BER mainly processesthymine glycol, AP sites, ring-alkylated purines and 8-oxoGua , while hypoxanthine, Ura, and 1,N6-ethenoadenine are processed by both SP- and LP-BER (Bennett et al. 2001; Akbari et al. 2004; Fortini & Dogliotti 2007; Dantzer et al.2003).

Finally, in short-patch repair backbone is sealed after single base insertion, while long-patch repair requires several additional steps. After single base insertion, Polβ is dissociated from DNA

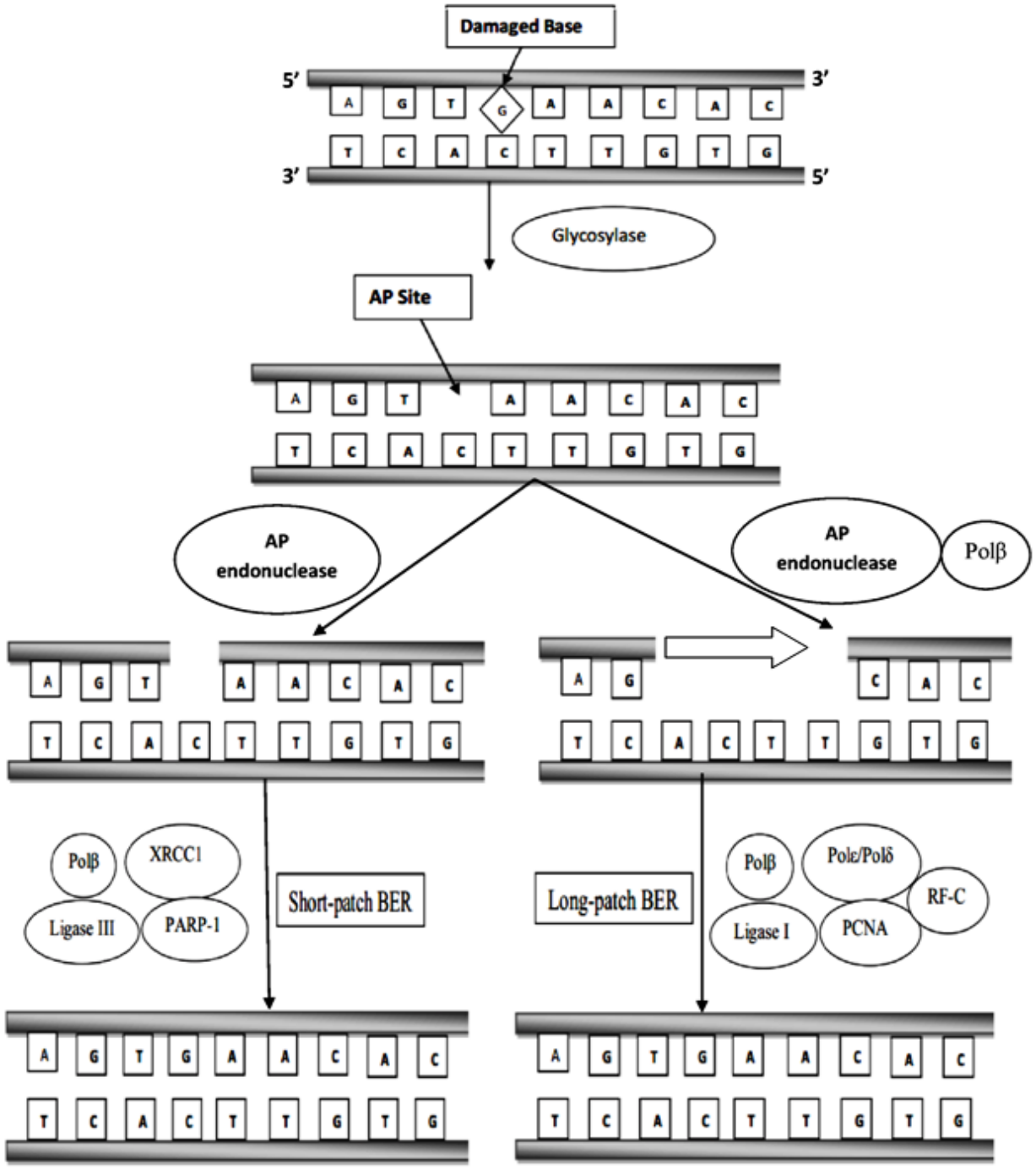


Fig. 1.4 Base excision repair pathway is shown schematically (adapted from Sancar et al. 2004). Incision into the phosphodiester bond of the AP site occurs by AP endonuclease. Single base insertion by Polβ is followed by sealing of nick by interaction of DNA ligase III with XRCC1, Polβ and PARP-1 in short-patch BER. In long-patch repair, single base insertion takes place by Polβ, then strand displacement and further DNA synthesis is made by Polε or Polδ together with PCNA and RF-C which is followed by sealing of nick by interaction of Ligase I, PCNA and Polβ.

backbone. DNA synthesis is initiated by Polε or Polδ together with PCNA and RF-C (Stucki et al.1998), which results in sealing upto 10 nucleotides. The nick is then sealed by DNA ligases I and III (Tomkinson et al. 2001). Ligase I interacts with Polβ and PCNA. This complex nicks DNA during long-patch BER pathway (Prasad et al. 1996; Srivastava et al. 1998). In short-patch BER pathway, DNA ligase III interacts with XRCC1, PARP-1 [poly (ADP-ribose) polymerase-1] and Polβ and nicks DNA (Kubota et al. 1996).

1.4 Evolutionary studies on AP endonucleases

AP endonuclease protein family is an important group of proteins that involves during BER pathway which is one of the busiest repair pathways. AP endonucleases are responsible for recognition and nicking of AP sites. However, limited evolutionary studies have been foundamong the AP endonuclease protein family. Most studies had considered AP endonuclease family as a part of higher super family. For example, Denver et al (Denver et al. 2003) had investigated endonuclease III protein family as a part of Helix-Hairpin-Helix Superfamily of DNA repair Glycosylases. In this study, a few endonuclease III homologs from bacteria, archaea and eukaryotes were placed in Nth protein family as separate clade within helix-hairpin-helix (HhH) Superfamily proteins.Very few AP endonuclease homologs were considered during evolutionary study of exonuclease III protein family (Eisen & Hanawalt 1999). Similarly, only few mammalian exonuclease III homologs were considered during evolutionary studies of redox function (Georgiadis et al 2008). However, evolutionary study of endonuclease IV proteins from AP endonuclease protein family is unexplored. In the present study, we aim at exploring the evolutionary scenario of AP endonuclease protein family. The evolution of AP endonuclease protein family may also be cross examined by protein structure modeling. These structural

models will be investigated for the various changes at protein domains and fold, various insertion/deletion /substitution in the protein core which can provide us important clue regarding evolution process. Our understanding about protein evolution will enhance our knowledge in functional divergence of proteins by combining both genomics and proteomics. This evolutionary study may be exploited for the prediction of the functions of uncharacterized genes/proteins in many unannotated gene/protein families of organisms which are completely/partially sequenced. Following are the specific objective of this study

1.5 Objectives of current study

- To understand the evolution process of AP endonuclease protein family (endonuclease III, endonuclease IV and exonuclease III protein family) through sequence and structure analysis.
- To develop unified evolutionary model of the AP endonuclease protein family across three domains of life.

Chapter II

Methodology

Chapter II

2.1 Overview

The DNA repair protein sequences of *E. coli* were used as a query to retrieve homologs of different DNA repair proteins. The homologs of these proteins were searched by different database searching tools. Based on pairwise sequence identity and presence of common domain, final protein homologs were selected. Multiple sequence alignment and sequence logo were utilized to identify structurally and functionally important residues. Both maximum likelihood (ML) and neighbor joining (NJ) methods were used to construct all phylogenetic trees. NJ is a reliable predictor when the rate of evolution among taxa varies. ML is also a true predictor of evolutionary relationship when the variance is high among different lineages.

16S ribosomal RNA gene sequences of bacterial, archaeal and eukaryotic species were retrieved and were used to generate species tree. cDNA of homologous protein sequences were retrieved and NJ as well as ML Phylogenetic tree based on cDNA were generated. Phylogenetic tree based on 16S ribosomal RNA genes, genes and proteins were compared among them to investigate the evolution of AP endonuclease protein/gene.

Homology models of representative homologs were generated which were further refined and evaluated for structural quality and finally compared to each other to examine possible structural changes of proteins during the course of evolution.

2.2 Retrieval of AP endonuclease protein homologs

AP endonucleases from *E.coli* were used as query to retrieve homologs of this group of protein from the National Center for Biotechnology Information (NCBI) protein sequence database. The

Protein database is a collection of sequences from several sources, including translations from annotated coding regions in GenBank and RefSeq as well as records from SwissProt, Protein Information Resource (PIR) and PDB. BLASTP (Basic Local Alignment Search Tool) (Altschul et al. 1990) was used for retrieving homologous sequences of these proteins from the NCBI non-redundant (NR) protein database (with E-value cut off of $\leq 1 \text{ X } 10^{-5}$) (Nagamune & Sibley 2006). We conducted BLASTP using the default parameters. BLASTP is a sequence similarity programme that generates results quickly. Blast compares a query protein to protein sequences in a target database to identify regions of local alignment and report those alignments that score above a given score threshold. There are versions of BLAST that compare protein queries to protein databases, nucleotide queries to nucleotide databases, as well as versions that translate nucleotide queries or databases in all six frames and compare to protein databases or queries. BLAST also calculates an expect value (E-value) that estimates how many matches would have occurred at a given score by chance, which helps in judging statistical significance of an alignment. In addition to BLASTP, two rounds of PSI-BLAST search with a default parameter (Altschul et al. 1997) were also conducted using the same query sequence to identify additional homologs.

Position-Specific Iterative-BLAST (PSI-BLAST) is another protein similarity search method that builds the alignments generated by BLASTP program then generates a multiple alignment of the highest scoring pairs of the BLASTp above a certain preset score or *e*-value threshold and calculates a profile or a position-specific score matrix (PSSM) from the multiple alignment. The PSSM captures the conservation pattern in alignment and stores it as a matrix of scores for each position in the alignment. This profile is used in place of the original substitution matrix for a further search and newly detected sequences, which are above the specified score (E-value)

threshold are added to alignment the profile is refined for another round of database search. This process is iteratively continued until desired or until convergence, i.e., the state where no new sequences are detected above the defined threshold. The iterative profile generation process makes PSI-BLAST far more capable or sensitive of detecting distant sequence than a single query alone in BLASTp (Altschul et al. 1997).

2.3 Retrieval of cDNA sequences and G+C content calculation

cDNA sequences of AP endonuclease protein homologs were retrieved from the nucleotide NCBI sequence database (http://www.ncbi.nlm.nih.gov/nucleotide/). Mean G+C content and G+C content at 3rd codon position of the cDNA sequences of AP endonuclease gene were calculated by CAIcal server (Puigbo et al. 2008). G+C content for the genomes of the species were retrieved from the NCBI genome database (http://www.ncbi.nlm.nih.gov/genome). 16S ribosomal RNA gene sequences of bacterial and archaeal species are retrieved from ribosomal sequences database (Cole et al. 2009). 18S ribosomal RNA gene sequences of eukaryotic species were taken from NCBI nucleotide database (http://www.ncbi.nlm.nih.gov/nucleotide).

2.4 Sequence analysis

Pairwise sequence alignments among AP endonuclease proteins were performed by the EMBOSS Needle programme (Rice et al. 2000). EMBOSS is a free open source software analysis package which provides extensive libraries to analyze sequence.

All multiple sequence alignments were conducted with ClustalW (Thompson et al. 1994: Larkin et al. 2007) and T-Coffee server (Moretti et al. 2007) with default parameters. ClustalW is a general purpose progressive multiple sequence alignment method for DNA and protein sequences. It uses matrix of pairwise alignments distances and theses distances are used to generate the guided tree. Within the T-coffee distribution, M-Coffee is a multiple sequence alignment package. The specificity of M-Coffee is that rather than computing a multiple sequence alignment on its own, it uses other packages to compute the alignments. It then uses T-Coffee to combine all these alignments into one unique final alignment. (http://www.tcoffee.org/) Domains and motifs are identified by the conserved domain (CD) search tools available at NCBI (Marchler & Bryant, 2004). CD-Search uses BLAST heuristics and searches a comprehensive collection of domain models quickly. Domain models are imported from Pfam (Bateman et al. 2004), SMART (Letunic et al. 2004) and Clusters of Orthologous Groups (COGs) (Tatusov et al. 2003). CD-Search tool graphically summaries all the search results. Finally a tabular list of hits and individual pairwise alignments between the query and the model sequences are prepared.

The conservation patterns of amino acids in protein sequence within the protein structure were analyzed by sequence logos. Weblogos 3.2 (Crooks et al. 2004) generates sequence logos without compositional adjustment. Sequence logos are a graphical representation of multiple sequence alignment of protein and/or DNA sequences. Each logo consists of stacks of symbols, one stack for each position in the sequence. The overall height of the stack indicates the sequence conservation at that position, while the height of the symbols within the stack indicates the relative frequency of each amino or nucleic acid at that position (Crooks et al. 2004).

2.5 Phylogenetic tree construction

Phylogenetic tree is the two dimensional graph which shows evolutionary relationship among organisms. Maximum Parsimony, Distance and Maximum likelihood are three common algorithms which are generally used to construct the evolutionary tree. These methods are used based on the sequence similarity among the sequences.

2.5.1 Maximum parsimony

Maximum parsimony predicts the evolutionary tree based on the requirement of minimized number of steps to generate the observed variation in the sequences. For this reason, the method is also called minimum evolution tree. Generally, if a strong sequence similarity among the sequences is present then maximum parsimony is most preferred method.

2.5.2 Distance based methods

Distance based methods are often used for sequences with a recognizable sequence similarity. Distance methods are based on the number of changes between each pair in a group of sequences to generate a evolutionary tree. Among the distance based methods Fitch, Kitch, Neighbor and UPGMA are popular methods. Fitch estimates the phylogenetic tree assuming additivity of branch lengths using Fitch-Margoliash algorithm and doesn't assume molecular clock (rates of evolution among branches vary) while Kitch estimates phylogenetic tree by Fitch-Margoliash algorithm which also assumes the molecular clock. Neighbor estimates the phylogenetic tree using neighbor-joining method and doesn't assume a molecular clock and produces unrooted

tree. UPGMA (Unweighted group method with Arithmatic mean) estimates phylogenetic tree by assuming molecular clock and generates a rooted tree.

2.5.2.1 Neighbor Joining (NJ)

NJ is the most popular distance based tree building algorithm. The most important fact about this method is that it is relatively and performs well when the rate of evolution varies in separate lineages. Neighbor joining uses the evolutionary model to estimate distances between sequences and then use methodology similar to hierarchical clustering to build the tree. The first step in the algorithm is converting the DNA or protein sequences into a distance matrix that represents evolutionary distances between sequences.

2.5.2 Maximum Likelihood (ML)

Maximum likelihood method is better method in case of less sequence similarity. Maximum likelihood method is one of the important methods to estimate the evolutionary relationship when the variance is high among different lineages. ML is a method for the inference of phylogeny in terms of the probability where a hypothesis about evolutionary history is evaluated to give rise to observed data. In ML, a hypothesis is judged by how well it predicts the observed data; the tree that has the highest probability of producing the observed sequences is preferred. To use this approach, we must be able to calculate the probability of a data set given a phylogenetic tree. Standard differential-equation techniques are used to convert the model into a statement of the probability of any two sequences if we have a model of sequence evolution that

describes the relative probability of various events (for example, the chance of a TRANSITION relative to the chance of a TRANSVERSION).

Neighbor joining and maximum likelihood methods are usually accompanied with bootstrapping which is a computer based technique for obtaining confidence limits on phylogenetic trees. Phylogenetic tree uses character based molecular (DNA or Protein sequence) data to form matrix which is replaced to produce bootstrap data sets. Each of which is analyzed plylogenetically and a consensus tree is constructed to summarize the results of all replicates (Soltis & Soltis 2003). Bootstrapping offers a measure of which parts of the tree are weakly or strongly supported. Bootstrap values are conservative measures of support, so a value of 70% might indicate strong support for a group (Zharkikh & Li 1992).

In our studies, we estimated phylogenetic trees for all selected AP endonuclease homologs by using Neighbor Joining and maximum likelihood method as implemented in Mega 5.1 (Tamura et al. 2011). The reliability of interior branches was assessed with 1000 bootstrap re-samplings of amino acid/cDNA/16S-18S sequences in phylogenetic trees (Felsenstein 1985). Phylogenetic tree based on amino acid/cDNA/16S-18S sequences of AP endonuclease family was constructed from alignment using both maximum likelihood (ML) and neighbor joining (NJ) method. The topologies of both trees generated using the ML and NJ methods are very similar to each other. Bootstrap analysis reveals that most of the clades in both trees are robust and majority of clades are supported by ≥50% bootstrap value and therefore, only one tree (ML tree) is discussed further.

2.5.4 Evolutionary distance calculation and generation of final evolutionary tree

Evolutionary distances for gene based tree was constructed by using Tamura-Nei (Tamura & Nei 1993) while protein based tree was constructed by using Jones Taylor Thronton method (Jones et al. 1992). Jones Taylor Thronton (JTT) method is a faster method and is used it to produce a replacement matrix from a much larger database as that of Dayhoff PAM matrix.

2.6 Homology modelling

AP endonuclease proteins 3D structure models were build for those homologs for which 3D structure (the target) were not known. Template search for modelling of AP endonuclease proteins was done by searching the target AP endonuclease proteins query sequences with the sequence of each of the structures in the PDB database by BLASTP by keeping default parameters. We selected the suitable template structure based on the highest sequence similarity and highest coverage with the target sequence AP endonuclease proteins as well the structure quality of the X ray crystal structure (Kesheri et al. 2015).

After selection of the of the suitable template, we checked the alignment between target AP endonuclease protein sequences and template sequence in core regions for errors and we found good alignment especially for core region residues so the alignment was kept intact and was used for model building.

We generated 3D models of AP endonuclease proteins by Modeler 9v14 (Sali & Blundell 1993) which is based on satisfaction of spatial restraints. We used standalone version of Modeler 9v14 for model building. Model was generated using single template. From each clade, one representative AP endonuclease protein homologs are chosen to model the 3D structure, except

the clade which contains experimental three dimensional structures for any endonuclease III homolog available in protein databank.

Refinement of 3D Models of AP endonuclease protein family was finally done by ModRefiner (Xu & Xhang 2011) for corrections either in structural packing of side chains/loops or secondary structural elements in the target model by energy minimization. ModRefiner is an algorithm for atomic-level, high-resolution protein structure refinement, which can start from either C-alpha trace, main-chain model or full-atomic model. Both side-chain and backbone atoms are completely flexible during structure refinement simulations, where conformational search is guided by a composite of physics- and knowledge-based force field.

AP endonuclease proteins 3D model structure validation was done by protein structure superimposition between target and template structure and calculations of the root mean square deviation (RMSD) values of the predicted structures compared to their respective PDB templates were performed by fitting carbon backbones of target and template using Swiss-PdbViewer (Guex & Peitsch 1997) by selecting <Calculate RMS> for the Cα/backbone option after <Iterative Magic Fit>. The 3D structure of the model had been evaluated for its stereochemical quality by Ramachandran plot generated by Procheck, (Laskowski et al. 1993) and Errat ver 2.0 (Colovos & Yeates 1993), available at http://nihserver.mbi.ucla.edu/SAVES/. The predicted 3D model of the protein was also validated by ERRAT 2.0. This quality score by ERRAT, being less than the minimum of 90 expected for well-refined, high-resolution structures, indicates that considerable model errors remain, which will require data at higher resolution to correct. Recently, Model quality validation was also done by Qmean score (Benkert et al. 2008) and Qmean Z-score (Benkert et al. 2011) which were generated by Qmean server

(Benkert et al. 2009) which showed the closeness of the computationally predicted models with the experimentally validated structures. Qmean score is mainly consists of a linear combination of six terms (Torsion, Pairwise, Solvation, All_atom, SSE_agree and ACC_agree) and it is basically a global score of the whole model reflecting the predicted model reliability ranging from 0 to 1. QMEAN Z-score (Benkert et al. 2011) provided an estimate of the absolute quality of a model by relating it to reference structures solved by X-ray crystallography. For a good quality model QMEAN score near to 1 was expected while |Z-score| <1 was expected for good predictions, 1<|Z-score|<2 for medium predictions, and |Z-score|>2 for bad predictions.

Chapter III

Evolution of endonuclease III protein family

Chapter III

3.1 Introduction

Endonuclease III or nth family proteins belongs to the helix-hairpin-helix (HhH) superfamily along with OggI, MutY/Mig, AlkA, MpgII, and OggII gene family. These specifically recognize and excise varying spectra of damaged bases and base pairing mismatches (Yang et al. 2000; Denver et al. 2003). Proteins from this superfamily have a helix-hairpin-helix (Thayer et al. 1995) structural element followed by a Gly/Pro-rich loop and a conserved aspartic acid residue (Nash et al. 1996; Labahn et al. 1996). Among these gene families, endonuclease III protein is a bi-functional enzyme having AP endonuclease as well as glycosylase activity. Endonuclease III has a broad specificity for DNA base excision repair, removing numerous forms of modified thymine and cytosine bases from DNA (Thayer et al. 1995). Endonuclease III family proteins repair damaged base and mismatched base wherein DNA glycosylases hydrolyses the N-glycosylic bond between the target base and the sugar moiety thus releasing the free damaged base. This process is giving rise to an AP site. These enzymes have been also classified as mono-functional and bi-functional DNA glycosylases. Mono-functional glycosylases have only glycosylase activity while bi-functional DNA glycosylases have AP lyase as well as glycosylase activity. Enzymatic mechanism of endonuclease III is reviewed in detail by Dodson et al. (Dodson et al. 1994).

This study aims to provide an insight into the evolution of endonuclease III gene/protein family among all lineages of life. We have focused on the insertion/deletion of domains/motifs and the evolutionary conservation of important amino acid during the course of evolution. Furthermore, phylogenetic analysis, based on gene and protein sequences of endonuclease III, examines the

evolution of endonuclease III genes/proteins homologs in all five kingdoms of life and then compared with 16S/18S rRNA sequences based species evolution. Evolutionary studies were also done to find horizontal gene transfer events and based on those we propose a model of the evolutionary history of the entire endonuclease III protein family. Homology models of different homologous sequences show possible structural evolution.

3.2 Material and methods

3.2.1 Retrieval and selection of endonuclease III protein homologs

The endonuclease III protein sequence (NCBI Acc. no. NP_416150.1) of *E.coli* was used as a query to retrieve homologs. Blastp (Altschul et al. 1990) and two rounds of PSI-Blast with default parameters were used for retrieving homologous sequences**.** After removing redundant hits, global pairwise sequence alignment was performed with remaining hits from BLAST and PSI-BLAST search. The Protein sequences with less than 15% identity against corresponding *E.coli* endonuclease III protein were removed. Finally, hits with at least one common domain of the endonuclease III protein family were selected. A set of 463 homologs of the endonuclease III family proteins was considered for evolutionary study (shown in table 3.1). Of the total 463 non redundant endonuclease III homologs proteins, 49 homologs were DNA-(apurinic or apyrimidinic site) lyase proteins, 26 homologs were HhH-GPD family protein, 13 homologs were MutY/A:G specific adenine glycosylase, 1 homolog was DNA-3-methyladenine glycosylase III protein and rest 374 were endonuclease III proteins. Among these proteins, 44 homologs belonged to eukaryotes, 42 to archaea and 377 to bacteria (Table 3.2).

Table 3.1 List of different proteins (as it was annotated in the database) retrieved as endonuclease III homologs.

Serial No.	Protein Names	No. of sequences
1.	DNA (apurinic or apyrimidinic site) lyase	49
2.	HhH-GPD family Protein	26
3.	MutY/ A:G specific adenine glycosylase	13
4.	DNA-3-methyladenine glycosylase III	1
5.	Endonuclease III	374
	Total	**463**

Table 3.2 Division and kingdom wise breakup of endonuclease III homologs. All bacterial species is further divided into 22 different categories (Garrity & Holt 2001).

Serial No.	Division	Phylum	Total no. from Each Family	Grand Total
1.	Archaea	Monera	42	42
	Bacteria		377	377
2.	Eukaryotes	Fungi	11	44
		Protista	11	
		Plantae	5	
		Animalia	17	
Total			**463**	**463**

All homologs of endonuclease III protein were divided into five kingdoms namely, Monera (comprising both bacteria and archaea), Protista, Fungi, Plantae and Animalia (Whittaker 1969). Due to a large number of bacterial homologs, sequences from bacterial species were divided into twenty two divisions (Garrity & Holt 2001). Phylogenetic trees were generated for endonuclease

III homologs of nine bacterial divisions (Bacteridetes, Chlorobi, Chloroflexi, Cyanobacteria, Actinobacteria, Firmicutes, Thermotogae, Verrucomicrobia and Proteobacteria). All phylogenetic trees were inspected visually and representative homolog from different clades of each tree was chosen (Fig. 3.1(A-H) A large number of homologs were within Proteobacteria division and hence for clarity it is not shown). For other bacterial divisions, single representative homolog was selected (Table 3.3).

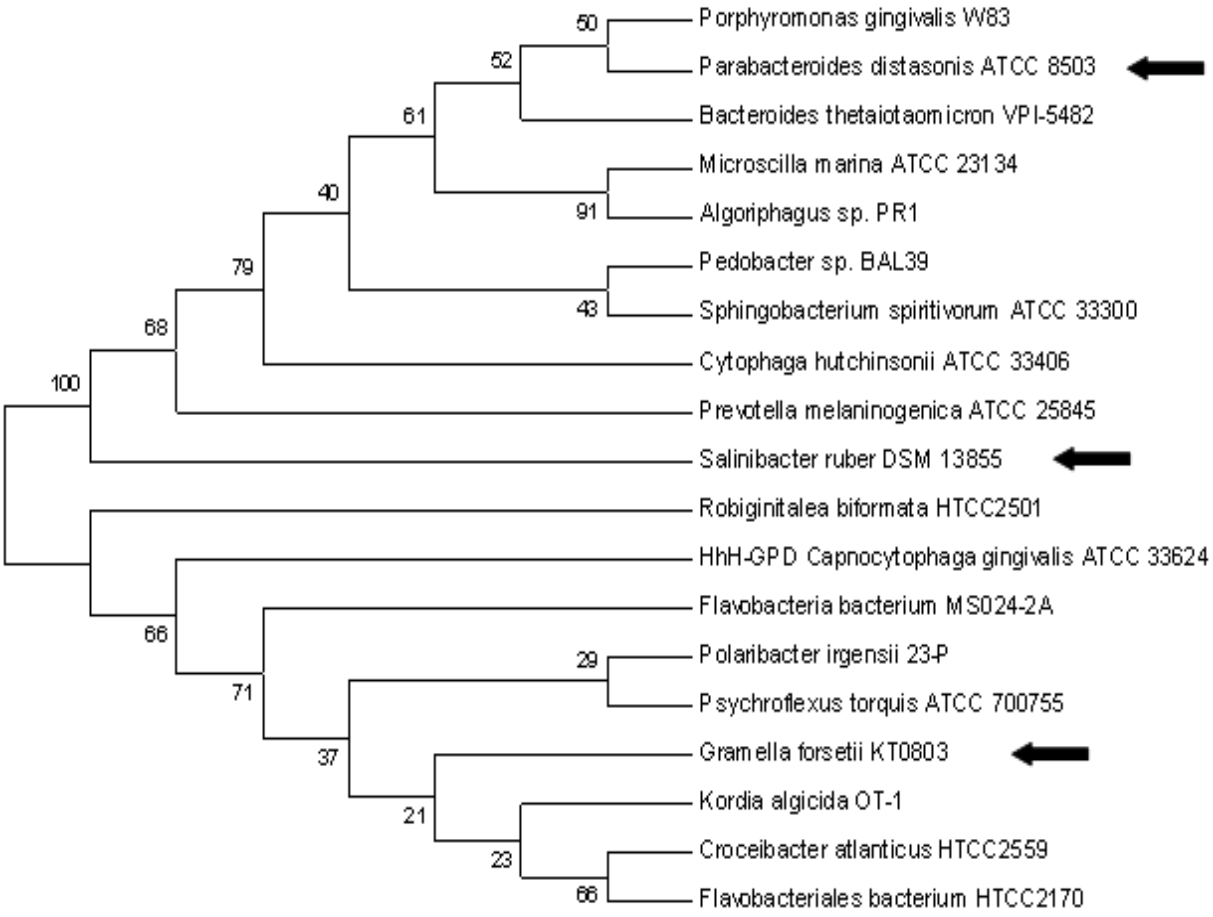


Fig. 3.1(A) NJ tree of 19 Bacteridetes species. Two major clades were observed in the tree. Selected species from Bacteridetes are shown by arrow.

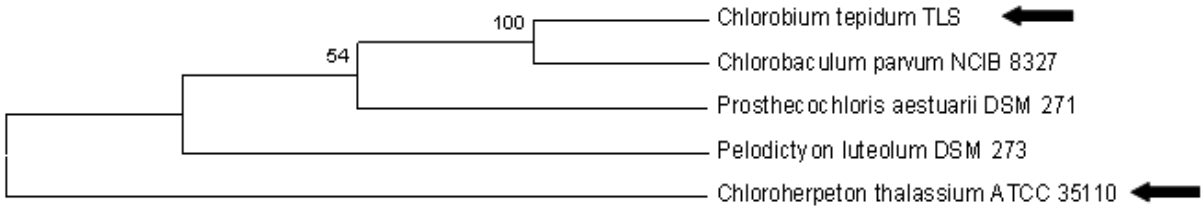


Fig. 3.1(B) Two species were selected from the NJ tree of 5 Chlorobi species which were shown by arrow. Only one clade was observed in the tree.

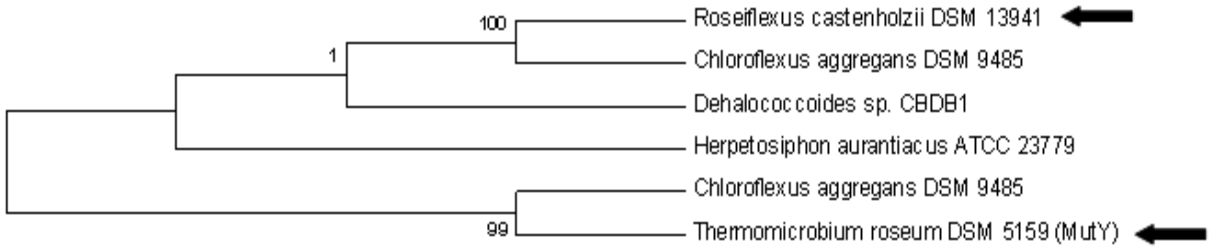


Fig. 3.1(C) The NJ tree of 6 Chloroflexi species. Two major clades were observed in the tree. Selected species of Chloroflexi are shown by arrow.

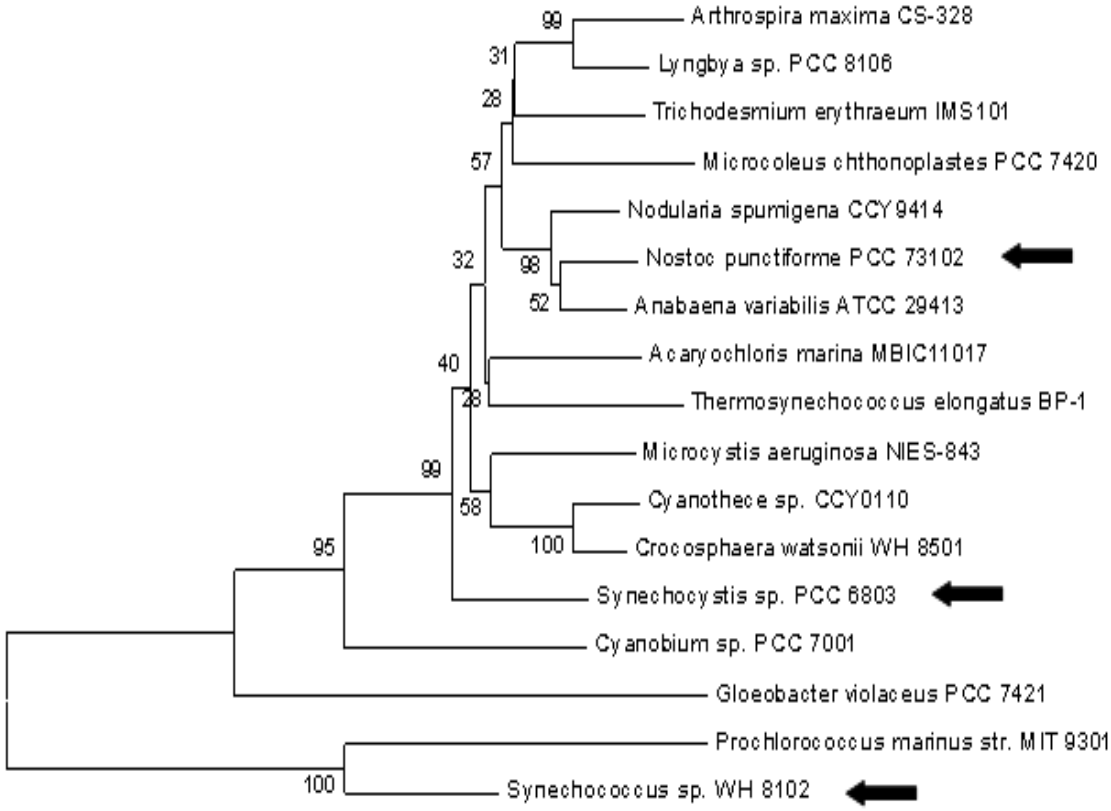


Fig. 3.1(D) The NJ tree of 17 Cyanobacteria produced two major clades. Selected species from Cyanobacteria are shown by arrow.

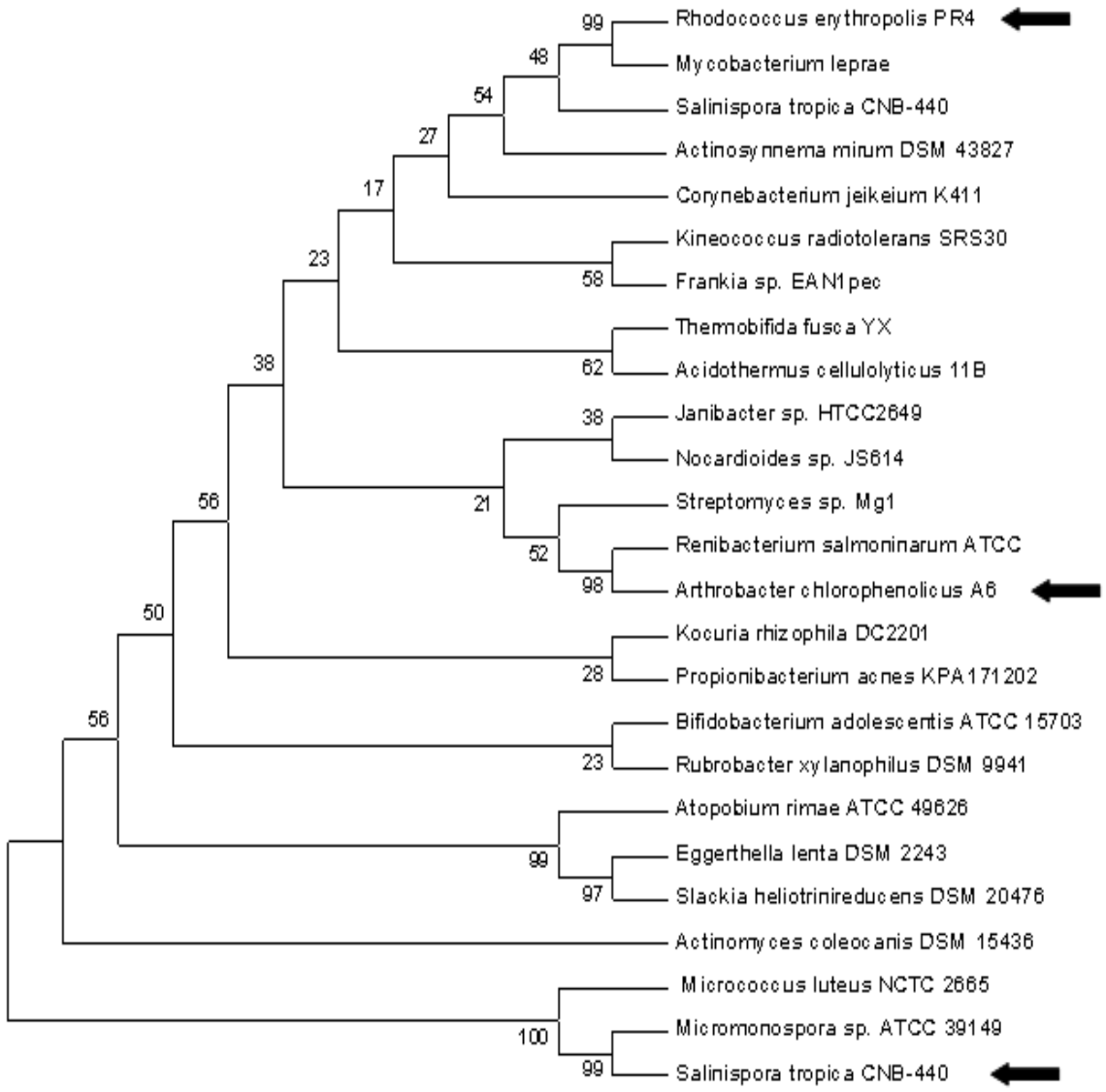


Fig. 3.1(E) NJ tree of 25 Actinobacteria species. Three species were selected which were shown by arrow.

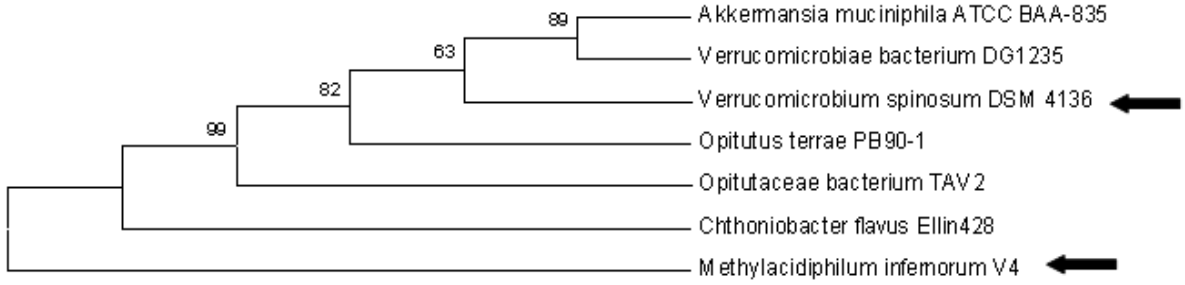


Fig. 3.1(F) NJ tree of 7 Verrucomicrobia species. Only one clade was observed in the tree. Selected species were shown by arrow.

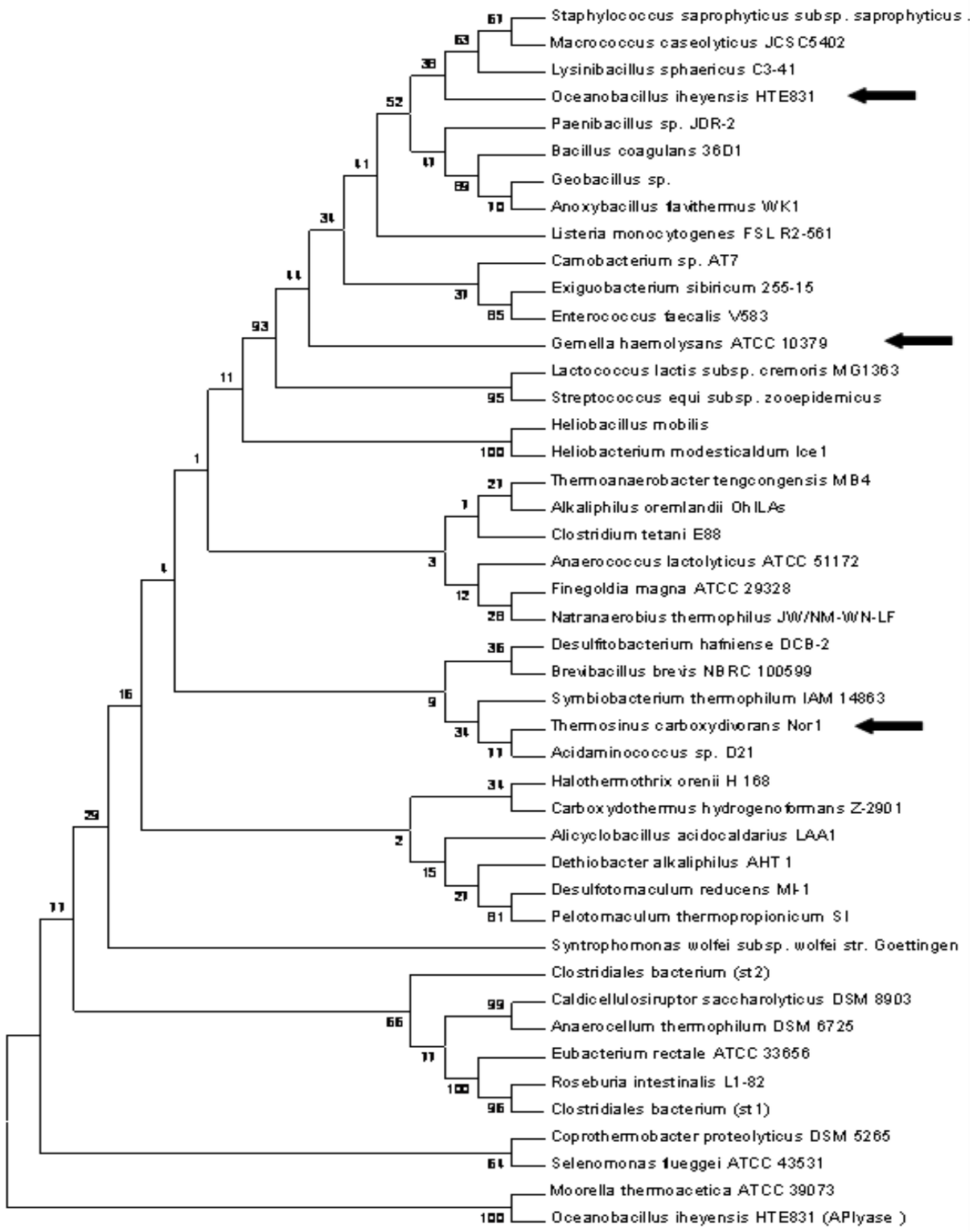


Fig. 3.3(G) Six major clades were observed in the NJ tree of 45 Firmicutes species. Selected species from Firmicutes are shown by arrow.

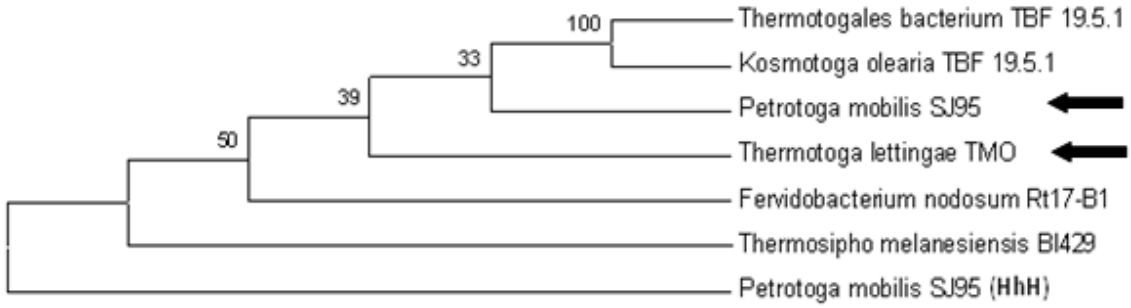


Fig. 3.1(H) NJ tree of 7 Thermotogae species. Only one clade was observed in the tree. Selected species were shown by arrow.

Four NJ trees are constructed for other four kingdoms (Fungi, Protista, Plantae and Animilia) and total 10 representative homologs were chosen after visual inspection (Table 3.3 & Fig. 3.2(A-D).

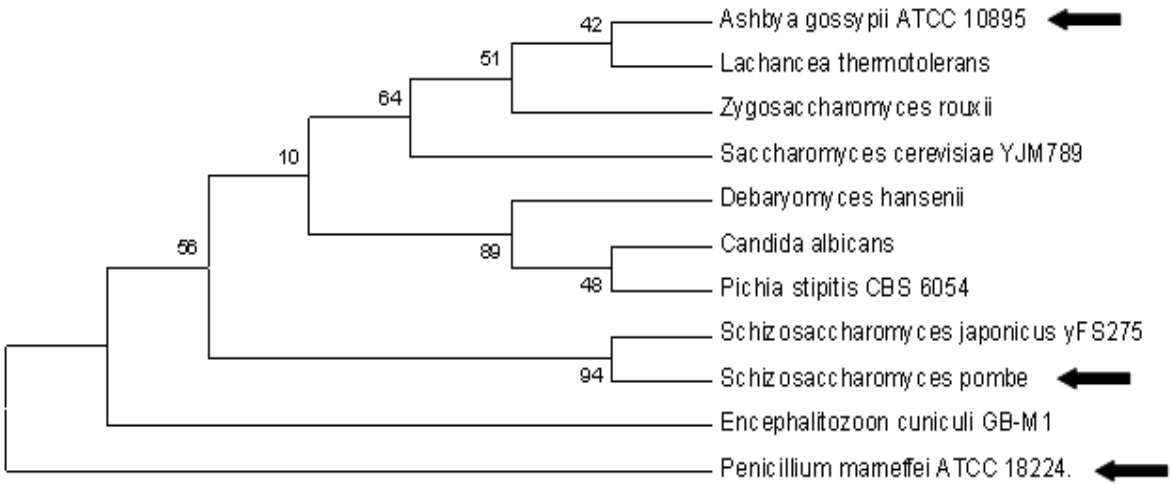


Fig. 3.2(A) The NJ tree of 11 Fungi species. Selected species from Fungi are shown by arrow.

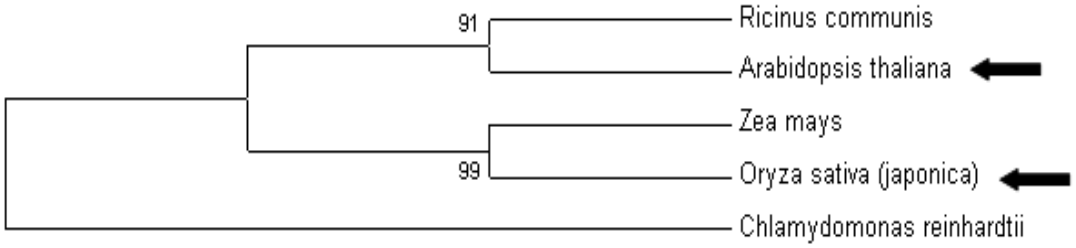


Fig. 3.2(B) The NJ tree of 5 Plantae species. Selected species from Plantae are shown by arrow.

