Evolution of endonuclease IV protein family: an in silico analysis

Swarna Kanchan¹ · Parva Sharma¹ · Shibasish Chowdhury¹

Received: 13 November 2018 / Accepted: 28 March 2019 / Published online: 6 April 2019
© King Abdulaziz City for Science and Technology 2019

Abstract

DNA repair is one of the key cellular events which balances between evolvability and integrity of the genome. Endonuclease IV enzymes are class II AP endonucleases under base excision repair pathway which act on abasic site and break the phosphodiester bond at the 5' side. The role and activity of endonuclease IV proteins vary among different organisms; even it is absent in higher eukaryotes. The evolution of this protein family was studied by analyzing all homologs of the endonuclease IV protein family through different in silico techniques including phylogenetic tree generation and model building. The sequence analysis revealed four consensus sequence motifs within the AP2EC domain which are functionally important and conserved throughout the evolution process. It was also observed that the species and endonuclease IV gene evolution shape up differently in most of the organisms. Presence of the mitochondria-targeted signal peptides in fungal species *Saccharomyces* and *Coccidioides* suggest a possible endosymbiotic transfer of *endonuclease IV* genes to lower eukaryotes. Evolutionary changes among various clades in the protein-based phylogenetic tree have been investigated by comparison of homology models which suggests the conservation of overall fold of endonuclease IV proteins except for few alterations in loop orientation in few clades.

Keywords AP endonuclease · Endonuclease IV · Evolution · Phylogenetic analysis · Horizontal gene transfer

Introduction

AP (apurinic/apyrimidinic) endonucleases act on abasic site and break the phosphodiester bond at the 5' side which is considered as the first step of base excision repair pathway (BER). Subsequently, enzymes like DNA deoxyribose phosphodiesterase, DNA polymerase and DNA ligase restore original DNA sequence and complete the repair process (Chatterjee and Walker 2017). Based on the position of the phosphodiester cleavages, all AP endonucleases can be classified into two major classes. Class I AP endonucleases cleaves DNA at the 3' side of the AP site while class II AP endonucleases enzymes cleave at the 5' side. Class II AP

endonucleases contain exonucleases III and endonucleases IV enzyme families (Daley et al. 2010).

The role and activity of the endonucleases IV enzyme vary among different organisms. For example, it is demonstrated that in *E. coli*, exonuclease III accounts for 90–95% of total cellular AP endonuclease activity while endonuclease IV accounts for 5–10% AP activity of a cell (Mol et al. 2000). However, in *Mycobacterium tuberculosis*, the majority of AP endonuclease activity is performed by endonucleases IV (Puri et al. 2013) while this protein is absent in higher eukaryotes. Both endonucleases IV and exonuclease III enzymes were observed to work as mutual back-ups as gene mutation studies have shown that entire AP activity was performed by one group of enzymes in the 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). However, it is also shown that endonuclease IV is the only known enzyme that is active against damaged nucleotides with deoxyadenosine base (Ide et al. 1994) and can recognize a set of oxidized pyrimidines. It is

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s13205-019-1696-6>) contains supplementary material, which is available to authorized users.

✉ Shibasish Chowdhury
shiba@pilani.bits-pilani.ac.in

¹ Department of Biological Sciences, Birla Institute of Technology and Science, Pilani, Pilani, Rajasthan 333031, India

demonstrated that the expression of endonuclease IV could be enhanced up to 20 fold in the presence of superoxide-generating agents, such as paraquat (Chan and Weiss 1987). Endonuclease IV protein is encoded by the *nfo* gene of bacteria (Richardson and Kornberg 1964), *APN-I* gene of fungi (Ramotar et al. 1998) and *CeAPN1* gene of nematode (Masson et al. 1996).

Zn²⁺ dependent single domain endonuclease IV directly participates in phosphodiester bond cleavage (Banner et al. 1975; Hosfield et al. 1999). During the enzymatic activity, five loops of enzyme recognize dsDNA and are considered to be crucial for enzymatic activity.

This study aims to provide insight into the evolution process of endonuclease IV gene/protein family. In this study, we have focused on the 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. Evolutionary analysis can also be further used for improved prediction of the functions of uncharacterized genes in endonuclease IV gene/protein family.

