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

THESIS SYNOPSIS

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

CHAPTER 1: INTRODUCTION

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. Mutations, migration, natural selection and genetic drift are four major elements that drive evolution. However, in the process of evolution, genetic variation is essential within the population. Mutation is one of the main source of genetic variation. Most of the mutations that related to evolution have happened naturally during DNA replication process of cell division. 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. However, cellular repair system always tries to restore the integrity of genome. 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. The molecular mechanisms of DNA repair can be classified into four main categories: direct repair (DR), nucleotide excision repair (NER), mismatch repair (MMR) and base excision repair (BER) (Dmitry & Kosuke 1997). 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. AP endonuclease protein family is an important group of proteins that involves during BER pathway which is one of the busiest repair pathway (Wilson & Barsky 2001). AP endonucleases are responsible for recognition and nicking of AP sites. However, limited evolutionary studies have been found among the AP endonuclease protein family (Denver et al. 2003; Georgiadis et al. 2008; Eisen & Hanawalt 1999). Most studies had considered AP endonuclease family as a part of higher super family. In the present study, we want to explore 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 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:

- 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 2: METHODOLOGY

AP endonuclease homologs are retrieved through BLAST (Altschul et al. 1990) and PSI-BLAST (Altschul et al. 1997) programmes using *E.coli* AP endonuclease protein as query sequence. Distant homologs was further searched by BLASTP programme using homolog with E value=e^{-5} as query. Protein sequences for all the homologs were downloaded from databases and then global pairwise alignment was carried out between *E.coli* AP endonuclease protein sequence with all other protein homologs and sequences having identity ≥15% were kept for further analysis. Protein sequences for homologs were analyzed for domains/motifs by CD search tool (Marchler & Bryant 2004). and those which had at least one domain common to *E.coli* AP endonuclease were kept for next set of analysis. Neighbor joining (NJ) trees were generated from the protein homologs of each bacterial, archaeal and eukaryotic divisions and divergent sequences from different clades/clusters were selected. Multiple sequence alignment is performed by Clustalw (Thompson et al. 1994) as well as MCoffee server (Moretti et al. 2007) and sequence logos are created by Weblogo 3.2 (Crooks et al. 2004). NJ and Maximum likelihood (ML) trees were generated by MEGA 5.1. Homology models of AP endonuclease homologs were built by Modeller 9v14.16S/18SrRNA gene sequences of bacterial, archaeal and eukaryotic species were retrieved from ribosomal sequence databases (Cole et al. 2009) and were used to generate species tree. cDNA of AP endonuclease homologs was retrieved from the NCBI sequence database.

CHAPTER 3: EVOLUTION OF ENDONUCLEASE III PROTEIN FAMILY

This chapter focuses on the evolution of the endonuclease III gene/protein family, which plays a key role in the base excision repair pathway. We analyzed 463 homologs of the endonuclease III protein and compared them with the corresponding gene and 16S/18S rRNA sequences to understand the evolutionary processes of this protein family. The sequence analysis reveals three consensus sequence motifs within the ENDO3c and one consensus sequence in iron–sulfur cluster loop motif that are functionally and structurally important.

Comparison of endonuclease III gene based and 16S/18S rRNA gene based phylogenetic tree suggests the evolution of endonuclease III genes shaped up differently from species evolution in most of the cases. Endonuclease III gene based phylogeny tree suggests that GC content of gene contribute significantly towards the position of taxa within the tree due to synonymous nucleotide substitution in GC content at third codon position as well as the overall GC content of gene.

We propose an evolutionary model of the endonuclease III protein family. Horizontal gene transfer was identified as the key event among bacteria, archaea, and eukaryotic organisms that occurred during the evolution of the endonuclease III gene family among bacteria, archaea, and eukaryotic organisms. Moreover, insertion of signal peptide at N terminal region of endonuclease III proteins of *Oryza sativa* and *Arabidopsis thaliana* seems to have taken place during the evolution of plant homologs which suggest a possible endosymbiotic transfer events from bacteria (Cyanobacteria) to plants. The basic fold of endonuclease III protein is conserved across different clads.