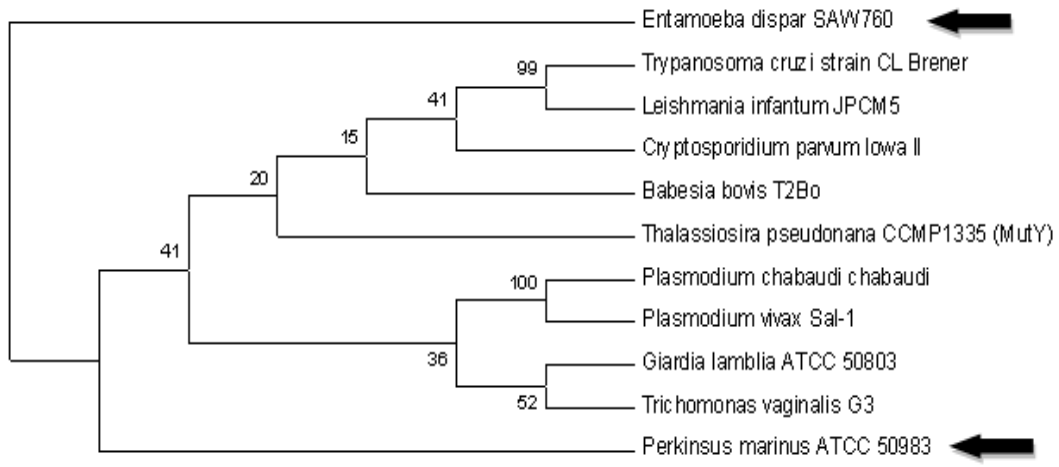


Fig. 3.2(C) The NJ tree of 11 Protista species. Two selected species are shown by arrow.

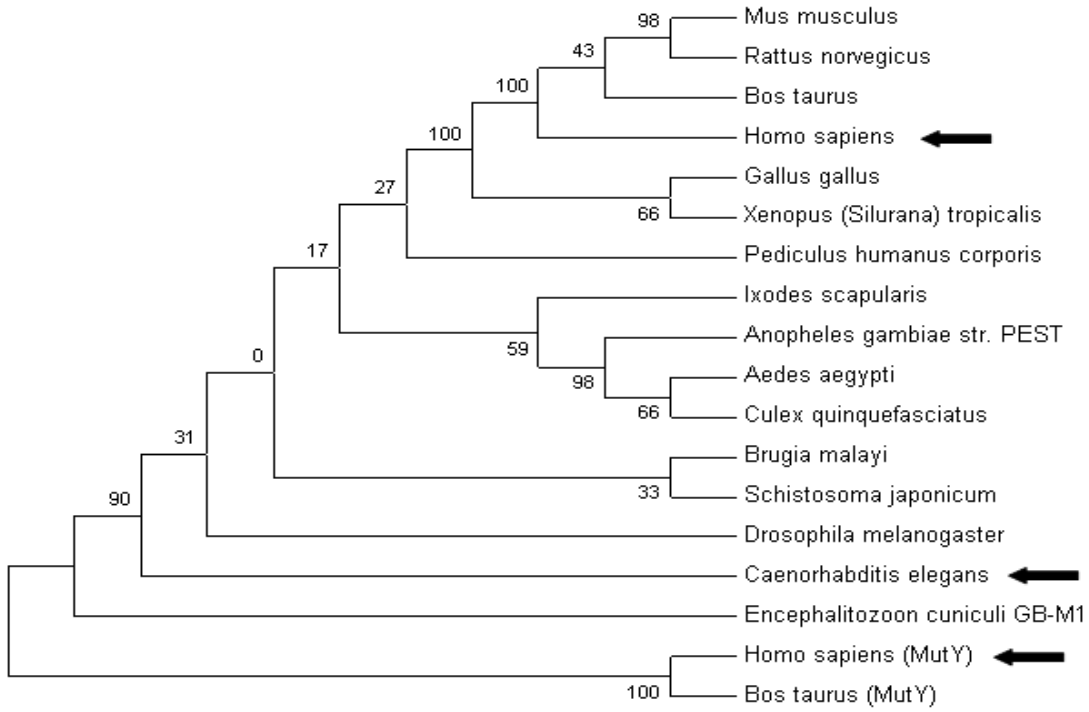


Fig. 3.2 (D) The NJ tree of 18 Animalia species. Selected species from Animalia are shown by arrow.

The NJ tree of 42 archaeal homologs produces four major clades and four representative homologs were chosen for further studies (shown in Fig. 3.3).

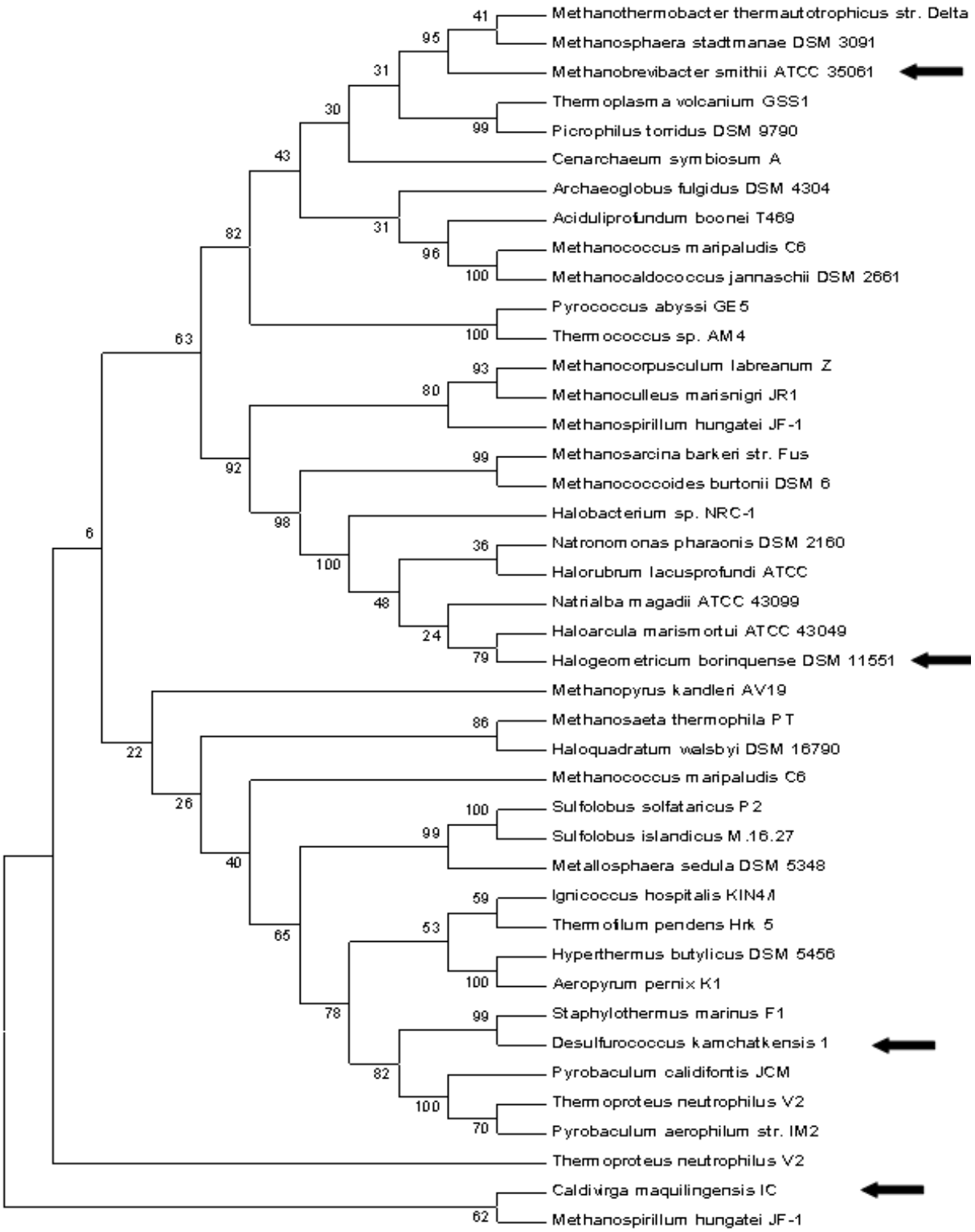


Fig 3.3 The NJ tree of 42 Archaeal homologs. Four clades were observed in the tree, selected species were indicated by arrow.

Finally, a set of 54 homologs of endonuclease III were selected for evolutionary study of the endonuclease III protein family. This set of protein homologs are from 22 bacterial, 2 archaeal and 4 eukaryotic divisions (Table 3.3). NCBI accession number of all 54 homologs of endonuclease III is listed in Table 3.4.

Table 3.3 List of no. of homologs selected from total number of homologs from each division of archaea, bacteria and eukaryotes.

Serial No.	Phylum	Division	Total no. of homologs	No. of homologs selected finally
1.	Archaea		42	4
2. 1	Bacteria	Bacteroidetes	19	3
2.2		Chlorobi	5	2
2.3		Chlamydiae	2	1
2.4		Chloroflexi	6	2
2.5		Cyanobacteria	17	3
2.6		Deinococcus	4	1
2.7		Dictyoglomi	1	1
2.8		Elusimicrobia	1	1
2.9		Actinobacteria	25	3
2.10		Aquificae	5	1
2.11		Lentisphaerae	2	1
2.12		Nitrospirae	2	1
2.13		Planctomycetes	4	1
2.14		Spirochaetes	3	2
2.15		Firmicutes	45	3
2.16		Thermotogae	7	2
2.17		Verrucomicrobia	7	2
2.18		Proteobacteria	215	6
2.19		Gemmatimonadetes	1	1
2.20		Tenericutes	1	1
2.21		Fusobacteria	1	1

2.22		Acidobacteria	3	1
3.1	Eukaryotes	Fungi	11	3
3.2.		Protista	11	2
3.3.		Plantae	5	2
3.4.		Animalia	18	3
		Total	**463**	**54**

Table 3.4 The NCBI accession numbers of 54 endonuclease III protein homologs with their carrier organism as well the length of the protein sequences is shown.

S. No.	**Name of the Organism**	**Accession no. (Protein seq.)**	**Endonuclease III homologs protein name**	**Protein sequence length**
1.	*Ashbya gossypii ATCC 10895*	NP_983680.1	ACR278Wp (Endonuclease III)	367
2.	*Aeromonas salmonicida subsp. salmonicida A449*	YP_001142271.1	Endonuclease III	213
3.	*Arthrobacter chlorophenolicus A6*	YP_002489216.1	Endonuclease III	291
4.	*Borrelia recurrentis A1*	YP_002223188.1	Endonuclease III	205
5.	*Chlamydia muridarum str. Nigg*	NP_296453.1	Endonuclease III	210
6.	*Chlorobium tepidum TLS*	NP_662592.1	Endonuclease III	213
7.	*Deinococcus radiodurans R1*	NP_294012.1	Endonuclease III	225
8.	*Dictyoglomus turgidum DSM 6724*	YP_002352582.1	Endonuclease III	210
9.	*Schizosaccharomyces pombe 972h-*	NP_593210.1	Endonuclease III	355
10.	*Elusimicrobium minutum Pei191*	YP_001875052.1	Endonuclease III	215
11.	*Arabidopsis thaliana*	CAC16135.1	Endonuclease III	354
12.	*Chloroherpeton thalassium ATCC 35110*	YP_001995475.1	Endonuclease III	213
13.	*Halogeometricum borinquense DSM 11551*	ZP_03997896.1	Endonuclease III	227
14.	*Halorhodospira halophila*	YP_001002308.1	Endonuclease III	213
15.	*Oryza sativa*	AAX96284.1	Endonuclease III	373
16.	*Sulfurihydrogenibium sp. YO3AOP1*	YP_001931649.1	Endonuclease III	209
17.	*Entamoeba dispar*	XP_001733691.1	Endonuclease III	241
18.	*Escherichia coli str. K-12 substr. MG1655*	NP_416150.1	Endonuclease III	211
19.	*Gemella haemolysans ATCC 10379*	ZP_04777149.1	Endonuclease III	214
20.	*Gemmatimonas aurantiaca T-27*	YP_002760336.1	Adenine glycosylase	221
21.	*Gramella forsetii KT0803*	YP_861821.1	Endonuclease III	218

22.	*Penicillium marneffei ATCC 18224*	XP_002143355.1	HhH-GPD family protein	449
23.	*Lentisphaera araneosa HTCC2155*	ZP_01872881.1	Endonuclease III	212
24.	*Thermomicrobium roseum (MutY)*	YP_002524150.1	MutY	358
25.	*Methanobrevibacter smithii ATCC 35061*	YP_001272845.1	Endonuclease III	210
26.	*Methylacidiphilum infernorum V4*	YP_001940005.1	A/G-specific DNA glycosylase	355
27.	*E.coli (MutY)*	1KG4|A	MutY	225
28.	*Homo sapiens (MutY)*	CAI21715.1	MutY	291
29.	*Nostoc punctiforme PCC 73102*	YP_001867358.1	Endonuclease III	229
30.	*Caenorhabditis elegans*	NP_497859.1	Nth-1 (Endonuclease III)	259
31.	*Homo sapiens*	AAH00391.2	Nthl1 (Endonuclease III)	305
32.	*Oceanobacillus iheyensis HTE831*	NP_691963.1	DNA-lyase	222
33.	*Parabacteroides distasonis ATCC 8503*	YP_001302528.1	Endonuclease III	221
34.	*Perkinsus marinus ATCC 50983*	EER15889.1	Endonuclease III	292
35.	*Petrotoga mobilis SJ95*	YP_001568415.1	Endonuclease III	210
36.	*Planctomyces maris DSM 8797*	ZP_01852813.1	Endonuclease III	240
37.	*Rhizobium leguminosarum bv. trifolii*	ZP_02297317.1	Endonuclease III	298
38.	*Rhodococcus erythropolis PR4*	YP_002763929.1	Endonuclease III	261
39.	*Roseiflexus castenholzii DSM 13941*	YP_001432854.1	DNA-(AP) lyase	219
40.	*Saccharophagus degradans 2-40*	YP_528076.1	Endonuclease III	227
41.	*Salinibacter ruber DSM 13855*	YP_446456.1	Endonuclease III	324
42.	*Salinispora tropica CNB-440*	YP_001161125.1	Endonuclease III	276
43.	*Synechococcus sp. WH 8102*	NP_897435.1	Endonuclease III	217
44.	*Synechocystis sp. PCC 6803*	NP_441082.1	Endonuclease III	219
45.	*Leptospirillum sp. Group II '5-way CG'*	EDZ38087.1	Endonuclease III	241
46.	*Caldivirga maquilingensis IC-167*	YP_001540481.1	HhH-GPD family protein	233
47.	*Verrucomicrobium spinosum*	ZP_02928493.1	Endonuclease III	217
48.	*Acholeplasma laidlawii PG-8A*	YP_001620232.1	Endonuclease III	214
49.	*Fusobacterium nucleatum subsp. polymorphum ATCC 10953*	YP_002164365.1	DNA-(AP) lyase	216
50.	*Candidatus Solibacter usitatus Ellin6076*	YP_825039.1	Endonuclease III	219
51.	*Thermotoga lettingae TMO*	YP_001471128.1	Endonuclease III	217
52.	*Thermosinus carboxydivor Nor1*	ZP_01666970.1	Endonuclease III	213
53.	*Treponema pallidum subsp. pallidum*	NP_219212.1	Endonuclease III	211
54.	*Arthrobacter chlorophenolicus A6*	YP_002489216.1	Endonuclease III	291

3.3 Results and Discussions

3.3.1 Domains and motifs of endonuclease III family

Domains and motifs are the two most important entities that are known to be conserved during evolution and are considered important parameter for gauging the evolution process (Kanchan et al. 2014). By applying conserved domain (CD) search, we have identified ENDO3c (smart00478) and FES (smart00525) as two major domains. ENDO3c is the main domain that spans residues 30 to 183 of *E.coli* endonuclease III protein. The ENDO3c domain spans almost entire protein in all the homologs of endonuclease III. A 19-residue Helix-hairpin-Helix (HhH) segment within the ENDO3c domain is responsible for non specific DNA binding through hydrogen bonds between N-atoms of the protein backbone and the phosphate groups of DNA. Multiple sequence analysis (MSA) of 54 homologs suggests the presence of a consensus $L_{111}X_2LP_{115}GVG_{118}XK_{120}TA_{122}$ sequence (Fig. 3.4) within HhH motif. Among these highly conserved residues, K120 is the main catalytic residue that is responsible for AP lyase activity (Thayer et al. 1995).

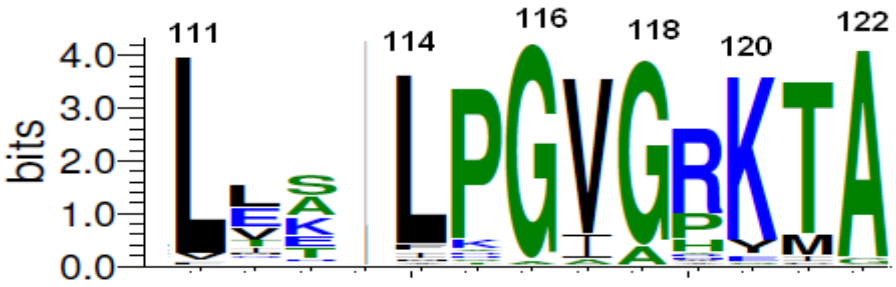


Fig. 3.4 Sequence conservation within HhH motif between residues 111 to 122 is shown by sequence logo. A bits score of 3.2 and above corresponds to more than 80% sequence conservation.

Mutation of K120 in seven AP endonulease III homologs (AP endonucleases of *Gemmatimonas aurantiaca, Methylacidiphilum infernorum, Oceanobacillus iheyensis, Caldivirga maquilingensis* and MutY of *E.coli, Homo sapiens* and *Thermomicrobium roseum*) suggests a loss of AP lyase activity. All these homologs are within mono-functional AP endonuclease protein family with only glycosylase activity. Conserved L111, L114 and A122 residues are a part of the hydrophobic core structure, which stabilizes the fold of HhH motif. Beta turn is formed by residues from P115 to G118 and is a key structure in HhH motif. V117 side chain within the beta turn region is oriented towards the hydrophobic core. Interaction among the hydrophobic residues is key to the overall architecture of the HhH motif which, in turn, helps in positioning the catalytically important K120 residue towards the active site cavity (Fig. 3.5). The D138 residue within the ENDO3c domain is also conserved among 53 homologs. This residue

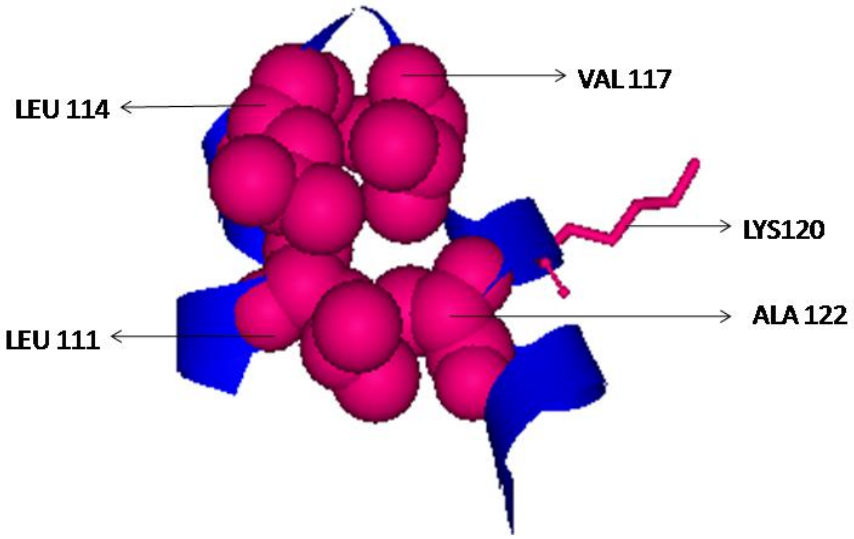


Fig. 3.5 HhH motif from *E.coli* AP endonuclease III crystal structure (PDB ID: 2ABK) by Discovery Studio software package is shown by ribbon diagram. Hydrophobic core forming residues are shown by spacefill model while orientation of catalytic L120 is shown by stick representation.

initiates nucleophilic attack during the glycosylase activity (Manuel et al. 2004). Absence of both lysine and aspartic acid residues at position 120 and 138 in *Caldivirga* indicates that this protein may not have glycosylase or lyase activity, though it contains ENDO3c domain. Around 100 amino acid N-terminus insertion is observed in endonuclease III protein of plant species. The detail sequence comparison predicts (Emanuelsson et al. 2007) the existence of a possible signal peptide within the inserted region, which is targeted to the chloroplast.

The FES domain (smart00525) at C-terminal end of the AP-endonuclease III protein is present in most of the homologs that contains a 21 residues iron–sulfur cluster loop (FCL) motif. Iron–sulfur cluster plays an important role in orienting the positively charged residues of the FCL motif for DNA binding (Lukianova & David 2005). MSA of 54 homologs suggests the presence of a consensus G_{183}-X_3-Cys_{187}-X_6-Cys_{194}-X_2-Cys_{197}-X_5-Cys_{203} sequence (Fig. 3.6) in the FES domain. It has been demonstrated that in case of *E.coli* MutY, the mutation of highly conserved cysteine residues reduces the stability of iron-sulfur cluster and the extent of destabilization is position dependent (Golinelli et al. 1999). Out of the total 54 selected homologs, all four cysteine residues are conserved in 47 homologs and all four cysteine residues are absent in *Gemella haemolysans* and *Gramella forsetii*, while the other five homologs contain at least two conserved cysteine residues. This data suggests that FCL motif may be absent in *Gemella haemolysans* and *Gramella forseti* and for the formation as well as stabilization of Fe-S cluster two cysteine residues are sufficient as indicated in previous mutational studies (Golinelli et al. 1999).

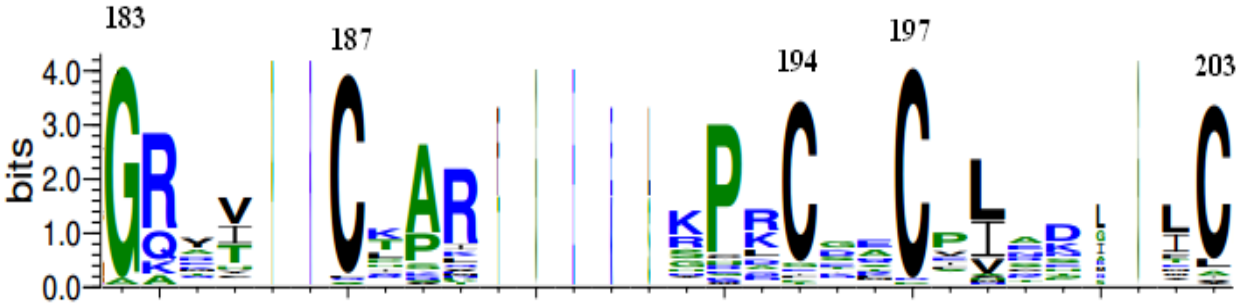


Fig. 3.6 Consensus sequence between residues 183 and 203 is shown by sequence logo representation. G183, C187, C194, C197 and C203 are seen conserved for more than 80% sequences.

The structure of the *E .coli* endonuclease III protein (Thayer et al. 1995) contains six helix barrel and FCL cluster domains. Crystal structure (PDB ID: 2ABK) also reveals that helices 2-7 are within the helix barrel domain; whereas, helix-1 and helices 8-10 belong to the FCL cluster. Multiple sequence analysis reveals that substantial sequence conservation exists within helix 2 and helix 3 containing helix-turn-helix region (shown in Fig. 3.7(A)). Among the conserved residues, Q41 is known to play a key role in substrate recognition as it penetrates through major groove of DNA and interacts with the AP site of DNA (Fromme & Verdine 2003) whereas, S39 and D44 form polar interactions with other segments of protein which give segmental flexibility. The L33 and L38 are part of the hydrophobic core that stabilizes this particular fold. In the case of FeS domain, residues within helix 8 (Fig. 3.7(B)) along with four cysteine residues are highly conserved. The conserved T139, H140, R143 and R147 residues form a hydrophilic/charge surface within helix 8 that interacts with DNA backbone and gives stability during DNA binding.

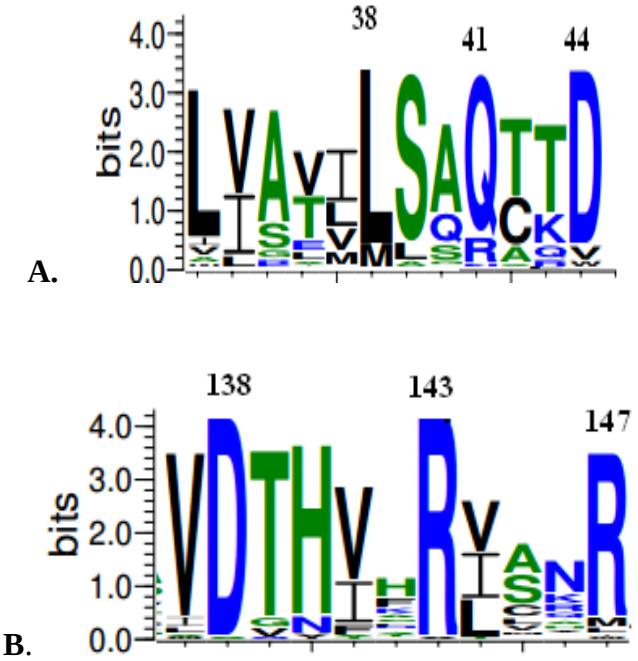


Fig. 3.7(A) Consensus sequence between residues 33 and 44 is shown by sequence logo representation. L33, L38, S39, Q41 and D44 are seen conserved for more than 80% sequences. **(B).** Conservation of residues from 137 to 147 is shown by sequence logo. V137, D138, T139, H140, R143 and R147 are conserved for more than 80% sequences.

In the endonuclease III protein family, the efficiency and specificity of enzyme are highly dependent on architecture of DNA binding HhH and FCL motifs. These motifs bind to DNA substrate with the help of positively charged amino acid residues. Sequence analysis of all 54 homologs demonstrates that the important basic residues are conserved throughout the evolution. We found that among all the homologs, on an average 3.74 positively charged residues are present within FCL motif and are strategically positioned to interact with negatively charged DNA backbone. In case of *Gemella haemolysans* and *Gramella forseti,* DNA binding activity of

the enzyme is entirely dependent on the HhH motif as this protein does not contain FCL motif. Hence, we speculate that enzyme activity and specificity will decrease for these species.

3.3.2 16S/18S rRNA gene based species tree

Complete 16S/18S r-RNA gene sequences of 47 species were retrieved while for rest of the species partially sequenced 16S/18S r-RNA gene were available in the databases. Partially sequences 16S/18S r-RNA genes were not considered to generate the species tree. Species tree based on 16S/18S rRNA gene sequences was shown in Fig 3.8.

3.3.3 Endonuclease III gene based phylogenetic tree

Phylogenetic tree of gene sequences of *endonuclease III* family was constructed from alignment using both maximum likelihood (ML) and neighbor joining (NJ) method. The topologies of both trees generated using the ML and NJ methods are very similar to each other. Bootstrap analysis reveals that most of the clades in both trees are robust and majority of clades are supported by ≥50% bootstrap value and therefore, only one tree (ML tree) is discussed. *Endonuclease III* gene tree was compared with the 16S/18S rRNA gene sequence-based species tree (as shown in Fig. 3.8). A total of 14 distinct clades were formed by 16S/18S r-RNA gene sequences of these species. Out of these 14 clades, 8 clades contain species from the same phylum/division. Given that only a few bacterial divisions contain single species, the remainder of the clades contains species from different bacterial divisions. Details of the 16S/18S rRNA sequence-based species

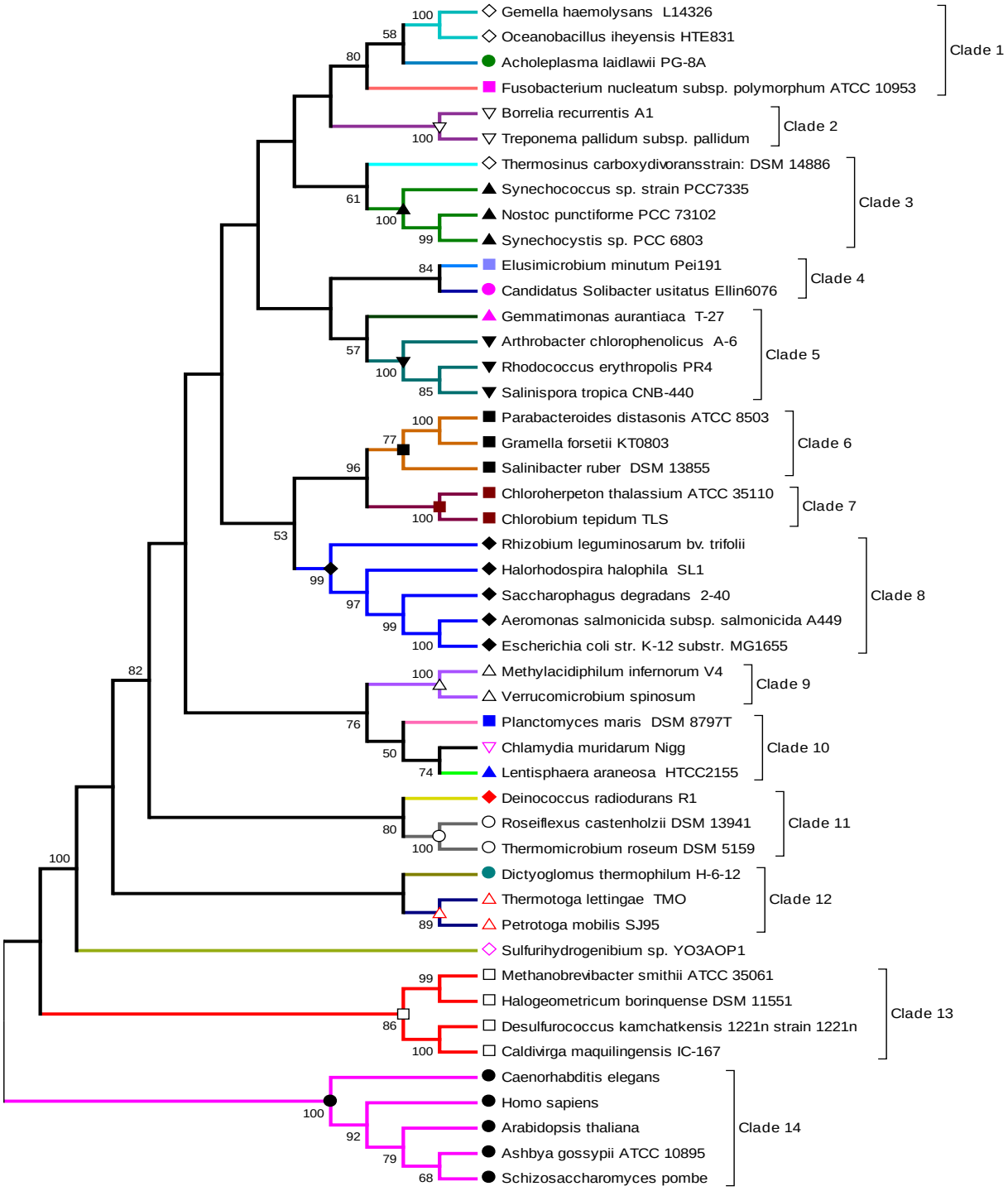


Fig. 3.8 16S/18S rRNA based Maximum likelihood tree. Bootstrap support values are presented next to the tree branches for each clade with >50. The tree is generated from a clustalw based multiple sequence alignment.

tree reveal that almost all species in different bacterial divisions, as well as species from archaea and eukaryotes form distinct clade, with the sole exception of *Thermosinus carboxydivorans*. As a Firmicutes bacterium, *Thermosinus carboxydivorans* shares a clade with the cyanobacterial species. A total of 11 distinct clades are formed by the homologs of the endonuclease III genes (Fig. 3.9). All eukaryote homologs (Clade 8) and homologs from three bacterial divisions (Cyanobacteria-clade 4, Thermotogae-clade 11 and Chlorobi-clade 1) stay close to each other within the phylogenetic clade. Interestingly, four out of five mono-functional *endonuclease III* genes of different species are clustered together (clade 9), indicating a co-evolution of mono-functional *endonuclease III* genes. Each of the other seven distinct clades contains homologs from different bacterial divisions. *Endonuclease III* gene-based phylogeny tree suggests that the gene and species evolution of Cyanobacterial, Thermotogal, Chlorobial and eukaryotic organisms may have similar pattern, as these classes of organisms stay within a distinct clade in both species and gene based phylogenetic tree. It is interesting to note that the *endonuclease III* genes of four archaeal species are within different lineages (shown as red branch in Fig. 3.9), although in species tree, all archaeal species are within the same clade (Fig. 3.8).

This observation indicates that the endonuclease III gene of these archaeal species is evolved differently, suggesting that environmental pressure might have played an important role in shaping *endonuclease III* gene. For example, *endonuclease III* gene of *Methanobrevibacter smithii* (archaea) shares the same clade (Clade 11) with two bacterial species *Petrotoga mobilis* and *Thermotoga lettingae*, which are anaerobic in nature and are associated with fermentation. *Endonuclease III* gene based phylogeny analysis shows that a large number of species share a clade with species from different divisions. In order to explain this anomaly of sharing different

clades by the species belonging to the same division, we have calculated mean GC content of the gene, GC content at the third codon position and GC content of the species, which could be an appropriate tool to explain the clustering pattern in phylogenetic tree (Brochier et al. 2000).

Table 3.5 lists out the GC content of endonuclease III gene, 3rd codon position of the endonuclease III gene and genome of the corresponding organism. Except for *Homo sapiens,* GC content of endonuclease III gene and GC content of all other organisms are close to each other. In *Homo sapiens* GC content of species is about 24% less than the GC content of endonuclease III gene. The average GC content and GC content at the third codon position of *Endonuclease III* gene of *Dictyoglomus thermophilum, Acholeplasma laidlawii, Gemella haemolysans, Borrelia recurrentis* and *Fusobacterium nucleatum* (clade 7) are 30.8 % and 18.1 % respectively (standard deviation of 3.3 and 6.3) which could possibly explain why these species share the same clade. Similarly, archaeal species *Methanobrevibacter smithii*, two species from Thermotogae division and one species each from Aquificae, Firmicutes, Chlamydiae and mono-functional Verrucomicrobia share the same clade (Clade 11) in the tree due to similar GC content at gene level and GC content at the third codon position (36.7% and 45.1% respectively with standard deviation of 4.8 and 5.1). This observation also points out that the mono-functional *Endonuclease III* gene from a species of Verrucomicrobia division shares different clades unlike other mono-functional genes (Clade 9) in which GC content and GC content at third base of *Endonuclease III* gene are 62.4% and 66.8% respectively. *Gramella forsetii, Parabacteroides distasonis* and *Salinibacter ruber* from Bacteroidetes division share

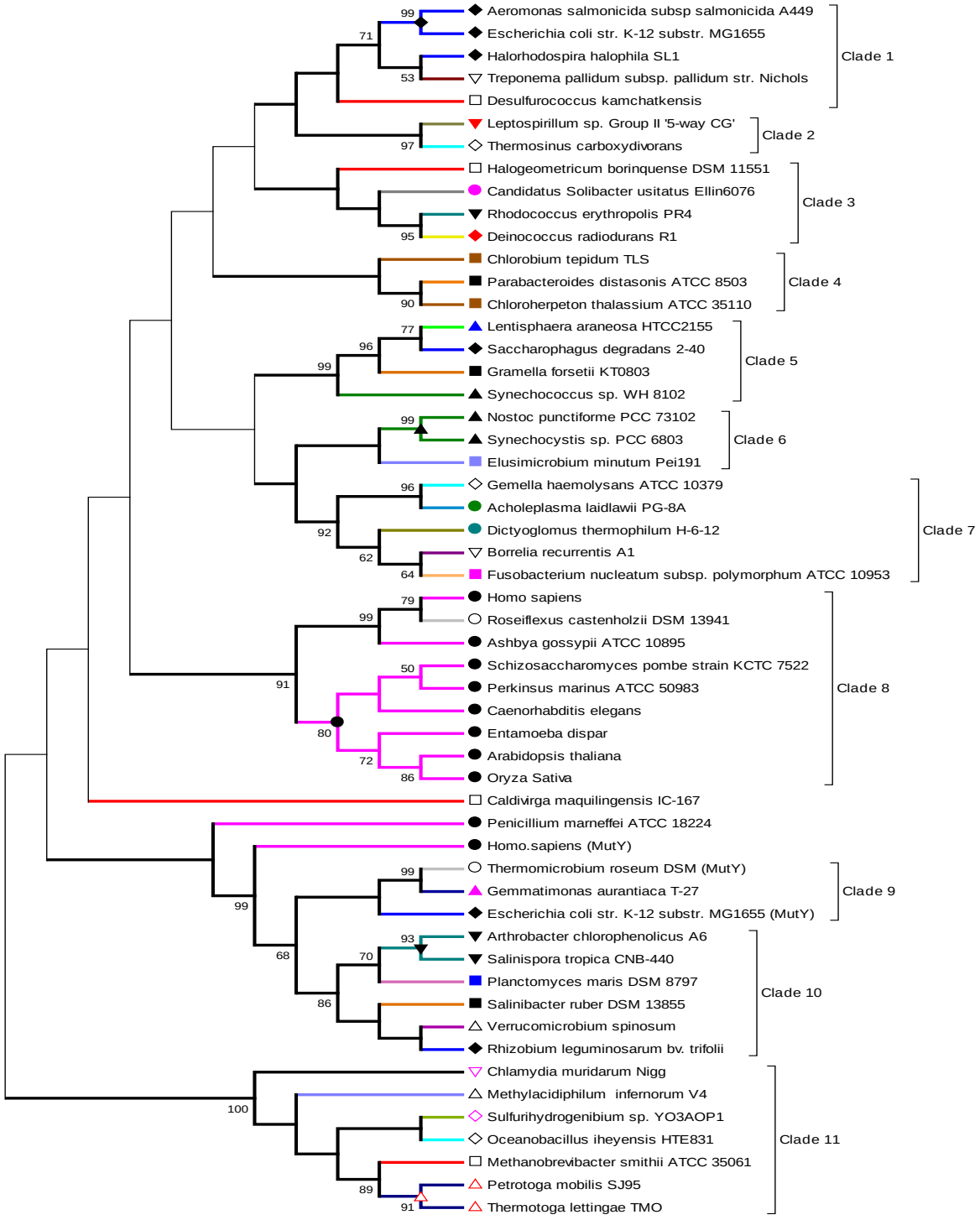


Fig. 3.9 AP endonuclease III gene based Maximum likelihood tree. Bootstrap support values are presented next to the tree branches for each clade with ≥50. The evolutionary distances were computed using the Tamura-Nei substitution model. The bootstrap values are given as Fig. 3.8.

three different clades as their GC contents and GC content at the third codon position of *Endonuclease III* gene differ significantly (36.5% and 30.6%; 46.1% and 50.5%; 65.6% and 68.3% respectively). These species reside within the clade that contains species with similar GC content. Similar trend is observed in case of species within the Spirochaetes and Chloroflexi division. Being a member of bacterial division, *Roseiflexus castenholzii* is found to share a clade with human AP endonuclease III gene because of its GC content (62%), which is very close to that of human (64%). *Endonuclease III* gene based phylogeny tree of 54 taxa suggests that GC content of gene contribute significantly towards the position of taxa within the tree and species evolution, and *endonuclease III* gene evolution shape up differently in most of the cases. Since AP endonuclease III gene in all living organisms diverges over a long evolutionary period, synonymous nucleotide substitution GC content at third codon position as well as the mean GC content of gene makes gene sequence based phylogenetic tree construction noisy.

Table 3.5 GC content based on endonuclease III gene, 3rd position of the endonuclease gene and genome of the organism as well average and standard deviation for three types of GC content are listed.

Clade No.	Organism Name	Division/Kingdom	Mean Gene GC content (%)	GC at IIIrd position (%)	Mean Genomic GC content (%)	Average (Standard Deviation) of GC content		
						Mean Gene	IIIrd position	Genomic
1	*Aeromonas salmonicida subsp. almonicida A449*	Proteobacteria	57.79	77.10	58.2	56.17 (8.90)	64.83 (21.11)	55.02 (8.60)
	Escherichia coli str. K-12 substr. MG1655	Proteobacteria	49.21	51.89	50.8			
	Halorhodospira halophila	Proteobacteria	70.40	95.33	68			
	Treponema pallidum subsp. pallidum str. Nichols	Spirochaetes	55.19	56.60	52.8			
	Desulfurococcus_kamchatkensis	Archaea	48.28	43.23	45.3			
2	*Leptospirillum sp. Group II '5-way CG'*	Nitrospirae	61.43	69.01	54.4	57.50 (5.55)	63.71 (7.50)	52.95 (2.05)

	Thermosinus carboxydivor Nor1	Firmicutes	53.58	58.41	51.5			
3	*Halogeometricum borinquense DSM 11551*	Archaea	58.33	71.05	60			
	Candidatus Solibacter usitatus Ellin6076	Acidobacteria	59.32	44.06	61.9			
	Rhodococcus erythropolis PR4	Actinobacteria	63.49	77.10	62.3	62.51	69.51	62.7
	Deinococcus radiodurans R1	Deinococcus	68.88	85.84	66.6	(4.80)	(18.02)	(2.78)
4	*Parabacteroides distasonis ATCC 8503*	Bacteroidetes	46.10	50.45	45.1			
	Chloroherpeton thalassium ATCC 35110	Chlorobi	44.86	45.33	45			
	Chlorobium tepidum TLS	Chlorobi	56.70	75.23	56.5	49.22 (6.51)	57.00 (15.99)	48.87 (6.61)
5	*Synechococcus sp. WH 8102*	Cyanobacteria	61.62	74.77	59.4			
	Lentisphaera araneosa HTCC2155	Lentisphaerae	41.31	31.92	40.9			
	Saccharophagus_degradans	Proteobacteria	48.68	50.88	45.8	47.03	47.04	45.68
	Gramella forsetii KT0803	Bacteroidetes	36.53	30.59	36.6	(10.9)	(20.67)	(9.89)
6	*Nostoc punctiforme PCC 73102*	Cyanobacteria	43.62	36.96	41.3			
	Synechocystis sp. PCC 6803	Cyanobacteria	51.21	53.64	47.3	45.34	44.09	42.87
	Elusimicrobium minutum Pei191	Elusimicrobia	41.20	41.67	40	(5.22)	(8.60)	(3.89)
7	*Dictyoglomus turgidum DSM 6724*	Dictyoglomi	34.28	27.96	33.7			
	Acholeplasma_laidlawii	Tenericutes	33.02	18.14	31.9			
	Gemella haemolysans ATCC 10379	Firmicutes	31.78	16.28	30.8			
	Borrelia recurrentis A1	Spirochaetes	26.38	10.68	27.5	30.79	18.08	30.12
	Fusobacterium nucleatum	Fusobacteria	28.49	17.38	26.7	(3.27)	(6.25)	(2.96)
8	*Homo sapiens*	Eukaryota	64.16	77.78	40.9			
	Roseiflexus castenholzii DSM 13941	Chloroflexi	61.52	77.27	60.7			
	Ashbya gossypii ATCC 10895	Eukaryota	47.64	53.26	51.4			
	Schizosaccharomyces pombe 972h-	Eukaryota	38.58	29.21	36			
	Arabidopsis thaliana	Eukaryota	44.32	36.06	36.1			
	Oryza sativa	Eukaryota	45.45	41.18	43.7			
	Perkinsus marinus ATCC 50983	Eukaryota	50.63	53.24	47.4			
	Entamoeba dispar	Eukaryota	29.34	12.40	24.1	47.26	46.58	41.73
	Caenorhabditis elegans	Eukaryota	43.72	38.85	35.3	(10.7)	(21.43)	(10.5)
9	*Thermomicrobium roseum (MutY)*	Chloroflexi	67.87	72.98	64.3			
	Homo sapiens (MutY)	Animalia	61.05	64.60	40.9			
	Gemmatimonas aurantiaca T-27	Gemmatimonadetes	66.07	69.37	64.3			
	Escherichia coli str. K-12 (MutY) (MutY)	Proteobacteria	54.47	60.24	50.8	62.36 (6.00)	66.79 (5.56)	55.07 (11.9)
10	*Arthrobacter chlorophenolicus A6*	Actinobacteria	69.41	71.23	66			
	*Salinispora_tropica_*	Actinobacteria	71.36	70.76	69.5			
	Planctomyces maris DSM 8797	Planctomycetes	53.39	56.02	50.5			
	Salinibacter ruber DSM 13855	Bacteroidetes	65.64	68.31	66.1			
	Verrucomicrobium_spinosum DSM	Verrucomicrobia	58.10	62.84	58.5			
	Rhizobium_leguminosarum_bv._trifolii WSM1325	Proteobacteria	62.99	61.20	60.8	63.48 (6.83)	65.06 (6.04)	61.9 (6.85)
11	*Chlamydia muridarum Nigg*	Chlamydiae	42.34	50.24	40.3			
	Methylacidiphilum infernorum V4	Verrucomicrobia	44.10	49.44	45.5			
	Sulfurihydrogenibium sp.	Aquificae	30.48	41.43	32	36.67 (4.83)	45.13 (5.70)	36.76 (5.11)

	YO3AOP1							
	Oceanobacillus iheyensis HTE831	Firmicutes	35.79	50.69	35.7			
	Petrotoga mobilis SJ95	Thermotogae	33.81	35.07	34.1			
	Methanobrevibacter smithii ATCC 35061	Archaea	34.76	45.97	31			
	Thermotoga lettingae	Thermotogae	35.47	43.12	38.7			

3.3.4 Endonuclease III protein based phylogenetic tree

Overall, seven distinct branches have been identified from the AP-endonuclease III protein based phylogenetic tree. Among them three branches are divided further to form ten clades (Fig. 3.10). Bi-functional AP-endonuclease III homologs of all eukaryotic organisms (except *Penicillum marneffei*), its mono-functional homologs, and homologs from Thermotogal, Actinobacterial division form four homogeneous distinct clades. We observed that organisms from Cyanobacterial, Chlorobial and Proteobacterial (except *Saccharophagus degradans*) division also stay together in the phylogenetic tree. Other clades contain proteins from species belonging to different divisions. Proteins from *Desulfurococcus kamchatkensis* and *Penicillum marneffei* as well as *Oceanobacillus iheyensis* and *Caldivirga maquilingensis* form two distinct clades which are out-grouped from rest of the species. A closer inspection of multiple sequence alignment of AP endonuclease III protein shows that along with the ~150-amino acid extra N-terminal region, there are three insertions within the helix barrel domain of *Penicillum marneffei*. This insertion (of ~57 amino acids) distinguishes *Penicillum marneffei* from other eukaryotic organisms. Comparison of protein sequences from *Caldivirga maquilingensis* and *Oceanobacillus iheyensis* reveals that the catalytic K120 residue is absent in both the species and D138 is absent in case of

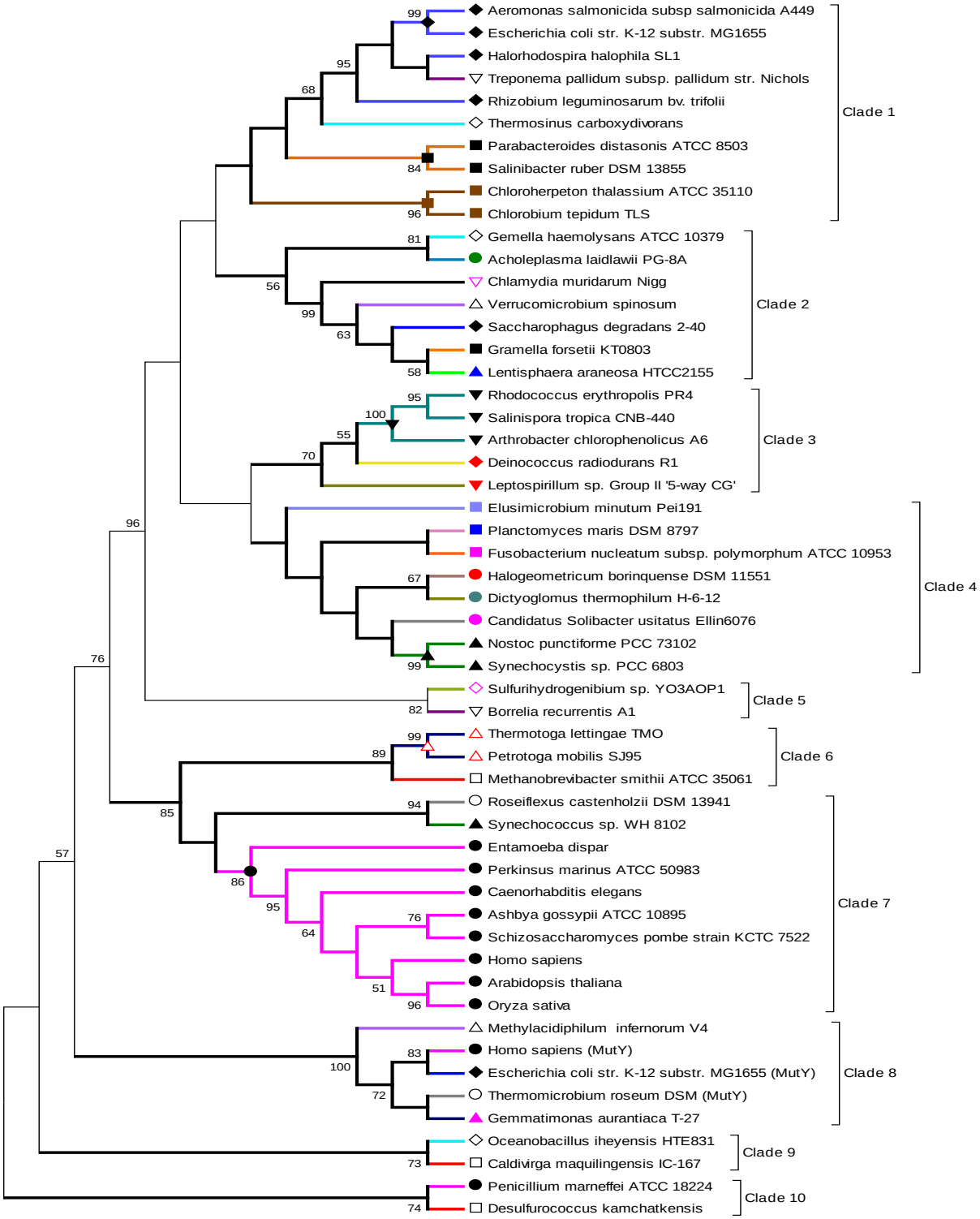


Fig. 3.10 AP endonuclease III protein sequences based Maximum likelihood tree. The evolutionary distances were computed using the JTT as substitution model. The bootstrap values are given as Fig. 3.8.