Materials and methods

Retrieval of endonuclease IV protein homologs

Escherichia coli endonuclease IV protein sequence (NCBI Acc. No. YP_541433.1) was used as a query to retrieve homologs from the National Center for Biotechnology Information (NCBI) protein sequence database. Blastp (Altschul et al. 1990) with default parameters was used for retrieving homologous sequences of these proteins from NCBI non-redundant (NR) protein database (with *E* value cut off of $\leq 1 \times 10^{-5}$). In addition to BlastP, homologous protein sequences were also searched by two round of PSI-Blast (Altschul et al. 1997) search using the same query sequence.

cDNA sequences of endonuclease IV protein homologs were retrieved from the NCBI sequence database and were used to construct a gene-based phylogenetic tree. Mean G+C content, and G+C content at 3rd codon position of the cDNA sequences of endonuclease IV was calculated by the CAIcal server (Puigbo et al. 2008). G+C contents for the genomes of the species were retrieved from the NCBI genome database. 16S ribosomal RNA gene sequences of bacterial and archaeal species were retrieved from the ribosomal sequences database (Cole et al. 2009). 18S ribosomal RNA gene sequences of eukaryotic species were downloaded from the NCBI nucleotide database.

Sequence alignment, conservation, and domain finding

Pairwise sequence alignments between *E. coli* endonuclease IV and other selected endonuclease IV homologs were executed by EMBOSS Needle programme (Rice et al. 2000). Multiple protein sequence alignment of endonuclease IV protein homologs were conducted with Mcoffee server (Moretti et al. 2007) by keeping default parameters. Endonuclease IV homologs were analyzed for their domains and motifs by conserved domain (CD) search tools available at NCBI (Marchler-Bauer and Bryant 2004).

The conservation patterns of amino acids positions based on protein sequence in the endonuclease IV protein family were analyzed by sequence logos. Sequence logos are a graphical representation of each amino acid position in the multiple sequence alignment of an endonuclease IV protein family. Sequence logo was generated by Web logo 3.2 (Crooks et al. 2004) without compositional bias. 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).

Phylogenetic analysis

Maximum likelihood (ML) and neighbour-joining (NJ) phylogenetic trees were generated by Mega5.1. Evolutionary distances for the gene-based tree were constructed using the maximum composite likelihood method (Tamura et al. 2004) while protein based tree is constructed using Jones Taylor Thornton method (Jones et al. 1992). The reliability of interior branches is assessed with 1000 bootstrap re-samplings of amino acid/cDNA/16S-18 sequences in phylogenetic trees (Felsenstein 1985). The speciation and gene duplication events were analyzed with Mega X server using Zmasek and Eddy algorithm (Zmasek and Eddy 2001).

Results and discussion

Data curation

All non-redundant endonuclease IV homologs of BLAST and PSI-BLAST search were subjected to global pairwise sequence alignment by EMBOSS Needle programme (Rice et al. 2000). Protein sequences (endonuclease IV homologs) 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 are selected. A set of 402 selected homologs of endonuclease IV protein family were divided into five kingdoms, e.g., monera (bacteria and archaea both come under this kingdom), protista, fungi, plantae, and animalia (Whittaker 1969). Since the numbers of bacterial homologs were large in number, sequences from bacterial species were further divided into various divisions (Garritty and Holt 2001).

Phylogenetic trees were generated for homologs of each of the bacterial divisions and four other kingdoms using the neighbour-joining algorithm implemented in the MEGA 5.1 package, (Tamura et al. 2011). Phylogenetic tree of each group is inspected visually, and most diverse homologs from different clusters of each tree were chosen. Finally, a set of 52 homologs of endonuclease IV were selected to study the evolution of endonuclease IV protein family (shown by supplementary Table 1). Out of total 52 non-redundant endonuclease IV homologs from 52 different species, 25 homologs were endonuclease IV proteins, 26 homologs were AP endonuclease/AP endonuclease 1 (APN1) proteins, and one homolog is deoxyribonuclease IV protein. This set of protein homologs were from 42 bacterial, four archaeal and six eukaryotic species. The details of selected homologs are listed in Tables 1 and 2.