CHAPTER 4: EVOLUTION OF ENDONUCLEASE IV PROTEIN FAMILY

The evolution process of the endonuclease IV gene/protein family was examined in chapter 4. We have analyzed 402 homologs of the endonuclease IV protein family. Our analysis reveals four consensus sequence motifs within the AP2EC domain which are functionally important. We also observed that the species and endonuclease IV gene evolution shape up differently in most of the homologs. Endonuclease IV gene based phylogeny tree suggests that GC content at third codon position of endonuclease IV gene as well as overall GC content of endonuclease IV gene plays major role in positioning species from same division/kingdom in different clades of the phylogenetic tree due to synonymous nucleotide substitution in GC content at third codon position as well as overall GC content.

On the basis of our analysis, we propose an evolutionary model of the endonuclease IV protein family. Horizontal gene transfer was identified as the key event that occurred during the evolution of the endonuclease IV gene family among bacteria, archaea, and eukaryotic organisms. Absence of AP endonuclease (endonuclease IV) homologs in higher eukaryotes (beyond nematode) suggests the possibility of loss of endonuclease IV gene in higher eukaryotes during the course of evolution. 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. Evolutionary changes among various clades in protein based phylogenetic tree were investigated by comparison of homology models which suggested the conservation of overall fold of endonuclease IV proteins except the minor changes in fold in endonuclease IV in *Thermococcus*.

CHAPTER 5: EVOLUTION OF EXONUCLEASE III PROTEIN FAMILY

The evolution of the exonuclease III gene/protein family is discussed in chapter 5. We analyzed 404 homologs of the exonuclease III protein family and compared them with the corresponding gene and 16S/18S rRNA sequences to understand the evolutionary processes of this protein family. The sequence analysis reveal four consensus sequence motifs within the Xtha domain that is functionally and structurally important. Two motifs called as Zinc finger and SAP motif, which were new to exonuclease protein family were also observed in exonuclease III gene/protein family homologs. Exonuclease III gene based phylogeny tree suggests that GC content of gene contribute significantly towards the position of taxa within the tree due to synonymous nucleotide substitution in GC content at third codon position as well as overall GC content of exonuclease III gene tree. Comparison of exonuclease III gene based and 16S/18S rRNA gene based phylogenetic tree suggests the evolution of Exonuclease III genes shaped up differently from species in most of the cases.

The evolutionary model suggests that horizontal gene transfer was a key event among bacteria, archaea, and eukaryotic organisms which is taken place during the evolution of the exonuclease III gene family. In higher organisms, two homologs of exonuclease III, called Ape1 (Hap1) and Ape2 have been found in *Homo sapiens* and *Mus musculus* suggested the gene duplication in exonuclease III gene family. 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. To map the evolutionary changes among all clades in protein based phylogenetic tree we generated the homology models and comparison of these homology models suggested the conservation of overall fold of exonuclease III proteins except the minor changes in fold in Ape2 in *Homo sapiens* due to insertion of loop regions at three places.

CHAPTER 5: CONCLUSIONS AND FUTURE PERSPECTIVES

Gene loss, gene duplication, horizontal gene transfer, endosymbiotic transfer were observed as important evolutionary event within AP endonuclease protein family. Insertion of DNA binding motifs are also observed in few homologs AP endonuclease protein family. We also noticed that the evolution of species and AP endonuclease genes shape up differently in most of the cases. The average GC content and GC 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 GC content and GC content at the third codon were biased to stay together in a clade. Gene sequence based phylogenetic analysis is much noisier than protein sequence based analysis. The later analysis provides more reliable picture of evolution. Preliminary protein modeling results suggests that homologs of AP endonucleases are structurally more conserved than sequence. Mainly folds and functionally active sites are conserved during evolution. 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

This study is focused on sequence based analysis 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 sapien*. 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, we observed that the duplicated six MUTS paralogs in eukaryotic organisms designated as MSH1–MSH6 have different combinations of domains where different combinations of paralogs (heteromers) takes part in various DNA mismatch and non-mismatch repair functions.

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).

REFERENCES:

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., 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.

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

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.

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

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

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

Marchler, BA., Bryant, SH. (2004). CD Search: protein domain annotations on the fly. Nucleic Acids Research, 32(Web Server issue), W327-331.

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.

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.

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.

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.

LIST OF PUBLICATIONS:

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, Chapter 19, Springer-Verlag, Germany, pp. 541-569.
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]