Caldivirga maquilingensis which supports the idea that the proteins from these two organisms belong to the HhH super-family (as both of these proteins contain HhH motif) but not in endonuclease III sub-family. All archaeal species occupy different clades in the protein-based tree just as observed in case of gene-based tree. Four archaeal species namely, *Halogeometricum borinquense, Caldivirga maquilingensis, Desulfurococcus kamchatkensis* and *Methanobrevibacter smithii* share clades with *Dictyoglomus thermophilum* (Dictyoglomi), *Oceanobacillus iheyensis* (Firmicutes), *Penicillum marneffei* (Fungi) and *Petrotogamobilis* (Thermotogae), respectively. This observation reaffirms that evolution of the endonuclease III protein within archaeal species is highly influenced by the environmental factors. It is interesting to note that the endonuclease III proteins of bacterial species *Roseiflexus castenholzii* and *Synechococcus sp.* stay very close to those of eukaryotic species, indicating a probable horizontal gene transfer (HGT) process. From this analysis it is evident that mean GC content and GC content at third base have influenced the location of species within a tree. Thus, compared to a gene-based tree, protein-based analysis gives a better picture of the evolutionary history of a protein.

3.3.5 Evolution of three dimensional structures of the endonuclease III homologs

Among the homologs of endonuclease III protein family, the structures of endonuclease III protein (PDB ID: 2abk) and MutY (PDB ID: 1KG2) protein from *E.coli* are solved experimentally. The C-alpha (1.85 Å which involved 146 atoms) and mainchain (1.87 Å which involved 592 atoms) RMSDs between these two structures are quite small (shown in Fig 3.11)

indicating that the overall fold of these two proteins are quite similar though the sequence identity between these two sequence is 15.1% only. Although, minor structural differences are observed between these two structures within helix 1, loop 1, helix 2, loop 2 and at loop between helix 10 and helix 11.

Since, we have identified 10 distinct clades within endonuclease III protein sequences based phylogenetic tree, the structural evolution of endonuclease III protein homologs are compared by constructing structural models of representative protein homologs of each clade. As endonuclease III and MutY structures belong to Clade 1 and Clade 8 respectively, the representative three dimensional structures of other clades are constructed through homology modellingusing endonuclease III structure of *E.coli* as a template. The summary of each model is shown in table 3.6.

Fig. 3.11 Superimposed crystal structure of *E.coli* endonuclease III (Red) and *E.coli* MutY (sky blue). 3D diagrams of these models were generated by Accelrys Discovery studio ViewerPro 5.0.

Interestingly, though the sequence identities among all sequences are within the range 25-39%, the qualities of model structures are reasonably good indicating folds of all homologs are same. The RMSD values of modeled structures with their template structure are in the range of 0.56-2.20 Å as listed in Table 3.6. 3D model structure of representative endonuclease III homologs from each clade are also validated though Ramachandran plot which shows that these protein structures are stereo chemically stable as maximum of 1% amino acids are outliers, while rest 99% amino acids are within favored regions of Ramachandran plot. The quality of predicted 3D models is checked by ERRAT plot and overall quality factor are better than 78 (shown in table 3.6). The qualities of model structures are also validated through Qmean server. The QMEAN scores and QMEAN Z-scores also indicate that the quality of models is reasonably good.

Table 3.6 Predicted model quality of endonuclease III homologs from each of the representative clade for which 3D model were generated.

	% identity with the Template	**RMSD values (in Å) with respect to template structure**	**Ramchandran plot statistics**	**Errat 2 Score**	**Qmean Score/Z-Score**
Chlamydia muridarum (Clade no. 2)	34	C_alpha- 1.50 Mainchain- 1.49	Favoured - 97.5% Allowed- 1.5% Outlier- 1.0%	82.65	0.655/-1.23
Arthrobacter chlorophenton (Clade no. 3)	39	C_alpha- 0.46 Mainchain- 0.52	Favoured- 98.9% Allowed- 0.5% Outlier- 0.5%	93.44	0.717/-0.51
Nostoc punctiforme (Clade no. 4)	39	C_alpha- 1.37 Mainchain-1.40	Favoured - 98.5% Allowed- 1.0% Outlier- 0.5%	98.45	0.734/-0.37

Borrelia recurrentis (Clade 5)	36	C_alpha- 2.13 Mainchain- 2.20	Favoured- 97.8% Allowed- 2.2% Outlier- 0.0%	78.29	0.600/-1.70
Methanobrevibacter Smithii (Clade 6)	35	C_alpha- 0.56 Backbone- 0.62	Favoured- 98.4% Allowed- 1.1% Outlier- 0.5%	93.41	0.715/-0.53
Homo sapiens (Clade 7)	31	C_alpha- 0.61 Mainchain- 0.68	Favoured- 97.1% Allowed- 2.9% Outlier- 0.0%	90.53	0.573/-1.96
Oceanobacillus iheyensis (Clade 9)	25	C_alpha- 1.43 Mainchain-1.45	Favoured- 95.7% Allowed- 3.8% Outlier- 0.5%	78.92	0.547/-2.37
Desulfurococcus kamchat (Clade 10)	32	C alpha- 1.37 Mainchain- 1.38	Favoured- 96.3% Allowed- 3.2% Outlier- 0.5%	78.09	0.495/-2.96

The modeled structures of endonuclease III protein homologs of *Arthrobacter*, *Methanobrevibacter* and *Homo sapience* are very close to crystal structure (PDB id: 2abk). The pairwise rms deviations of these four structures are well below 1 Å (Table 3.6 and Fig 3.12).

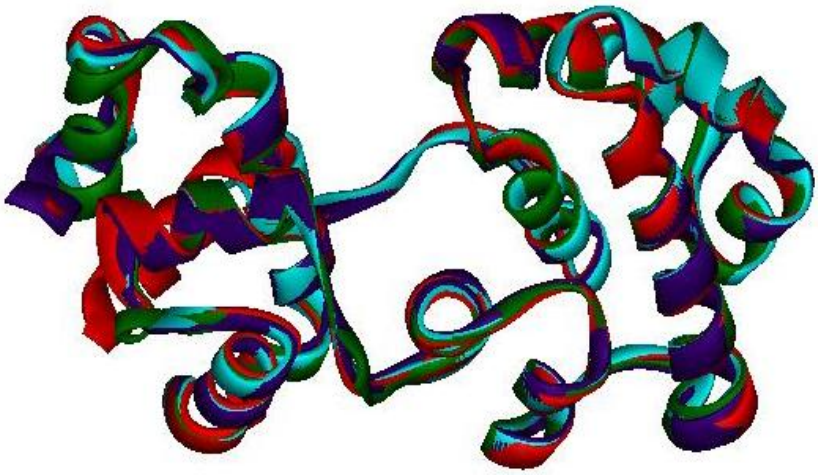


Fig. 3.12 Superimposition of endonuclease III homologs of *Arthrobacter* (sky blue), Human (green), *Methanobrevibactor* (deep blue) with crystal structure (PDB ID:2abk, Red). The mainchain conformation of three modeled structures and crystal structure is shown by ribbon diagram. 3D diagrams of models were generated by Accelrys Discovery studio ViewerPro 5.0.

The endonuclease III protein model of *Borrelia recurrentis* shows maximum deviation from the crystal structure as C-terminal region of both structures do not superimpose well. The conformations of 6-helix barrel region of all model structures are quite similar whereas small conformational deviation is observed towards the C-terminal region.

3.4 A model for the evolutionary history of endonuclease III protein family

We seek to reconstruct the possible evolutionary history of endonuclease III family on the basis of sequence and phylogenetic tree analysis. Proteins of the endonuclease III family belong to the HhH-GPD superfamily and are close to proteins of the MutY family. As shown in Table 1, proteins from the HhH-GPD superfamily and MutY family are retrieved along with endonuclease III family proteins. These proteins are grouped together and are placed as out-group relatives to other clades in the phylogenetic tree. The proposed model (Fig. 3.13) suggests that the eukaryotic lineage of *endonuclease III* gene family diverged from bacteria and archaea during early evolution. One of the important features in this evolutionary model is that a few Horizontal Gene transfer (HGT) events have occurred in endonuclease III gene family from bacteria to eukaryotes and/or archaea. HGT event can be inferred from the observation that the clades of *Roseiflexus* (Chlorofexi) and eukaryotes are closely associated in gene based phylogenetic tree (Kanchan et al. 2015). Similarly, the clade for Thermotogae/other bacterial division homologs shared with archaeal species. Phylogenetic tree of all bacterial and archaeal species reveals a mixed association of endonuclease III proteins. Another possible event of

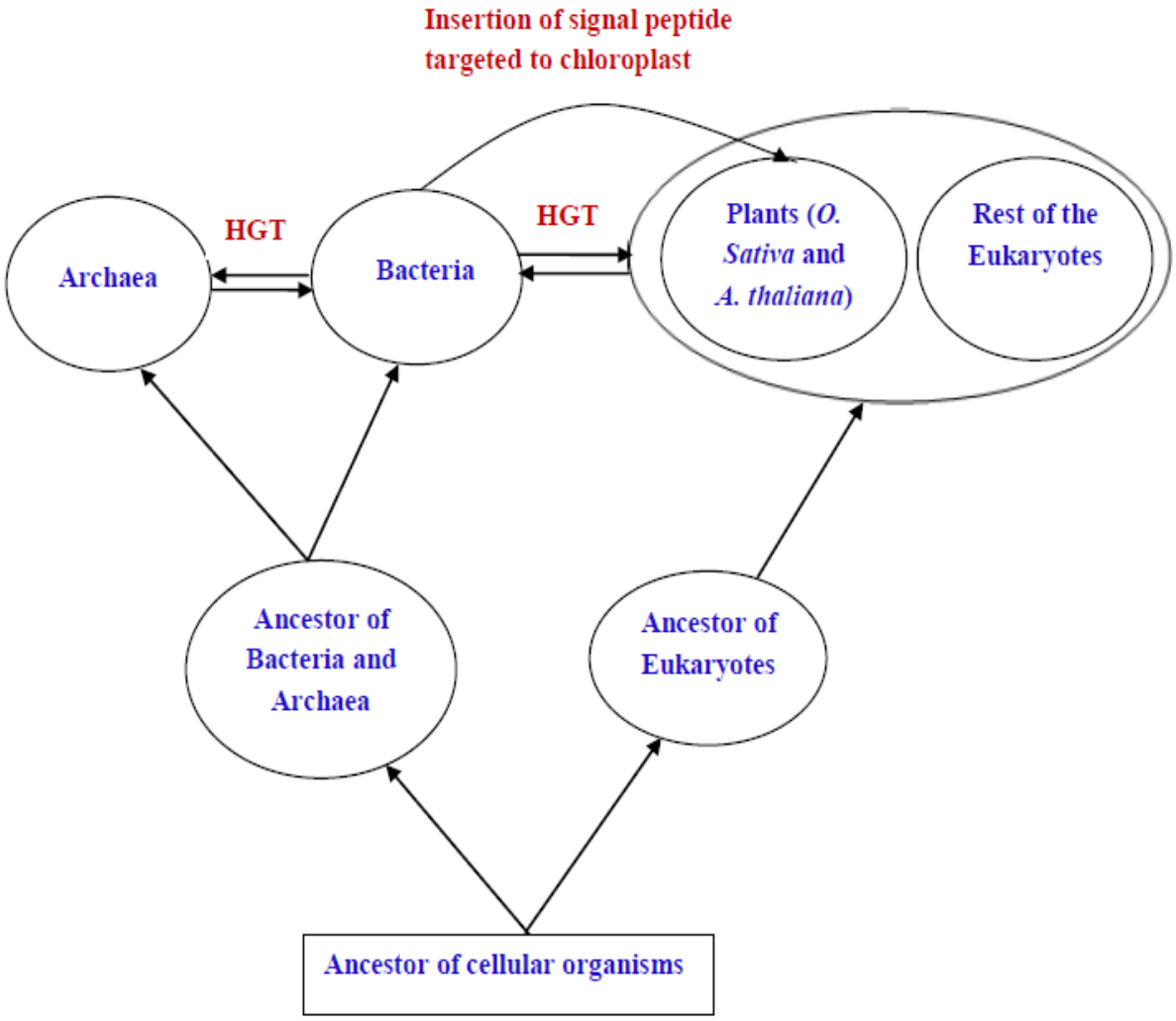


Fig. 3.13 A Model for the evolutionary history of the *endonuclease III* Gene family is schematically shown. The eukaryotic endonuclease III genes were likely originated from bacterial homologs through HGT. HGT events were also observed between bacterial and archaeal species. Eukaryotic endonuclease III in plants were found to have insertion at N terminal which are targeted to chloroplast.

importance in endonuclease III protein family is the endosymbiotic transfer of endonuclease III genes from bacteria (Cyanobacteria) to plants. It is a well-known fact that plant chloroplasts are derived from an ancestral endosymbiosis related to Cyanobacteria. *Synechococcus* clade

(Cyanobacteria) is placed near the eukaryotic clade along with *Oryza sativa* and *Arabidopsis thaliana* suggesting a possible endosymbiotic transfer events from bacteria (Cyanobacteria) to plants (Kanchan et al. 2015). Moreover, insertion of signal peptide at N terminal region of endonuclease III protein seems to have taken place during the evolution of plant homologs.

3.5 Conclusions

This study provides an overall picture of the evolutionary history of endonuclease III gene family that plays a crucial role in base excision repair of DNA. Based on conservation of amino acid positions, two consensus sequences have been identified for helix-hairpin motif and FES domain. Sequence analysis has also identified that quite a few residues within helix 2 and helix 8 are crucial for the structure and function. *Endonuclease III* gene based phylogenetic tree reveals few HGT events. Endosymbiotic transfer of *endonuclease III* gene events are also predicted from this evolutionary model. This evolutionary analysis may be exploited to understand the functions of uncharacterized genes in species.

Chapter IV

Evolution of endonuclease IV protein family

4.1 Introduction

Endonuclease IV enzymes are class II AP endonucleases which act on abasic site and break the phosphodiester bond at the 5' side. In *E.coli,* these enzymes accounts for 5-10% AP activity of cell (Mol et al. 2000) while, exonuclease III, another class II AP endonucleases accounts for rest of the total cellular AP endonuclease activity. Both enzymes are expected to work as mutual back-ups as gene mutation studies have shown that entire AP activity was performed by one group of enzyme in absence of the other group (Cunningham et al. 1986). Endonuclease IV enzyme also removes 3' DNA-blocking groups such as 3' phosphates, 3' phosphoglycolates and 3' α, β-unsaturated aldehydes that arise from either oxidative base damage or by the combined activity of glycosylase/lyase enzymes (Ramotar et al. 1991; Demple & Harrison 1994). However, it is also shown that endonuclease IV is the only known enzyme that is active against damaged nucleotides with α-deoxyadenosine base (Ide et al. 1994) and is able to recognize a set of oxidized pyrimidines. It is demonstrated that expression of endonuclease IV could be enhanced upto 20 fold in presence of superoxide-generating agents, such as paraquat (Chan & Weiss 1987). Endonuclease IV protein is encoded by the *nfo* gene of bacteria (Richardson & Kornberg 1961), *APN-1* gene of fungi (Ramoter et al. 1998) and *CeAPN1* gene of nematode (Masson et al. 1996).

Endonuclease IV is a Zn^{2+} dependent, single domain αβ and TIM barrel fold containing protein which directly participates in phosphodiester bond cleavage (Banner et al. 1975; Hosfield et al. 1999). Endonuclease IV active site contains a trinuclear Zn center, which is ligated by conserved amino-acid side chains. Importantly, the most conserved sequence region at 171–189 amino acid positions in endonuclease IV provides hydrogen bonding and packing interactions among structural elements (Hosfield et al. 1999). Endonuclease IV recognizes AP sites by flipping both

the abasic site and its partner nucleotide out of duplex DNA and bending the DNA at 90 degree at the flipped-out nucleotides (Hosfield et al. 1999). Such binding of endonuclease IV enzyme to dsDNA is mediated via five DNA recognition loops (R loops). Among these five R loops, three conserved regions at 7-12, 147-152 and 220-239 amino acid positions are called phosphate binding loop motif while two conserved regions at 33-45 and 69-78 amino acid positions are called Minor groove DNA binding motif (Hosfield et al. 1999). The Endonuclease IV AP site pocket is formed by His-7, Phe-32, Tyr-72, Trp-268, and the Zn observed in the Endo IV: DNA complex, together with ligand Glu-261 (Hosfield et al. 1999).

This study aims to provide insight into the unified evolution process of the endonuclease IV gene/protein family which leads to loss of this gene/protein in higher organisms. In this study, we have focused on insertion/deletion of domains/motifs, evolutionary conservation of important amino acid residues and gene duplication/loss during the course of evolution. Furthermore phylogenetic analysis based on gene sequences and protein sequences of endonuclease IV examines the evolution of endonuclease IV genes/proteins homologs in all five kingdoms of life which is then compared with 16S/18S rRNA sequences based species evolution. An evolutionary model has been proposed based on these studies.

4.2 Material and methods

4.2.1 Retrieval and selection of endonuclease IV protein homologs

E.coli endonuclease IV protein sequence (NCBI Acc. No. YP_541433.1) was used as a query to retrieve homologs. BLASTp (Altschul et al. 1990) two rounds of PSI-BLAST with default parameter was used for retrieving homologous sequences, After removing redundant hits, global

pairwise sequence alignment was performed with remaining hits from BLAST and PSI-BLAST search. Protein sequences with less than 15% identity against corresponding *E.coli* endonuclease IV protein were removed. Finally, hits with at least one common domain of endonuclease IV protein family were selected. A set of 402 homologs of the endonuclease IV family proteins was considered for evolutionary study. Of the 402 endonuclease IV homologs, 317 homologs belonged to bacteria, 48 to archaea, 26 to fungi, 8 to protista and 3 to animalia. Due to the large number of bacterial homologs, sequences from bacterial species were further divided into various divisions (Garrity & Holt 2001). Phylogenetic trees were generated for endonuclease IV homologs of nine bacterial divisions (Bacteridetes, Chlorobi, Chloroflexi, Cyanobacteria, Actinobacteria, Firmicutes, Thermotogae, Verrucomicrobia and Proteobacteria. All phylogenetic trees were inspected visually and representative homolog from different clades of each tree was chosen (Fig. 4.1 A-J). A large number of homologs were within proteobacteria division and clarity it is not shown for clarity).

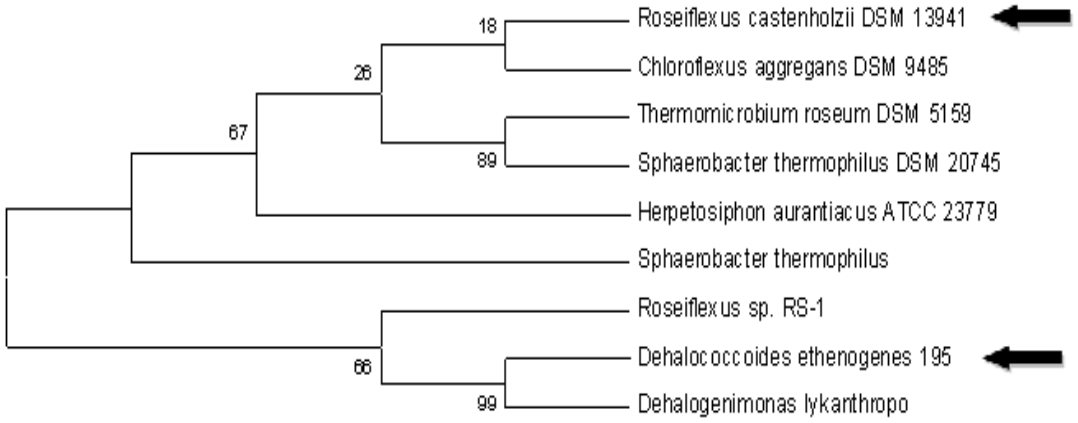


Fig 4.1(A) NJ tree of 9 Chloroflexi species. Two species from two major clades were selected. Selected species are shown by arrow.

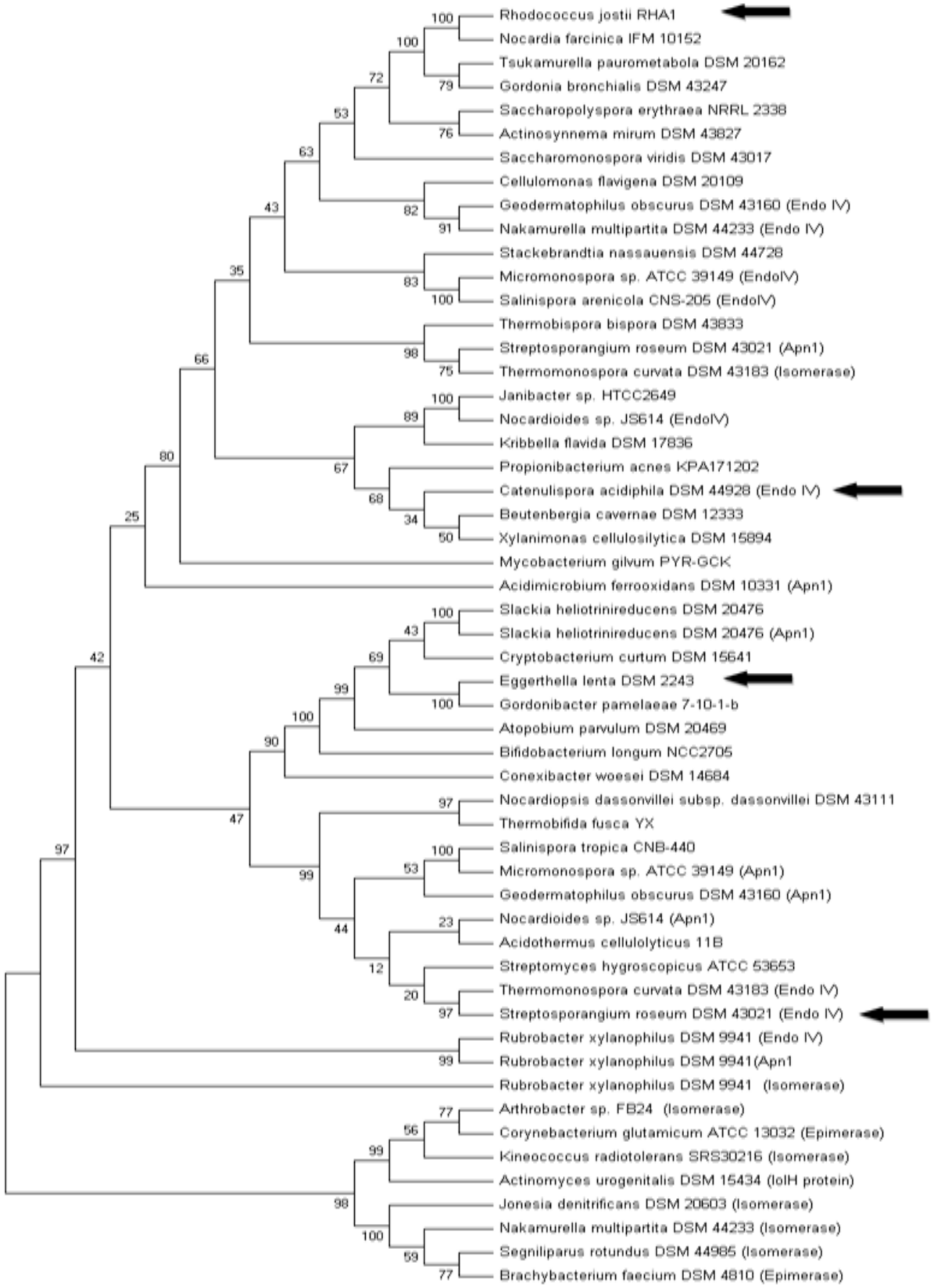


Fig. 4.1(B) NJ tree of 54 Actinobacteria species. Selected species are indicated by arrow.

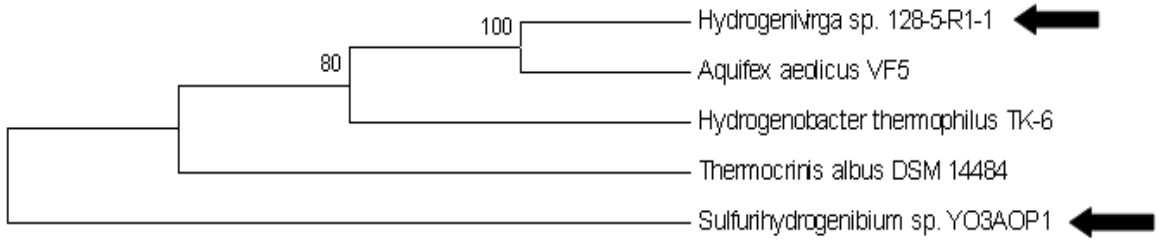


Fig. 4.1(C) NJ tree of 5 Aquificae species. Two species (shown by arrow) are chosen from the tree.

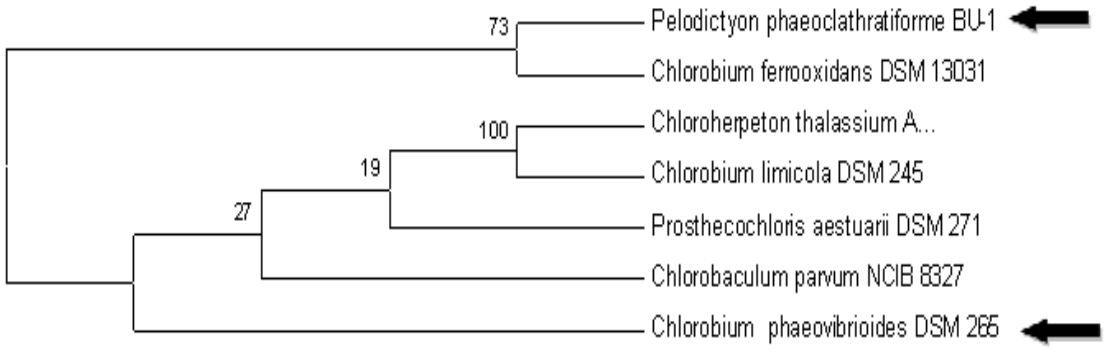


Fig. 4.1(D) NJ tree of 7 Chlorobi species which forms two clades. Two Chlorobi species are selected (shown by arrow).

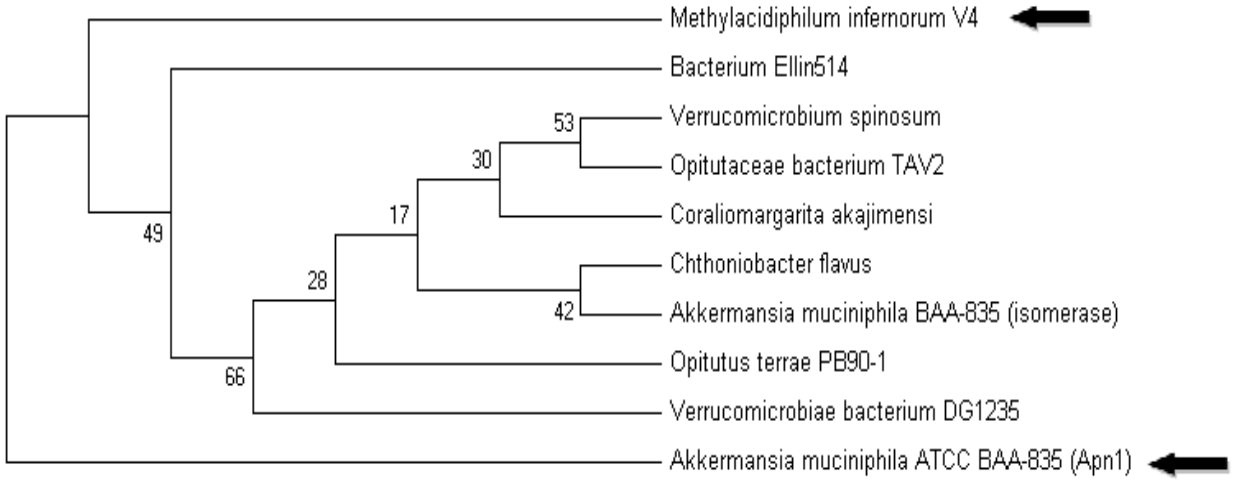


Fig 4.1(E) NJ tree of 10 Verrucomicrobium species. Two selected species from are shown by arrow.

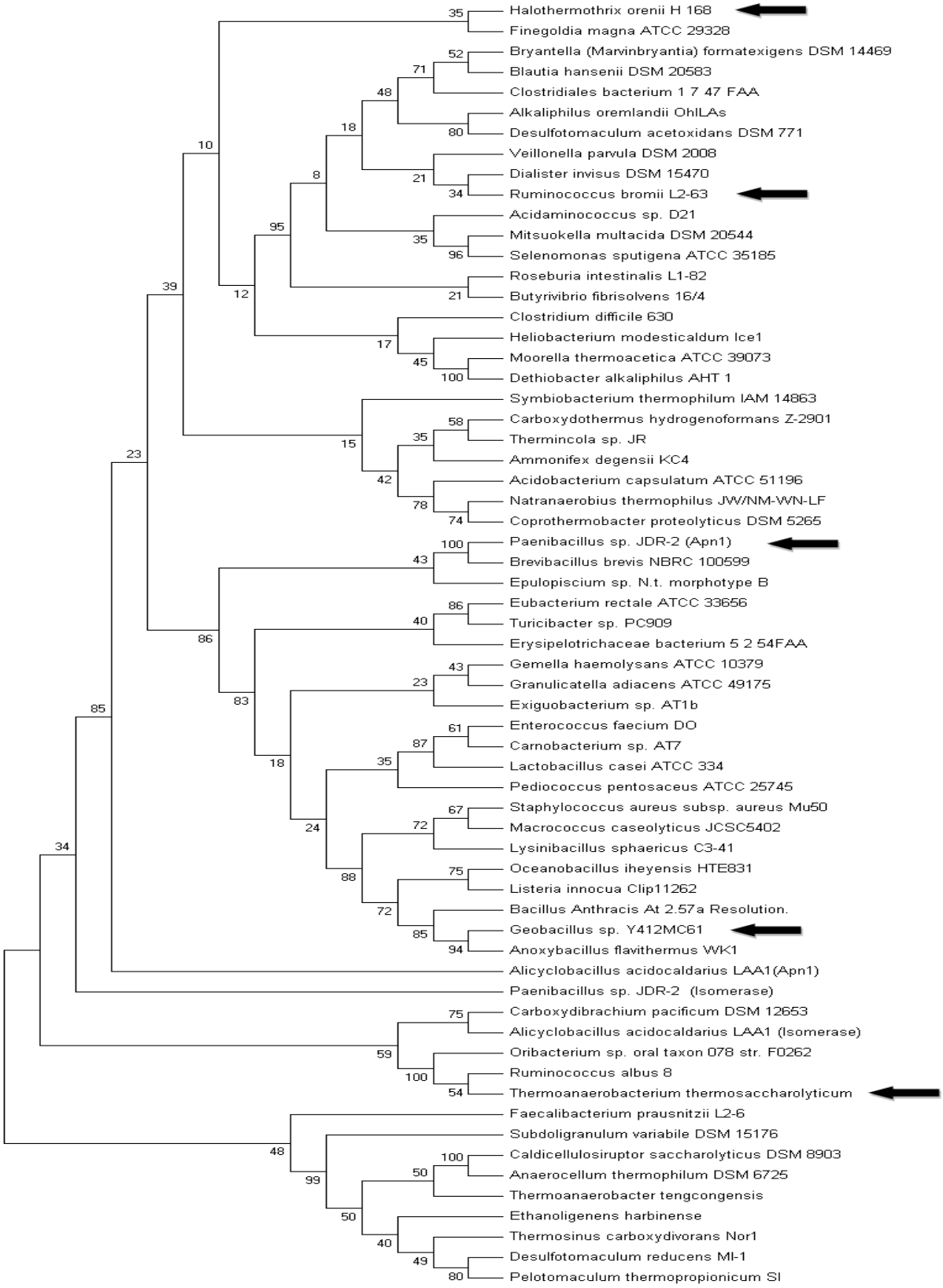


Fig. 4.1(F) NJ tree of 63 Firmicutes species. . Selected species are shown by arrow.

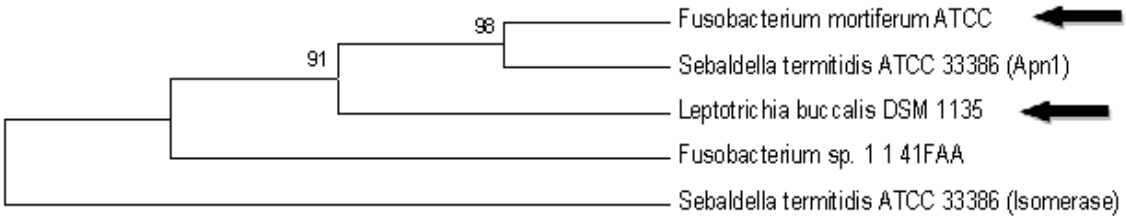


Fig. 4.1(G) NJ tree of 5 Fusobacteria species. One clade was observed in tree. Selected species are shown by arrow.

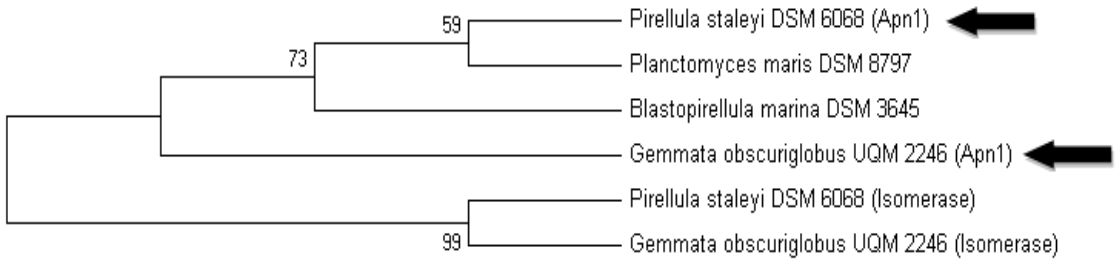


Fig. 4.1(H) NJ tree of 6 Plancetomyces species. Selected species from Plancetomyces are shown by arrow.

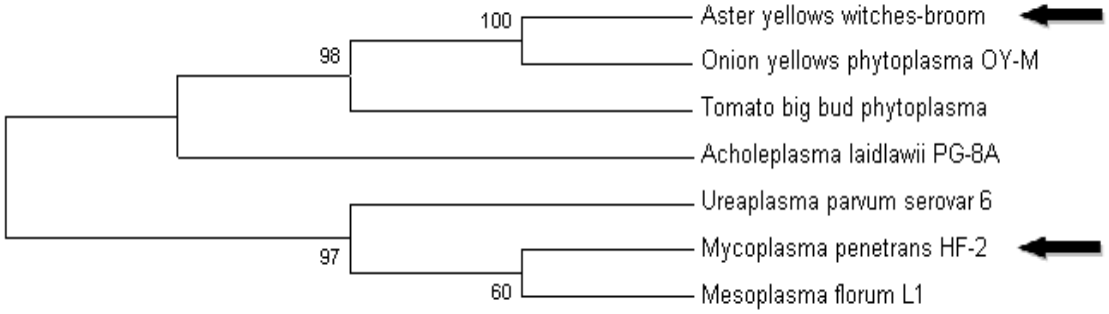


Fig. 4.1(I) NJ tree of 7 Tenericutes species. Two major clades were observed in tree. Selected species of endonuclease IV homologs are shown by arrow.

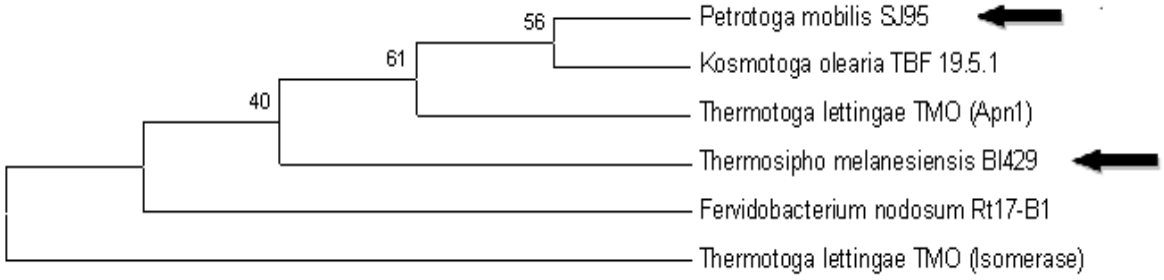


Fig. 4.1(J) NJ tree of 6 Thermotogae species. Two selected species from Thermotogae are shown by arrow.

Two separate NJ trees were constructed for 26 fungal and 46 archaeal species (Fig. 4.2 and 4.3) from which three fungal and four archaeal species were selected.

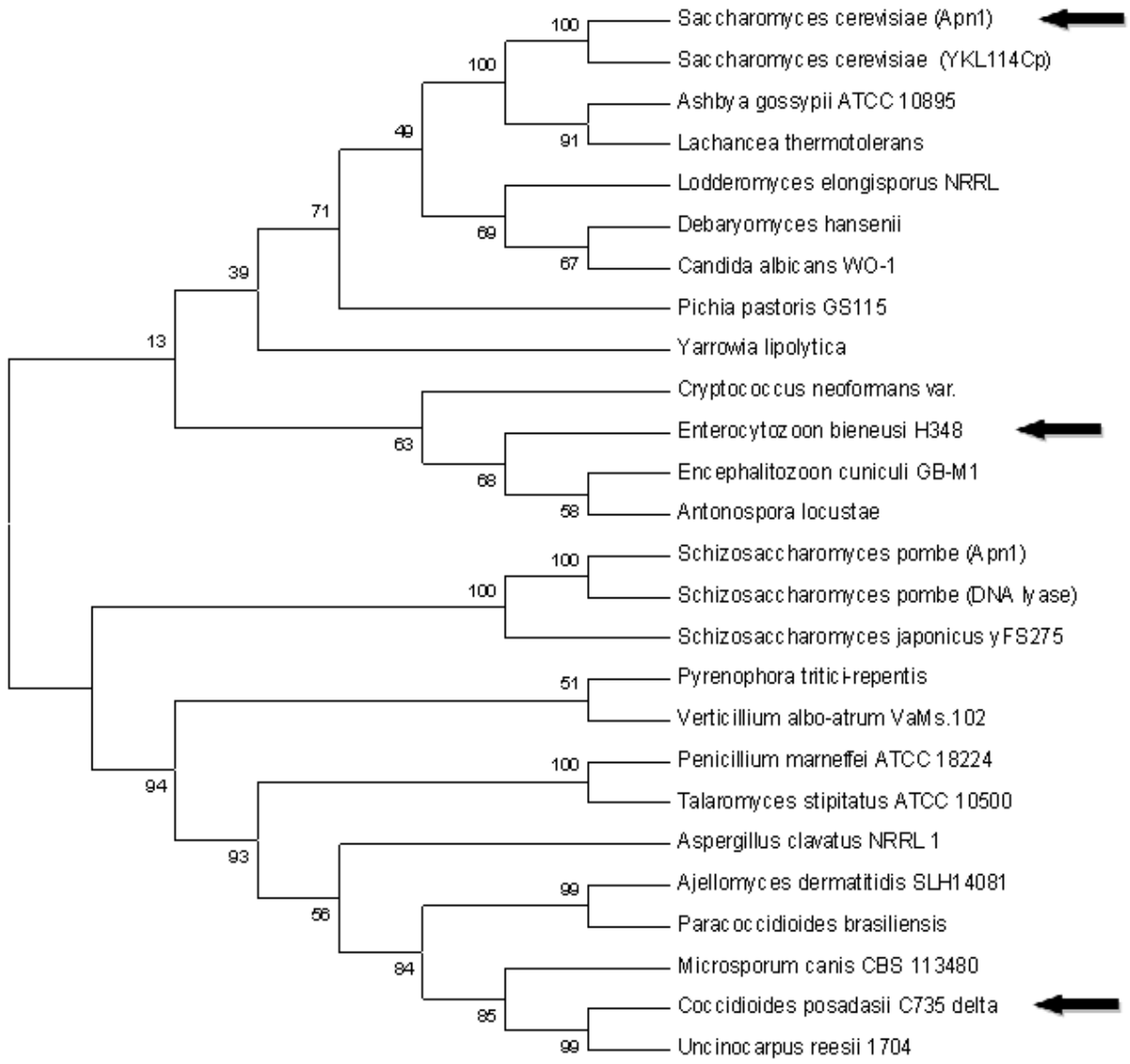


Fig. 4.2 NJ tree of 26 Fungal species produces two major clades. Representative species from each clade is selected for further study (indicated by arrow).

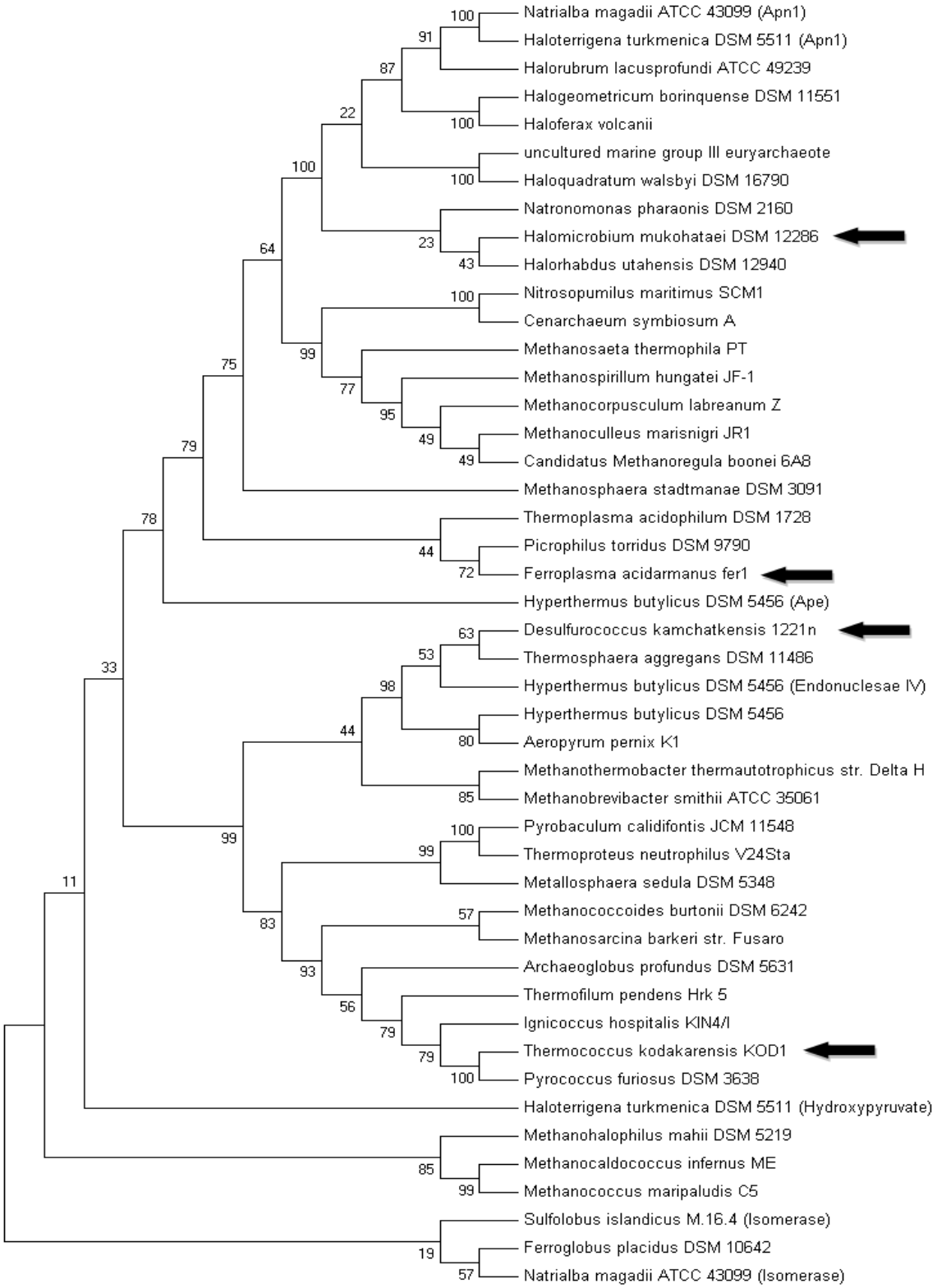


Fig. 4.3 NJ tree of 46 Archaeal species. Four major clades were observed in tree. Selected species from Archaea are shown by arrow.