Sequence, motif, and phylogeny

AP2Ec (cd00019) is the 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 interacting with phosphate groups of DNA substrate. Multiple sequence alignment (MSA) of 52 homologs shows that H₇ of the first loop; A₁₄₈, G₁₄₉, G₁₅₁ of second loop and S₂₂₆, D₂₂₉, R₂₃₀, H₂₃₁, G₂₃₅, G₂₃₇ of the third loop are highly conserved (more than 80% occasion). The

Table 1 List of different proteins (as it was annotated in the database) retrieved as endonuclease IV homologs

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

Table 2 Division and kingdom wise breakup of endonuclease IV homologs

Division	Kingdom	Non-redundant homologs	Homologs considered for phylogenetic tree
Archaea	Monera	4	48
Bacteria		42	317
Eukaryotes	Fungi	3	26
	Protista	1	8
	Animalia	2	3
Total		52	402

All bacterial species is further divided into 22 different categories (Garritty and Holt 2001)

phosphate group interacting loops of *E. coli* endonuclease IV protein is shown in Fig. 1. Overall, residues of the third phosphate-binding loop are relatively more conserved than those of first and second phosphate-binding loop. Glycines of phosphate binding motifs are crucial for loop conformation which is responsible for non-specific interaction with the 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 H₆₉-X₂-Y₇₂-X₂-N₇₅ within minor groove binding motif at 69–78 amino acid positions where each amino acid position is conserved in more than 80% homologs (Figs. 1, 2a). 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

Fig. 1 The mainchain of three phosphate binding motifs (between residues 7–12, 147–152 and 220–239) within *E. coli* endonucleases IV crystal structure (PDB: 1QUM) is shown by red ribbon representation while the minor groove binding motifs (between 33–45 and 69–78) is shown by blue ribbon. The side chain of conserved residues within these two regions are shown by ball and stick representation

Fig. 2 The sequence conservation within a different region of endonuclease IV protein is shown by sequence logo representation. Each of the numbered amino acids is conserved in more than 80% homologs. **a** Consensus sequence $H_{69}-X_2-Y_{72}-X_2-N_{75}$ within minor groove binding motif at 69–75 amino acid. **b** Consensus sequence $D_{179}-X_2-H_{182}-X_4-G_{187}$ in 9 residues long motif at 179–187 amino acid positions. **c** Consensus sequence $H_{216}-X-N_{218}-D_{219}$ within 216–219 amino acid positions. **d** Consensus sequence $D_{229}-X-H_{231}-X_3-G_{235}-X-G_{237}$ in phosphate binding loop motif at 229–237 amino acid positions

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 than 80% species. Tri-nuclear Zn centre near the active site of the enzyme plays an important role in phosphodiester cleavage by holding catalytically active amino acid residues close together (Hosfield et al. 1999). The 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-nuclear 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 has also identified three consensus sequence segments between 179th and 187th amino acid, 216th and 219th amino acid and between 229th and 237th 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. 2b–d).

Complete 16S/18S rRNA gene sequences of 48 species were retrieved while partial 16S/18S rRNA gene sequence was available for rest of the four species which were not considered to generate the species tree. A total of 15 distinct clades are formed in species tree (Fig. 3) with eukaryotes are the farthest clade within the tree. As it was expected, species from the same bacterial division as well as species from the same kingdom stayed close to each other in a clade. Single species from the closely related bacterial division are

clustered together. For example, clade 5 contains one species each from *Deinococcus*, *Bacteroidetes* and *Chlorobibacterial* division. Similarly, *Tenericutes*, *Chloroflexi*, *Synergistetes*, *Thermotogae* and *Proteobacteria* share clusters with species from *Fusobacteria*, *Thermobaculum*, *Verrucomicrobia*, *Dictyoglomi*, and *Chlamydiae* respectively.

Both maximum likelihood (ML) and neighbour-joining (NJ) trees of endonuclease IV gene sequences generated ten distinct clades with almost similar topology which possibly indicates that phylogenetic incongruence owing to tree-building errors (stochastic or systematic) are minimal. Thus, only the ML tree is discussed in detail.