Finally a set of 52 homologs of endonuclease IV are selected to study evolution of endonuclease IV protein family. Out of total 52 non-redundant endonuclease IV homologs from 52 different species, 25 homologs are endonuclease IV proteins, 26 homologs are AP endonuclease/ AP endonuclease 1 (APN1) proteins, and 1 homolog is deoxyribonuclease IV protein. This set of protein homologs are from 42 bacterial, 4 archaeal and 6 eukaryotic species. The detail of data mining is listed in Table 4.1. (A) & (B). The detailed breakup of species from 42 bacterial, 6 eukaryotic divisions and 4 archaeal are listed in Table 4.2. NCBI accession number of all 52 homologs of endonuclease IV is listed in Table 4.3.

Table 4.1(A) List of different proteins (as it was annotated in the database) retrieved as endonuclease IV homologs

Serial No.	**Protein Names**	**No. of sequences selected**
1.	Endonuclease IV	25
2.	AP endonuclease/ AP endonuclease 1 (APN1)	26
3.	Deoxyribonuclease IV	1
Total		**52**

. **Table 4.1(B)** Division and kingdom wise breakup of endonuclease IV homologs. All bacterial species is further divided into 22 different categories (Garrity & Holt 2001).

Serial. No.	**Division**	**Kingdom**	**Total number of non-redundant homologs**	**Homologs considered for evolution analysis**

1.	Archaea	Monera	48	4
2.	Bacteria		317	42
3.	Eukaryotes	Fungi	26	3
		Protista	8	1
		Animalia	3	2
Total			**402**	**52**

Table 4.2 Total number of endonuclease IV homologs retrieved from different bacterial division eukaryotes and archaea is listed along with final set of homologs selected for the present study.

Serial No.	**Phylum**	**Division**	**Total no. of Homologs**	**No. of homologs selected**
1.	Archaea		46	3
2. 1	Bacteria	Bacteroidetes	12	2
2.2		Chlorobi	7	3
2.3		Chlamydiae	4	1
2.4		Chloroflexi	9	3
2.5		Deinococcus	4	1
2.6		Dictyoglomi	1	1
2.7		Actinobacteria	54	4
2.8		Aquificae	5	2
2.9		Nitrospirae	2	1
2.10		Planctomycetes	6	2
2.11		Firmicutes	63	3
2.12		Thermotogae	6	2
2.13		Verrucomicrobia	10	2
2.14		Proteobacteria	110	4
2.15		Gemmatimonadetes	1	1
2.16		Tenericutes	7	2
2.17		Fusobacteria	5	2
2.18		Acidobacteria	2	1
2.19		Fibrobacteres	1	1
2.20		Synergistetes	6	1
2.21		Deferribacteres	2	2
2.22		Thermobaculum	1	1
3.	Viruses		2	0
4.1	Eukaryotes	Fungi	26	3
4.2		Protista	7	2
4.3		Animalia	3	2

		Total	**402**	**52**

Table 4.3 The NCBI accession numbers of the 52 endonuclease IV protein homologs and the length of the protein sequences.

Serial No.	**Name of the Organism**	**NCBI accession number**	**Protein name**	**Sequence length**
1.	*Halothermothrix orenii H 168*	YP_002508543.1	Apurinic endonuclease Apn1	278
2.	*Thermosipho melanesiensis BI429*	YP_001306634.1	Apurinic endonuclease Apn1	283
3.	*Desulfurococcus kamchatkensis*	YP_002428679.1	Endonuclease IV	288
4.	*Dehalococcoides ethenogenes 195*	YP_181137.1	Apurinic endonuclease	276
5.	*Citreicella sp. SE45*	ZP_05783870.1	Apurinic endonuclease family 2	262
6.	*Pelodictyon_phaeoclathratiforme BU-1*	ZP_00589943.1	Apurinic endonuclease, family 2	280
7.	*Ixodes scapularis*	XP_002401866.1	Apurinic endonuclease,	314
8.	*Saccharomyces cerevisiae*	AAA34429.1	Apurinic endonuclease (APN1)	367
9.	*Halomicrobium mukohataei*	YP_003176029.1	Apurinic endonuclease Apn1	276
10.	*Akkermansia muciniphila ATCC BAA-835,*	YP_001877459.1	Apurinic endonuclease Apn1	277
11.	*Thermanaerovibrio acidaminovorans*	ZP_04468726.1	Apurinic endonuclease APN1	277
12.	*Denitrovibrio acetiphilus DSM 12809*	ZP_03907386.1	Apurinic endonuclease APN1	273
13.	*Sulfurihydrogenibium sp.*	YP_001931889.1	Apurinic endonuclease Apn1	279
14.	*Petrotoga mobilis SJ95,*	YP_001568826.1	Apurinic endonuclease Apn1	287
15.	*Gemmata obscuriglobus UQM 2246*	ZP_02736399.1	Apurinic endonuclease Apn1	291
16.	*Leptotrichia buccalis*	YP_003164169.1	Apurinic endonuclease Apn1	307
17.	*Geobacillus sp. Y412MC61 ctg31*	ZP_03558891.1	Apurinic endonuclease Apn1	299

18.	*Fibrobacter succinogenes subsp. succinogenes S85*	ZP_04786621.1	Apurinic endonuclease Apn1	275
19.	*Thermobaculum terrenum ATCC BAA-798*	ZP_03857848.1	Apurinic endonuclease APN1	296
20.	*Roseiflexus castenholzii DSM 13941*	YP_001430697.1	Apurinic endonuclease Apn1	289
21.	*Coccidioides posadasii*	EER23862.1	Apurinic endonuclease family protein	556
22.	*Theileria parva*	XP_764900.1	Apurinic endonuclease	422
23.	*Caenorhabditis elegans*	NP_495687.1	Apurinic endonuclease 1	396
24.	*Hydrogenivirga sp*	ZP_02176824.1	Deoxyribonuclease IV	279
25.	*Fusobacterium sp. 1_1_41FAA*	ZP_04567756.1	Endonuclease IV	289
26.	*Eggerthella lenta*	ZP_03894977.1	Endonuclease IV	277
27.	*Methylacidiphilum infernorum V4*	YP_001939894.1	Endonuclease IV	321
28.	*Chlamydia trachomatis*	NP_220142.1	Endonuclease IV	288
29.	*Candidatus Solibacter usitatus Ellin6076*	YP_827210.1	Endonuclease IV	290
30.	*Meiothermus silvanus*	ZP_04034706.1	Endonuclease IV	282
31.	*Deferribacter desulfuricans*	BAI80868.1	Endonuclease IV	281
32.	*Rhodococcus sp.*	YP_701860.1	Endonuclease IV	256
33.	*Aster yellows witches'-broom*	YP_456513.1	Endonuclease IV	292
34.	*Parabacteroides distason*	YP_001302585.1	Endonuclease IV	281
35.	*Prosthecochloris vibrioformis (Chlorobium phaeovibrioides) DSM 265)*	YP_001130003.1	Endonuclease IV	279
36.	*Enterocytozoon bieneusi*	XP_002649640.1	Endonuclease IV	273
37.	*Leptospirillum sp.*	EAY57355.1	Endonuclease IV	317
38.	*Dictyoglomus thermophilus*	YP_002251795.1	Endonuclease IV	271
39.	*Gemmatimonas aurantiaca*	YP_002761675.1	Endonuclease IV	293
40.	*Mycoplasma penetrans*	NP_757507.1	Endonuclease IV	311
41.	*Escherichia coli*	YP_541433.1	Endonuclease IV	285
42.	*Pirellula staleyi DSM*	YP_003370769.1	Apurinic endonuclease Apn1	321
43.	*Paenibacillus sp JDR-2*	ZP_02850179.1	Apurinic endonuclease Apn1	376
44.	*Ruminococcus bromii*	CBL15255.1	Endonuclease IV	277
45.	*Thermoanaerobacterium thermosaccharolyticum*	ZP_05336398.1	Apurinic endonuclease Apn1	277
46.	*Thermococcus kodakarensis*	YP_182583.1	Endonuclease IV	281
47.	*Ferroplasma acidarmanus*	ZP_05411925.1	Endonuclease IV	283
48.	*Streptosporangium roseum DSM 43021*	ZP_04477472.1	Endonuclease IV	283

49.	*Catenulispora acidiphila*	ZP_04368610.1	Endonuclease IV	265
50.	*Aliivibrio salmonicida*	YP_002263900.1	Endonuclease IV	282
51.	*Desulfurivibrio alkaliphilus*	ZP_05709857.1	Apurinic endonuclease Apn1	289
52.	*Aeromonas salmonicida*	YP_001142010.1	Endonuclease IV	281

4.3 Results and Discussion

4.3.1 Domains and motifs of endonuclease IV family

AP2Ec (cd00019) is main domain in endonuclease IV protein homologs which spans almost entire protein in most of the homologs. AP2Ec domain contains functional motifs like phosphate binding, minor groove binding loops and metal binding active site residues (Hosfield et al. 1999). These active site residues are responsible for positioning tri-nuclear metal center which is directly involved in phosphodiester bond cleavage. Crystal structure of endonuclease IV *E.coli* homolog (PDB ID: 1QUM) has demonstrated that three different loops (within residues 7-12, 147-152 and 220-239) are interacted with phosphate groups of DNA substrate. Multiple sequence alignment (MSA) of 52 homologs shows that H_7 of first loop; A_{148}, G_{149}, G_{151} of second loop and S_{226}, D_{229}, R_{230}, H_{231}, G_{235}, G_{237} of third loop are highly conserved (more than 80% occasion). Overall, residues of third phosphate binding loops are relatively more conserved than those of first and second phosphate binding loops. Glycines of phosphate binding motifs are crucial for loop conformation which is responsible for non-specific interaction with phosphate backbone of DNA substrate. In contrary to phosphate binding motifs, residues of minor groove binding motifs (residues between 33-45 and 69-78) penetrate deep into minor groove and interact with DNA bases. MSA of 52 homologs suggests the presence of a consensus sequence H_{69}-X_2-

Y_{72}-X_2-N_{75} within minor groove binding motif at 69-78 amino acid positions where each amino acid position is conserved in more than 80% homologs.

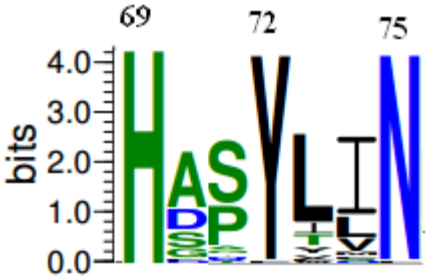


Fig. 4.4 Sequence conservation within minor groove binding motif between residues 69 to 75 is shown using sequence logo. A bits score of 3.2 and above corresponds to more than 80% sequence conservation.

In this regard N, Q, R and Y residues of minor groove binding motif interact specifically with DNA bases. MSA analysis reveals that residues (specifically between 35-39 and 72-76) of minor groove binding motif contain a substantial number of N, Q, R and Y residues. Among these residues, Y_{72} and N_{75} are conserved for more that 80% species. Tri-nuclear Zn center near the active site of the enzyme plays an important role in phophodiester cleavage by holding catalytically active amino acid residues close together (Hosfield et al. 1999). Crystal structure also reveals that histidine residues and negatively charged aspartates and glutamates are coordinated with three Zn ions to shape up active site. The tri-nulear Zn center coordinated with histidine 69, 109, 182, 216 and 231, glutamic acid 145 and 261 and aspartic acid 179 and 229. MSA shows that more than 90% of 52 species have conserved Zinc binding residues. MSA of 52 homologs also identifies three consensus sequence segments between 179^{th} and 187^{th} amino acid, 216^{th} and 219^{th} amino acid and between 229^{th} and 237^{th} amino acid which are D_{179}-X_2-H_{182}-X_4-

G_{187}, H_{216}-X-N_{218}-D_{219} and D_{229}-X-H_{231}-X_3-G_{235}-X-G_{237} where each amino acid position is conserved in more than 80% homologs. (Fig. 4.5(A-C)).

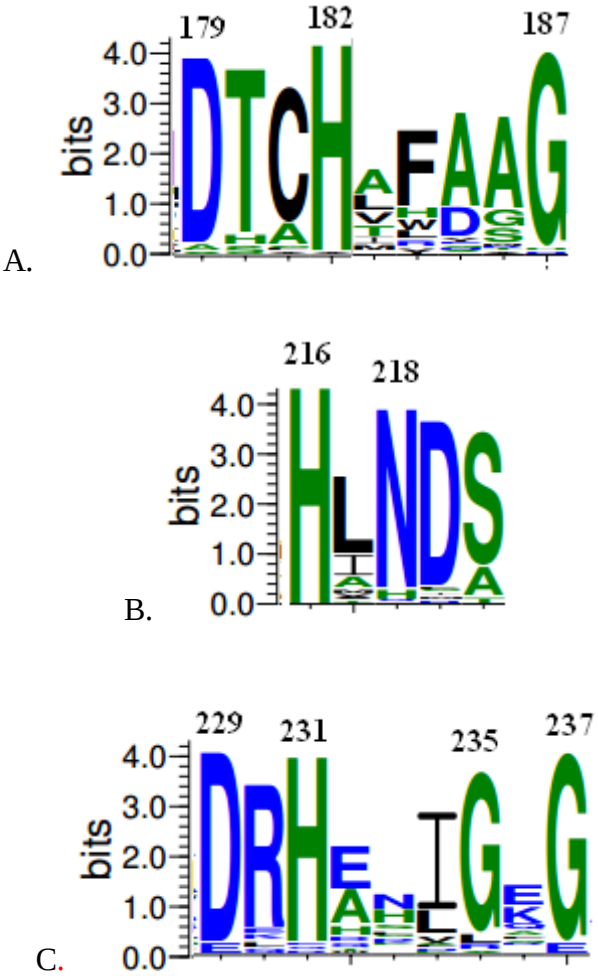


Fig. 4.5(A-C) Three conserved sequence segments between residues (A) 179-187, (B) 216-220 (C) 229-237 are shown by sequence logo.

4.3.1 16S/18S rRNA gene based species tree

Complete 16S/18S r-RNA gene sequences of 48 species are retrieved while partial 16S/18S r-RNA gene sequence are available for rest of the 4 species which are not considered to generate the species tree. A total of 15 distinct clades are formed in species tree (Fig. 4.6) with eukaryotes

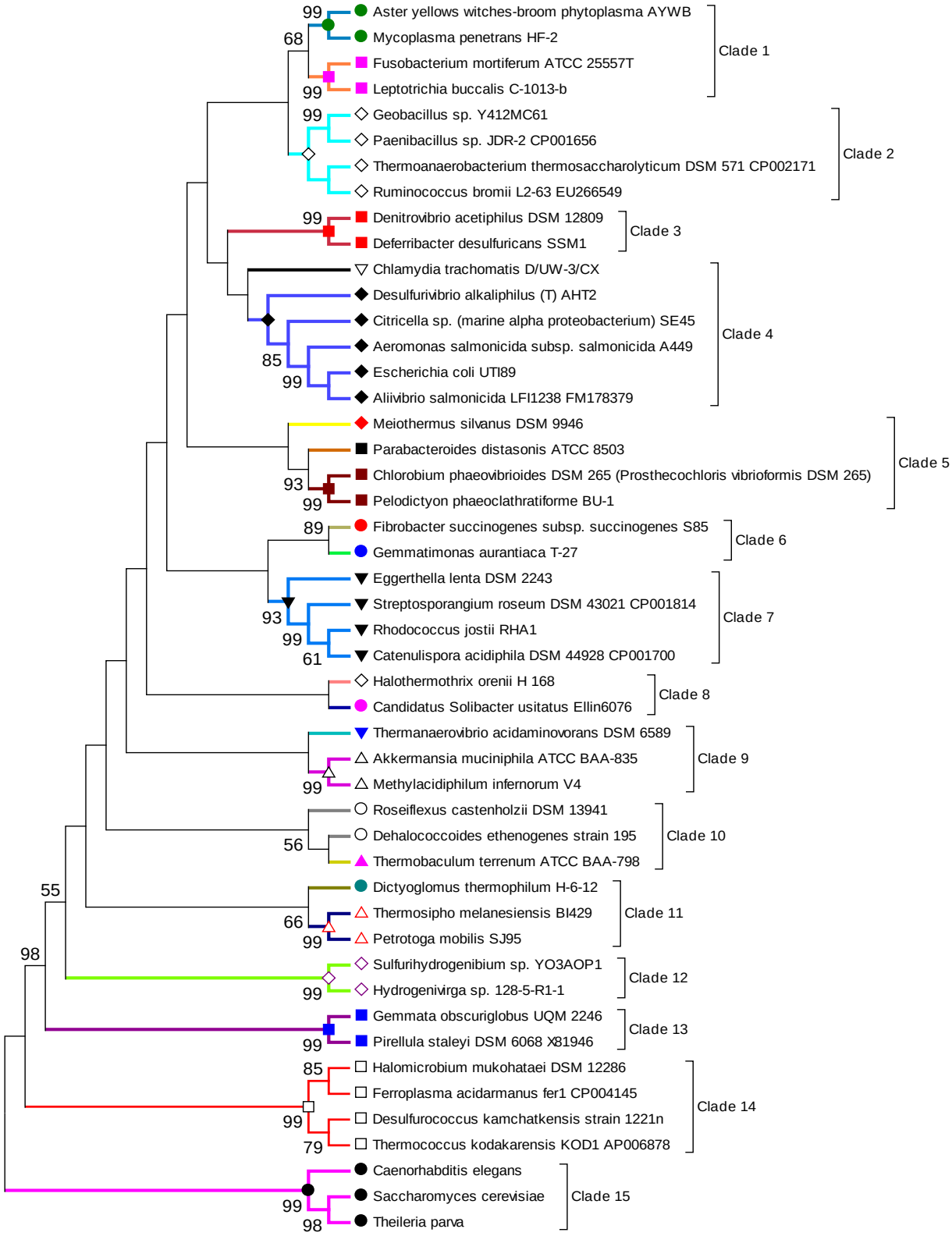


Fig. 4.6 16S/18S rRNA sequences of species based maximum likelihood tree**.** The evolutionary distances were computed using the Tamura Nei substitution model (Tamura & Nei 1993)**.** Bootstrap support values are presented next to the tree branches for each clade with ≥50.

were the farthest clades within the tree. As it is expected, species from same bacterial division as well as species from same kingdom stay close to each other in a clade. Single species from closely related bacterial divisions are clustered together. For example, clade 5 contains species from Deinococcus, Bacteroidetes and Chlorobi bacterial divisions. Similarly, Tenericutes, Chloroflexi, Synergistetes, Thermotogae and Proteobacteria share clusters with species from Fusobacteria, Thermobaculum, Verrucomicrobia, Dictyoglomi and Chlamydiae respectively.

4.3.3 Endonuclease IV gene based phylogenetic tree

Both the ML and NJ trees of gene sequences of endonuclease IV family generates 10 distinct clads with almost similar topology (Fig. 4.7). Thus, only ML tree is discussed for clarity. Interestingly, none of the clusters are pure archaeal, bacterial (which belong to the same division) or eukaryotic cluster. All 10 clusters have a mixed population. However, 16S/18S rRNA sequence based species tree shows that most of the species in different bacterial subdivisions as well as species from archaea and eukaryote forms distinct clad. It is interesting to note that endonuclease IV genes of three archaeal species were positioned within three different clades, although all archaeal species are within the same clad in species tree. This observation indicates that endonuclease IV gene of these archaeal species are evolved differently due to the different environmental pressure. Endonuclease IV gene based

show that in most of the cases, large number of species share a clad with species from different division. The content of endonuclease IV gene and GC content of 3rd codon position of the endonuclease gene are listed in Table 4.2. It has been observed that average GC content and GC content at the 3rd codon position of endonuclease IV gene in archaeal species

Ferroplasma as well bacterial species *Halothermothrix, Dictyoglomus, Thermobaculum,* and *Methylacidiphilum* (clade 2) were 39.3 % and 34% (with standard deviation of 4.46 and 7.5) respectively, which explains possible reason for these species to share same clad in gene based phylogeny tree. Similarly, the average GC and GC content at the third codon position of *Halomicrobium* (archeal species)of gene (are 65.7% and 70.0% respectively which differs significantly from other three archaeal species *Desulfurococcus* (44.3% and 51.9%) , *Thermococcus* (52.8% and 54.2) and *Ferroplasma* (38.4% and 68.6%). Hence, it is placed with bacterial species (clade 7) having similar GC content. From these observations, it is clear that these species reside within the clad that contain species with similar GC content. Being a member of bacterial division, *Chlamydia, Mycoplasma and Escherichia* shares clad 1 with most of the eukaryotic homologs because of its GC content (39.74%) and GC content at the third codon position (35.21%) of endonuclease IV genes. Interestingly, one of the eukaryotic species *Ixodes scapularis* shares clade 4 with bacterial species *Rhodococcus Candidatus , Meiothermus* and *Gemmatimonas* due to the similar GC content at gene level and the GC content at the third codon position (63.13% and 77.49% respectively with standard deviation of 3.29 and 6.25). Endonuclease IV gene based phylogeny tree of 52 taxa suggests that GC content of gene contribute significantly towards the position of the taxa within the tree. The species and endonuclease IV gene evolution shape up differently in most of the cases. From this analysis it is clear that the GC content at third codon position of endonuclease IV gene as well as mean GC content of endonuclease IV gene plays major role in positioning different species in phylogenetic tree. It has been observed that average GC content and GC content at the third codon position of endonuclease IV genes in archaeal species *Ferroplasma* as well in different bacterial divisional

Table 4.4 The GC content of the endonuclease IV gene, the 3rd codon position of the endonuclease IV gene of the corresponding organism are listed along with the average value of each clade.

Clade No.	Name of the Organism	Division/Kingdom	Mean Gene GC content (%)	GC at 3rd position (%)	Average	
					Mean Gene	3rd position
1.	*Enterocytozoon bieneusi H348*	Fungi	27.49	16.79	39.74 (8.83)	35.21 (13.68)
	Mycoplasma penetrans HF-2	Tenericutes	28.63	17.63		
	Chlamydia trachomatis D/UW-3/CX	Chlamydiae	43.83	32.18		
	Caenorhabditis elegans	Animalia	40.05	36.78		
	Escherichia coli UTI89	Proteobacteria	52.45	55.59		
	Coccidioides posadasii C735 delta SOWgp	Fungi	49.13	50.09		
	Saccharomyces cerevisiae	Fungi	38.50	38.59		
	Theileria parva strain Muguga	Protista	37.83	34.04		
2.	*Halothermothrix orenii H 168*	Firmicutes	35.36	25.45	39.25 (4.46)	33.95 (7.5)
	Dictyoglomus thermophilum H-6-12	Dictyoglomi	34.80	26.47		
	Thermobaculum terrenum ATCC BAA-798	Thermobaculum	44.67	41.75		
	Ferroplasma acidarmanus fer1	Archaea	38.38	36.62		
	Methylacidiphilum infernorum V4,	Verrucomicrobia	43.06	39.44		
3.	*Fibrobacter succinogenes subsp. succinogenes S85*	Fibrobacteres	49.15	51.81	49.01 (4.79)	54.01 (10.46)
	Geobacillus sp. Y412MC61 ctg31	Firmicutes	52.78	64.00		
	Paenibacillus sp. JDR-2	Firmicutes	51.90	59.95		
	Ruminococcus bromii L2-63	Firmicutes	42.21	40.29		
4.	*Thermanaerovibrio acidaminovorans DSM 6589*	Synergistetes	61.39	76.62	63.13 (3.29)	77.49 (6.25)
	Ixodes scapularis	Animalia	58.41	73.65		
	Rhodococcus jostii RHA1	Actinobacteria	67.57	88.72		
	Candidatus Solibacter	Acidobacteria	65.64	80.41		

	usitatus Ellin6076					
	Meiothermus silvanus DSM 9946	Deinococcus	61.72	72.08		
	Gemmatimonas aurantiaca T-27	Gemmatimonadetes	64.06	73.47		
5.	*Gemmata obscuriglobus UQM 2246*	Planctomycetes	70.82	75.85	70.52 (0.42)	74.42 (2.01)
	Citreicella sp. SE45	Proteobacteria	70.22	73.00		
6.	*Dehalococcoides ethenogenes 195,*	Chloroflexi	51.62	57.40	49.58 (4.62)	54.52 (2.76)
	Desulfurococcus kamchatkensis 1221n	Archaea	44.29	51.90		
	Thermococcus kodakarensis KOD1	Archaea	52.84	54.26		
7.	*Leptospirillum rubarum sp. Group IV 'UBA BS*	Nitrospirae	59.33	64.78	63.78 (5.6)	67.59 (3.99)
	Desulfurivibrio alkaliphilus AHT2	Proteobacteria	62.99	67.24		
	Halomicrobium mukohataei DSM 12286,	Archaea	65.70	70.04		
	Pirellula staleyi DSM 6068	Planctomycetes	55.80	60.25		
	Roseiflexus castenholzii DSM 13941	Chloroflexi	61.61	68.97		
	Streptosporangium roseum DSM 43021	Actinobacteria	69.01	69.72		
	Catenulispora acidiphila DSM 44928	Actinobacteria	72.06	72.18		
8.	*Aeromonas salmonicida subsp. salmonicida A449*	Proteobacteria	60.64	60.99	52.26 (7.47)	58.46 (4.29)
	Aliivibrio salmonicida LFI1238	Proteobacteria	40.28	52.65		
	Parabacteroides distasonis ATCC 8503	Bacteroidetes	51.89	55.32		
	Chlorobium phaeovibrioides (Prosthecochloris vibrioformis)DSM 265	Chlorobi	55.12	62.86		
	Pelodictyon_phaeoclathratiforme BU-1	Chlorobi	53.38	60.50		
9.	*Thermoanaerobacterium thermosaccharolyticum DSM 571*	Firmicutes	36.93	49.28	44.41 (14.97)	53.31 (10.51)
	Aster yellows witches phytoplasma AYWB	Tenericutes	29.81	39.93		
	Leptotrichia buccalis C-1013-b	Fusobacteria	35.28	49.68		

	Eggerthella lenta DSM 2243	Actinobacteria	65.35	66.19		
	Akkermansia muciniphila ATCC BAA-835,	Verrucomicrobia	54.68	61.51		
10.	*Sulfurihydrogenibium sp. YO3AOP1*	Aquificae	30.00	41.43	35.98 (6.90)	45.79 (4.71)
	Hydrogenivirga sp. 128-5-R1-1	Aquificae	48.93	53.21		
	Thermosipho melanesiensis BI429	Thermotogae	30.87	42.25		
	Petrotoga mobilis SJ95,	Thermotogae	37.50	44.79		
	Denitrovibrio acetiphilus DSM 12809	Deferribacteres	43.07	50.00		
	Deferribacter desulfuricans SSM1	Deferribacteres	32.39	45.04		

species *Halothermothrix, Dictyoglomus, Thermobaculum,* and *Methylacidiphilum* (clade 2) were 39.25% and 33.95% (standard deviation of 4.46 and 7.5) respectively, which explains possible reason for these species to share same clad in gene based phylogeny tree. Similarly, one of the archaeal species *Halomicrobium* where the average GC (65.70%) and GC content at the third codon position of gene (70.04%) differs significantly from rest three archaeal species *Desulfurococcus, Thermococcus* and *Ferroplasma* were placed with bacterial species from different divisions having similar GC content. From these observations, it is clear that these species reside within the clad that contain species with similar GC content. Being a member of bacterial division, *Chlamydia, Mycoplasma* and *Escherichia* shares a clad with eukaryotic homologs because of similar average GC content (39.74%) and GC content at the third codon position (35.21%) of endonuclease IV genes. Interestingly, one of the eukaryotic species *Ixodes scapularis* shares a clad with many bacterial species *Rhodococcus, Candidatus, Meiothermus* and *Gemmatimonas* share the same clade (Clade 4) in the tree due to similar GC content at gene level and GC content at the third codon position (63.13% and 77.49%

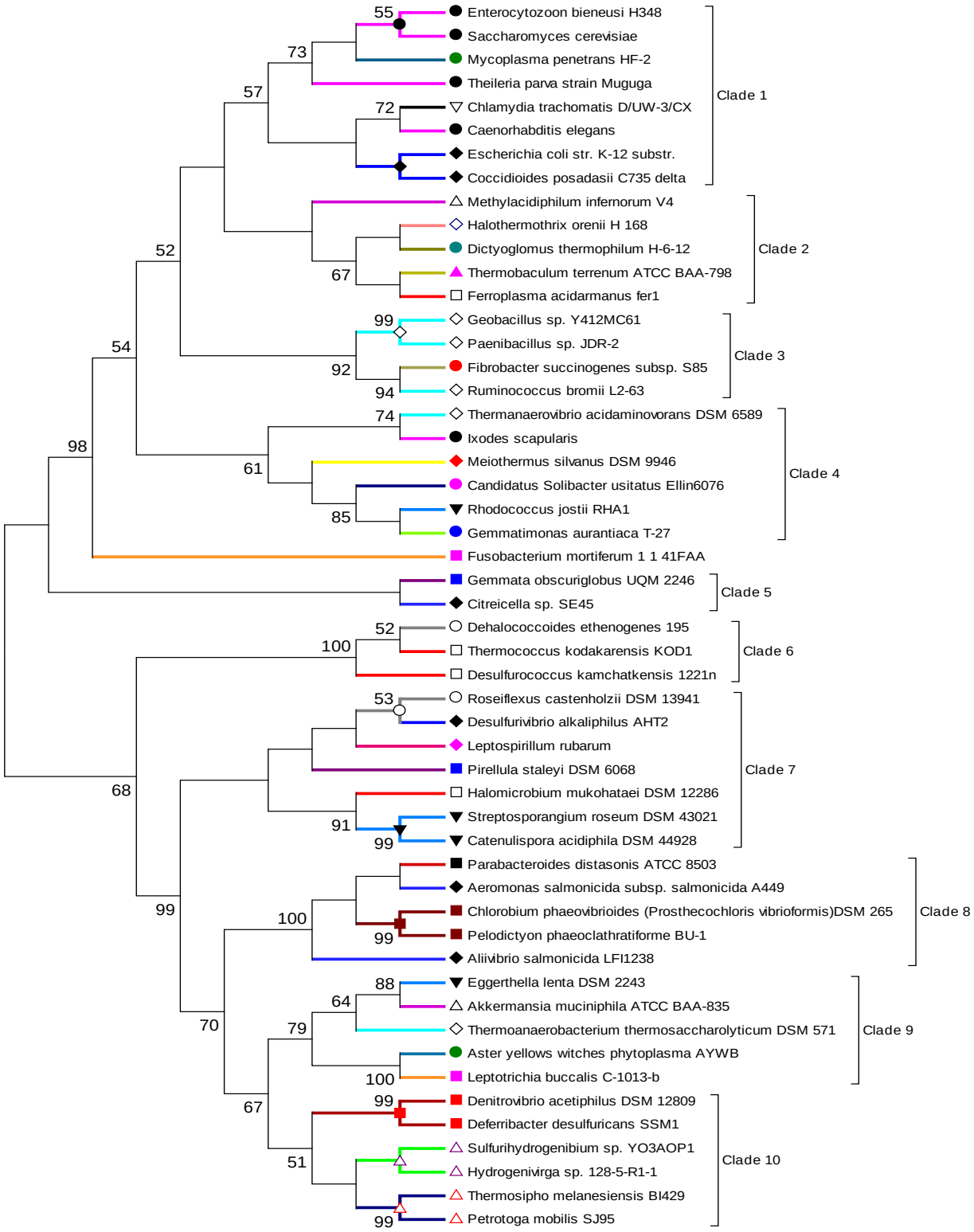


Fig. 4.7 Endonuclease IV gene sequences based maximum likelihood tree**.** The evolutionary distances were computed using the Tamura Nei as substitution model (Tamura & Nei 1993)**.** The bootstrap values are given as Fig. 4.6.

respectively with standard deviation of 3.29 and 6.25). *Endonuclease IV* gene based phylogeny tree of 52 taxa suggests that GC content of gene contribute significantly towards the position of taxa within the tree and species evolution, and *endonuclease IV* gene evolution shape up differently in most of the cases. Since AP endonuclease IV gene in all living organisms diverges over a long evolutionary period, synonymous nucleotide substitution GC content at third codon position as well as mean GC content of gene makes gene sequence based phylogenetic tree construction noisy.

4.3.4 Endonuclease IV protein based phylogenetic tree

Overall eleven distinct clades were identified from endonuclease IV protein based phylogenetic tree (Fig. 4.8). Interestingly, topology of gene based and protein based trees are quite different

indicating strong influence of synonymous nucleotide substitution in gene level phylogenetic tree. All six eukaryotic species (*Caenorhabditis elegans*, *Ixodes scapularis*, *Saccharomyces cerevisiae*, *Coccidioides possadasi*, *Enterocytozoon bieneusi* and *Theileria parva*) remain within a single clade (clade 2) indicating close sequence similarity of endonuclease IV protein homologs among eukaryotes. However, archaeal homologs of endonuclease IV protein occupy different clades in the protein based tree. Similarly, homologs of different bacterial division form heterogeneous clade though the protein homologs from Thermotogal, Aquificae, Chlorobial division remain close with each other within a clade (clade 5, clade 8 and clade 7 respectively).

A closer inspection of multiple sequence alignment of homologs of endonuclease IV protein showed a sizable insertion of ~137 amino acids in *Theileria parva* (Protista), ~116

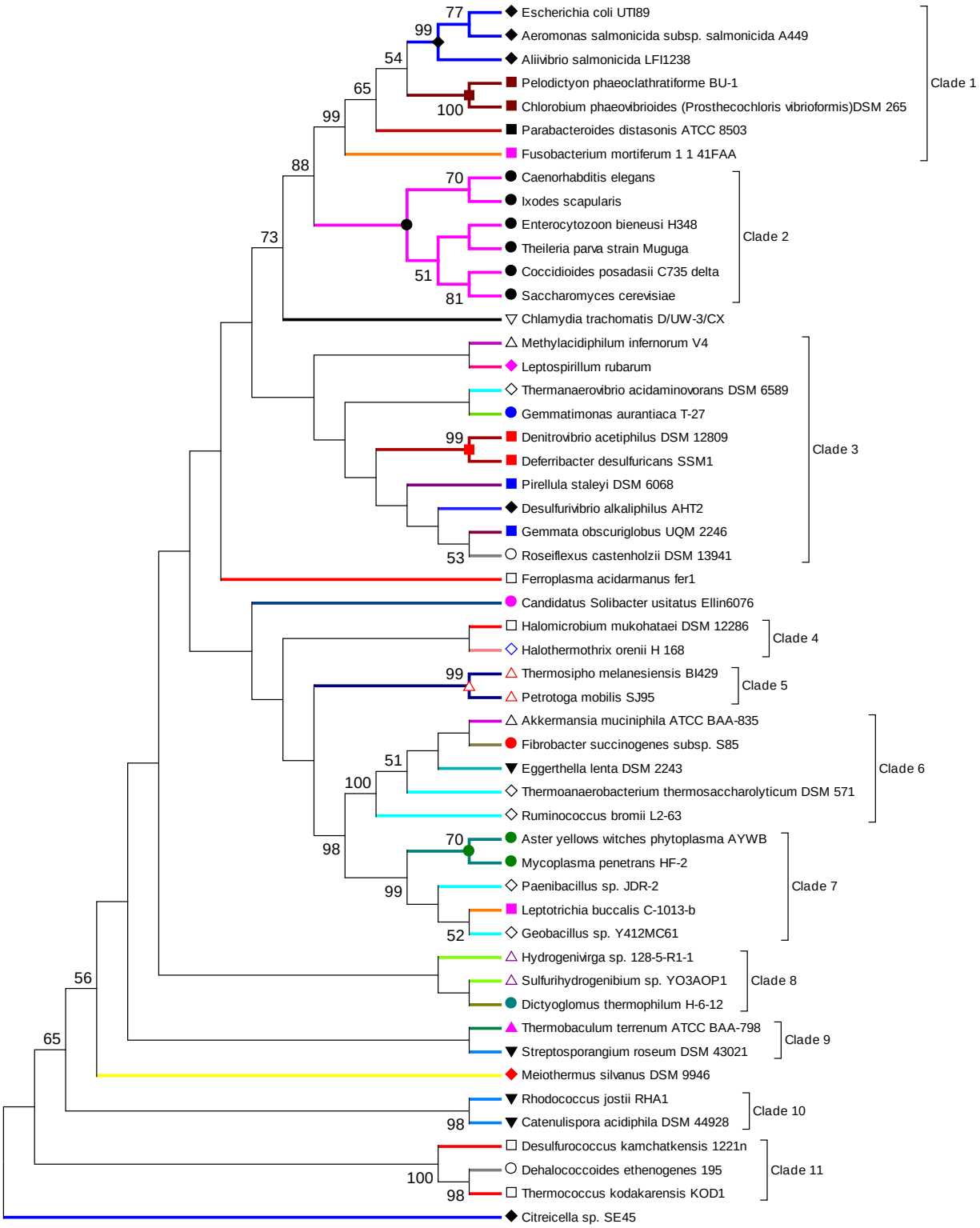


Fig. 4.8 Endonuclease IV protein sequences based maximum likelihood tree. The evolutionary distances were computed using the JTT as substitution model**.** The bootstrap values are given as Fig. 4.6.

amino acids in *C.elegans* (Nematoda), ~118 amino acids in *Coccidioides posadasii* (Animalia), at N-terminal region.

All the endonuclease IV proteins were also examined for likely transit peptide and its sub-cellular localization using the TargetP program (http://www.cbs.dtu.dk/services/TargetP). A 12-mer and 18-mer signal peptides were predicted within N-terminal region of *Coccidioides* and *Saccharomyces cerevisae* respectively. TargetP has also predicted that these signal peptides were localized at mitochondria.

4.4 3. Structural evolution of endonuclease IV homologs

Among the homologs of endonuclease IV protein family, the structures of *E.coli* (PDB ID: 1qum) and *Geobacillus sp.* (PDB ID: 3aal) are solved experimentally. Both the structures belonged to $\alpha_8\beta_8$ TIM barrel fold with mainchain rms deviation of around 1.39Å (shown in Fig. 4.9) although the sequence identity between these two structures is only 32.3% which indicates that structural fold plays important role in enzymatic activity. The conformations of loops in these two structures mostly deviate from each other. As it is observed in the superimposed structure (Fig 4.9), the loop between β strand 8 and helix 13 of *E.coli* is much shorter than that of *Geobacillus sp*. However, these differences in loop conformation do not affect the active site as well as Zn binding pocket of both structures.

Since, within the protein ML tree, endonuclease IV protein of *E.coli* and *Geobacillus sp* belong to clade 1 and clade 7 respectively, we have modeled representative sequence of remaining 9 distinct clades. Homology model of nine representative endonuclease IV protein sequences are

generated using *E.coli* endonuclease IV crystal structure as template structure. The structure quality each model is shown in Table 4.5. The RMSD values of modeled structures with their template structure are in the range of 0.47-1.63 Å. Among all the modeled structures homolog of

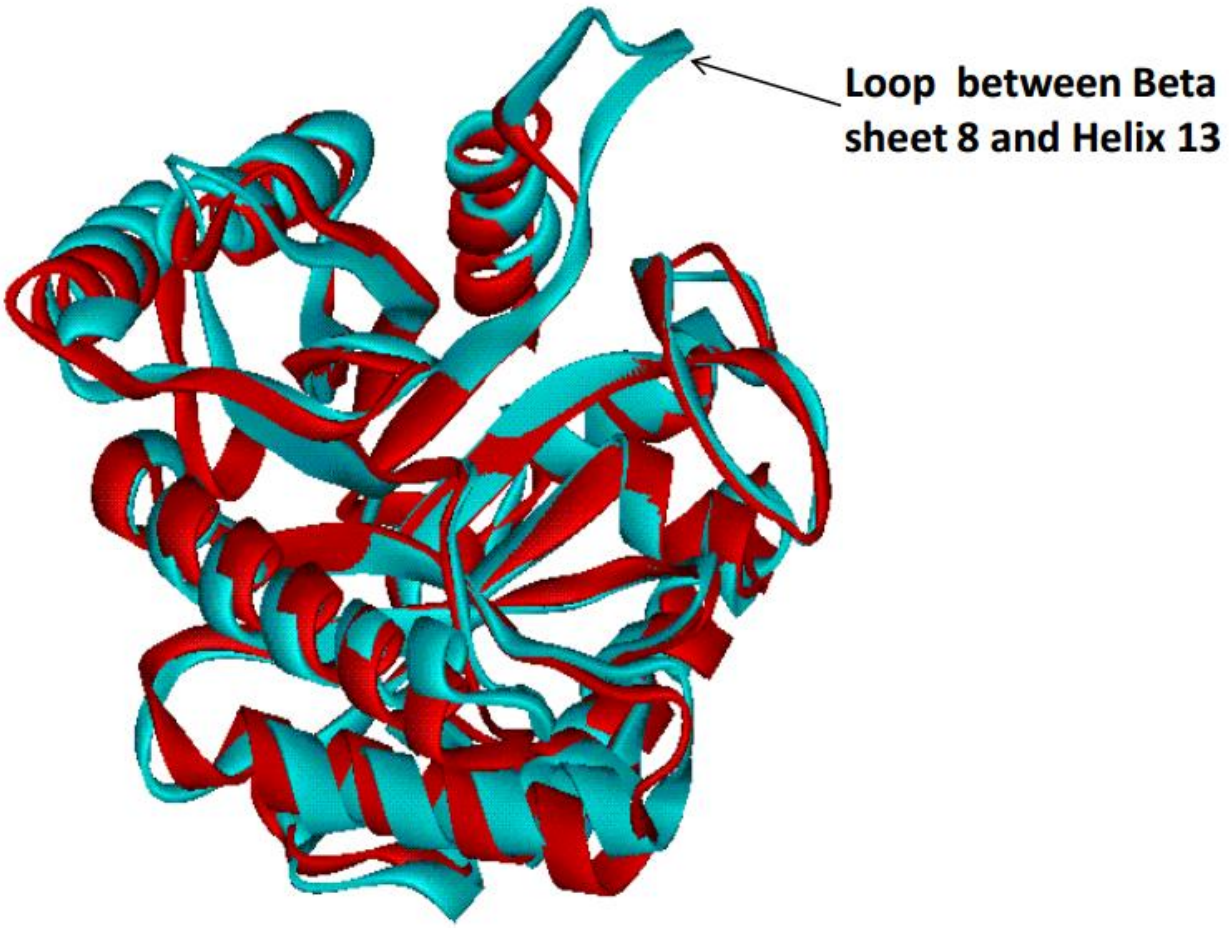


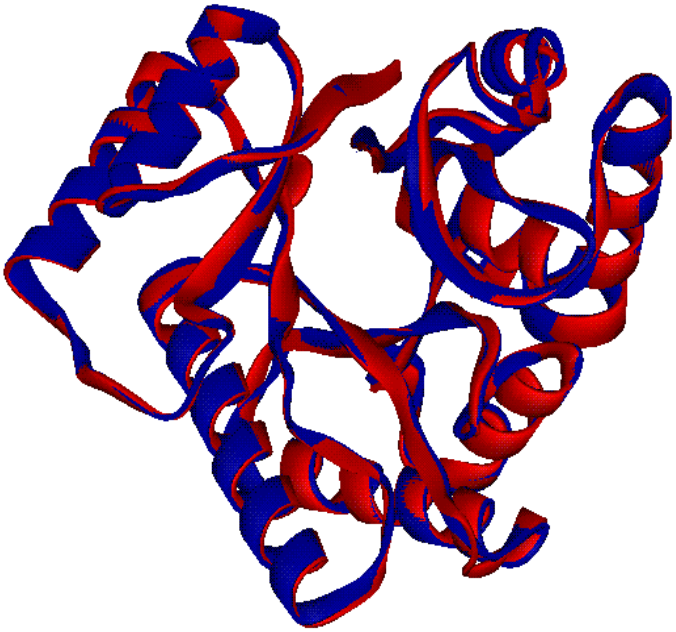
Fig. 4.9 Superimposition of the crystal structure of endonuclease IV protein of *E.coli* (Red) and *Geobacillus* (Cyan). The mainchain of both the structures are shown by ribbon representation. 3D diagrams of these models *were* generated by Accelrys Discovery studio ViewerPro 5.0.

Halothermothrix (Fig. 4.10 (A)) shows closest match with template structure whereas *Thermococcus* shows (Fig 4.10 (B)) largest deviation from template crystal structure. Each model is also validated by Ramachandran plot which shows that these protein structures are

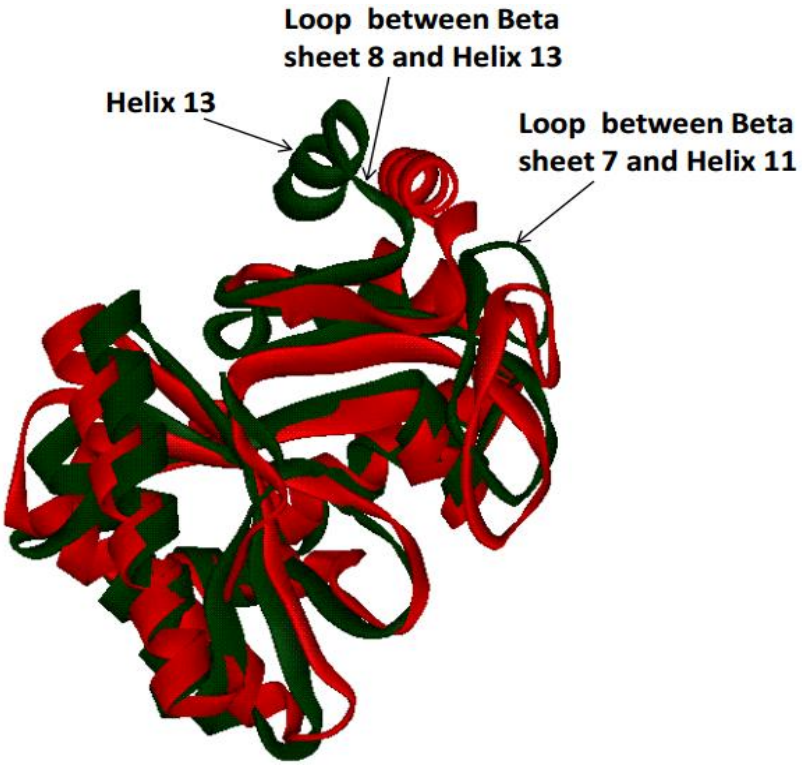
stereo chemically stable as maximum of 1% amino acids are outliers, while rest 99% amino acids are within favored regions of Ramachandran plot.

Table 4.5 Quality of the modeled endonuclease IV homologs.

Species Name (Clade No.)	**% identity with the template**	**RMSD with template structure (in angstrom)**	**Ramachandran Plot Statistics**	**Errat 2 Score**	**Qmean Score/Z-Score**
Gemmata (Clade no. 3)	44	C alpha- 0.9 Mainchain- 0.92	Favoured - 98.5% Allowed- 1.5% Outlier- 0%	88.14	0.782 (Z-score: 0.14)
Halothermothrix (Clade no. 4)	38%	C alpha- 0.47 Mainchain -0.53	Favoured - 98.8% Allowed- 1.2% Outlier- 0%	95.18	0.807 (Z-score: 0.41)
Petrotoga (Clade 5)	37%	C alpha- 0.83 Mainchain- 0.86	Favoured- 98.9% Allowed- 0.4% Outlier- 0 %	88.41	0.703 (Z-score: 0.72)
Ruminococcus (Clade no. 6)	35%	C alpha- 0.57 Mainchain- 0.63	Favoured- 99.6%) Allowed- 0.4% Outlier- 0 %	88.97	0.819 (Z-score: 0.53)
Saccharomyces (Clade 7)	42%	C alpha- 1.09 Mainchain- 1.07	Favoured- 98.9% Allowed- 1.1% Outlier- 0.0%	84.19	0.776 (Z-score: 0.05)
Hydrogenivirga (Clade 8)	36%	C alpha- 0.69 Mainchain- 0.73	Favoured- 98.8% Allowed- 0.8% Outlier- 0.4%	81.67	0.733 (Z-score: -0.39)
Thermobaculum (Clade 9)	34%	C alpha- 1.22 Mainchain- 1.22	Favoured- 98.4% Allowed- 1.2% Outlier- 0.4%	87.09	0.737 (Z-score: -0.34)
Rhodococcus (Clade 10)	30%	C alpha- 1.01 Mainchain- 1.04	Favoured- 96.4% Allowed- 2.6% Outlier- 1.0%	90.32	0.654 (Z-score: -1.18)
Thermococcus (Clade 11)	25%	C alpha- 1.61 Mainchain- 1.63	Favoured- 95.6% Allowed- 3.9% Outlier- 0.5%	62.00	0.542 (Z-score: -2.43)



A.



B.

Fig. 4.10(A-B) Crystal structure of endonuclease IV in *E.coli* (Red) A. Model structure of endonuclease IV in *Halothermothrix* (Closest, Blue) B. Model structure of endonuclease IV in *Thermococcus* (Farthest, Green). 3D diagrams of these models were generated by Accelrys Discovery studio ViewerPro 5.0.

The QMEAN scores and QMEAN Z-scores also indicate that the qualities of models are reasonably good. In all the homology modeled structures, only orientation of majority of loops differ marginally indicating possible rigidness of this enzyme structure over long evolutionary period.

4.4 A model for the evolutionary history of endonuclease IV protein family

To understand the possible evolutionary history of endonuclease IV family, we propose a model on the basis of sequence and phylogenetic tree analysis. The proposed model (Fig. 4.9) suggests Horizontal gene transfer (HGT) event which may have occurred between the endonuclease IV protein homologs of archaeal and bacterial species and vice-versa as in the gene tree. We observed that species from different bacterial division shared same clade with archaeal species. Similarly, endonuclease IV protein homologs of eukaryotic species in gene tree shared same clade with bacterial species suggest the possible HGT from bacteria to eukaryotes.

Another possible event of importance in endonuclease IV protein family is the endosymbiotic transfer of endonuclease IV genes from bacterial species to eukaryotes. It has been postulated that mitochondria might be incorporated into eukaryotic cells from proteobacteria progenitors (Liu et al. 1993). Prediction of signal peptide and their targeting to mitochondria in fungal species *Saccharomyces* and *Coccidioides* suggests possible endosymbiotic transfer of endonuclease IV genes to eukaryotes.

The absence of endonuclease IV homologs in higher eukaryotes (beyond nematode) suggests that possibly of loss of endonuclease IV gene higher eukaryotes during the course of evolution.

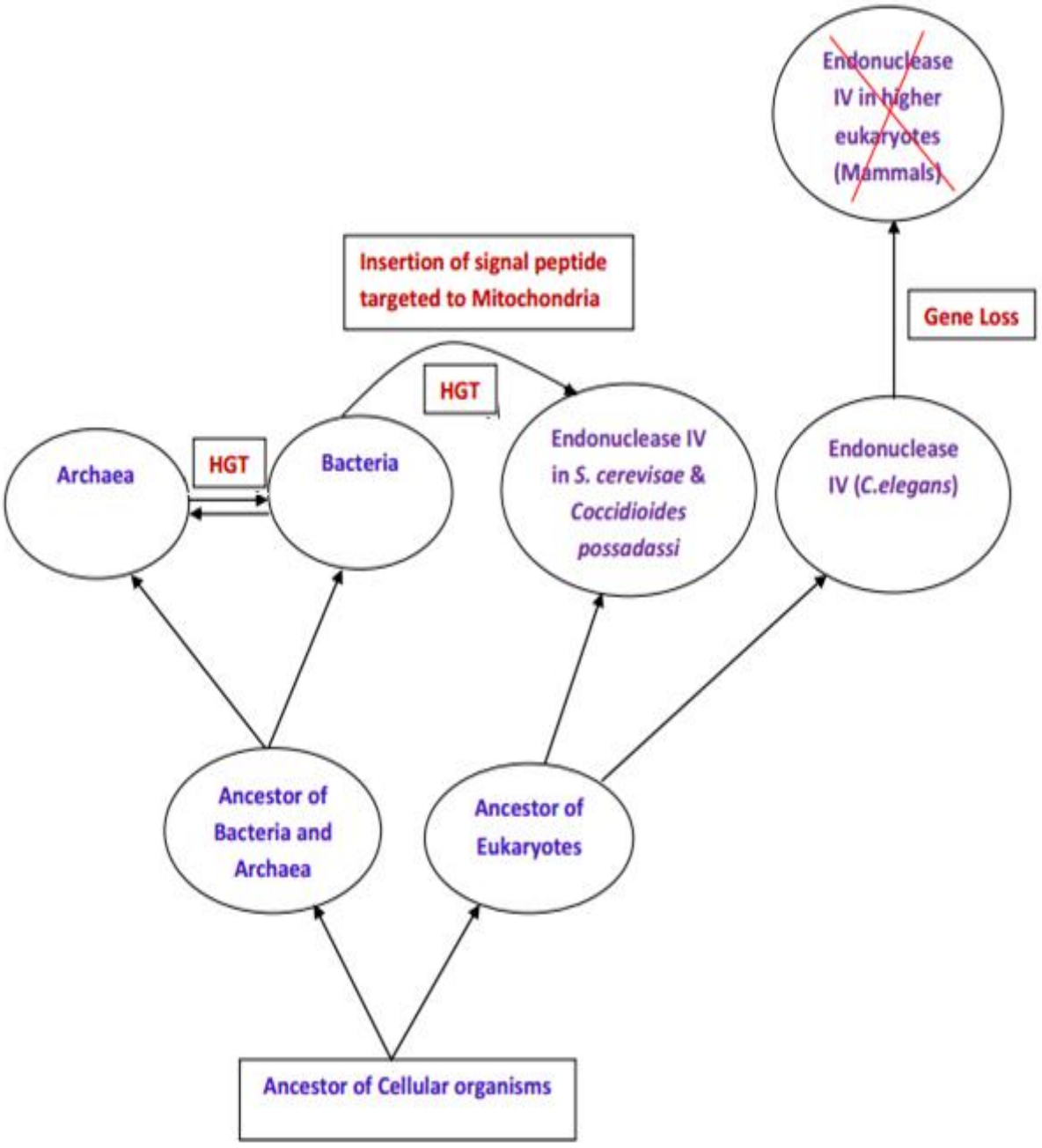


Fig. 4.11 A model for the evolutionary history of the endonuclease IV gene family is schematically shown. The eukaryotic endonuclease IV genes were likely originated from bacterial homologs through HGT. HGT events were also observed between bacterial and archaeal species. Gene loss was also observed from higher eukaryotes after *C.elegans*. An N-terminal insertion is noticed in fungal endonuclease IV protein which is targeted to mitochondria.

4.7 Conclusions

This study provides an overall picture of the evolutionary history of endonuclease IV gene/protein family that plays a crucial role in base excision repair of DNA. Based on conservation of amino acid positions, four consensus sequences have been identified for minor groove motif and for other regions of importance. Endonuclease IV protein based phylogenetic tree reveals few HGT events from bacteria to archaea. Endosymbiosis events are also predicted from the evolutionary model.

Chapter V

Evolutionary study of exonuclease III protein family

5.1 Introduction

Exonuclease III is a monomeric protein with four catalytic activities (multifunctional). (i) 3′-5′exonuclease specific for bihelical DNA (exodeoxyribonuclease activity); (ii) Ribonuclease H activities for the RNA strand in a RNA-DNA hybrid duplex (iii) DNA 3′-phosphatase activity which hydrolyzes 3’ phosphomonoesters from DNA and (iv), AP endonuclease activity which cleave DNA endonucleolytically at Apurinic/Apyrimidinic site creating base-free deoxyribose 5’-phosphate end group (Saporito et al. 1988; Kow & Wallace 1985; Hoheisel 1993; Rogers & Weiss 1980; Weiss 1981). Exonuclease III enzyme basically belong to Class II family of AP endonucleases which mainly act on abasic site during base excision repair mechanism and break the phosphodiester bond at the 5’ side of an apurinic/apyrimidinic (AP) site.

Homologs of exonuclease III are major constituents of this class. Proteins within this class of enzyme belong to Apn1 and Apn2 subfamilies. Homologs of all belong to In humans, Ape1 is the major exonuclease III protein (also known as Apex, HAP1, or Ref-1) which accounts for >90% of the cellular AP endonuclease activity (Dimple & Harrison, 1994). In case of lower eukaryotes like *Schizosaccharomyces pombe*, Apn2 provides the major AP-endonuclease activity while the Apn1 serves only as a back-up activity (Ribar et al. 2004). In addition to its major role as an AP endonuclease during BER, Ape1 also possesses 3’-5’-exonuclease activity, a weak 3’-phosphatase activity and a 3,-phosphodiesterase activity (Dimple & Harrison 1994). These activities are required for the removal of 3’-blocking groups created by ionizing radiation, oxygen free radicals, radiomimetic anti-tumor drugs, and the 3’-AP lyase activities of bifunctional DNA glycosylases (Demple & DeMott 2002; Mitra et al. 2002). Unlike exonuclease III in *E.coli,* Ape1 can activate transcription factors via a redox mechanism (Evans et al. 2000).

In contrast to Ape1 in human, the Apn2 protein has only a weak AP-endonuclease activity but exhibits a strong 3' exonuclease-phosphodiesterase activity. In addition, Ape2 has been shown to localize not only to the nucleus but also to some extent to the mitochondria, in which it may help to maintain the function and integrity of mitochondrial DNA. The difference in enzymatic activities of Apn1 and Apn2 is indicated by that the apn1Δstrain displays a much higher level of sensitivity to the alkylating agent methyl-methanesulfonate (MMS) than theapn2Δstrain (Unk et al. 2001).

Exonuclease III participates in phosphate bond cleavage at AP site through a nucleophilic attack which is facilitated by single bound metal ion Mn2+. Various active site residues are responsible for phosphate bond cleavage at AP site. Crystal structure of exonuclease III in E coli homolog (PDB ID: 1AKO) has demonstrated that active site pocket is surrounded by Asn7, Tyr109, Gln112, Asn153, Trp212, His259. During the catalysis ASP229 forms a hydrogen bond with His 259, and stabilizes the positive charge that develops when acting as general base. His259 abstracts a proton from water molecule. Then resultant hydroxide ion attacks the phosphate group and forms a penta-covalent transition state. The metal ion bound by Glu34 interacts with negatively charged phosphate group and aids the nucleophilic attack, stabilizes the penta-covalent transition state and also polarize the P-O bond for phosphate bond cleavage (Mol et al. 1995).

The unified view of evolution of exonuclease III proteins are presented in this study which provide an insight into the evolution of the exonuclease III gene/protein family among all lineages of life. We propose a model of the evolutionary history of the entire exonuclease III protein family.

5.2 Material and methods

5.2.1 Retrieval and selection of exonuclease III protein homologs

E.coli exonuclease III protein sequence (NCBI Acc. No. NP_416263.1) was used as a query to retrieve homologs. BLASTP (Altschul et al. 1990) two rounds of PSI-BLAST with default parameter was used for retrieving homologous sequences After removing redundant hits, global pairwise sequence alignment was performed with remaining hits from BLAST and PSI-BLAST search. Protein sequences (exonuclease III homologs) with less than 15% identity against corresponding *E.coli* exonuclease III protein were removed after pairwise global alignment. Then, exonuclease III homologs were evaluated for having at least one domain which was common in all exonuclease III protein homologs, which were selected finally.

A set of 404 homologs of exonuclease III family proteins were considered finally for evolutionary study and were divided into five kindoms e.g. monera (bacteria and archaea both comes under this kingdom), protista, fungi, plantae and animalia (Whittaker 1969).

Out of 404 homologs of exonuclease III family proteins, 355 were from bacteria, 15 were from archaea and rest 34 were from eukaryotes. Since the numbers of bacterial homologs were large in number, so sequences from various divisions were taken into account (Garrity & Holt 2001). Phylogenetic trees were generated for exonuclease III homologs of nine bacterial divisions (Bacteridetes, Aquificae, Chlorobi, Cyanobacteria**,** Actinobacteria, Firmicutes, Spirochaetes**,** Verrucomicrobia and Proteobacteria).

All phylogenetic trees were inspected visually and representative homolog from different clades of each tree was chosen (Fig. 5.1 A-H). A large number of homologs were within Proteobacteria

division and hence for clarity it is not shown). For other bacterial divisions, single representative homolog was selected.

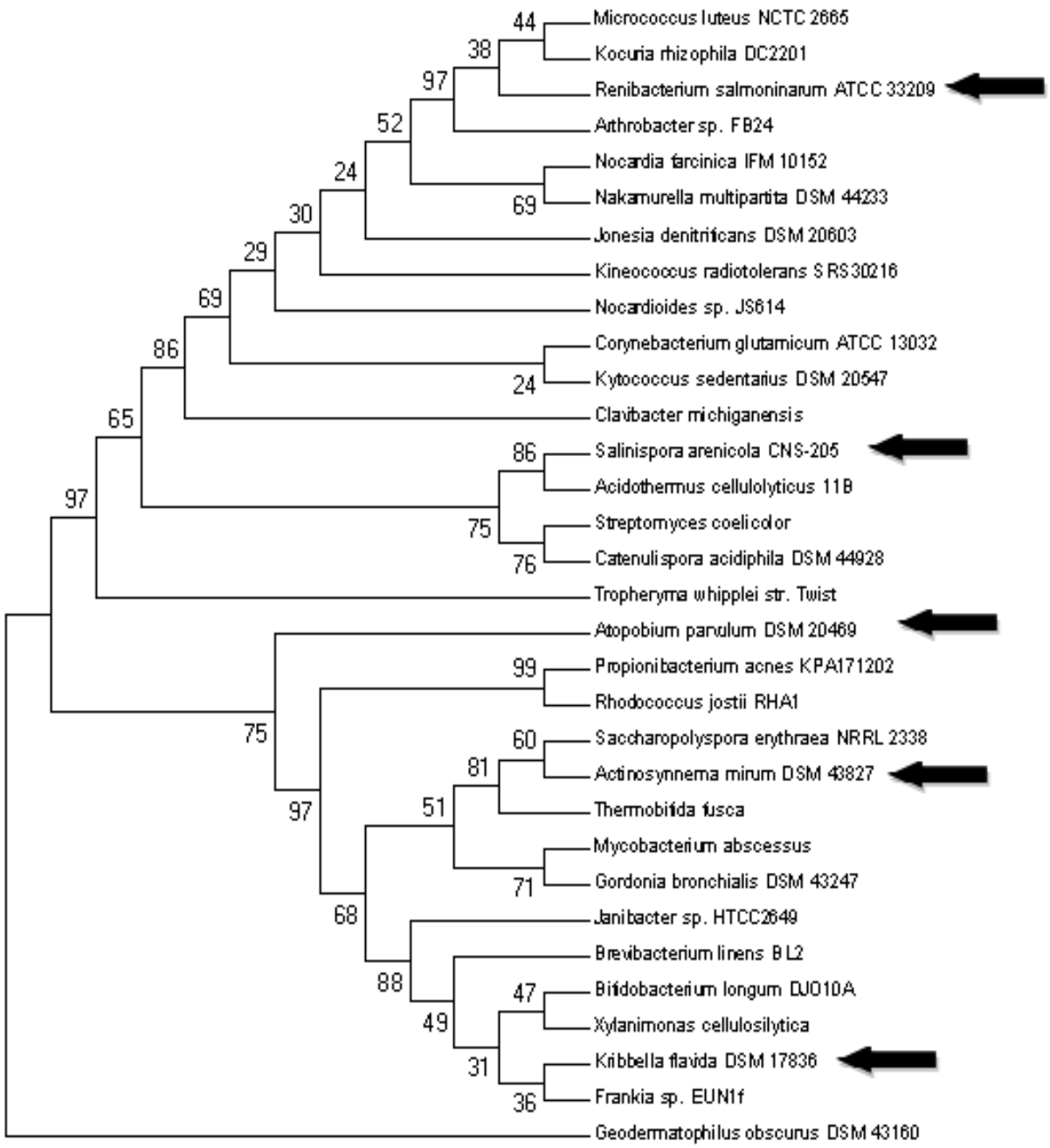


Fig 5.1(A) The NJ tree of 32 Actinobacteria species is shown. Selected species from Actinobacteria are shown by arrow.

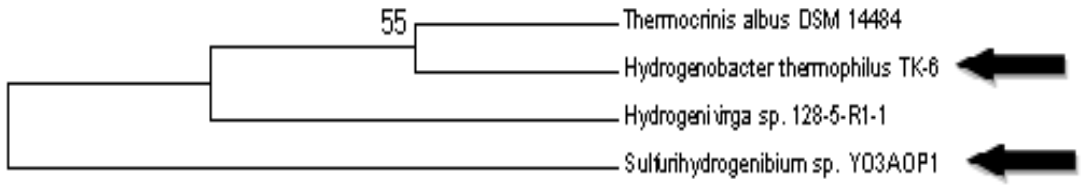


Fig 5.1(B) Two selected Aquificae species from NJ tree are indicated by arrow.

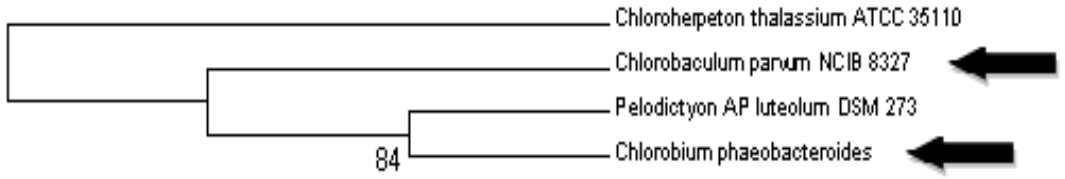


Fig. 5.1(C) NJ tree of 4 Chlorobi species is shown. Selected species from Chlorobi are shown by arrow.

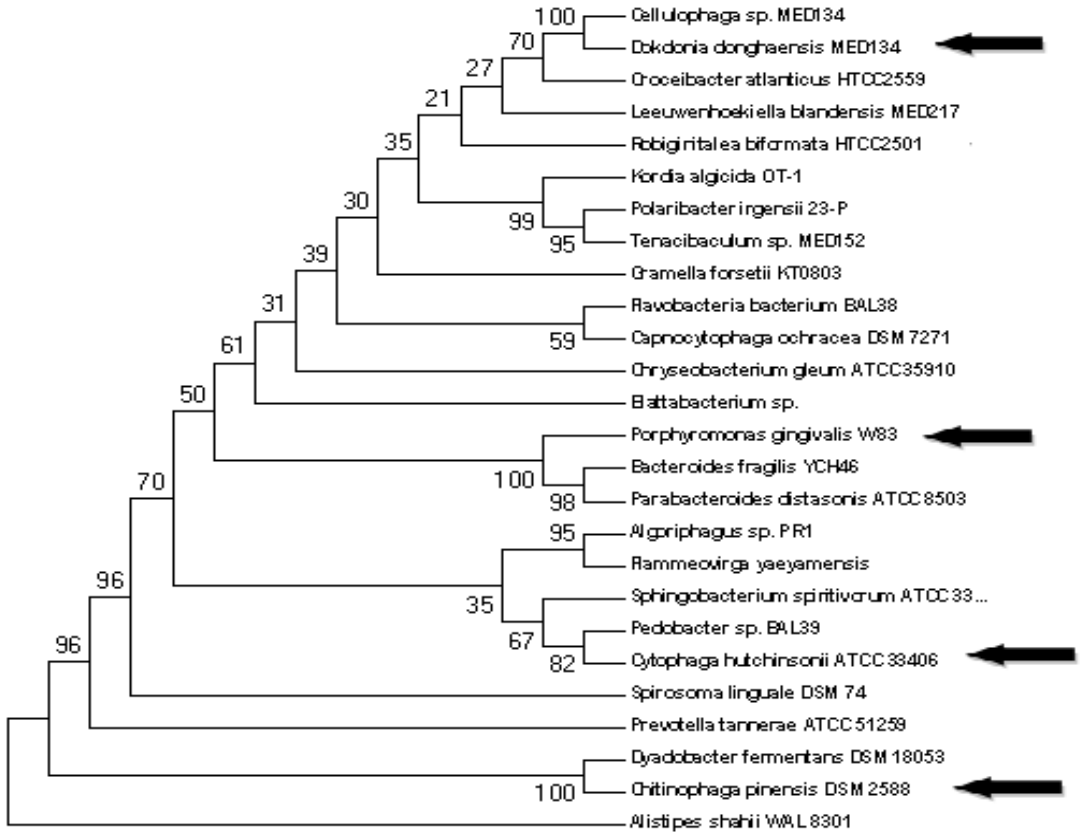


Fig. 5.1(D) Four Bacteroidetes species are selected from NJ tree (indicated by arrow).

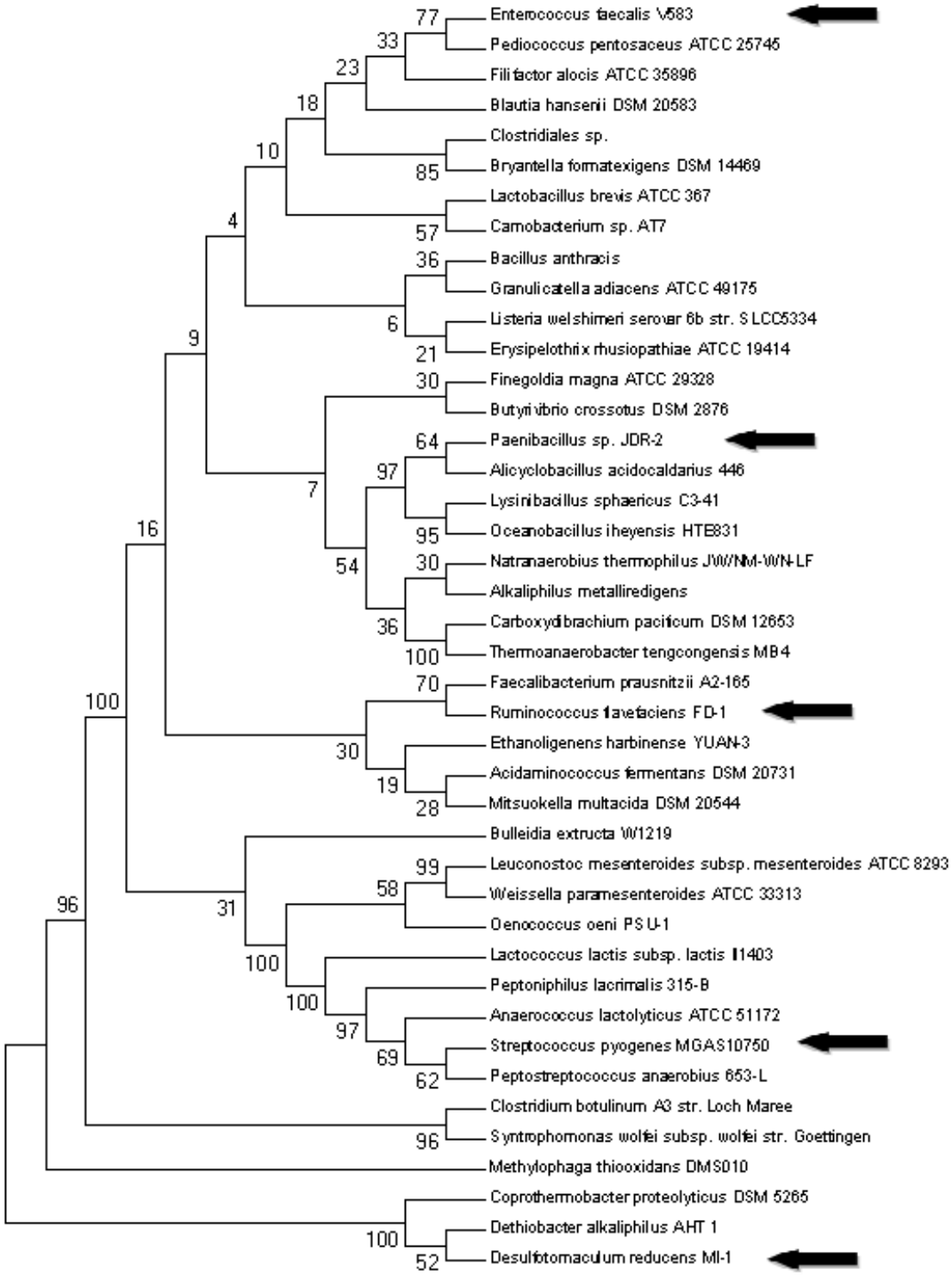


Fig. 5.1(E) NJ tree of 42 Firmicutes species produces 6 clusters. Selected species from Firmicutes are shown by arrow.

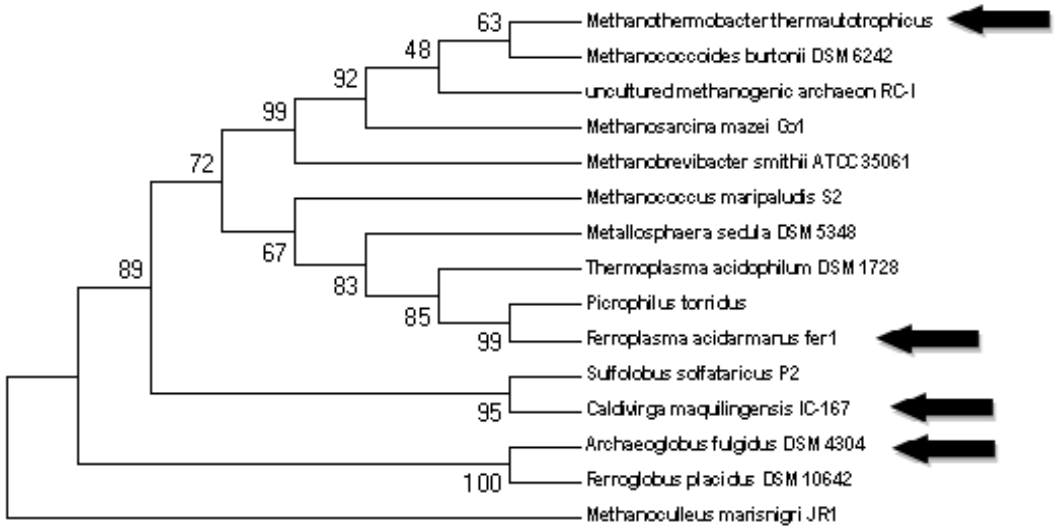


Fig. 5.1(F) NJ tree of 17 Cyanobacteria. Selected species from Cyanobacteria are shown by arrow.

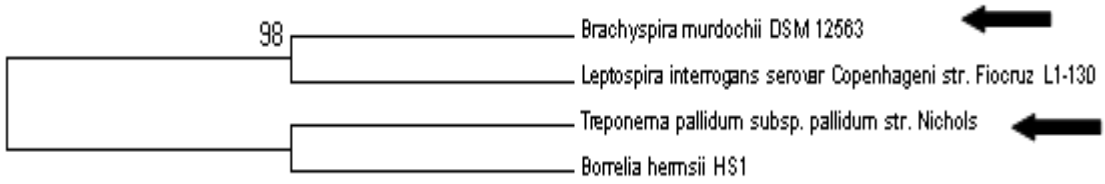


Fig. 5.1(G) Two Spirochaetes species are selected from the NJ tree.

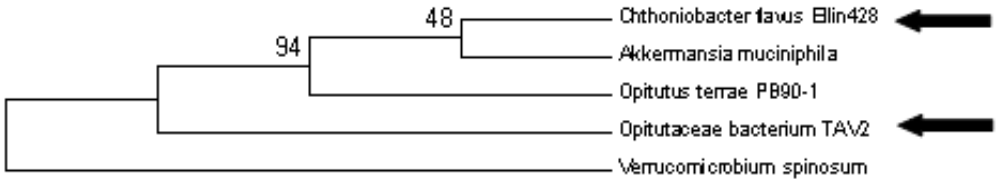


Fig. 5.1(H) Selected species from Verrucomicrobium division are shown by arrow.

NJ trees are constructed for other three eukaryotic kingdoms (Protista, Plantae and Animalia) and total 9 representative homologs were chosen after visual inspection (Figure 5.2 (A-C).

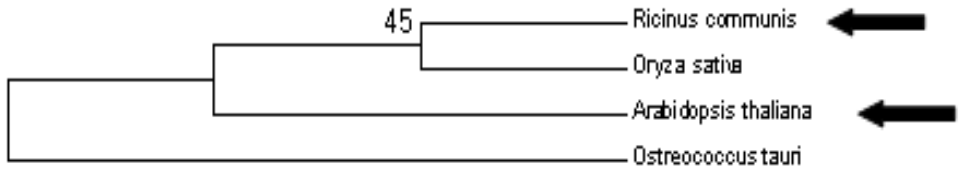


Fig. 5.2(A) NJ tree of 4 Plantae species are shown. Selected species from Plantae are shown by arrow.

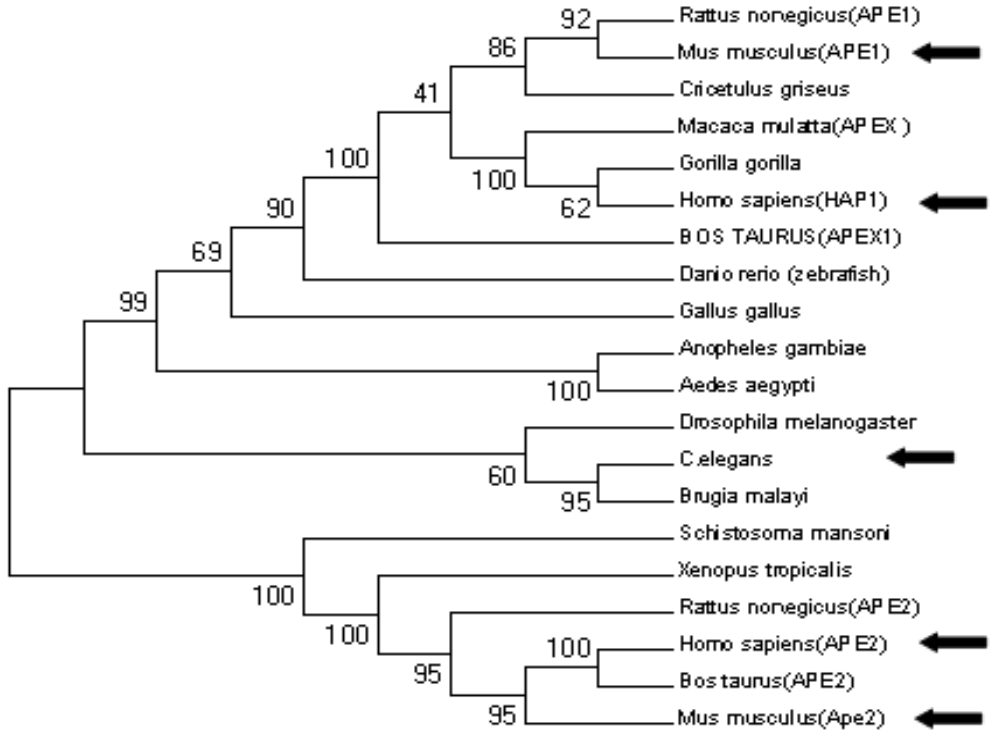


Fig. 5.2(B) NJ tree of 20 Animalia species produces three major clades. Five species are selected from three major clades (shown by arrow).

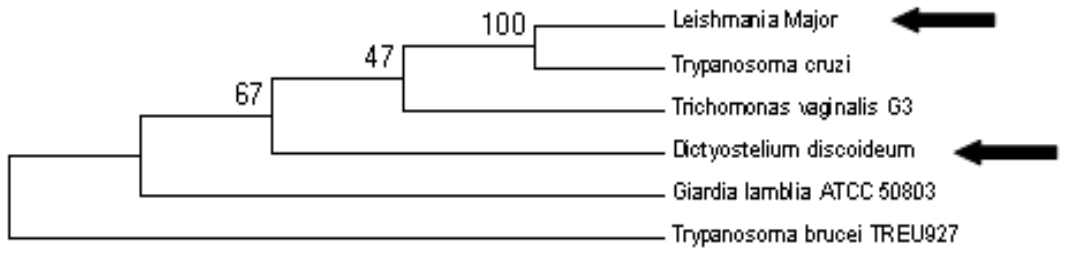


Fig. 5. 2(C) Two Protista species is selected from the NJ tree of 6 Protista species.

NJ tree was constructed for 15 exonuclease III homologs which belong to taxa Archaea. Finally, exonuclease III homolog in *Methanothermobacter*, *Ferroplasma Caldivirga* and *Archaeoglobus* were selected (Fig. 5.1).

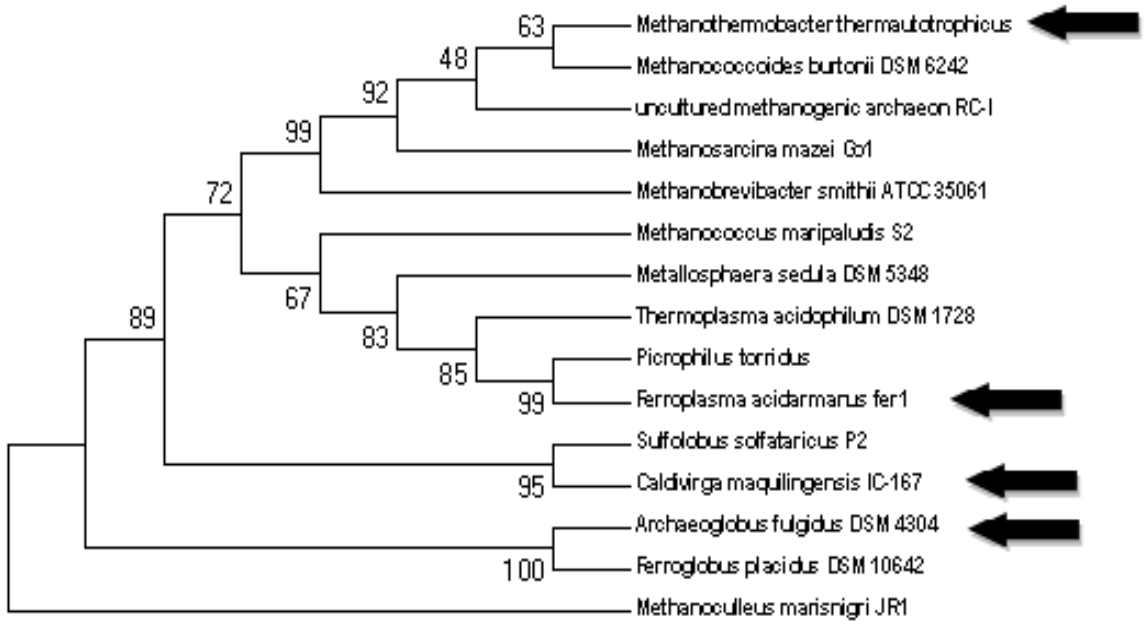


Fig 5.3. NJ tree is constructed with 15 archaeal species. Four archaeal species are selected from three major clades.

A set of 55 homologs of exonuclease III were finally selected to study evolution of exonuclease III protein family. Out of total 55 non redundant exonuclease III homologs, 41 homologs were exodeoxyribonuclease III/exoA proteins, 9 homologs were AP endonuclease/ AP endonuclease 1 proteins, 4 homologs were exonuclease III proteins and 1 homolog is apurinic endonuclease-redox protein. This set of protein homologs are from 41 bacterial, 4 archaeal and 11 eukaryotic divisions. The detail of data mining is listed in Table 5.1(A) and (B). The NCBI accession number of all 55 homologs of exonuclease III is listed in Table 5.2.

Table 5.1(A) List of different proteins (as it was annotated in the database) retrieved as exonuclease III homologs.

Serial No.	**Protein Names**	**No. of sequences**
1.	Exodeoxyribonuclease III/exoA	41
2.	AP endonuclease	4
3.	AP endonuclease 1 (Ape1)	2
4.	AP endonuclease 2 (Ape2)	3
5.	Exonuclease III	4
6.	Apurinic endonuclease-redox protein	1
Total		**55**

Table 5.1(B) Division and kingdom wise breakup of exonuclease III homologs. All bacterial species is further divided into 22 different categories (Garrity & Holt 2001).

Serial No.	**Division**	**Kingdom**	**Non-redundant Homologs for initial analysis**	**Homologs considered for Phylogenetic tree**

1.	Archaea	Monera	15	4
2.	Bacteria		355	41
3.	Eukaryotes	Fungi	4	1
		Protista	6	2
		Plantae	4	2
		Animalia	20	5
Total			**404**	**55**

Table 5.2 List of no. of homologs selected from total number of homologs from each division of archaea, bacteria and eukaryotes.

Serial No.	**Phylum**	**Division**	**Total no. of homologs**	**Numbers of homologs selected**
1.	Archaea		15	4
2. 1	Bacteria	Bacteroidetes	26	4
2.2		Chlorobi	4	2
2.3		Deinococcus	1	1
2.4		Actinobacteria	32	5
2.5		Aquificae	4	2
2.6		Nitrospirae	1	1
2.7		Firmicutes	42	5
2.8		Verrucomicrobium	5	2
2.9		Proteobacteria	209	5
2.10		Fusobacteria	2	1
2.11		Acidobacteria	1	1
2.12		Synergistetes	3	1
2.13		Deferribacteres	1	1
2.14		Cyanobacteria	17	4
2.15		Spirochaetes	4	2
2.16		Elusimicrobia	1	1
2.17		Lentisphaerae	1	1
2.18		Planctomycetes	1	1
3.1	Eukaryotes	Fungi	4	2
3. 2		Protista	6	2
3.3		Animalia	20	5

3.4		Plantae	4	2
Total			**404**	**55**

Table 5.3 The NCBI accession numbers of 55 exonuclease III protein homologs with their carrier organism as well the length of the protein sequences is shown.

S. No.	Name of the Organism	Exonuclease III Protein sequence accession no.	Protein name	Sequence length
1.	*Renibacterium salmoninarum ATCC 33209*	YP_001626209.1	exodeoxyribonuclease III	267
2.	*Atopobium parvulum DSM 20469*	YP_003179137.1	exodeoxyribonuclease III Xth	259
3.	*Actinosynnema mirum DSM 43827*	YP_003104122.1	exodeoxyribonuclease III Xth	264
4.	*Methanothermobacter thermautotrophicus str. Delta H*	NP_275355.1	exodeoxyribonuclease	258
5.	*Ferroplasma acidarmanus fer1*	ZP_05570798.1	exodeoxyribonuclease III	252
6.	*Dokdonia donghaensis MED134*	ZP_01049470.1	exodeoxyribonuclease	254
7.	*Porphyromonas gingivalis W83*	NP_904590.1	exodeoxyribonuclease III	254
8.	*Chitinophaga pinensis DSM 2588*	YP_003123046.1	exodeoxyribonuclease III Xth	259
9.	*Cyanothece sp. CCY0110*	ZP_01728343.1	Exodeoxyribonuclease III xth	262
10.	*Anabaena variabilis ATCC 29413*	YP_323069.1	Exodeoxyribonuclease III xth	260
11.	*Prochlorococcus marinus str. MIT 9313*	NP_895122.1	Exodeoxyribonuclease III	279
12.	*Enterococcus faecalis V583*	NP_816366.1	Exodeoxyribonuclease III	251
13.	*Paenibacillus sp. JDR-2*	ZP_02846960.1	Exodeoxyribonuclease III Xth	258
14.	*Desulfotomaculum reducens MI-1*	YP_001112539.1	Exodeoxyribonuclease III	264
15.	*Variovorax paradoxus S110*	YP_002943876.1	exodeoxyribonuclease III Xth	260
16.	*Chthoniobacter flavus Ellin428*	ZP_03132177.1	Exodeoxyribonuclease III Xth	216
17.	*Opitutaceae bacterium TAV2*	ZP_02011264.1	Exodeoxyribonuclease III Xth	255
18.	*Dictyostelium discoideum isolate NC4*	AAC47024.1	class II AP endonuclease	361
19.	*Leishmania Major strain Friedlin*	pdb|2J63|A	Ap Endonuclease Lmap	467
20.	*Homo sapiens* (HAP1)	CAA42437.1	HAP1	318
21.	*Homo sapiens* (APE2)	AAD43041.1	APE2	518

22.	*Caenorhabditis elegans*	NP_001021584.1	EXOnuclease family member (exo-3)	288
23.	*Candidatus Solibacter usitatus Ellin6076*	YP_827936.1	Exodeoxyribonuclease III Xth	258
24.	*Hydrogenobacter thermophilus TK-6*	YP_003432177.1	Exodeoxyribonuclease III	268
25.	*Chlorobium tepidum TLS*	NP_663010.1	AP endonuclease	253
26.	*Denitrovibrio acetiphilus DSM 12809*	YP_003503257.1	Exodeoxyribonuclease III	266
27.	*Chromobacterium violaceum*	NP_900547.1	exodeoxyribonuclease III	257
28.	*Ricinus communis*	XP_002530098.1	AP endonuclease,	486
29.	*Sebaldella termitidis ATCC 33386*	YP_003307192.1	Exodeoxyribonuclease III Xth	255
30.	*Anaerobaculum hydrogeniformans ATCC BAA-1850*	ZP_06440964.1	Exodeoxyribonuclease III	261
31.	*Brachyspira murdochii DSM 12563*	ZP_04048223.1	Exodeoxyribonuclease III	257
32.	*E.coli str. K-12 substr. MG1655*	NP_416263.1	Exonuclease III	268
33.	*Brucella suis 1330*	NP_697886.1	Exodeoxyribonuclease III	260
34.	*Sulfurovum sp. NBC37-1*	YP_001357960.1	Exodeoxyribonuclease III	260
35.	*Marinobacter algicola DG893*	ZP_01892520.1	Exonuclease III	270
36.	*Treponema pallium*	NP_218565.1	Exodeoxyribonuclease (exoA)	264
37.	*Mus musculus* (APE1)	AAH52401.1	Apurinic/apyrimidinic endonuclease 1	317
38.	*Mus musculus* (APE2)	NP_084219.1	Apurinic/apyrimidinic endonuclease 2	516
39.	*Caldivirga maquilingensis IC-167*	YP_001541552.1	Exodeoxyribonuclease III Xth	237
40.	*Deinococcus geothermalis DSM 11300*	YP_593992.1	Exodeoxyribonuclease III	254
41.	*Leptospirillum sp.*	gb|EAY55834.1	Exonuclease III	265
42.	*Chlorobaculum parvum NCIB 8327*	YP_001997736.1	Exodeoxyribonuclease III	253
43.	*Arabidopsis thaliana*	gb|ACB29409.1	Apurinic endonuclease-redox protein	538
44.	*Roseobacter denitrificans*	YP_683364.1	Exodeoxyribonuclease III	260
45.	*Elusimicrobium minutum Pei191*	YP_001876134.1	Exodeoxyribonuclease III Xth	255
46.	*Lentisphaera araneosa HTCC2155*	ZP_01877254.1	Exodeoxyribonuclease (ExoA)	258
47.	*Rhodopirellula baltica SH1*	NP_868219.1	Exodeoxyribonuclease	256
48.	*Streptococcus pyogen*	YP_279796.1	Exodeoxyribonuclease III	303
49.	*Ruminococcus flavefaciens*	ZP_06142172.1	Exodeoxyribonuclease	248
50.	*Sulfurihydrogenibium sp*	YP_001931761.1	Exodeoxyribonuclease III Xth	264
51.	*Salinispora arenicola*	YP_001536599.1	Exodeoxyribonuclease III Xth	265
52.	*Kribbella flavida DSM 17836*	YP_003381272.1	Exodeoxyribonuclease III Xth	267

53.	*Cytophaga hutchinsonii*	YP_680073.1	Exodeoxyribonuclease III	254
54.	*Archaeoglobus fulgidus*	NP_069414.1	Exodeoxyribonuclease III (xthA)	257
55.	*Schizosaccharomyces pombe* (Apn2)	NP_595522.1	Apn2	523

5.3 Results and Discussion

5.3.1 Domains and motifs of exonuclease III family

We have identified XthA (COG0708) as main domain within exonuclease III protein homolog. XthA domain spans from residue 1 to 267 in *E.coli* exonuclease III and covers almost entire protein in most of the homologs. Multiple sequence alignment of all exonuclease III homologs shows that Asn229, His259 and Glu34 is fully conserved among all 55 homologs while Asn7, Tyr109 and Asn153 residues around active site pocket are highly conserved (more than 90% occasion). We have identified four consensus motifs between residues 28 and 36, 150 and 160, 225 and 229, and between 257 and 261. These motifs are shown by sequence logos (Fig 5.4 (A-D)). Sequences within the motifs are conserved in more than 80% homologs. For example, Lys36 (Fig 5.4A) along with Arg90 is conserved among almost all homologs and binds with the DNA substrate and interacts with the phosphodiester backbone (Mol et al. 1995).

Similarly, Asp151 and Asn153 (Fig 5.4B) are the residues of functional importance and their side chain interacts with the 5'-phosphate group (Mol et al. 1995). Asp229 (Fig 5.4 C) is the catalytic residue which acts as proton donor and thus initiates the catalytic activity while His259 (Fig 5.4D) acts as proton acceptor. Through multiple sequence alignment we have predicted the presence of two motifs within few eukaryotic species which was not reported so far. The first motif is predicted as a DNA binding Zinc finger ribbon with around 45-50 amino acids which

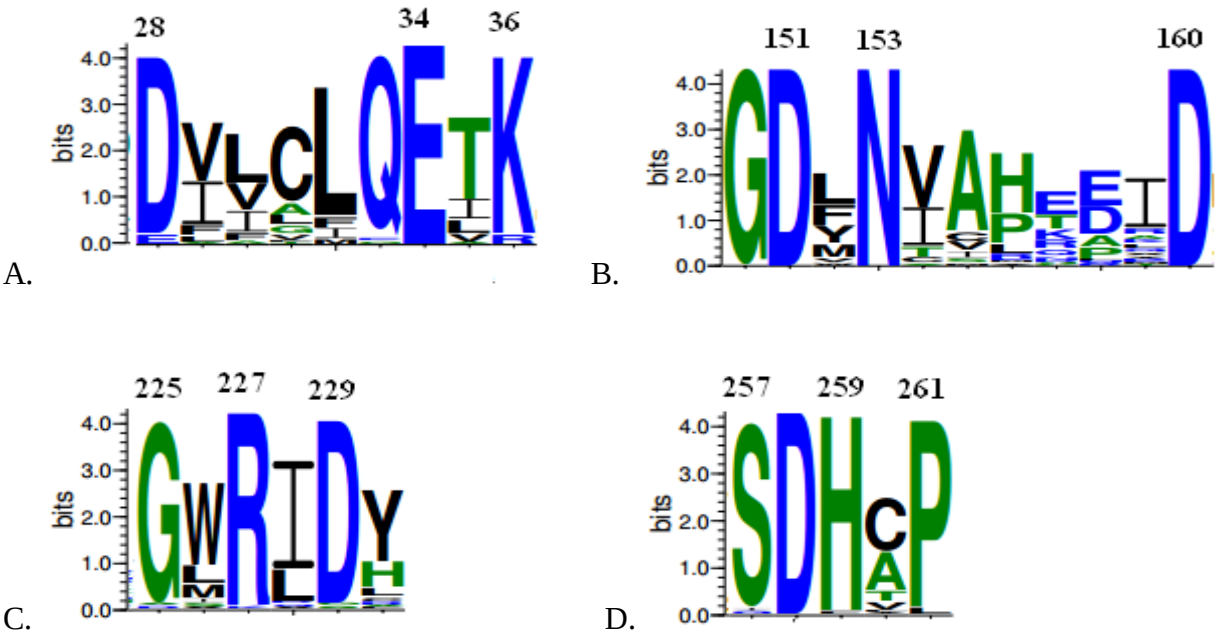


Fig. 5.4(A-D) Sequence conservation between residues A. 28-36 (DNA binding regions) B. 150-160 (DNA binding regions) C. 225-229 and D. 257-261 are shown by sequence logo representation.

was found towards C terminal end of Ape2 proteins of *Homo sapiens, Mus musculus* and *Schizosaccharomyces pombe*. The secondary structure prediction tool (Rost & Sander, 1993) predicts the presence of three β-strands within Zinc finger motif which has 30% sequence identity with 4th Lim domain of pinch protein (PDB ID: 1NYP). The other motif was identified as DNA-binding SAP motif with 20-25 amino acids. This motif was inserted at N terminal end of Ape proteins of *Arabidopsis thaliana* and *Ricinus communis.* This is involved in chromosomal re-organization to target the DNA repair proteins to transcriptionally active chromatin and helps in coupling of transcription with splicing and repair. The predicted SAP motif has 38% sequence identity with N-Terminal Sap Domain of Sumo E3 Ligases from *Saccharomyces Cerevisiae* (PDB ID: 2RNN).

5.3.2 16S/18S rRNA gene based species tree

Complete 16S/18S r-RNA gene sequences of 48 species are retrieved while partial 16S/18S r-RNA gene sequence are available for rest of the 7 species which are not considered to generate the species tree. A total of 11 distinct clads are formed in species tree (Fig. 5.5), among which archaea (Clade 10) and eukaryotes (Clade 11) are the farthest clades within the tree. As expected, species from same bacterial division as well as species from same kingdom stay close to each other in a clade, such clades are frequent in species tree. For example, in clade 6, *Ruminococcus*, *Desulfotomaculum*, *Paenibacillus*, *Enterococcus* and *Streptococcus* are placed together in and all of these species belong to bacterial division, firmicutes. Similarly, *Hydrogenobacter* and *Sulfurihydrogenibium* species belong to bacterial division aquificae and placed together in clade 9. In some clades, species from different bacterial divisions are clustered together. For example, clade 7 contains species from acidobacteria, elusimicrobia, lentisphaerae and verrucomicrobia bacterial division. Similarly, proteobacteria and fusobacteria and deinococcus-Thermus share clusters with species from chlorobi, bacteroidetes and actinobacteria respectively.

5.3.3 Exonuclease III gene based phylogenetic tree

The exonuclease III gene-based tree (Fig. 5.6) is initially diverged into two major branches which is supported by strong bootstrap values. Upper major branch is hierarchically divided into total eight clade. Among these, clade 1 is composed of exodeoxyribonuclease III, exonuclease III and AP endonucleases proteins which are majorly from six eukaryotic species. However, it also

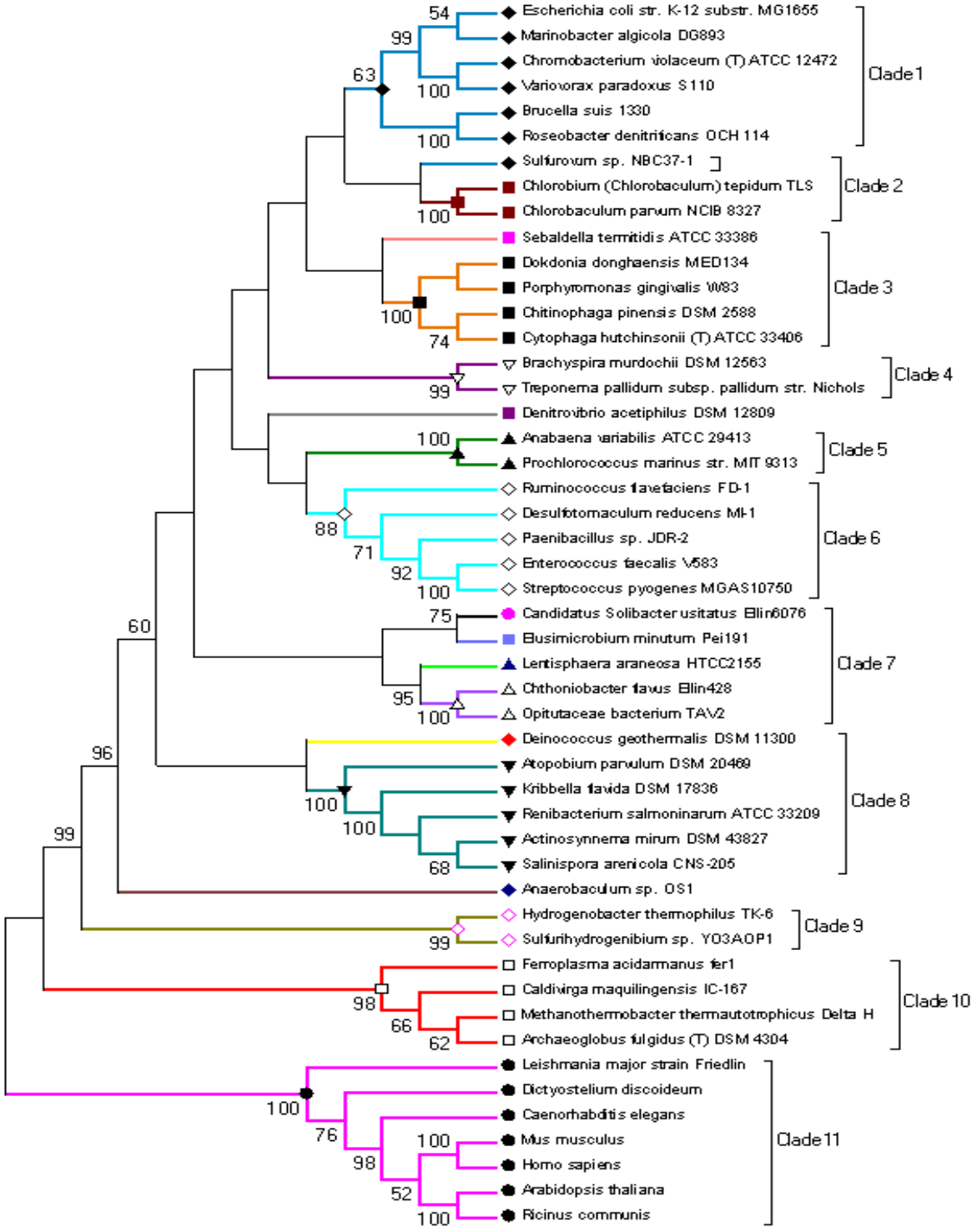


Fig. 5.5 16S/18S rRNA sequences of species based maximum likelihood tree**.** The evolutionary distances were computed using the Tamura Nei as substitution model**.** Bootstrap support values are presented next to the tree branches for each clade with ≥50. The tree is generated from a clustalw based multiple sequence alignment.

contains exonuclease III genes from two firmicutes and from one archaeal species. Clade 2 is hosts two species from Bacteroidetes which contains exodeoxyribonuclease III proteins. Clade 3 and clade 4 contains gene of exodeoxyribonuclease III proteins from different bacterial divisions. Clade 5, clade 7 and clade 8 contain species from different bacterial division with different class of exonuclease III proteins. Clade 6 hosts gene of exodeoxyribonuclease III proteins from two cyanobacterial and one other bacterial species along with one archaeal species.

The second major branch is further divided into five clades (clade 9-clade 13). Among theses clades, clade 9 contains Ape2 homologs from mammals. Clade 10 contains four bacterial species among which two are from aquificae. Similarly three actinobacterial species with two other bacterial species forms clade 11. Two proteobacterial species form Clade 12. Clade 13 has a mixed population of archaeal as well bacterial species. The branch containing AP endonuclease 2 (Ape2) of *S. pompe* may be considered as outgroup in the gene tree. *Exonuclease III* gene based phylogeny tree suggests that gene (*Exonuclease III*) and species evolution of aquificae may have similar pattern as these classes of organisms stay within a distinct clade in both species and gene based phylogenetic tree. It is interesting to note that the *Exonuclease III* genes of four archaeal species are within three different lineages and (Fig. 5.6**)**, although in species tree, all archaeal species are within the same clade. This observation indicates that *exonuclease III* gene of these archaeal species are evolved differently in which environmental pressure might have played an important role in shaping *exonuclease III* gene.

Exonuclease III gene based phylogeny analysis shows that a large number of species share a clade with species from different divisions. We try to explain this anomaly of sharing different clades by the species belonging to the same division through mean GC content of the gene, GC

content at the third codon position. GC content based on exonuclease III gene, 3rd position of the exonuclease III gene are listed in Table 5.4. In clade 1, mean GC content and GC content at the third codon position of *exonuclease III* gene of six eukaryotic organisms (*Dictyostelium, Ricinus, Arabidopsis, Caenorhabditis, Homo sapiens* and *Mus musculus*) with bacterial (*exonuclease III* homologs in *Enterococcus* and *Ruminococcus* in Firmicutes) and archaeal (*Ferroplasma*) species are 43.28 (Standard division of 6.97) and 42.28 (Standard division, of 13.59) respectively. In clade 5, we observed that average GC content and GC content at the third codon position of *exonuclease III* gene of eukaryotic organism *Leishmania* and bacterial species *Chlorobaculum* are 62.17 (standard deviation of 1.62) and 79.41 (standard deviation of 0.40) which explains the sharing of clade in the phylogenetic tree. *Denitrovibrio* and *E.coli* species shares same clade (clade 7) because their average GC content and GC content at the third codon position of *exonuclease III* gene are 48.55 (standard deviation of 6.86) and 51.83 (standard deviation of 13.97) respectively. We have observed that average GC content of *exonuclease III* gene of *Deinococcus*, *Renibacterium*, *Chromobacterium*, *Leptospirillum*, *Roseobacter*, *Chitinophaga*, *Candidatus* and *Variovorax* (clade 8) are 61.29 (with standard deviation of 5.7) which is possible reason behind the sharing of same clade by these species (from different bacterial divisions). *Exonuclease III* gene based phylogeny tree of 55 taxa suggests that GC content of gene contribute significantly towards the position of taxa within the tree and species evolution and e*xonuclease III* gene evolution shape up differently in most of the cases. Since, AP exonuclease III gene in all living organisms diverges over a long evolutionary period, synonymous nucleotide substitution, GC content at third codon position as well as mean GC content of gene makes gene sequence based phylogenetic tree construction noisy.

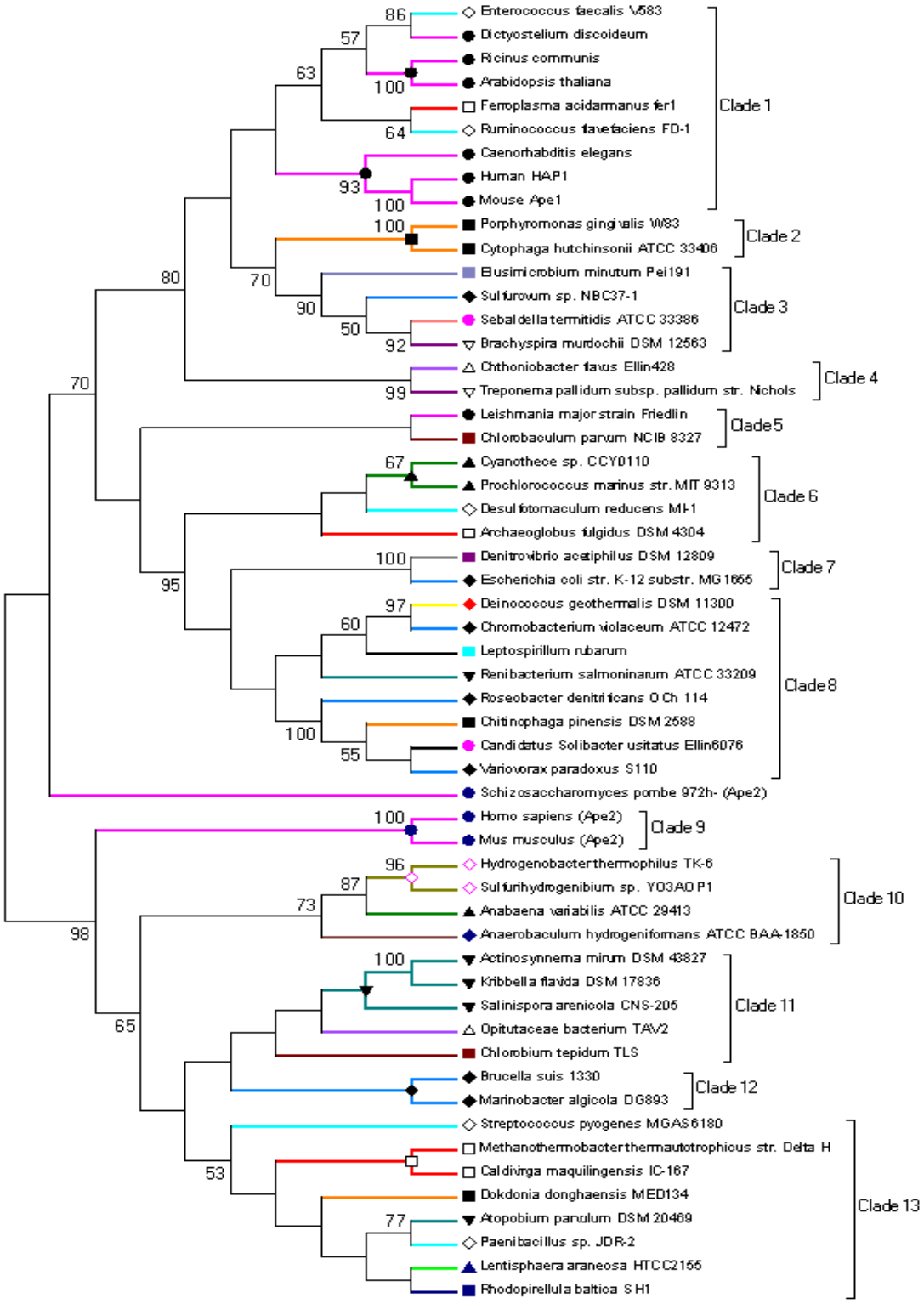


Fig. 5.6 Exonuclease III gene sequences based maximum likelihood tree**.** The evolutionary distances were computed using the Tamura Nei as substitution model**.** The bootstrap values are given as Fig. 5.5.

5.3.4 Exonuclease III protein based phylogenetic tree

The exonuclease III protein based tree (Fig. 5.7) is initially diverged into two major branches among which one is supported by strong bootstrap values. Upper major branch is further divided into total eight clades. Among these, clade 1 is composed of Exonuclease III, Apurinic endonuclease redox, AP endonucleases and AP endonuclease 1 (Ape1) proteins which are from seven eukaryotic species. Clade 2 hosts exodeoxyribonuclease III proteins of four species and one species each from Firmicutes and actinobacteria. Clade 3 contains exodeoxyribonuclease III proteins from different bacterial divisions. Clade 4 is composed of exodeoxyribonuclease III proteins from archaeal species. Clade 5 contains species from different bacterial division with III protein from two species belonging to bacterial division Chlorobi occupy Clade 7. Clade 8 is composed of AP endonuclease 2 proteins (Ape2) from three eukaryotic species. The second major branch is hierarchically divided into six clades (clade9-clade14). Among theses clades, clade 9 contains exodeoxyribonuclease III protein from mixed population of archaeal as well bacterial species. Clade 10 is composed of exodeoxyribonuclease III proteins from two aquificae

Table 5.4 GC content based on exonuclease III gene, 3rd position of the exonuclease III gene and genome of the organism as well average and standard deviation for three types of GC content are listed.

Serial no.	Name of the Organism	Mean GC content based on Exonuclease III gene	Average	GC content at 3rd bases based on Exonuclease III gene	Average	Clade No.

1.	*Enterococcus faecalis V583*	37.43	43.28 (6.97)	29.37	42.28 (13.59)	Clade 1
2.	*Dictyostelium discoideum isolate NC4*	32.50		20.44		
3.	*Ricinus communis*	40.38		34.50		
4.	*Arabidopsis thaliana*	42.30		39.70		
5.	*Caenorhabditis elegans*	48.56		59.17		
6.	*Homo sapiens* (Hap1)	52.66		54.86		
7.	*Mus musculus* (Ape1)	51.68		53.46		
8.	*Ferroplasma acidarmanus fer1*	37.55		34.78		
9.	*Ruminococcus flavefaciens FD-1*	46.45		54.22		
10.	*Porphyromonas gingivalis W83*	46.80	43.53 (4.6)	49.80	42.94 (9.7)	Clade2
11.	*Cytophaga hutchinsonii ATCC 33406*	40.26		36.08		
12.	*Elusimicrobium minutum Pei191*	42.71	37.142 5 (6.3)	45.31	35.58 (11.68)	Clade3
13.	*Sulfurovum sp. NBC37-1*	41.89		42.41		
14.	*Sebaldella termitidis ATCC 33386*	34.51		30.86		
15.	*Brachyspira murdochii DSM 12563*	29.46		19.77		
16.	*Chthoniobacter flavus Ellin428*	59.60	58.1 (2.12)	79.26	69.44 (13.88)	Clade4
17.	*Treponema pallidum subsp. pallidum*	56.60		59.62		
18.	*Leishmania major strain Friedlin*	63.32	62.17 (1.62)	79.69	79.41 (0.40)	Clade5
19.	*Chlorobaculum parvum NCIB 8327*	61.02		79.13		
20.	*Cyanothece sp. CCY0110*	32.45	44.59 (9.64)	19.01	44.22 (20.91)	Clade 6
21.	*Prochlorococcus marinus str . MIT 9313*	52.86		56.43		
22.	*Desulfotomaculum reducens MI-1*	41.26		35.85		
23.	*Archaeoglobus fulgidus DSM 4304*	51.81		65.60		
24.	*Denitrovibrio acetiphilus DSM 12809*	43.70	48.55 (6.86)	41.95	51.83 (13.97)	Clade 7
25.	*Escherichia coli str. K-12 substr. MG1655*	53.41		61.71		
26.	*Deinococcus geothermalis DSM 11300*	67.32	61.285 (5.7)	83.92	75.05 (12.34)	Clade 8
27.	*Renibacterium salmoninarum ATCC 33209*	56.22		61.19		
28.	*Chromobacterium violaceum ATCC 12472*	66.93		92.64		
29.	*Leptospirillum rubarum*	57.77		70.30		
30.	*Roseobacter denitrificans*	60.41		75.48		

	OCh 114					
31.	*Chitinophaga pinensis DSM 2588*	52.56		56.92		
32.	*Candidatus Solibacter usitatus Ellin6076*	61.13		73.36		
33.	*Variovorax paradoxus S110*	67.94		86.59		
34.	*Homo sapiens* (Ape2)	57.48	55.885 (2.25)	67.05	62.64 (6.4)	Clade 9
35.	*Mus musculus* (Ape2)	54.29		58.03		
36.	*Hydrogenobacter thermophilus TK-6*	43.37	41.11 (7.582)	46.10	48.48 (4.80)	Clade 10
37.	*Sulfurihydrogenibium sp . YO3AOP1*	30.06		43.02		
38.	*Anabaena variabilis ATCC 29413*	43.68		53.64		
39.	*Anaerobaculum hydrogeniformans ATCC*	47.33		51.15		
40.	*Actinosynnema mirum DSM 43827*	71.95	67.134 (5.6)	71.70	68.106 (3.8)	Clade 11
41.	*Kribbella flavida DSM 17836*	70.40		69.78		
42.	*Salinispora arenicola CNS-205*	70.68		70.68		
43.	*Chlorobium tepidum TLS*	58.79		62.60		
44.	*Opitutaceae bacterium TAV2*	63.85		65.77		
45.	*Marinobacter algicola DG893*	58.43	56.67 (2.48)	61.25	60.32 (1.31)	Clade 12
46.	*Brucella suis 1330*	54.92		59.39		
47.	*Methanothermobacter thermautotrophicus*	50.90	46.235 (4.54)	48.45	50.81 (4.31)	Clade 13
48.	*Caldivirga maquilingensis IC-167*	42.58		42.02		
49.	*Dokdonia donghaensis MED134*	40.13		52.94		
50.	*Atopobium parvulum DSM 20469*	47.82		52.31		
51.	*Paenibacillus sp. JDR-2*	50.45		54.44		
52.	*Lentisphaera araneosa HTCC2155*	44.14		51.74		
53.	*Rhodopirellula baltica SH 1*	51.75		55.64		
54.	*Streptococcus pyogenes MGAS6180*	42.11		49.01		

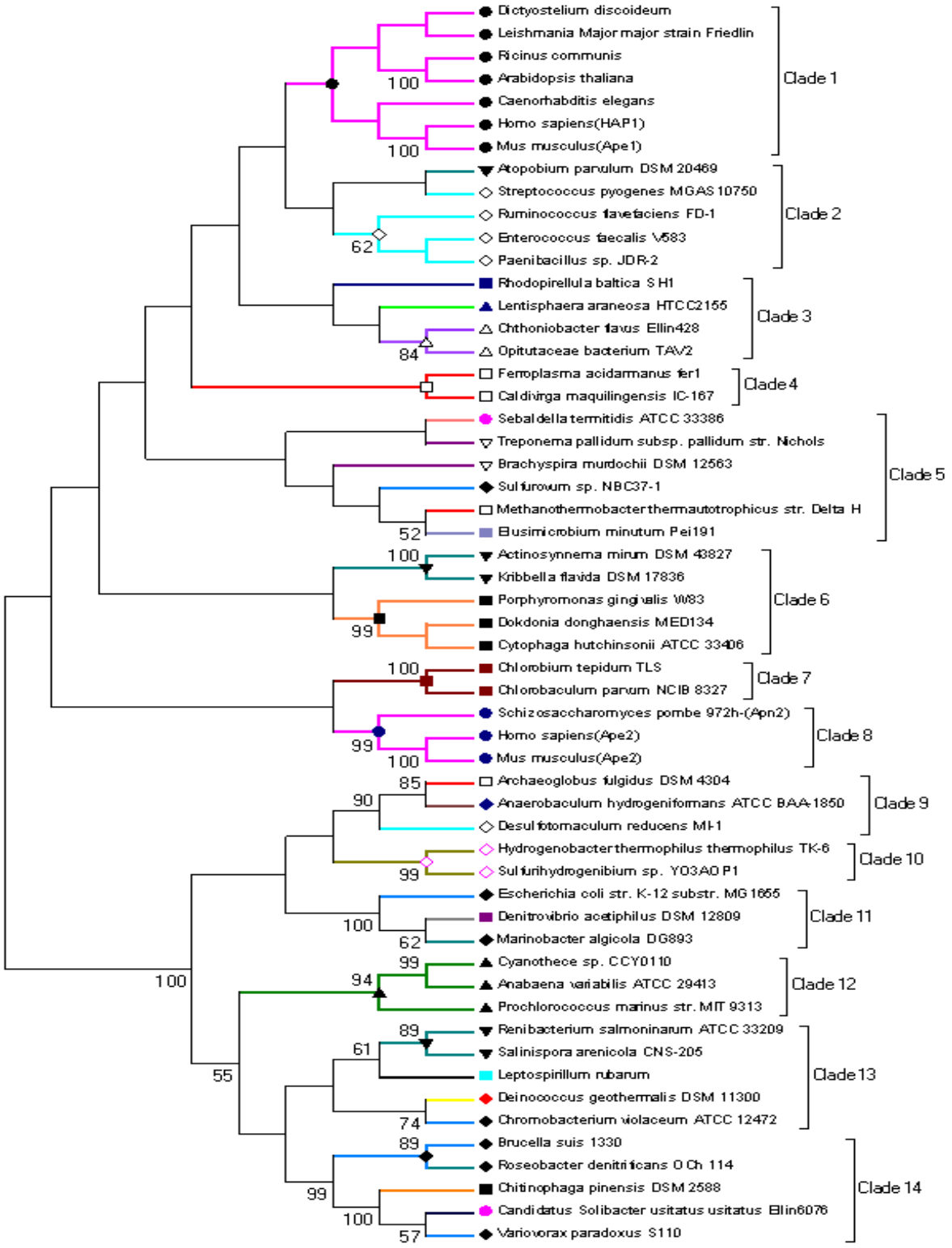


Fig. 5.7 Exonuclease III protein sequences based maximum likelihood tree**.** The evolutionary distances were computed using the JTT as substitution model**.** The bootstrap values are given as Fig. 5.5.

species. Clade 11 and Clade 14 hosts exonuclease III and exodeoxyribonuclease III proteins form mixed population of bacterial species. Clade 12 contains exodeoxyribonuclease III proteins from three cyanobacterial species. Clade 13 is composed of exodeoxyribonuclease III and exonuclease III proteins from mixed population of bacterial species.

5.4 Structural evolution of exonuclease III homologs

Among the homologs of exonuclease III, proteins from *E.coli* (PDB ID:1AKO), human (PDB ID: 1DE8) and *Archaeoglobus fulgidus* (PDB ID: 2VOA) are solved experimentally. The superimposition of exonuclease III homolog of human (Hap1) and *Archaeoglobus fulgidus* against *E.coli* is shown in figure 5.11 which indicates very little variation among these three experimental structures though sequence identity was 25.3% and 33.2% respectively (Fig. 5.8).

Since, we have identified 14 distinct clades within exonuclease III protein sequences based phylogenetic tree, the structural evolution of exonuclease III protein homologs are examined through constructing structural models of representative protein homologs of each clade. As experimental exonuclease III structures belong to clade 1, clade 9 and clade 11, the representative three dimensional structures of other clades are constructed through homology modelling using exonuclease III structure of *E.coli* as template.

The summary of each model is shown in table 5.5. Interestingly, though the sequence identities among all sequences are within the range 27-32%, the qualities of model structures are reasonably good indicating all homologs belong to same fold family. The RMSD values of

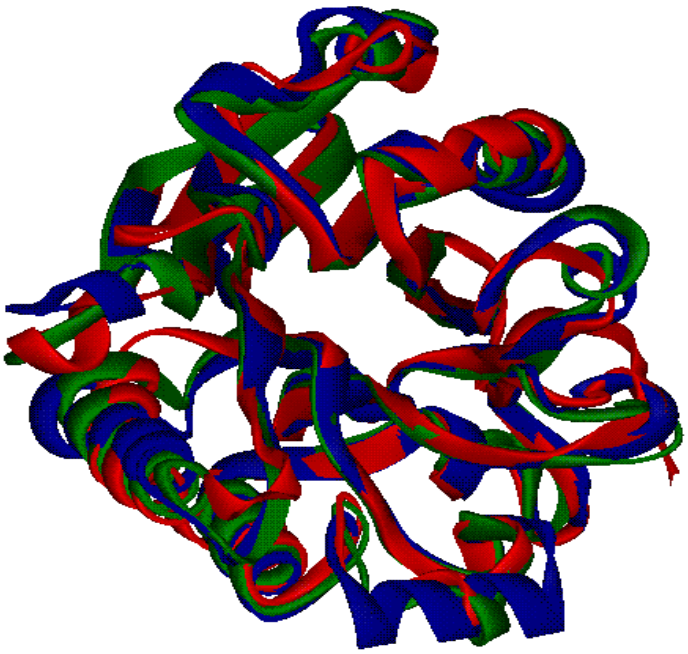


Fig. 5.8 Superimposed structure of exonuclease III in *E.coli* (Blue) against Hap1 in *Human* (Red) and AP endonuclease in *Archaeoglobus fulgidus* (Green). The Cα rmsd values are 1.14 Å and 1.13 Å respectively. Protein main-chains are shown by ribbon representation. 3D diagrams of these models *were* generated by Accelrys Discovery studio ViewerPro 5.0.

modeled structures with their template structure are in the range of 0.85-1.52 Å as listed in table 5.5. 3D model structure of representative exonuclease III homologs from each clade are also validated by Ramachandran plot which shows that these protein structures are stereo chemically stable as maximum of 1.6% amino acids are outliers, while rest 98.4% amino acids are within favored regions of Ramachandran plot.

The quality of predicted 3D models is checked by ERRAT plot and overall quality factor are better than 68 (shown in Table 5.5). The qualities of model structures are also validated through Qmean server. The QMEAN scores and QMEAN Z-scores also indicate that the qualities of models are reasonably good.

Table 5.5 Predicted model quality of exonuclease III homologs from each of the representative clade for which 3D model were generated.

Species Name (Clade No.)	**Identity with template**	**RMS (in Å) with respect to template structure**	**Ramachandran Plot Statistics**	**Errat 2 Score**	**Qmean Score/Z-Score**
Ruminococcus (Clade no. 2)	28%	C alpha- 1.39 Mainchain- 1.39	Favored - 94.2% Allowed- 5% Outlier- 0.8%	74.90	0.615 (Z-score: -1.65)
Chthoniobacter (Clade no. 3)	29%	C alpha- 0.85 Mainchain- 0.90	Favored - 95.1% Allowed- 4.0% Outlier- 0.9%	75.0	0.645 (Z-score: -1.33)
Ferroplasma (Clade 4)	31%	C alpha- 1.49 Mainchain- 1.52	Favored- 96.0%) Allowed- 3.6% Outlier- 0.4 %	72.31	0.651 (Z-score: -1.27)
Elusimicrobium (Clade 5)	30%	C alpha- 1.34 Mainchain- 1.34	Favored- 96.3% Allowed- 2.9% Outlier- 0.8%	68.15	0.600 (Z-score: -1.81)
Actinosynnema (Clade 6)	32%	C alpha- 1.08 Mainchain- 1.09	Favored- 94.6% Allowed- 5.0% Outlier- 0.4%	77.95	0.583 (Z-score: -2.02)
Chlorobium (Clade 7)	28%	C alpha- 1.14 Mainchain- 1.20	Favored- 94.7% Allowed- 4.5% Outlier- 0.8%	84.17	0.622 (Z-score: -1.57)
Human (Ape2) (Clade 8)	27%	C alpha- 1.20 Mainchain- 1.22	Favored- 95.8% Allowed- 3.6% Outlier- 0.6%	76.89	0.498 (Z-score: -3.12)
Hydrogenobacter (Clade 10)	31%	C alpha- 0.89 Mainchain- 0.93	Favored- 95.3% Allowed- 3.5% Outlier- 1.2%	86.29	0.671 (Z-score: -1.05)
Anabaena (Clade 12)	32%	C alpha- 1.34 Mainchain- 1.36	Favored- 97.6% Allowed- 2.9%) Outlier- 0.8%)	71.43	0.676 (Z-score: -0.99)

Renibacterium (Clade 13)	32%	C alpha- 0.85 Mainchain- 0.88	Favored- 95.8% Allowed- 3.1% Outlier- 1.1%	90.62	0.678 (Z-score: -0.99)
Candidatus (Clade 14)	32%	C alpha- 0.97 Mainchain- 1.00	Favored- 96.1% Allowed- 2.4% Outlier- 1.6%	71.08	0.679 (Z-score: -0.96)

The modeled structures of exonuclease III protein homologs in *Chthoniobacter*, *Hydrogenobacter*, *Renibacterium* and *Candidatus* are very close to crystal structure (PDB ID: 1AKO). The rms deviations of these four structures (with respect to *E.coli* crystal structure) are below 1 Å (Table 5.5). The rms deviation between the model of Ape2 structure (Human) and *E. coli* exonuclease III structure (crystal structure) is 1.20 Å. The superimposition (Fig. 5.9) of both the structures revealed that the orientation of three inserted loops regions. However, the overall fold of Human Ape2 remains same as that of exonuclease III structure.

The modeled three dimensional structure of *Ferroplasma* shows maximum deviation from the crystal structure where region around C-terminal shows maximum deviation (Fig. 5.10).

5.5 A model for the evolutionary history of exonuclease III protein family

The evolutionary history of exonuclease III family was proposed on the basis of sequence and phylogenetic tree analysis (Fig. 5.11). The in-silico phylogeny analysis suggested that during evolution process of exonuclease III gene, horizontal gene transfer (HGT) events occurred among the exonuclease III homologs of archaeal, bacterial and eukaryotic species and vice-versa as archaeal and bacterial species share the same clade as well as they also share clade with eukaryotic species in gene based phylogenetic tree.

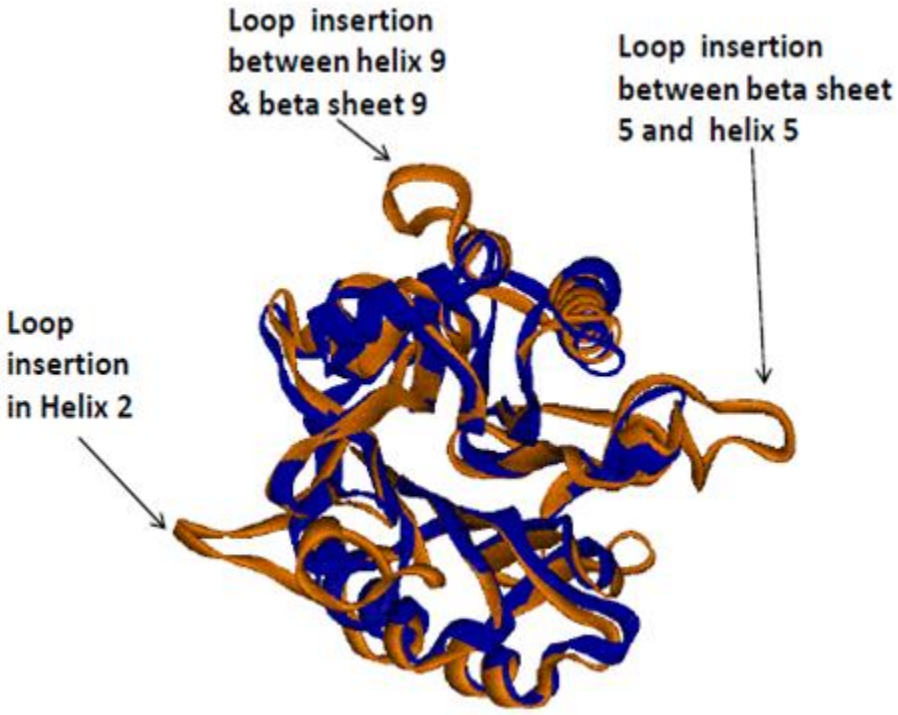


Fig. 5.9 Superimposed structure of Exonuclease III in *E.coli* (Blue) against Ape2 in Human (Brown colour). The Cα rmsd value is 1.20 Å. 3D diagram was generated by Accelrys Discovery studio ViewerPro 5.0 where protein main-chains are shown by ribbon representation.

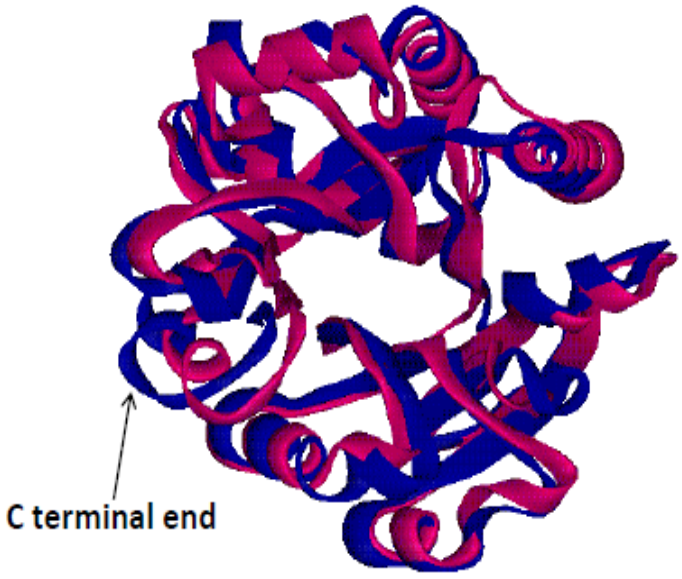


Fig. 5.10 Exonuclease III homolog of *E.coli* structure (Blue) is superimposed against modeled *Ferroplasma* (Pink) structure. The Çrmsd value is 1.49 Å.

Thus, the eukaryotic exonuclease III genes were likely originated from mixed population of different bacterial divisions as well as archaeal species by HGT.

In other evolutionary event, endosymbiotic transfer was observed as an important evolutionary event from bacteria to plant homologs. It is also well known, that mitochondria were incorporated from into eukaryotic cells from proteobacteria progenitors (Liu et al. 1993). Insertion of signal peptide at N terminal region and its targeting to mitochondria in Ape2 proteins of *Homo sapiens* and *Mus musculus* suggests the endosymbiotic transfer of exonuclease III genes from bacteria to eukaryotes during the evolution of eukaryotes.

In exonuclease III protein family, lower eukaryotes contain only one copy of AP endonuclease protein (homolog of exonuclease III) while in higher eukaryotes, like *Homo sapiens* and *Mus musculus,* two copies of AP endonuclease proteins, Ape1 and Ape2 are found which also have also overlapping function. Gene duplication event in higher eukaryotes provides the robustness of exonuclease III function.

5.6 Conclusions

This study provides an overall picture of the evolutionary history of exonuclease III gene/protein family that plays a crucial role in base excision repair of DNA. Based on conservation of amino acid positions, four consensus sequences have been identified among various regions of importance. Two new motifs were also found in exonuclease III gene/protein family homologs. Horizontal gene transfer (HGT) events occurred in the evolution process of exonuclease III gene among the species of bacteria, archaea and eukaryotes species and vice-versa. Endosymbiotic

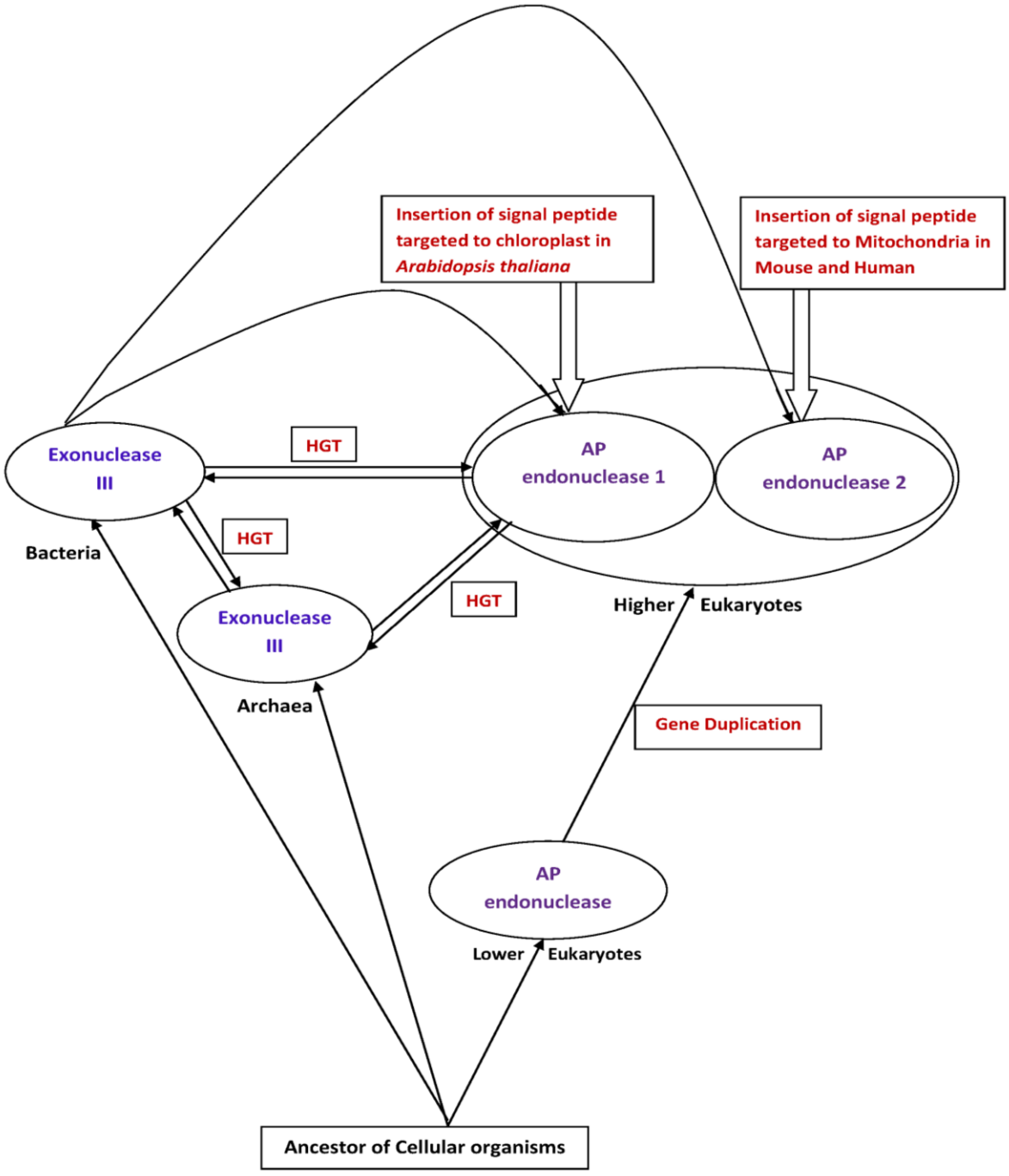


Fig. 5.11 A model for the evolutionary history of the exonuclease III Protein family is schematically shown. The eukaryotic exonuclease III genes were likely originated from mixed population of different bacterial divisions and archaea by HGT. Eukaryotic exonuclease III in plants were found to have insertion at N terminal which are targeted to chloroplast. Eukaryotic exonuclease III homologs Ape2 in *Mus musculus* and Human were found to have insertion at N terminal which are targeted to mitochondria. Gene duplication is also found in exonuclease III protein family during the course of evolution in eukaryotes.

events are also observed in plants (*Arabidopsis thaliana*) which were targeted to chloroplast and in Ape2 (*Mus musculus* and human) which were targeted to mitochondria. Gene duplication is also found in exonuclease III protein family during the course of evolution in higher eukaryotes. This evolutionary analysis may be exploited to understand the functions of uncharacterized genes in various species.

Chapter VI

Conclusion and Future perspective

Chapter VI

Biological evolution studies are one of the most challenging, unexplored areas of biological research. Evolution is the phenomenon which can be defined in simple terms as descent with modification. Evolution can also be defined as the gradual process in which simpler forms of life changes into a different and usually more complex form. This study is mainly focused on evolution of DNA repair proteins which restore the normal nucleotide sequence and DNA structure after damage. In spite of the critical need for DNA repair, ‘‘evolvability’’ i.e., the ability to generate a certain level of mutations also selected during the course of evolution where organisms with an optimal level of evolvability have the best chance to survive due to variation in its genome, which provides the raw material for natural selection. In most of the cases, evolution studies of proteins of interest are done in small groups of bacteria/archaea/eukaryotes i.e. the unified mechanism in all the three domains of life (Bacteria, Archaea and Eukaryotes) is still unexplored.

Among the different repair mechanisms, BER is one of the important pathways for repair of DNA lesions. AP endonuclease proteins are one of key role players in this pathway. AP endonuclease protein family mainly includes endonuclease III, endonuclease IV and exonuclease III proteins families. In the present study, the unified evolution mechanisms of endonuclease III, endonuclease IV and exonuclease III proteins in all the three domains of life are investigated. In addition to that we have analyzed the sequences of MGMT, XPD protein, G/T mismatch specific DNA glycosylases and MutS proteins from Direct repair, Nucleotide excision repair pathway, Base excision repair pathway and Mismatch repair pathway in six lineages *Escherichia coli (E.coli)*, *Pyrococcus kodakaraensis*, *Saccharomyces cerevisae (S.cerevisae)*, *Drosophila melanogester (D.melanogester)*, *Mus musculus* and *Homo sapiens.* These sequences were investigated to understand the unified evolution mechanism of different repair systems.

Searching of homologous sequences from databases based on sequence similarity is one of the traditional methods used in evolution of gene/protein families. In this study, homologous sequences from databases were further investigated for conservation and interaction of various residues for DNA binding motifs like HhH, FCL, minor groove binding and metal binding motifs which is the key to the overall architecture of thereof DNA binding and observed that theses helped in positioning the catalytically important residue towards the active site cavity, stabilizing the protein fold as well as participating in directly catalysis. Insertion of new DNA binding motifs in few homologs of AP endonuclease protein family is also observed in AP endonuclease family.

Gene loss, gene duplication, Horizontal gene transfer (HGT) Endosymbiotic transfer of AP endonuclease were observed as important evolutionary event in AP endonuclease protein family homologs. Absence of AP endonuclease (endonuclease IV) homologs in higher eukaryotes (beyond nematode) suggests that function of endonuclease IV is carried by homologs of exonuclease III protein (Hap1 in human) and possibly the endonuclease IV gene has been lost in higher eukaryotes during the course of evolution. In higher organisms, presence of two homologs of exonuclease III, called Ape1(Hap1) and Ape2 have been found in *Homo sapiens* and *Mus musculus* suggested the gene duplication in AP endonuclease family. MGMT, XPD protein, G/T mismatch specific DNA glycosylases and MutS proteins from Direct repair, Nucleotide excision repair pathway, Base excision repair pathway and Mismatch repair pathway in six lineages *Escherichia coli (E. coli)*, *Pyrococcus kodakaraensis*, *Saccharomyces cerevisae (S. cerevisae)*, *Drosophila melanogester (D. melanogester)*, *Mus musculus* and *Homo sapiens* were investigated. It is interesting to note that the AP endonuclease genes of bacterial species stay together with archaeal species frequently in gene based phylogenetic trees while in few cases stay close to those of eukaryotic species (exonuclease III), suggest a probable horizontal gene transfer (HGT)

process during the evolution of AP endonuclease family. Moreover, Endosymbiotic transfer events were also investigated in AP endonuclease family where genes from bacterial species to eukaryotes have taken place by incorporation of mitochondria into eukaryotic cells from proteobacteria progenitors. We have also predicted the insertion of signal peptide at N terminal region and their targeting to chloroplast and mitochondria in AP endonuclease protein seems to have taken place during the evolution of plant and eukaryotic homologs.

We also observed that the evolution of species and AP endonuclease genes shape up differently in most of the cases because in AP endonuclease gene based tree, species from different lineages share the same clade. The average G+C content and G+C content at the third codon position of AP endonuclease gene could possibly explain why these species from different lineages share the same clade. Species having similar average G+C content and G+C content at the third codon were biased to stay together in a clade. This also suggests that G+C content of gene contribute significantly towards the position of taxa within the tree and evolution.

To map the evolutionary changes among various clusters in protein based phylogenetic tree we generated the homology models of representative AP endonuclease protein from each cluster. These models were refined and validated by superimposition of target and template structures, Ramachandran statistics, Errat score, Qmean and Z-Scores which suggested the conservation of overall fold of AP endonuclease proteins. But in few cases, minor distortions in fold structure were identified mainly due to insertion/deletion of few residues which might affect the DNA binding and the catalytic residues in active site pocket which needs to be verified by experiments.

We observed that in spite of large sequence variation within MGMT, XPD, G/T mismatch specific DNA glycosylases and MutS proteins across different model organisms, the main

functional domains with important catalytic residues are conserved during evolution. During the course of evolution, the duplicated six MUTS paralogs in eukaryotic organisms designated as MSH1–MSH6 have different combinations of domains and different combinations of paralogs (heteromers) which takes part in various DNA mismatch repair function.

Crosstalk among the proteins of direct repair, nucleotide excision repair, base excision repair and mismatch repair pathway were also observed. NER pathway often interacts with proteins of MMR and BER pathways and physical interaction of BER protein with NER proteins significantly stimulates its activity. Crosstalk between MMR and NER proteins modulate and boost the repair mechanism and cross talk between DR proteins and NER protein is rather complex as most cases DR mechanism does not require cleavage of phosphodiester bond. Overall, DNA repair process in four major repair pathways mainly involves either repair of modified bases (DR) or repair of DNA lesions by removing the damaged base (MMR, BER) followed by cleavage of phosphodiester bond (MMR, BER and NER).

Although we have addressed the filtration of redundant proteins obtained during homology search based on their sequence identity and presence of at least one domain, which is common among all. The sequence identity can be increased or decreased by manually checking so that true homologs should not miss. In multidomain protein family, rather than keeping the presence of one domain among all homologs, more number of domains can be used for exploring the true homologs in many unannotated/hypothetical gene/protein family members. Insertion/deletion of domains/motifs as well as duplication/loss of genes during evolution are one of the important evolutionary events which could be useful for exploring the requirement of new function or loss of existing function due to environmental pressure in especially in particular domain of life, which could be useful in exploring many unannotated

gene/protein families. Moreover, we have addressed the evolution of species based on phylogenetic analysis of 16S and 18S rRNA gene sequences in prokaryotes as well in eukaryotes respectively which could be further compared to evolution based genes and protein sequences to map the differences in evolutionary pattern of genes/proteins in different species for many unannotated gene/protein families. Differences between evolutionary patterns could be also investigated by GC content in genes in many unexplored/annotated gene/protein families. Crosstalk among the proteins of various repair pathways could be more helpful in understanding the interaction among these repair proteins and understand the complex DNA repair process in all the three domains of life.

Appendix

Evolutionary study of four representative DNA repair proteins among six model organisms

A.1 Introduction

The molecular mechanisms of DNA repair can be classified into four main categories: direct repair (DR), nucleotide excision repair (NER), base excision repair (BER), and mismatch repair (MMR). O6-methyl guanine alkyltransferase (MGMT), Xeroderma pigmentosum group D (XPD) protein, G/T mismatch-specific DNA glycosylases, and MutS proteins are class of proteins that involved in DR, NER, BER, and MMR pathways, respectively. MGMT catalyzes transfer of methyl groups from O6-methylguanine and other methylated moieties of the DNA to its own molecule, thereby repairing the toxic lesions in a single-step reaction (Pegg 2000). If this damage is not repaired, G:C pair can convert into A:T pair (Jane et al. 2000). Ada protein (product of the *ada* gene) and the Ogt protein (product of the *ogt* gene) are two types of DNA methyl transferases present in *E.coli*. Although both O6-methyl guanine and O4-methyl thymine are substrates of *E.coli* alkyltransferase, Ogt repairs O4-methyl thymine more efficiently than Ada (Pegg 2000). Eukaryotic alkyltransferase repairs O6-methyl guanine more efficiently than O4-methyl thymine, thus, resembles close to Ada protein. XPD is one of the subunits of TFIIH protein complex which is involved in RNA polymerase II transcription and nucleotide excision repair. It has been shown that mutations in XPD affects DNA repair which result in the skin cancer disorder, Xeroderma pigmentosum (XP) (Winkler et al. 2000). It is observed that helicase activity of 80 kD XPD subunit is required for the formation of both the 5’ and 3’ incisions around a site of DNA damage (Winkler et al.2000). G/T mismatch-specific DNA glycosylase protein belongs to the uracil-DNA glycosylase superfamily which removes thymine/Uracil moieties from G/T or G/U mismatches by hydrolyzing the carbon–nitrogen bond between the sugar-phosphate backbone of DNA and the mispaired thymine/uracil (Hardeland et al. 2001). MutS protein recognizes and binds to a mismatch or insertion–deletion loop (IDL) within DNA molecule. Almost all mismatches are recognized and repaired by MutS. However, there is some variation in the

affinity and efficiency of MutS protein which depends on the nature of the mismatch and sequence of DNA moiety (Lamers et al. 2000). Due to its similarity in function, DNA repair systems are highly conserved among different species, ranging from prokaryotes to eukaryotes (Dmitry & Kosuke 1997). In spite of this conservation, insertion/deletion of domains and motifs play critical roles in evolution of these proteins. In this study, we examined the evolution of various domains and motifs of MGMT, XPD, G/T mismatch-specific DNA glycosylases, and MutS proteins across six different model organisms belonging to different lineages of life. Motivation for this study came from the general idea that historically the requirement for DNA repair may have been greater than now due to environmental conditions that as a result there may have been some degeneration of repair mechanisms.

A.2 Material and methods

A.2.1 Sequence retrieval and data curation

MGMT, XPD, G/T mismatch specific thymine DNA glycosylase and MutS proteins from *Escherichia coli (E.coli)*, *Pyrococcus kodakaraensis*, *Saccharomyces cerevisae (S. cerevisae)*, *Drosophila melanogaster (D.melanogaster)*, *Mus musculus* and *Homo sapiens* are retrieved from whole genome database of respective organism (http://www.ncbi.nlm.nih.gov/genome). If multiple entry of same protein in an organism exists in the database, we have selected the protein sequence with largest sequence length. The details of retrieved proteins with their sequence length are tabulated in Table A.1.

The conserved domains and motifs in these proteins were analyzed by conserved domain (CD) search (Marchler-Bauer & Bryant 2004). The domains having the least expected value

were considered as potent domains within a protein. MGMT, XPD, G/T mismatch specific thymine DNA glycosylase and MutS proteins from six model organisms were compared using multiple sequences alignment tool of ClustalW (Chenna et al. 2003) (www.ebi.ac.uk/Tools/msa/clustalw2/).

Table A.1 List of retrieved proteins selected for this study with their accession numbers is shown. Sequence length of each protein is shown within parenthesis.

Name of the organism	**Accession no. of XPD proteins (Sequence length)**	**Accession no. of G/T mismatch specific proteins (Sequence length)**	**Accession no. of MGMT proteins (Sequence length)**	**Accession no. of MutS Protein (Sequence length)**
Pyrococcus kodakarensis	YP_183197.1 (637)	YP_184556.1 (196)	YP_184384.1 (174)	YP_183095.1 (576)
Escherichia coli	NP_415320.1 (716)	NP_417540.1 (168)	NP_415851.1 (354)	NP_417213.1 (853)
Saccharomyce cerevisae	NP_011098.3 (778)	NP_013691.1 (359)	NP_010081.2 (188)	NP_011988.1 (MSH1) (959)
				NP_014551.1 (MSH2) (964)
				NP_010016.2 (MSH3) (1018)
				NP_116652.1 (MSH4) (878)
				NP_010127.1 (MSH5) (901)
				NP_010382.3 (MSH6) (1242)
Drosophila melanogester	NP_726036.2 (769)	NP_001259060.1 (1738)	NP_477366.1 (194)	NP_523565.2 (MSH2) (917)
				NP_648755.1 (MSH6) (1190)
Mus musculus	NP_031975.2 (760)	NP_766140.2 (491)	NP_032624.1 (211)	NP_032654.1 (MSH2) (935)
				NP_034959.2 (MSH3) (1095)
				NP_114076.1 (MSH4) (958)
				NP_038628.2 (MSH5) (833)
				NP_034960.1 (MSH6) (1358)
Homo sapiens	NP_000391.1 (760)	NP_003202.3 (410)	AAA59596.1	NP_000242.1 (MSH2) (934)
				NP_002430.3 (MSH3) (1137)

			(207)	NP_002431.2 (MSH4) (936)
				NP_002432.1 (MSH5) (834)
				NP_000170.1 (MSH6) (1360)

The evolutionary history was inferred through the maximum likelihood (ML) tree. The evolutionary distances were computed using the JTT matrix-based method (Jones et al. 1992). The branch support of the ML phylogenetic tree was estimated using a bootstrap test with 1000 replicates (Felsenstein 1985). Evolutionary analyses were conducted In case of gaps or missing data in sequences, partial deletion of the sequences parameters was chosen. Other parameters were kept as default.

A.2.2 Conservation patterns

The conservation patterns of protein structure in MGMT, XPD and G/T mismatch specific thymine DNA glycosylase proteins were analyzed using ConSurf server (Ashkenazy et al. 2010). The conservation scores at each amino acid position were calculated using the same server. We have calculated the evolutionary conservation of amino acid positions in proteins using an empirical bayesian inference starting from protein structure and sequence. The conservation analysis of ConSurf used the evolutionary conservation scores of the residues to identify functional regions from proteins with known three-dimensional structures. The conservation of the various residues was projected onto the molecular surface of the 3D structure of proteins to reveal the patches with highly conserved residues which are important for biological function.

A.3 Results and Discussion

A.3.1 O6-methyl guanine alkyltransferase

The sequence size of MGMT proteins of model organisms (*Escherichia coli*, *Pyrococcus kodakaraensis*, *Saccharomyces cerevisae*, *Drosophila melanogester*, *Mus musculus* and *Homo sapiens*) considered in the present study varied from 174 to 354 amino acid residues. Due to the insertion of around 130 residues at N-terminal end, the MGMT protein of *E.coli* is significantly larger than the MGMT proteins of other organisms. Pairwise as well as multiple sequence analysis of the MGMT proteins of all model organisms have revealed that the protein sequence of *Pyrococcus kodakaraensis* is maximally different from the *E.coli* protein (28.6% identity) whereas the proteins of *Homo sapience* and *Mus musculus* are most similar (68.9% identity). The MGMT protein of *E.coli* is least similar to that of any other five organisms. However it is surprising to note that, MGMT protein of *Pyrococcus kodakaraensis* is more similar to MGMT protein of *Mus musculus* than that of any other organisms. Detailed sequence analysis of O6-methyl guanine alkyltransferase from *Pyrococcus kodakaraensis to Homo sapiens* has indicated fare amount of sequence conservation within the C-terminal ATase/Ogt domain. Multiple sequence alignment of the conserved C-terminal ATase/Ogt domain is shown in Fig. A.1.

It contains conserved active-site cysteine motif (PCHR), O6-alkyl guanine binding channel and the helix-turn-helix (HTH) DNA binding motif. ATase/Ogt domain is responsible for reversing O6-alkylation DNA damage by transferring O6-alkyl adducts to an active site cysteine, without inducing DNA strand breaks (Madeleine et al. 1994; Jane et al. 2000). This domain contains all the residues that are necessary for DNA binding as well as catalytic alkyl transfer activity and hence could be considered as the main functional unit of protein which alkyl transfer activity and hence could be considered as the main functional unit of protein

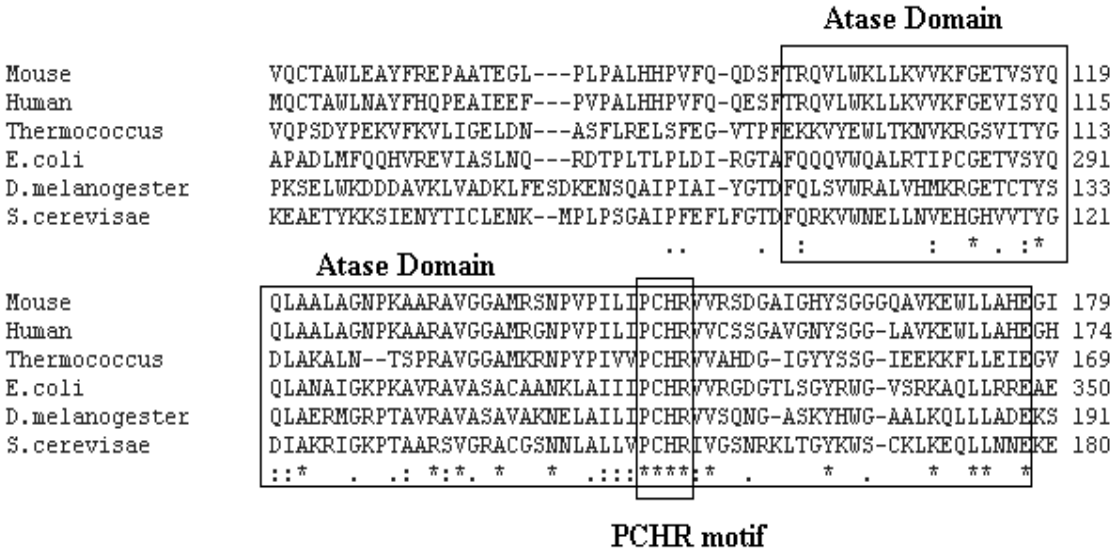


Fig. A.1 Conserved ATase/Ogt domain as well as the PCHR motif is shown in multiple sequence alignment of MGMT protein.

which is evident from conservation of domain as well as crucial residues across six lineages of life.

The conservation pattern of amino acid residues of MGMT proteins are mapped on the 3D structure of Human MGMT protein (PDB ID: 1EH6) and are analyzed using ConSurf server. The ConSurf result has revealed that catalytic residues and residues in the vicinity of catalytic pocket e.g. P124, R128, A129, N137, C145, H146, Y158 and E172 are highly conserved (Fig. A.2).

Other conserved residues are within the core of protein structure and are helped in formation of hydrophobic core. The details of all domains within MGMT protein are mentioned in Table A.2. N terminal, Methyltransf_1N [pfam02870] domain is other important domain which is a ribonuclease-like domain with nearly 78-91 amino acid residues. This N-terminal domain of the human MGMT protein contains a bound zinc atom which is however not essential for its repair activity, but its presence increases the rate of DNA repair

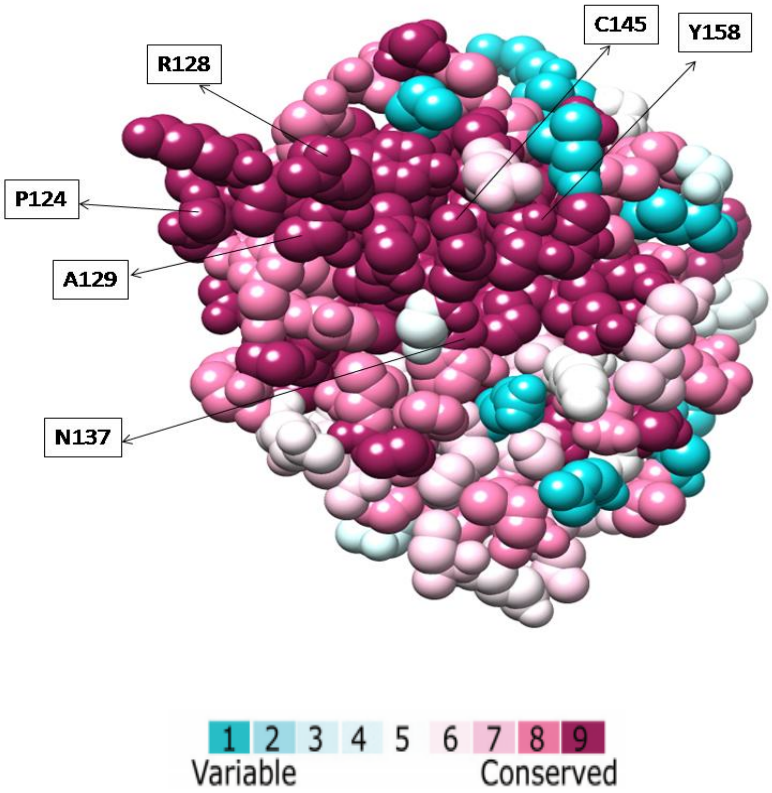


Fig. A.2 Conservation pattern of catalytic residues and residues in the vicinity of catalytic pocket which are P124, R128, A129, N137, C145 and Y158, showed on the 3D structure of AGT Human (PDB ID: 1EH6). Amino acid conservation scores were classified into 9 levels. Color code 1 shows the highly variable while color code 9 shows the highly conserved amino acid residue. The color scale for residue conservation is indicated in the figure.

(Rasimas et al. 2003). N- terminal domain plays a critical structural role in maintaining the C-terminal domain in an active configuration. It is interesting to note that the N-terminal domain alone had a weak AGT (alkyl guanine transferase) activity in vitro that was totally zinc dependent (Fang et al. 2005). Domain search shows that this domain is present in all model organisms except *Drosophila melanogester*. Our analysis has also identified helix-turn-helix AraC domain and Ada_Zn_binding domains within MGMT protein of *E.coli* which is deleted during course of evolution. These domains mainly increase specificity of DNA binding and hence do not take part in direct repair mechanism. Thus, during the process

Table A.2 List of domains in MGMT proteins, XPD proteins and G/T mismatch specific DNA glycosylases protein covering six different lineages of life. '+' sign indicates the presence of particular domain in individual lineage of life. The domain length is given within parenthesis.

Protein Name	**Domains (Accession no.) [Length range]**	***Pyrococcus kodakaraensis***	***E. coli***	***S. cerevisae***	***Drosophila melanogester***	***Mus musculus***	***Homo sapiens***
MGMT Protein	ATase [cd06445] [65-79]	+	+	+	+	+	+
	Methyltransf_1N[pfam02870] [78-91]	+	+	+		+	+
	Ada_Zn_binding [pfam02805] [65]		+				
	HTH_AraC [pfam00165] [32]		+				
XPD Protein	DEAD_2 [pfam06733] [63-186]	+	+	+	+	+	+
	Helicase_C_2 [pfam13307] [162-176]	+	+	+	+	+	+
	DUF1227 (pfam06777) [145-146]			+	+	+	+
G/T mismatch specific DNA glycosylases protein	UDG Family 2 [cd10028] [157-170]		+		+	+	+
	UDG Family 4 [cd10030] [171]	+					
	UDG Family 1 [cd10027] [217]			+			

of evolution deletion of these domains does not affect DNA repair process.

Phylogenetic tree reconstruction of MGMT proteins in six organisms shows (Fig. A.3) the formation of two clusters. In first cluster, mouse and human are found closer to *Pyrococcus*

kodakaraensis with high bootstrap support which suggests that MGMT proteins in mouse and human may have archaeal origin. In the second cluster, *Saccharomyces cerevisae* and *Drosophila melanogester* were placed with *E.coli* which is supported by high bootstrap values. This observation suggests that the MGMT protein of *D.melanogester* may have bacterial origin.

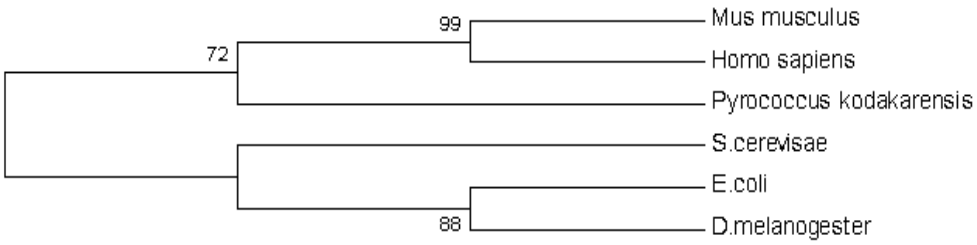


Fig. A.3 Maximum Likelihood tree of MGMT proteins based on the JTT matrix-based model. The numbers indicates the bootstrap support values. The tree suggests archaeal origin of MGMT protein of human and mouse

Direct DNA repair mechanism is a single step lesion process in which MGMT catalyzes transfer of methyl groups from O6-methylguanine as well as other methylated moieties of the DNA to its own molecule. In this repairing process, identifying the DNA defect, binding of MGMT protein to DNA molecule and transfer of methyl group within DNA base to its own Cys residues are the crucial steps.

Our analysis has shown that domain and residues participating in these steps are conserved throughout the evolution process. In MGMT protein, catalytic ATase domain is conserved in all the homologs whereas Methyltransf_1N, Ada_Zn_binding and HTH_AraC domains are present only in certain lineages of life which mainly alters rate of DNA repair, substrate specificity and binding efficiency of the enzyme.

A.3.2 Xeroderma Pigmentosum group-D (XPD) protein

Sequence analysis shows that damage inducible G (DinG) protein in *E.coli*, Rad 3 in *Pyrococcus kodakaraensis* and *Saccharomyces cerevisae*, ERCC2 in *Mus musculus* and *Homo sapiens* are the closest homolog of XPD group protein. In these proteins, sequence length varies from 637 to 778 amino acid residues which indicate that quite a significant amount of insertion/deletion is taken place during the evolution of this protein. Pairwise and multiple sequence analysis has clearly demonstrated that proteins from *E.coli* and *Pyrococcus kodakaraensis* are distinctly diverged from that of higher organisms including *Saccharomyces cerevisae*. In case of eukaryotic organisms, it has been found that DUF1227 [pfam06777], an uncharacterized domain is inserted in between N-terminal DEAD_2 [smart00489] domain (Fig. A.4(A) and C-terminal Helicase_C_2 [smart00492] (Fig. A.4(B) domain during the course of evolution (Table A.2) which might adopts a novel fold. The crystal structure of XPD related protein from *T. acidophilum* is divided into four distinct domains (Wolski et al. 2008). Multiple as well as pairwise sequence analysis of XPD protein of eukaryotic and prokaryotic organisms have revealed that DUF1227 region is aligned well with domain 3 of XPD protein of *T. acidophilum* with significant sequence similarity. This domain majorly stabilizes domains at N-terminal and C-terminal ends. During the course of evolution, a large sequence variation is observed within this domain. N-terminal DEAD_2 domain which is also called motor domain 1, is involved in DNA helicase activity. These proteins utilize its helicase activity for damage verification. Helicase_C_2 domain at the C-terminal end also belongs to helicase superfamily which is present within a wide variety of helicase and helicase related proteins that contains composite ATP binding site (Wolski et al. 2008; Singleton 2007). Structural mapping of conserved residues on *Sulfolobus acidocaldarius* crystal structure (PDB ID: 3CRW) revealed that residues like Pro43, Lys48, Thr76, Pro191, Asp234, Glu235, His237, Asn238, Thr460, Val618, Gln662,

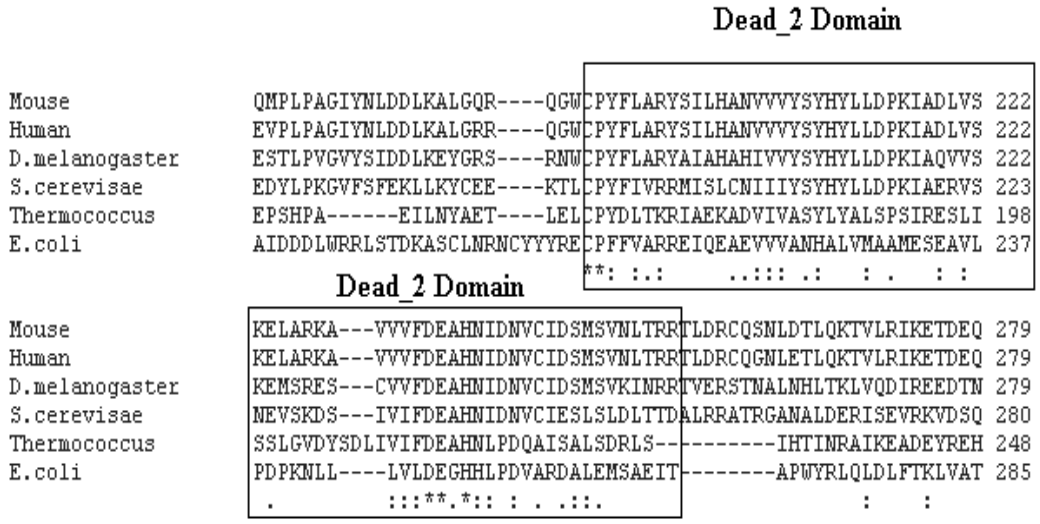


Fig. A.4 (A) Multiple sequence alignment of Xpd proteins where conserved. Dead_2 domain is shown within box.

Arg666 and Arg669 played a key role in structural integrity of protein and the residues like Asp513, Ser541, Glu606, Ala656, Asp681, and Arg683 in human are present within the catalytic pocket which either bind with DNA or helps in DNA binding are found conserved throughout the evolution process (Fig. A.5).

Phylogenetic tree of Xpd proteins (Fig. A.6) demonstrated that this protein evolved progressively from bacteria to eukaryotes where mouse and human Xpd proteins are close to each other.

The insertion of large domain in between two functionally important domains with large number of variable residues makes eukaryotic organisms separated from prokaryotic organism in phylogenetic tree which also indicates that the structure and detail mode of action of Xpd proteins of eukaryotes and prokaryotes could be different. Nucleotide excision repair is performed by different proteins in different lineages of life, *Xeroderma pigmentosum* group D protein is one of NER DNA repair protein. The functional DEAD_2 and HELICc2

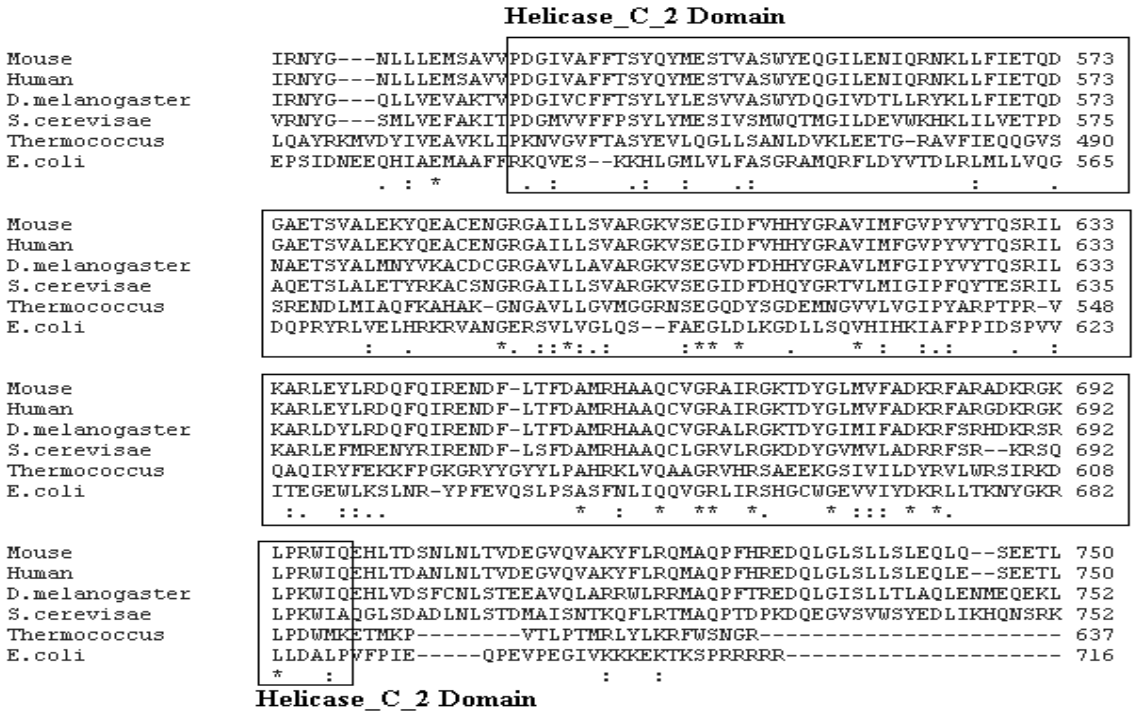


Fig. A.4 (B) Multiple sequence alignment of Xpd proteins where conserved Helicase_C_2 domain is shown within box.

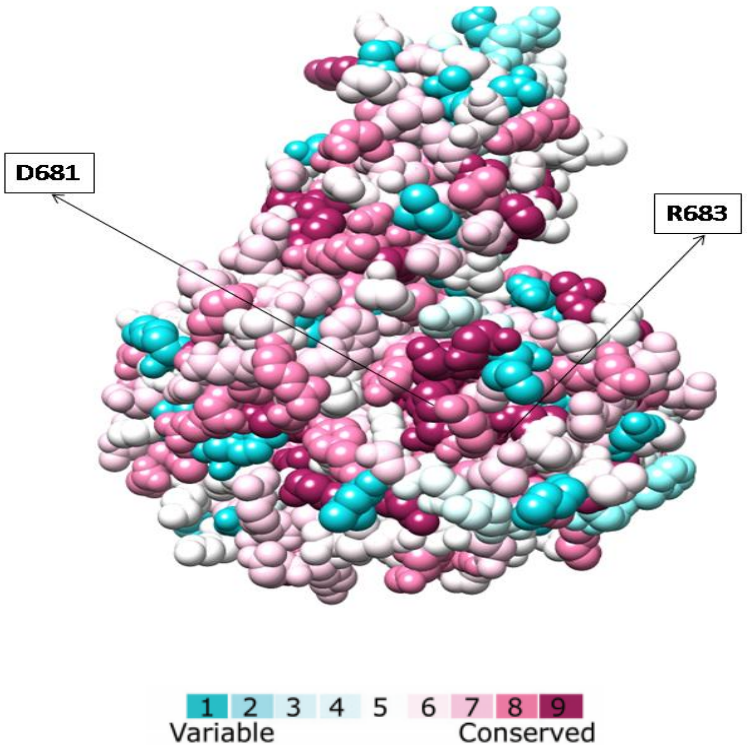


Fig. A.5 Conservation pattern of amino acids where D681 and R531 are present within the catalytic pocket which either bind with DNA or helps in DNA binding are shown on 3D structure of TDG Human (PDB ID: 3UO7). Amino acid conservation scores and color coding was done as in Fig. A.2.

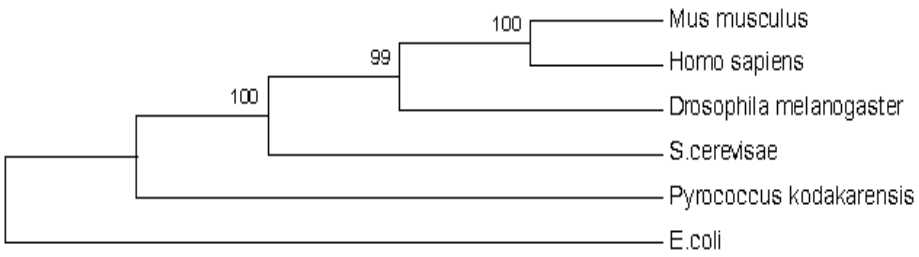


Fig. A.6 The Maximum Likelihood (ML) evolutionary tree of XPD proteins shows progressive evolution of this protein with human and mouse protein being close together. The numbers indicate the bootstrap support values.

domains of this group of proteins belong to helicase super family. In between N-terminal and C-terminal functional domain, a domain with around 146 amino acid is inserted in eukaryotic organisms whose precise function is still unknown suggesting this region might represents a novel fold and inserted during course of evolution

A.3.3 G/T mismatch specific thymine DNA glycosylase protein

G/T mismatch specific DNA glycosylases protein belongs to monofunctional uracil DNA glycosylases (UDGs) superfamily that share a common alpha/beta fold structure. The size of the protein varies from 168 to 1738 amino acid residues with single functional domain (Table A.2). These proteins have simple domain architecture of a conserved core that constitutes the active site and non-conserved N and C-terminal extensions of variable lengths. This enzyme plays a central role in cellular defense against genetic mutation caused by the spontaneous deamination of 5-methylcytosine and cytosine. Large variation in protein size is due to the fact that G/T mismatch specific DNA glycosylases of different organisms belongs to different UDG families (UDG family 1, family 2 and family 4), which are mainly classified according to base excision repair mechanism (Pearl 2000). DNA glycosylase of *Saccharomyces cerevisae* belongs to UDG family 1 whereas that of *E. coli, Drosophila melanogester, Mus*

musculus and *Homo sapiens* belongs to family 2 and DNA glycosylase of *Pyrococcus kodakaraensis* is classified as family 4 UDG which generally contains DNA glycosylase of thermophilic organisms. However, catalytic mechanism of family 1 and family 4 UDG are not very different as enzymes of both families removes mismatched uracil from double stranded as well as single stranded DNA. This is also evident from pairwise sequence analysis of G/T mismatch repair proteins of all six organisms. Protein of family 1 UDG has maximum sequence identity (~30%) with that of family 4 UDG whereas only around 20% sequence identity is observed between family 1 and family 2 UDG proteins. It is observed that in comparison to family 1 or family 4, family 2 UDG proteins recognize wider range of substrates. Family 2 UDG can excise guanine mismatched with uracil, thymine or 3, N(4)-ethenocytosine (Hardeland 2001; Saparbaev & Lavel 1998) while, family 1 and family 4 UDG protein only recognize and repair guanine mismatched with uracil. The conservation pattern of G/T mismatch specific proteins is mapped on the 3D structure of Human G/T mismatch specific proteins (PDB ID: 3UO7) and it is noticed that the residues I139, N140, P141, G142, S200 and R275 are within the active site pocket and found as highly conserved throughout the evolution (Fig. A.7).

Phylogenetic tree of G/T mismatch specific proteins of six organism shows (Fig. A.8) that mouse, human and *D.melanogester* are placed closer to *E. coli* with good bootstrap support. This is expected as G/T mismatch specific proteins of these organisms belong to family 2 UDG. Moreover it confirms that family 2 UDG proteins have bacterial origin. G/T mismatch-specific thymine DNA glycosylase proteins of every lineages of life belong to uracil DNA glycosylases (UDGs) super family. Though this protein of all lineages contains single functional domain and during the course of evolution substrate specificities has broaden as verity of substrates are accommodated by the protein of higher organisms.

A.3.4 MutS/MutS1 group of proteins

Mismatched repair within an organism is explicitly performed by MutS family of proteins.

Protein sequence database search revealed that prokaryotic organisms *E.coli*

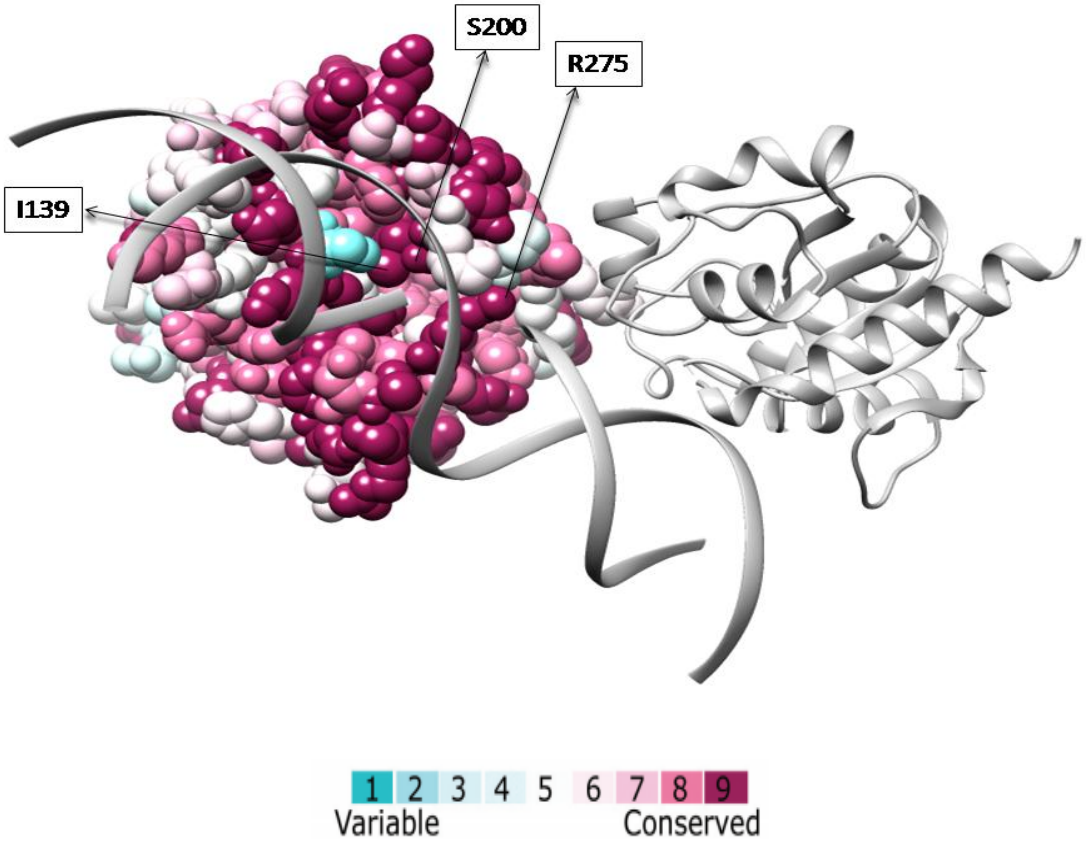


Fig A.7 Conservation pattern of amino acids where I139, S200 and R275 which are present in the active site cavity and are shown on 3D structure of TDG Human (PDB ID: 3UO7). Amino acid conservation scores and color coding was done as in Fig. A.2.

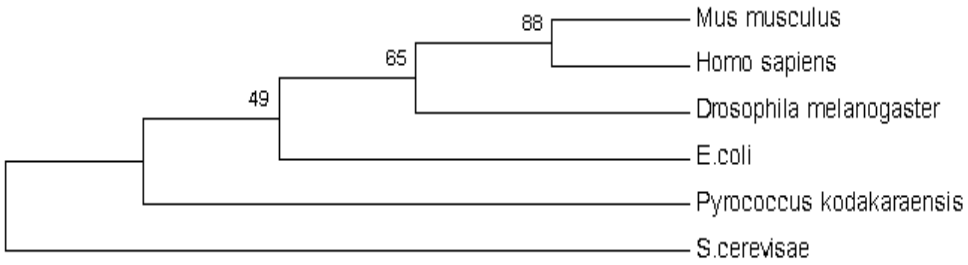


Fig. A.8 The evolutionary relationship of G/T mismatch specific glycosylase proteins is shown by constructing Maximum Likelihood tree. The numbers around branches indicate the bootstrap support values.

including archaea *Pyrococcus kodakaraensis* contains single MUTS protein whereas eukaryotic model organism possesses one or more MUTS paralogs (MSH1 to MSH6). Sequence length of Muts protein of *E.coli* and *Pyrococcus kodakaraensis* are 853 and 567 amino acid residues respectively with around 21% of sequence identity between them. Domain analysis shows that Muts of *E. coli* contains MutS_I, MutS_II, MutS_III and MutS_V/MutSac domain whereas that of *Pyrococcus kodakaraensis* contains only MutS_V/MutSac domain. Among the four eukaryotic model organisms, *S.cerevisae* contains all six (MSH1 to MSH6) MUTS paralogs whereas *D. melanogester* has MSH2 and MSH6. MUTS proteins of Mouse and human contain all MSH paralogs except MSH1. Among the MSH paralogs, MSH1, MSH2, MSH3 and MSH6 contain all four MUTS domains whereas MutS_I domain is absent in MSH4 and both MutS_I and MutS_II are absent in MSH5 paralogs (Table A.3).

Table A.3 List of domains present within MutS protein family in six organisms. *, #, @ and $ symbols indicate presence of MutS_I, MutS_II, MutS_III and MutS_V/MutSac domain respectively.

Name of the Organism	**Proteins Name**						
	MutS	MSH1	MSH2	MSH3	MSH4	MSH5	MSH6
Pyrococcus kodakaraensis	$						
E.coli	*#@$						
S.cerevisae		*#@$	*#@$	*#@$	#@$	@$	*#@$
D.melanogester			*#@$				*#@$
Mus Musculus			*#@$	*#@$	#@$	@$	*#@$
Homo Sapiens			*#@$	*#@$	#@$	@$	*#@$

The N-terminal, MutS_I [pfam01624] domain which is also called as mismatch binding domain with 115-125 amino acid residues is believed to interact closely with DNA mismatches (Dufner et al. 2000; Warren et al. 2007). Second domain is MutS_II [pfam05188] which is a connector domain which is involved in allosteric signaling mediate protein-protein interactions (Warren et al. 2007). Third domain is MutS_IV [pfam05190] is small clamp domain which is composed of largely beta strand. This domain is responsible for making contacts with nucleotide bases of DNA on both sides of the mispair (Warren et al. 2007). The last domain at C-terminal end is designated as MutS_V [pfam00488]/ ABC_MSH2_euk [cd03285]/ABC-ATPase/MutSac domain which is conserved throughout and present in bacterial, archaeal as well as eukaryotic paralogs (MSH1-MSH6). This domain contains helix–turn–helix and nucleotide/magnesium-binding (Walker box) sub-domains and is envisaged to interact with DNA and mediate ATP binding and hydrolysis (Allen et al. 1997; Fishel & Wilson 1997; Kolodner 1996; Eisen 1998).

It is worth mentioning that mismatch repair genes in eukaryotes perform both mismatch as well as non mismatch repair functions. MSH1, MSH2, MSH3, MSH6 in eukaryotes are involved in mismatch repair function while non mismatch repair genes/ proteins are MSH4 and MSH5. Different MUTS/MUTS1 and MSH1–MSH6 paralogs have different combinations of domains and these paralogs form different heteromers which performs different functions. In eukaryotes, the heterodimeric MSH2/MSH6 (MutS alpha) repairs short IDLs, whereas MSH2/MSH3 (MutS beta) repairs longer IDLs. The heterodimer of MSH4/MSH5 stimulate meiotic crossovers which is an example of non-mismatch repair function. Presence of all six *MSH* genes in multiple eukaryotic lineages suggests that they were generated by duplication of prokaryotic gene.

A.3.5 Crosstalk among proteins of four repair pathways

DNA repair process in four major repair pathways mainly involves either repair of modified bases (DR) or repair of DNA lesions by removing the damaged base (MMR, BER) followed by cleavage of phosphodiester bond (MMR, BER and NER). Thus, proteins of NER pathway often interact with proteins of MMR and BER pathways. In most cases, physical interaction of BER protein with NER proteins significantly stimulates its activity (Shimizu et al. 2003; Viswanathan & Doetsch 1998; Tornaletti et al. 2001). Similarly, crosstalk between MMR and NER proteins modulate and boost the repair mechanism (Bertrand et al. 1998). The cross talk between DR proteins and NER protein is rather complex as most cases DR mechanism does not require cleavage of phosphodiester bond. However, earlier studies (Tubbs et al. 2009; Mazon et al. 2009; Tubbs & Tainer 2010) have identified group of proteins that activate both base repair (DR) and NER pathways and initiate the cross talk between these two pathways. Functional interactions between proteins of BER and MMR pathways are frequently observed (Gellon et al. 2002; Millar et al. 2002; Rada et al. 2004; Kovtun & McMurray 2007; Bai & Lu 2007) as both pathways involve in repairing of damaged or mismatched based followed by breaking of phosphodiester bond. In some cases, evidences have suggested that proteins from both pathways might operate together to carry out repair process (Kovtun & McMurray 2007).

A.4 Conclusions

The sequence and domain analysis of O6-methyl guanine alkyltransferase (MGMT), *Xeroderma pigmentosum* group D (XPD) protein, G/T mismatch specific DNA glycosylases and MutS proteins demonstrated that in spite of large sequence variation within a protein across different model organisms, the main functional domains with important catalytic

residues are conserved during evolution. Our analysis suggested that MGMT protein involved in direct repair mechanism might follow retrogressive evolution whereas a progressive evolution may have taken place in case of *Xeroderma pigmentosum* group D and G/T mismatch-specific thymine DNA glycosylase protein. During the course of evolution, the *MUTS gene* is duplicated and exists as six MUTS paralogs in eukaryotic organisms. MUTS and paralogs of MUTS protein (designated as MSH1–MSH6) have different combinations of domains and these paralogs form different heteromers which takes part in various DNA mismatch repair function.

References

Aboussekhra, A., Biggerstaff, M., Shivji, MK., Vilpo, JA., Moncollin, V., Podust, VN., Protic, M., Hubscher, U., Egly, JM., Wood, RD. (1995). Mammalian DNA nucleotide excision repair reconstituted with purified protein components. *Cell*, 80, 859–868.

Adler, M., Anjum, M., Berg, OG., Andersson, DI., Sandegren, L. (2014). High fitness costs and instability of gene duplications reduce rates of evolution of new genes by duplication-divergence mechanisms. *Molecular Biology and Evolution*. 31(6), 1526-1535.

Akbari, M., Otterlei, M., PeÇa-Diaz, J., Aas, PA., Kavli, B., Liabakk, NB., Hagen, L., Imai, K., Durandy, A., Slupphaug, G. & Krokan, HE. (2004). Repair of U/G and U/A in DNA by UNG2-associated repair complexes takes place predominantly by short-patch repair both in proliferating and growtharrested cells. *Nucleic Acids Research*, 32, 5486 – 5498.

Allen, DJ., Makhov, A., Grilley, M., Taylor, J., Thresher, R., Modrich, P. & Griffith, JD. (1997). MutS mediates heteroduplex loop formation by a translocation mechanism. *The EMBO Journal*, 16, 4467–4476.

Altschul, SF., Gish, W., Miller, W., Myers, EW. & Lipman, D.J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215, 403-410.

Altschul, SF., Gish, W., Miller, W., Myers, EW. & Lipman, DJ. (1990). Basal local alignment search tool. *Journal of Molecular Biology*, 215, 403-410.

Altschul, SF., Madden, TL., Schaffer, AA., Zhang, J. & Zhang, Z. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*. 25, 3389–3402.

Andersson, DI., Jerlstrom-Hultqvist, J., Nasvall, J. (2015). Evolution of new functions de novo and from preexisting genes. *Cold Spring Harbor Perspectives in Biology*. 7, a017996.

Apel, K., Hirt, H. (2004). Reactive oxygen species: metabolism, oxidative stress, and signal transduction. *Annual Reviews Plant Biology,* 55, 373–399.

Araujo, SJ., Tirode, F., Coin, F., Pospiech, H., Syvaoja, JE., Stucki, M., Hubscher, U., Egly, JM., Wood, RD. (2000). Nucleotide excision repair of DNA with recombinant human proteins: definition of the minimal set of factors, active forms of TFIIH, and modulation by CAK. *Genes and Development,* 14, 349–359.

Aroul Selvam R., Hubbard, T. & Sasidharan, R. (2004). Domain Insertions in Protein Structures. *Journal of Molecular Biology*, 338, 633–641.

Ashkenazy, H., Erez, E., Martz, E., Pupko, T. & Ben-Tal, N. (2010). ConSurf 2010: Calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Research*, 38(Web Server issue), W529–533.

Aziz, S. (1996). DNA excision repair. *Annual Reviews Biochemistry*, 65, 43-81.

Aziz, S. (1995). DNA repair in humans, *Annual Reviews Genetics*, 29, 69-105.

Bai, H., Lu, AL. (2007). Physical and Functional interactions between *Escherichia coli* MutY Glycosylase and Mismatch Repair Protein MutS. *Journal of Bacteriology*, 189(3), 902–910.

Banner, DW., Bloomer, AC., Petsko, GA., Phillips, DC., Pogson CI., Wilson, IA., Coran, PH., Furth, AJ., Milman, JD., Offord, RE. , Priddle, JD., Waley, SG. (1975). Structure of chicken muscle triose phosphate isom erase determined crystallographically at 2.5 A˚ resolution using amino acid sequence data. *Nature*, 255, 609–614.

Bateman, A., Coin, L., Durbin, R., Finn, RD., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, EL., Studholme, DJ., Yates, C. & Eddy, SR. (2004). The Pfam protein families database. *Nucleic Acids Research*, 32, 138–141.

Benkert, P., Biasini, M. , Schwede, T. (2011). Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics*, 27(3), 343-350.

Benkert, P., Künzli, M., Schwede, T. (2009). QMEAN Server for Protein Model Quality Estimation. *Nucleic Acids Research*, 37(Web Server issue), W510-14.

Benkert, P., Tosatto, SCE. & Schomburg, D. (2008). "QMEAN: A comprehensive scoring function for model quality assessment." *Proteins: Structure, Function, and Bioinformatics*, 71(1), 261-277.

Bennett, SE., Sung, JS. & Mosbaugh, DW. (2001). Fidelity of uracil-initiated base excision DNA repair in DNA polymeraseb-proficient and -deficient mouse embryonic fibroblast cell extracts. *Journal of Biological Chemistry*, 276, 42588 – 42600.

Bertrand, P., Tishkoff, DX., Filosi, N., Dasgupta, R. & Kolodner, RD. (1998). Physical interaction between components of DNA mismatch repair and nucleotide excision repair, *Proceedings of the National Academy of Sciences, USA*, 95, 14278–14283.

Bjorklund, AK., Ekman, D., Light, S., Frey-Skott, J. & Elofsson, A. (2005). Domain Rearrangements in Protein Evolution. *Journal of Molecular Biology*, 353, 911–923.

Bogorad, L. (1975). Evolution of organelles and eukaryotic genomes. *Science*, 188, 891–898.

Bohr, VA., Smith, CA., Okumoto, DS., Hanawalt, PC. (1985). DNA repair in an active gene: removal of pyrimidine dimers from the DHFR gene of CHO cells is much more efficient than in the genome overall. *Cell*, 40, 359–369.

Boiteux, S. & Guillet, M. (2004). Abasic sites in DNA: repair and biological consequences in Saccharomyces cerevisae. *DNA Repair*, 3, 1–12.

Brochier, C., Philippe, H., Moreira, D. (2000). The evolutionary history of ribosomal protein RpS14: horizontal gene transfer at the heart of the ribosome. *Trends in Genetics*, 16(12), 529-533.

Burns, JL., Guzder, SN., Sung, P., Prakash, S., Prakash, L. (1996). An affinity of human replication protein A for ultraviolet-damaged DNA. *Journal of Biological Chemistry*, 271, 11607–11610.

Chan E. & Weiss B. (1987). Endonuclease IV of Escherichia coli is induced by paraquat. *Proceedings of the National Academy of Sciences, USA.* 84, 3189-3193.

Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, TJ., Higgins, DG. & Thompson, JD. (2003). Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Research*, 31, 3497–3500.

Cole, JR., Wang, Q., Cardenas, E., Fish, J, Chai, B., Farris, RJ., Kulam-Syed-Mohideen, AS., McGarrell, DM., Marsh, T., Garrity, GM. & Tiedje, JM. (2009). The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Research*, 37(1), D141-D145.

Colovos, C., Yeates TO., (1993). Verification of protein structures: patterns of nonbonded atomic interactions. *Protein Science*, 2, 1511-1519.

Crooks, GE., Hon, G., Chandonia, JM. & Brenner, SE. (2004). WebLogo: A sequence logo generator. *Genome Research*, 14, 1188-1190.

Cunningham RP., Saporito, SM., Spitzer, SG. & Weiss, B. (1986). Endonuclease IV (nfo) mutant of *Escherichia coli*. *Journal of Bacteriology*, 168, 1120–1127.

Dantzer, F., Bjørås, M., Luna, L., Klungland, A. & Seeberg, E. (2003). Comparative analysis of 8-oxoG:C, 8-oxoG:A, A:C and C:C DNA repair in extracts from wild type or 8-oxoG DNA glycosylase deficient mammalian and bacterial cells. *DNA Repair*, 2, 707 – 718.

Das, A., Wiederhold, L., Leppard, JB., Kedar, P., Prasad, R., Wang, H., Boldogh, I., Karimi-Busheri, F., Weinfeld, M., Tomkinson, AE, Wilson, SH, Mitra, S, Hazra, TK. (2006). NEIL2-initiated, APE-independent repair of oxidized bases in DNA: Evidence for a repair complex in human cells. *DNA Repair (Amst.),* 5, 1439-1448.

De Bont, R. & Van Larebeke, N. (2004). Endogenous DNA damage in humans: a review of quantitative data. *Mutagenesis*, 19(3), 169–185.

De Laat, WL., Jaspers, NG. & Hoeijmakers, JH. (1999). Molecular mechanism of nucleotide excision repair. *Genes & Development*, 13, 768–785.

Demple, B., and DeMott, M. S. (2002). Dynamics and diversions in base excision DNA repair of oxidized abasic lesions. *Oncogene,* 21, 8926–8934

Demple, B., Herman, T., & Chen, DS. (1991). Cloning and expression of APE, the cDNA encoding the major human apurinic endonuclease: definition of a family of DNA repair enzymes. *Proceedings of the National Academy of Sciences, USA*, 88, 11450–11454.

Demple, B., Johnson, A., Fung, D. (1986). Exonuclease III and endonuclease IV remove 3′ blocks from DNA synthesis primers in H2O2-damaged Escherichia coli. *Proceedings of the National Academy of Sciences, USA*, 83, 7731-7735.

Denver DR., Swenson SL. & Lynch M. 2003. An evolutionary analysis of the Helix-Hairpin Helix superfamily of DNA repair glycosylases. *Molecular Biology and Evolution*, 20(10), 1603–1611.

Denver, DR., Swenson, SL. & Lynch, M. (2003). An evolutionary analysis of the Helix-Hairpin-Helix Superfamily of DNA Repair Glycosylases. *Molecular Biology and Evolution*, 20(10), 1603–1611.

Dexheimer, TS. (2013). DNA Repair Pathways and Mechanisms. In: Mathews LA, Cabarcas SM, Hurt EM (eds) DNA Repair of Cancer Stem Cells. Springer Netherlands, pp. 19-32.

Dianov, G., Price, A. & Lindahl, T. (1992). Generation of single-nucleotide repair patches following excision of uracil residues from DNA. *Molecular and Cellular Biology*, 12, 1605–12.

Dianov, GL., Prasad R., Wilson, SH. & Bohr, VA. (1999). Role of DNA polymerase beta in the excision step of long patch mammalian base excision repair. *Journal of Biological Chemistry*, 274, 13741–743.

Dimple, B. & Harrison, LA. (1994). Repair of oxidative damage to DNA: enzymology and biology. *Annual Reviews Biochemistry*, 63, 915-948.

Dmitry, GV, & Kosuke, M. (1997). DNA repair enzymes. *Current Opinion in Structural Biology*, 7, 103-109.

DNA sequence data sets. *Molecular Biology and Evolution*, 14, 248–265.

Dodson, ML., Michaels ML. & Lloyd, RS. (1994). Unified Catalytic Mechanism for DNA Glycosylase. *Journal of Biological Chemistry*, 269, 32709-32712.

Dufner, P., Marra, G., Schle, MR., Jiricny, J. (2000). Mismatch Recognition and DNA-dependent Stimulation of the ATPase Activity of hMutSa Is Abolished by a Single Mutation in the hMSH6 Subunit. *Journal of Biological Chemistry*, 275(47), 36550–36555.

Eisen JA. & Hanawalt PC. 1999. A phylogenomic study of DNA repair genes, proteins, and Processes. *Mutation Research*, 435, 171–213.

Eisen, JA. (1998). A phylogenomic study of the MutS family of proteins, *Nucleic Acids Research*, 26, 4291–4300.

Eisthen, HL. (1997). Evolution of vertebrate olfactory systems. *Brain Behavior Evolution*, 50(4), 222-233.

Ellis, R J. (1981). Chloroplast proteins: synthesis, transport and assembly. *Annual Review Plant Physiology*, 32, 111–137.

Emanuelsson, O., Brunak, S., Von, HG. & Nielsen, H. (2007). Locating proteins in the cell using TargetP, SignalP and related tools. *Nature Protocols*, 2(4), 953-971.

Essen, LO. & Klar. T. (2006). Light-driven DNA repair by photolyases. *Cellular and Molecular Life Sciences*, 63(11), 1266-77.

Evans, AR., Limp-Foste,r M., Kelley, MR. (2000). Going APE over ref-1. *Mutation Research*, 461(2), 83–108.

Fang, Q., Kanugula, S. & Pegg, AE. (2005). Function of Domains of Human O6-Alkylguanine-DNA Alkyltransferase. *Biochemistry*, 44, 15396-15405.

Felsenstein J. Confidence Limits on Phylogenies: An Approach Using the Bootstrap. (1985). *Evolution*, 39, 783–791.

Finn, RD,. Tate, J., Mistry, J., Coggill, PC., Sammut, JS., Hotz. HR., Ceric, GKF., Eddy, SR. Sonnhammer, EL., Bateman, A. (2008). The Pfam protein families database. *Nucleic Acids Research*, 36(Database Issue), D281-D288.

Fishel, R. &Wilson, T. (1997). MutS homologs in mammalian cells. *Current Opinion Genetics & Development*, 7(1), 105-113.

Fishel, R. (1998). Mismatch repair, molecular switches, and signal transduction. *Genes and Development*, 12, 2096–101.

Fleming, K., Kelley, LA., Islam, SA., MacCallum, RB., Muller, A., Pazos, F. & Sternberg, MJE. (2006). The proteome: structure, function and evolution. *Philosophical Transactions of the Royal Society*. B, 361, 441–445.

Fortini, P. & Dogliotti, E. (2007). Base damage and singlestrand break repair: Mechanisms and functional significance of short- and long-patch repair subpathways. *DNA Repair*, 6, 398 – 409.

Friedberg, EC. (2003). DNA damage and repair. *Nature*, 421, 436-440.

Friedberg, EC. 2001. How nucleotide excision repair protects against cancer. *Nature Review Cancer*, 1, 22–33.

Fromme, JC. & Verdine, GL. (2003). Structure of a trapped endonuclease III-DNA covalent intermediate. *The EMBO Journal*, 22, 3461-3471.

Galburt, EA., Chevalier, B., Tang, W., Jurica, MS., Flick, KE., Monnat, RJ. Jr., & Stoddard, BL. (1999). A novel endonuclease mechanism directly visualized for I-PpoI. *Nature Structural Biology*. 6,1096–1099.

Garinis, GA., Jans, J. & Van der Horst, GT. (2006). Photolyases: capturing the light to battle skin cancer. *Future Oncology*, 2(2), 191-99.

Garrity, GM. & Holt, JG. (2001). The *Archaea* and the deeply branching and phototropic *Bacteria*, In: Boone DR, Castenholz RW & Garrity GM (eds.) Bergey's Manual of Systematic Bacteriology. 2nd edition, Vol . 1, Springer NewYork , pp. 119-166.

Gellon, L., Werner, M., Boiteux, S. (2002). Ntg2p, a Saccharomyces cerevisiaeDNA N-glycosylase/apurinic or apyrimidinic lyase involved in base excision repair of oxidative

DNA damage, interacts with the DNA mismatch repair protein Mlh1p. Identification of a Mlh1p binding motif, *Journal of Biological Chemistry*, 277, 29963–29972.

Genschel, J., Bazemore, LR. & Modrich, P. (2002). Human exonuclease I is required for 5'and 3'mismatch repair. *Journal of Biological Chemistry*, 277, 13302–311.

Genschel, J., Littman, SJ., Drummond, JT., Modrich, P., (1998). Isolation of MutSbeta from human cells and comparison of the mismatch repair specificities of MutSbeta and MutSalpha. *Journal of Biological Chemistry*, 273, 19895–901.

Georgiadis MM., Luo M, Gaur RK.,, Delaplane S.,, Li X., & Kelley MR. 2008. Evolution of the redox function in mammalian Apurinic/apyrimidinic endonuclease. *Mutatation Res*earch, 25, 643(1-2), 54–63.

Golinelli, MP., Chmiel, NH., David, SS. (1999). Site-Directed Mutagenesis of the Cysteine Ligands to the [4Fe 4S] Cluster of Escherichia coli MutY. *Biochemistry*, 38, 6997-7007.

Gray, MW. & Doolittle, WF. (1982). Has the endosymbiont hypothesis been proven *Microbiology Review*, 46, 1–42.

Gray, MW., Burger, G. & Lang, BF. (1999). Mitochondrial evolution. *Science*, 283, 1476–1481.

Grishin, NV. (1997). Estimation of evolutionary distances from protein spatial structures, *Journal of Molecular Evolution*, 45, 359–369.

Grishin, NV. (2001). Fold Change in Evolution of Protein Structures. *Journal of Structural Biology*, 134, 167–185.

Guex, N., Peitsch, MC. (1997). SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis*, 18(15), 2714-23

Habraken, Y., Sung, P., Prakash, L., Prakash, S. (1994). A conserved 5' to 3'exonuclease activity in the yeast and human nucleotide excision repair proteins RAD2 and XPG. *Journal of Biological Chemistry*, 269, 31342–31345.

Hall, C., Brachat, S., Dietrich, FS. (2005). Contribution of Horizontal Gene Transfer to the Evolution of *Saccharomyces cerevisae*. *Eukaryotic Cell*, 4(6), 1102–1115.

Hardeland, U., Bentele, M., Lettieri, T., Steinacher, R., Jiricny, J., Schär, P. (2001). Thymine DNA glycosylase. *Progress in Nucleic Acid Research and Molecular Biology*, 68, 235-253.

Hazra, TK., Das, A., Das, S. Choudhury, S., Kow, YW. & Roy, R. (2007). Oxidative DNA damage repair in mammalian cells: a new perspective. *DNA Repair (Amst)* 6, 470-480.

Heck, DE., Vetrano, AM., Mariano, TM. & Laskin, JD. (2003). UVB light stimulates production of reactive oxygen species. *Journal of Biological Chemistry*, 278, 22432–22436

Hegde, ML., Hazra, TK. & Mitra, S. (2008). Early steps in the DNA base excision/single-strand interruption repair pathway in mammalian cells. *Cell Research*, 18, 27-47.

Hey, T., Lipps, G., Sugasawa, K., Iwai, S., Hanaoka, F., Krauss, G. (2002). The XPC–HR23B complex displays high affinity and specificity for damaged DNA in a true-equilibrium fluorescence assay. *Biochemistry*, 41, 6583–6587.

Hoheisel, JD. (1993). On the activities of *Escherichia coli* exonuclease III. Analytical Biochemistry, 209, 238–246.

Hosfield, DJ., Guan, Y., Haas, BJ., Cunningham, RP., Tainer, JA. (1999). Structure of the DNA repair enzyme endonuclease IV and its DNA complex: double-nucleotide flipping at abasic sites and three-metal-ion catalysis. *Cell*, 98(3), 397-408.

Hughes, AL. (1994). The Evolution of Functionally Novel Proteins after Gene Duplication. *Philosophical Transactions of the Royal Society,* 256, 119-124.

Hughes, AL. (2005). Gene duplication and the origin of novel proteins. *Proceedings of the National Academy of Sciences, USA,* 102(25), 8791–8792.

Irigaray, P. & Belpomme, D. (2010). Basic properties and molecular mechanisms of exogenous chemical carcinogens. *Carcinogenesis,* 31(2), 135–148.

Itoh, M., Nacher, JC., Kuma, K., Goto, S. & Kanehisa, M. (2007). Evolutionary history and functional implications of protein domains and their combinations in eukaryotes. *Genome Biology*, 8(6), R121.

Jane, EAW., Anthony, EP., Peter, CEM. (2000), Crystal structure of the human O6-alkylguanine-DNA alkyltransferase. *Nucleic Acids Research*, 28(2), 393-401.

Jones, DT., Taylor, WR. & Thornton, JM. (1992). The rapid generation of mutation data matrices from protein sequences. *Computer Applications in Biosciences*, 8, 275-282.

Jordan, IK., Kondrashov, FA., Adzhubei, IA., Wolf, YI., Koonin, EV., Kondrashov, AS. & Sunyaev, S. (2005) A universal trend of amino acid gain and loss in protein evolution. *Nature,* 433, 633–638.

Jorg, S,. Frank, M., Peer, B., Chris, PP. (1998), SMART, a simple modular architecture research tool: Identification of signaling domains. *Proceedings of the National Academy of Sciences, USA*,95, 5857–5864.

Kanchan, S., Mehrotra, R. & Chowdhury, S. (2014). Evolutionary pattern of four representative DNA repair proteins across six model organisms: an *in silico* analysis. *Network Modeling Analalysis in Health Informatics and Bioinformatics*, 3, 70.

Kanchan, S., Mehrotra, R., Chowdhury, S. (2015). *In silico* study of endonuclease III protein family identifies key residues and processes during evolution. *Journal of* molecular *evolution*. 81, 54–67.

Kesheri, M., Kanchan, S., Chowdhury, S. Sinha, RP. (2015). Secondary and Tertiary Structure Prediction of Proteins: A Bioinformatic Approach. In: Zhu Q, Azar AT (eds) Complex system modelling and control through intelligent soft computations, Studies in Fuzziness and Soft Computing. Vol. 319, Springer-Verlag, Germany, pp. 541-569.

Kesheri, M., Kanchan, S., Richa, Sinha, RP. (2014). Isolation and in silico analysis of Fe-superoxide dismutase in Nostoc commune. *Gene***.** 553(2), 117-125.

Kinch, LN. & Grishin NV. (2002). Evolution of protein structures and functions. *Current Opinion Structural Biology*, 12, 400–408.

Kisaburo N. & David, SL. (2006). Comparative Genomic and Phylogenetic Analyses of Calcium ATPases and Calcium-Regulated Proteins in the Apicomplexa. *Molecular Biology and Evolution*, 23(8), 1613–1627.

Kolodner, R. (1996). Biochemistry and genetics of eukaryotic mismatch repair. *Genes and Development*, 10, 1433–1442.

Kolodner, RD, (1995), Mismatch repair: mechanisms and relationship to cancer susceptibility. *Trends in Biochemical Science*, 20, 397-401.

Kondrashov, AS. (1988). Deleterious mutations and the evolution of sexual reproduction. *Nature*, 336, 435–440.

Kovtun, IV. & McMurray, CT. (2007). Crosstalk of DNA glycosylases with pathways other than base excision repair. *DNA repair*, 6, 517–529.

Kow, YW. & Wallace, SS. (1985). Exonuclease III recognizes urea residues in oxidized DNA. *Proceedings of the National Academy of Sciences, USA*, 82, 8354–8358.

Kubota, Y., Nash, RA., Klungland, A., Schar, P., Barnes, DE., Lindahl, T. (1996). Reconstitution of DNA base excision-repair with purified human proteins: interaction between DNA polymerase beta and the XRCC1 protein. *EMBO Journal*, 15, 6662–6670.

Kunkel, TA., & Erie, DA. (2005). DNA Mismatch repair. *Annual Review Biochemistry*, 74, 681–710.

Labahn, J., Schärer, OD., Long, A., Ezaz-Nikpay, K., Verdine, GL. & Ellenberger, TE. (1996). Structural basis for the excision repair of alkylation-damaged DNA. *Cell*, 86, 321-329.

Lamers, MH., Perrakis, A., Enzlin, JH., Winterwerp, HH., De Wind, N. & Sixma, TK. (2000). The crystal structure of DNA mismatch repair protein MutS binding to a G x T mismatch**.** *Nature*, 407(6805), 711-717.

Larkin MA., Blackshields G., Brown N P., Chenna R,. McGettigan P A., McWilliam H., Valentin F., Wallace IM., Wilm A., Lopez R., Thompson JD., Gibson TJ. & Higgins DG. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics*, 23(21), 2947-2948.

Laskowski, RA., MacArthur, MW., Moss, DS., Thornton, JM. (1993). PROCHECK: A program to check the stereochemical quality of protein structure. *Journal of Applied Crystallography*, 26, 283-291.

Letunic, I., Copley, RR., Schmidt, S., Ciccarelli, FD., Doerks, T., Shultz, J., Ponting, CP. & Bork, P. (2004). SMART 4.0: towards genomic data integration. *Nucleic Acids Research*, 32, 142–144.

Li WH. (1997). Molecular Evolution, 2nd edition, Sinauer. USA.

Li, GM. & Modrich, P. (1995). Restoration of mismatch repair to nuclear extracts of H6 colorectal tumor cells by a heterodimer of human MutL homologs. *Proceedings of the National Academy of Sciences, USA*, 92, 1950–54.

Lindahl, T. (1993). Instability and decay of the primary structure of DNA, *Nature*, 362, 709–715.

Lindahl, T. &Wood, RD. (1999). Quality Control by DNA Repair. *Science*, 286, 1897-1905.

Lindahl, T. (1982). Dna repair enzymes. *Annual Reviews Biochemistry*, 51, 61-87.

Lindahl, T., Sedgwick, B., Sekiguchi, M. & Nakabeppu, Y. (1988). *Annual Review Biochemistry* , 57, 133-57.

Liu, S., Zhuang, Y., Zhang, P., Adams, KL. (2009). Comparative Analysis of Structural Diversity and Sequence Evolution in Plant Mitochondrial Genes Transferred to the Nucleus. *Molecular Biology and Evolution*, 26(4), 875–889.

Ljungquist, S. (1977). A new endonuclease from Escherichia coli acting at apurinic sites in DNA. *Journal of Biological Chemistry*, 252, 2808-2814.

Longley, MJ., Pierce, AJ. & Modrich, P. (1997). DNA polymerase delta is required for human mismatch repair in vitro. *Journal of Biological Chemistry*, 272, 10917–921.

Lukianova, OA., David, SS. (2005). A role for iron–sulfur clusters in DNA repair. *Current Opinion in Chemical Biology*, 9, 145–151.

Madeleine, HM., Jacqueline, M., Gulb, EJD., Bruce, D., Peter, MCE. (1994), Crystal structure of a suicidal DNA repair protein: the Ada 06-methylguanine-DNA methyltransferase from E.co1i. *The EMBO Journal*, 13(7), 1495-1501.

Magadum, S., Banerjee, U., Murugan, P., Gangapur, D., Ravikesavan R. (2013). Gene duplication as a major force in evolution. *Journal of Genetics*. 92, 155–161.

Maher, RL. & Bloom, L. B. (2007). Pre-steady-state kinetic characterization of the AP endonuclease activity of human AP endonuclease 1. *Journal of Biological Chemistry*, 282 ,30577–30585.

Manuel, RC., Hitomi, K., Arvai, AS., House, PG., Kurtz, AJ., Dodson, ML., McCullough, AK., Tainer, JA. & Lloyd, RS. (2004). Reaction intermediates in the catalytic mechanism of Escherichia coli MutY DNA glycosylase. *Journal of Biological Chemistry*, 279(45), 46930-46939.

Marchler, BA., Bryant, SH. (2004). CD Search: protein domain annotations on the fly. *Nucleic Acids Research*, 32(Web Server issue), W327-31.

Marchler-Bauer, A., Anderson, JB., Derbyshire, MK., DeWeese-Scott, C., Gonzales, NR., Gwadz, M., Hao, L., He, S., Hurwitz, DI., Jackson, JD., Ke, Z., Krylov, D., Lanczycki, CJ., Liebert, CA., Liu, C., Lu, F., Lu, S., Marchler, GH., Mullokandov, M., Song, JS., Thanki, N., Yamashita, RA., Yin, JJ., Zhang, D. & Bryant, SH. (2007). CDD: a conserved domain database for interactive domain family analysis. *Nucleic Acids Research*, 35(Database Issue), D237-240.

Martin, W., Rujan, T., Richly, E., Hansen, A., Cornelsen, S., Lins, T., Leister, D., Stoebe, B., Hasegawa, M., Penny, D. (2002). Evolutionary analysis of Arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proceedings of the National Academy of Sciences, USA*, 99, 12246–12251.

Masson, JY., Tremblay, S., & Ramotar, D. (1996). The Caenorhabditis elegans gene CeAPN1 encodes a homolog of Escherichia coli and yeast apurinic/apyrimidinic endonuclease. *Gene*, 179, 291–293.

Mazon, G., Philippin, G., Cadet, J., Gasparutto, D., Fuchs, RP. (2009). The alkyltransferase-like ybaZ gene product enhances nucleotide excision repair of O6-alkylguanine adducts in E. coli. *DNA Repair*, 8, 697–703.

McCulloch, SD. & Kunkel, TA. (2008), The fidelity of DNA synthesis by eukaryotic replicative and translesion synthesis polymerases. *Cell Research*, 18(1), 148–161.

Meijer, M., Karimi-Busheri, F., Huang, TY., Weinfeld, M., Young, D. (2002). Pnk1, a DNA kinase/phosphatase required for normal response to DNA damage by gamma-radiation or camptothecin in *Schizosaccharomyces pombe*. *Journal of Biological Chemistry*, 277, 4050-4055.

Mellon, I., Spivak, G., Hanawalt, PC. (1987). Selective removal of transcription-blocking DNA damage from the transcribed strand of the mammalian DHFR gene. *Cell*, 51, 241–249.

Millar, CB., Guy, J., Sansom, OJ., Selfridge, J., MacDougall, E., Hendrich, B., Keightley, PD., Bishop, SM., Clarke, AR. & Bird, A. (2002). Enhanced CpG mutability and tumorigenesis in MBD4-deficient mice. *Science*, 297, 403–405.

Mitra, S., Izumi, T., Boldogh, I., Bhakat, K. K., Hill, J. W., and Hazra, T. K. (2002). Choreography of oxidative damage repair in mammalian genomes**.** *Free Radical Biology & Medicine*, 33,15–28.

Modrich, P., Lahue, R., (1996). Mismatch repair in replication fidelity, genetic recombination, and cancer biology. *Annual Review Biochemistry*, 65, 101–33.

Mol, CD., Hosfield, DJ., & Tainer, J. A. (2000). Abasic site recognition by two apurinic/apyrimidinic endonuclease families in DNA base excision repair: the 3' ends justify the means. *Mutation Research*, 460, 211–229.

Moretti, S., Armougom, F., Wallace, IM. Higgins, DG., Jongeneel, CV. & Notredame, C. (2007). The M-Coffee web server: a meta-method for computing multiple sequence alignments by combining alternative alignment methods. *Nucleic Acids Research*, 35(Web Server issue), W645-648.

Mu, D., Park, CH., Matsunaga, T., Hsu, DS., Reardon, JT., Sancar, A. (1995). Reconstitution of human DNA repair excision nuclease in a highly defined system. *Journal of Biological Chemistry*, 270, 2415–2418.

Murzin, AG., Brenner, SE., Hubbard T. & Chothia C., (1995). SCOP: A structural classification of proteins database for the investigation of sequences and structures, *Journal of Molecular Biology*, 247, 536–540.

Nash, HM., Bruner, SD., Schärer, OD., Kawate T, Addona TA, Spooner E, Lane WS & Verdine GL. (1996). Cloning of a yeast 8-oxoguanine DNA glycosylase reveals the existence of a baseexcision DNA-repair protein superfamily. *Current Biology*, 6, 968-980.

Ohno, S. (1973). Ancient linkage groups and frozen accidents. *Nature*, 244, 259–262.

Orengo, CA. & Thornton, JM. (2005). Protein families and their evolution a structural perspective. *Annual Review Biochemistry*, 74, 867–900.

Pál, C., Papp B. & Lercher, MJ. (2006). An integrated view of protein evolution. *Nature Reviews Genetics*, 7, 337-348.

Panier, S. & Boulton, SJ. (2014). Double-strand break repair: 53BP1 comes into focus**.** *Nature Reviews Molecular Cell Biology*, 15(1), 7-18.

Pearl, FM., Lee, D., Bray, JE., Sillitoe, I., Todd, AE. & Harrison, AP., Thornton, JM. & Orengo CA. (2000). Assigning genomic sequences to CATH, *Nucleic Acids Research*. 28, 277–282.

Pearl, LH. (2000), Structure and function in the uracil-DNA glycosylase superfamily. *Mutation Research*, 460, 165–181.

Pegg, AE. (2000). Repair of O(6)-alkylguanine by alkyltransferases *Mutation Research*, 462, 83–100.

Pennisi, E. (2002). Jumbled DNA separates chimps and humans. *Science*, 298(5594), 719–721.

Pinheiro, MM., Galhardo, RD., Lage, C., Keronninn M., Bessa, L., Aires, KA. & Menck, CFM. (2004). Different patterns of evolution for duplicated DNA repair genes in bacteria of the Xanthomonadales group. *BMC Evolutionary Biology*, 4, 29.

Podlutsky, AJ., Dianova, II., Podust, VN., Bohr, VA. & Dianov, GL. (2001). Human DNA polymerasebinitiates DNA synthesis during long-patch repair of reduced AP sites in DNA. *EMBO Journal*, 20, 1477 – 1482.

Podlutsky, AJ., Dianova, II., Wilson, SH., Bohr, VA. & Dianov, GL. (2001). DNA synthesis and dRPase activities of polymerase beta are both essential for single-nucleotide patch base excision repair in mammalian cell extracts. *Biochemistry*, 40, 809–13.

Prasad, R., Singhal, RK., Srivastava, DK., Molina, JT., Tomkinson, AE. & Wilson, SH. (1996). Specific interaction of DNA polymerase beta and DNA ligase I in a multiprotein base excision repair complex from bovine testis. *Journal of Biological Chemistry*, 271, 16000–16007.

Puigbo, P., Bravo, IG. & Garcia-Vallve, S. (2008). CAIcal: a combined set of tools to assess codon usage adaptation. *Biology Direct*, 3, 38.

Qian, W. & Zhang, JG. (2014). Genomic evidence for adaptation by gene duplication. *Genome Research*, 24:1356–1362.

Rada, C., Di Noia, JM., Neuberger, MS. (2004). Mismatch recognition and uracil excision provide complementary paths to both Ig switching and the A/T-focused phase of somatic mutation. *Molecular Cell*, 16, 163–171.

Ramotar, D., Popoff, SC., Gralla, EB., & Demple, B. (1991). Cellular role of yeast Apn1 apurinic endonuclease/3'-diesterase: repair of oxidative and alkylation DNA damage and control of spontaneous mutation. *Molecular and Cellular Biology*, 11, 4537–4544.

Ramotar, D., Vadnais, J., Mason, JY. &Tremblay, S. (1998). Schizosaccharomyces pombe apn1 encodes a homologue of the Escherichia coli endonuclease IV family of DNA repair proteins. *Biochimica et Biophysica Acta*, 1396, 15–20.

Rasimas, JJ., Kanugula, S., Dalessio, PM., Ropson, IJ. & Pegg, AE. (2003). Effects of zinc occupancy on human O6-alkylguanine-DNA alkyltransferase. *Biochemistry*, 42, 980-990.

Rastogi, S. & Liberles, DA. (2005). Subfunctionalization of duplicated genes as a transition state to neofunctionalization. *BMC Evolutionary Biology*, 5, 28.

Ribar, B., Izumi, T. & Mitra, S. (2004). The major role of human AP-endonuclease homolog Apn2 in repair of abasic sites in Schizosaccharomyces pombe. *Nucleic Acids Research*, 32, 115–126.

Rice, P., Longden, I., Bleasby, A. (2000). EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics*, 16(6), 276-277.

Richard, DW. (2001). Human DNA repair genes. *Science*, 291, 1284-1289.

Richardson,CC. & Kornberg, A. (1961), A deoxyribonucleic acid phosphatase-exonuclease from Escherichia coli. Purification and characterization of the phosphatase activity. *Journal of Biological Chemistry*, 239, 242–255.

Rogers, SG. & Weiss, B. (1980). Exonuclease III of *Escherichia coli* K-12 an AP endonuclease, *Methods in Enzymology,* 65, 201–211.

Rost, B., Sander, C. 1993. Prediction of protein secondary structure at better than 70% accuracy. *Journal of Molecular Biology*, 20, 232(2), 584-599.

Ruvolo, M. (1997). Molecular phylogeny of the Hominoids: Inferences from multiple independent DNA sequence data sets. *Molecular Biology and Evolution,* 14, 248–265.

Sali, A., Blundell, TL. (1993). Comparative protein modeling by satisfaction of spatial restrains. *Journal of Molecular Biology*, 234, 779-815.

Sancar, A. (1995). DNA repair in humans. *Annual Reviews Genetics*, 29, 69-105.

Sancar, A., Lindsey-Boltz, LA, Ünsal-Kaçmaz, K. & Linn, S. (2004). Molecular mechanisms of mammalian dna repair and the dna damage checkpoints. *Annual Review Biochemistry*, 73, 39–85.

Sankoff, D. (2001). Gene and genome duplication. *Current Opinion in Genetics & Development*, 11, 681–684.

Santos, CL., Vieira, J. Tavares, F., Benson, DR., Tisa, LS., Berry, AM., Ferreira, PM. & Normand, P. (2008). On the nature of fur evolution: A phylogenetic approach in Actinobacteria. *BMC Evolutionary Biology*, 8, 185.

Saparbaev, M. & Laval, J. (1998). 3, N^4-ethenocytosine, a highly mutagenic adduct, is a primary substrate for *Escherichia coli* double-stranded uracil-DNA glycosylase and human mismatch-specific thymine-DNA glycosylase. *Proceedings of the National Academy of Sciences*, 95, 8508–8513.

Saporito, SM., Smith-White, BJ. & Cunningham, RP. (1988). Nucleotide sequence of the xth gene of *Escherichia coli* K-12. *Journal of Bacteriology*, 170, 4542–4547.

Scharer, OD. & Jiricny, J. (2001). Recent progress in the biology, chemistry and structural biology of DNA glycosylases. *Bioessays,* 23, 270–281.

Selby, CP., Drapkin, R., Reinberg, D., Sancar, A., (1997). RNA polymerase II stalled at a thymine dimer: footprint and effect on excision repair. *Nucleic Acids Research,* 25, 787–793.

Shida, T., Kaneda, K., Ogawa, T., Sekiguchi, J. (1999). Abasic site recognition mechanism by the Escherichia coli exonuclease III. Nucleic Acids Symposium Series, 42, 195-196.

Shimizu, Y., Iwai, S., Hanaoka, F., Sugasawa, K. (2003). Xeroderma pigmentosum group C protein interacts physically and functionally with thymine DNA glycosylase. *The EMBO Journal*, 22, 164–173.

Sijbers, AM., de Laat, WL., Ariza, RR., Biggerstaff, M., Wei, YF., Moggs, JG., Carter, KC., Shell, BK., Evans, E., de Jong, MC., Rademakers S, de Rooij J, Jaspers NG, Hoeijmakers JH, Wood RD. (1996). Xeroderma pigmentosum group F caused by a defect in a structure-specific DNA repair endonuclease. *Cell*, 86, 811–822.

Singleton, MR., Dillingham, MS. & Wigley, DB. (2007). Structure and mechanism of helicases and nucleic acid tanslocases. *Annual Reviews Biochemistry*, 76, 23–50.

Soltis, PS, & Soltis, DE. (2003). Applying the bootstrap in phylogeny reconstruction. *Statistical Sciencem*, 18(2), 256-267.

Soskine, M. & Tawfik, DS. 2010. Mutational effects and the evolution of new protein function. *Nature Reviews Genetics.* 11(8), 572-82.

Srivastava, DK., Berg, BJ., Prasad, R., Molina, JT., Beard, WA., Tomkinson, AE. & Wilson, SH., (1998). Mammalian abasic site base excision repair. Identification of the reaction

sequence and rate-determining steps. *Journal of Biological Chemistry*, 273, 21203–21209.

Stucki, M., Pascucci, B., Parlanti, E., Fortini, P., Wilson, SH., Hubscher, U. & Dogliotti, E. (1998). Mammalian base excision repair by DNA polymerases delta and epsilon. *Oncogene*, 17, 835–43.

Takahashi, H., Kamiya, A., Ishiguro, A., Suzuki, AC., Saitou, N., Toyoda A., & Aruga J. (2008). Conservation and Diversification of Msx Protein in Metazoan Evolution. *Molecular Biology and Evolution*, 25(1), 69-82.

Tamura, K. & Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution* 10, 512-526.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M. & Kumar, S. (2011). MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. *Molecular Biology and Evolution*, 28, 2731-2739.

Tang, J., Chu, G., 2002. Xeroderma pigmentosum complementation group E and UV-damaged DNA-binding protein. *DNA Repair (Amst.)*, 1, 601–616.

Tatusov, RL, Fedorova, ND., Jackson JD., Jacobs, AR., Kiryutin, B., Koonin, EV, Krylov, DM., Mazumder, R., Mekhedov, SL., Nikolskaya, AN., Rao, BS., Smirnov, S., Sverdlov, AV., Vasudevan, S., Wolf, YI., Yin, JJ. & Natale, DA. (2003). The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, 4, 41.

Thayer, MM., Ahern, H., Xing, D., Cunningham, RP., Tainer, JA. (1995). Novel DNA binding motifs in the DNA repair enzyme endonuclease Ill crystal structure. *The EMBO Journal*, 14(16), 4108-4120.

Thomas, AK. & Dorothy, AE. (2005). DNA mismatch repair. *Annual Reviews Biochemistry*, 74, 681–710.

Thompson, JD., Higgins, DG., Gibson, TJ. (1994). Clustalw: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22(22), 4673-4680.

Tomkinson, AE., Chen, L., Dong, Z., Leppard, JB., Levin, DS., Mackey, ZB. & Motycka, TA. (2001). Completion of base excision repair by mammalian DNA ligases. *Progress in Nucleic Acid Research and Molecular Biology*, 68, 151–164.

Tornaletti, S., Maeda, LS., Lloyd, DR., Reines, D., Hanawalt, PC. (2001). Effect of thymine glycol on transcription elongation by T7 RNA polymerase and mammalian RNA polymerase II. *Journal of Biological Chemistry*, 276, 45367–45371.

Tsutakawa, SE., Shin, DS., Mol, CD., Izumi, T., Arvai, AS., Mantha, AK., Szczesny, B., Ivanov, IN., Hosfield, DJ., Maiti, B., Pique, ME., Frankel, KA., Hitomi, K., Cunningham, RP., Mitra, S., Tainer, JA. (2013). Conserved structural chemistry for incision activity in structurally non-homologous apurinic/apyrimidinic endonuclease APE1 and endonuclease IV DNA repair enzymes. *Journal of Biological Chemistry*, 22, 288(12), 8445-55.

Tubbs, JL., Latypov, V., Kanugula, S,, Buttm A,, Melikishvilim M,, Kraehenbuehlm R,, Fleck, O., Marriott, A., Watson, AJ., Verbeek, B., McGown, G., Thorncroft, M., Santibanez-Koref, MF., Millington, C., Arvai, AS., Kroeger, MD., Peterson, LA., Williams, DM.,

Fried, MG., Margison, GP., Pegg, AE., Tainer, JA. (2009). Flipping of alkylated DNA damage bridges base and nucleotide excision repair. *Nature*, 459, 808–813.

Tubbs, JL., Tainer, JA. (2010). Alkyltransferase-like proteins: Molecular switches between DNA repair pathways. *Cellular and Molecular Life Sciences*, 67(22), 3749–3762.

Umar, A., Boyer, JC. & Kunkel, TA., (1994). DNA loop repair by human cell extracts. *Science*, 266, 814–16.

Unk, I., Haracska, L., Prakash, S. & Prakash, L. (2001). 3'-Phosphodiesterase and 3'→5' exonuclease activities of yeast Apn2 protein and requirement of these activities for repair of oxidative DNA damage. *Molecular and Cellular Biology*, 21, 1656–1661.

Vermeulen, W., de Boer, J., Citterio, E., van Gool, AJ., van der Horst, GT., Jaspers, NG., de Laat, WL., Sijbers, AM., van der Spek, PJ., Sugasawa, K., Weeda, G, Winkler, GS., Bootsma D., Egly, JM., Hoeijmakers, JH. (1997). Mammalian nucleotide excision repair and syndromes. *Biochemical Society Transactions*, 25, 309–315.

Vincent, WF. & Neale, PJ. (2000). Mechanisms of UV damage to aquatic organisms. In: Mora SD, Demers S, Vernet M (eds) The Effects of UV Radiation on Marine Ecosystems. Cambridge University Press, Cambridge, pp. 149–176.

Virel, A. & Backman, L. (2007). A Comparative and Phylogenetic Analysis of the a-Actinin Rod Domain. *Molecular Biology and Evolution*, 24(10), 2254–2265.

Vision, TJ., Brown, DG. & Tanksley, SD. (2000). The origins of genomic duplications in Arabidopsis. *Science*, 290, 2114–2117.

Viswanathan, A., Doetsch, PW. (1998). Effects of nonbulky DNA base damages on Escherichia coli RNA polymerase-mediated elongation and promoter clearance. *Journal of Biological Chemistry*, 273, 21276–21281.

Vogel, C., Bashton, M., Kerrison, ND., Chothia, C. & Teichmann, SA. (2004). Structure, function and evolution of multidomain proteins. *Current Opinion Structural Biology,* 14, 208–216.

Warren, JJ., Pohlhaus, TJ., Changela, A., Iyer, RR., Paul, LM., Beese, LS. (2007). Structure of the Human MutS alpha DNA Lesion Recognition Complex. *Molecular Cell*, 26, 579–592.

Weber S. (2005). Light-driven enzymatic catalysis of DNA repair: a review of recent biophysical studies on photolyase. *Biochimica Biophysica Acta*, 1707(1), 1-23.

Weeden, N F. (1981). Genetic and biochemical implications of the endosymbiotic origin of the chloroplast. *Journal of Molecular Evolution,* 17,133–139.

Weiss, B. (1981). Exodeoxyribonucleases of *Escherichia coli*. In: Boyer PD (ed), The enzymes. Third edition, Vol. 14, Academic press, New York, pp. 203-231

Whittaker, RH. (1969). New concepts of kingdoms or organisms. Evolutionary relations are better represented by new classifications than by the traditional two kingdoms. *Science*, 163(3863), 150-160.

Wilson 3rd DM. & Barsky, D. (2001). The major human abasic endonuclease: formation, consequences and repair of abasic lesions in DNA. *Mutation Research*, 485, 283–307.

Wilson, DM 3rd., Bohr, VA. (2007). The mechanics of base excision repair, and its relationship to aging and disease**.** *DNA Repair (Amst)*, 6(4), 544-59.

Winkler, GS., Araujo, SJ., Fiedler, U., Vermulen, W., Coin, F. (2000). TFIIH with inactive Xpd helicase functions in transcription initiation but it is defective in DNA repair. *Journal of Biological Chemistry*, 275, 4258-4266.

Wolski, SC., Kuper, J., Nzelmann, PH., Truglio, JJ., Croteau, DL., Houten, BV., Kisker, C. (2008). Crystal Structure of the FeS Cluster–Containing Nucleotide Excision Repair Helicase XPD. *Plos Biology*, 6(6), e149.

Xu, D. & Zhang, Y. (2011). Improving the Physical Realism and Structural Accuracy of Protein Models by a Two-step Atomic-level Energy Minimization. *Biophysical Journal*, 101, 2525-2534.

Yang, H., Fitz-Gibbon, S, Marcotte, EM. (2000). Characterization of a thermostable DNA glycosylase specific for U/G and T/G mismatches from the hyperthermophilic archaeon *Pyrobaculum aerophilum*. *Journal of Bacteriology*, 182, 1272-1279.

Zhang, J. (2003). Evolution by gene duplication: an update. *Trends in Ecology and Evolution*, 18(6), 292-298.

Zharkikh, A. & Li, W.-H. (1992). Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. I. Four taxa with a molecular clock. *Journal of Molecular Evolution*, 9, 1119–1147.

List of Publications

From Thesis:

1. **Kanchan, S.**, Mehrotra, R. & Chowdhury, S. (2015). *In silico* study of endonuclease III protein family identifies key residues and processes during evolution. *Journal of molecular evolution*, 81, 54–67.

2. **Kanchan, S.,** Mehrotra, R. & Chowdhury, S. (2014). Evolutionary pattern of four representative DNA repair proteins across six model organisms: an *in silico* analysis. *Network Modeling Analysis in Health Informatics and Bioinformatics*, 3(1), 70.

3. Kesheri, M., **Kanchan, S.,** Chowdhury, S. & Sinha, RP. (2015). Secondary and Tertiary Structure Prediction of Proteins: A Bioinformatic Approach. In: Zhu Q, Azar AT (eds) Complex system modelling and control through intelligent soft computations, Studies in Fuzziness and Soft Computing. Vol. 319, Springer-Verlag, Germany, pp. 541-569.

From other projects:

4. Priya P., Kesheri M., Sinha RP, **Kanchan S. (**2015). Molecular dynamics simulations for Biological Systems. In: Karâa W. B. A., Dey N. (eds.), Biomedical Image Analysis and Mining Techniques for Improved Health Outcomes, Advances in Bioinformatics and Biomedical Engineering (ABBE) Series. Chapter 14, IGI Global, USA, pp 286-313.

5. Kesheri, M., **Kanchan, S.**, Richa & Sinha, RP. (2014). Isolation and in silico analysis of Fe-superoxide dismutase in *Nostoc commune*. *Gene***,** 553(2), 117-125. .

6. Gahoi, S., Mandal, RS., Ivanisenko, N., Shrivastava, P., Jain, S., Singh, AK., Raghunandanan, MV., **Kanchan, S.**, Taneja, B., Mandal, C., Ivanisenko, VA., Kumar, A., Kumar, R., Open Source Drug Discovery Consortium & Ramachandran, S. Computational screening for new inhibitors of M. tuberculosis mycolyltransferases

antigen 85 group of proteins as potential drug targets. (2013). *Journal of Biomolecular Structure and Dynamics*, 31(1), 30-43.

7. Garg, S., Saxena, V., **Kanchan, S.**, Sharma, P., Mahajan, S., Kochar, D., & Das, A. (2009). Novel point mutations in sulfadoxine resistance genes of *plasmodium falciparum* from India**.** *Acta Tropica*, 110(1), 75-79.

8. Kesheri, M., **Kanchan, S.,** Richa, & Sinha, RP. (2015). Oxidative stress: Challenges and its mitigation mechanisms in cyanobacteria In: Sinha RP, Richa & Rastogi RP (eds) Biological Sciences: Innovations and Dynamics. New India Publishing Agency, New Delhi, pp. 309-324.

9. Kesheri, M., **Kanchan, S.** & Chowdhury, S. (2014). Cyanobacterial Stresses: An Ecophysiological, Biotechnological and Bioinformatic Approach. LAP Lambert Academy Publishing, Germany. [ISBN: 9783848438839]

Biography of Prof. Shibasish Chowdhury

Prof. Shibasish Chowdhury obtained master's degree in physical chemistry from Calcutta University. Then, he shifted to biophysics and obtained PhD degree from Molecular Biophysics Unit (MBU) at Indian Institute of Science, Bangalore on "Computer modelling studies on G-rich unusual DNA structure". Subsequently, entered into protein folding field and worked as postdoctoral research fellow in the Department of Chemistry and Biochemistry, University of Delaware, USA for three years. Then, He joined department of Biological Science, BITS Pilani as lecturer in 2004, after that promoted to Assistant Professor (2006-2012) and then Associate Professor at the same department (2013-Till date). Apart from performing his academic duties he is also heading over as Chief Warden of BITS, Pilani, Pilani campus. His broad research area is Protein folding, Modelling, Molecular evolution and Bioinformatics.

Biography of Swarna Kanchan

Mr. Swarna Kanchan has done his Bachelor's in Fisheries Sciences (Four years Degree Programme) from Rajendra Agricultural University, Pusa, Bihar and Master's in Bioinformatics from University Institute of Engineering and Technology, Chatrapati Sahuji Maharaj University, Kanpur. He has worked as Research Associate at Amity Institute of Biotechnology, Amity University, Noida before joining BITS, Pilani. His doctoral thesis is entitled "Understanding the protein evolution: a genomic and modeling study" which he is completing under the guidance of Dr. Shibasish Chowdhury. He has also qualified CSIR-UGC National Eligibility Test (NET) for Lectureship in 'Life sciences' category twice in 2008. During the period of Ph.D. research, he was awarded BITS Pilani Research Fellowship and BSR Fellowship from UGC, New Delhi. His research interest includes Molecular evolution, Molecular modeling, Genomics and Structural bioinformatics.