Endonuclease IV gene tree (shown in Fig. 4) is compared with 16S/18S rRNA gene sequences-based species tree. The bootstrap analysis reveals that most of the clusters in species trees and few clusters in gene tree are robust and these clades are supported by $\geq 50\%$ bootstrap value.

Interestingly, none of the clusters forms a pure archaeal, bacterial (which belong to the same division) or eukaryotic cluster. All the 10 clusters have a mixed population of archaea/bacteria/eukaryotes. However, 16S/18SrRNA sequence-based species tree shows that most of the species in different bacterial subdivisions, as well as species from archaea and eukaryote, formed distinct clades. It is interesting to note that endonuclease IV genes of three archaeal species are positioned within three different clades, although all archaeal species are within the same clade in the species tree.

Phylogenetic incongruence among gene trees is well-known and quite rampant. The observed differences between a gene tree and species tree are mainly attributed to genetic drift in natural populations, gene duplication, horizontal gene transfer or lineage sorting. Even, hidden paralogy (Nei and Rooney 2005), incomplete lineage sorting (Carstens and Knowles 2007) mean G+C content of the gene, as well as G+C content at the third codon position (Brochier et al. 2000), could contribute to the observed tree incongruence.

To understand the tree incongruence or this unusual clustering pattern, i.e., differences in sharing different clades by the species belonging to the same division, we have calculated the mean G+C content of the gene and G+C content at the third codon position which is a well-known tool to explain clustering pattern in the phylogenetic tree (Brochier et al. 2000). G+C content based on endonuclease IV gene, the 3rd position of the endonuclease genes are listed in Table 3.

It is observed that mean G+C content and G+C content at the third codon position of endonuclease IV genes in archaeal species *Ferroplasma* as and bacterial species *Halothermothrix*, *Dictyoglomus*, *Thermobaculum*, *Methyhalacidiphilum* (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 the same clade in

Fig. 3 16S/18S rRNA sequences of species-based maximum likelihood tree. The evolutionary distances were computed using the Tamura and Nei (1993) as substitution model. Bootstrap support values were presented next to the tree branches for each clade with

≥50. Organisms from the same division are represented with the same symbol and colour code. Details are provided in Supplementary Table S2

gene-based phylogeny tree. Similarly, the average G+C content and G+C content at the third codon position of *Halomicrobium mukohataei* (archaeal species in clade 7) are 65.7% and 70.0%, respectively, which differs significantly from rest of the three archaeal species, i.e., *Desulfurococcus kamchatkensis* in clade 6 (44.3% and 51.9%), *Thermococcus* in clade 6 (52.8% and 54.2%) and *Ferroplasma* in clade 2 (38.4% and 68.6%). Hence it is placed with bacterial species (clade 7) with similar G+C content. From these observations, it

is clear that these species reside within the clade that contains species with similar G+C content. Being a member of bacterial division *Chlamydia*, *Mycoplasma* and *Escherichia* shares clade 1 with eukaryotic homologs as its G+C content (39.74%) and GC content at the third codon position (35.21%) of endonuclease IV genes is similar to eukaryotes. Interestingly, one of the eukaryotic species *Ixodes scapularis* shares clade 4 with bacterial homologs *Rhodococcus*, *Candidatus*, *Meiothermus* and *Gemmatimonas* due to the

Fig. 4 Endonuclease IV gene sequences-based maximum likelihood tree. Details are similar to Fig. 2

similar GC content at the gene level and the GC content at the third codon position (63.13% and 77.49%, respectively, with a standard deviation of 3.29 and 6.25).

Endonuclease IV gene-based phylogeny tree of 52 taxa suggests that the GC content at third codon position of endonuclease IV gene as well as mean GC content of

endonuclease IV gene plays a major role in positioning different species in the phylogenetic tree. The species evolution and endonuclease IV gene evolution shapes up differently in most cases. Since endonuclease IV gene in all living organisms diverged over a long evolutionary period, which leads to synonymous nucleotide substitution particularly at third

Table 3 G+C content based on endonuclease IV gene and 3rd position of the endonuclease IV gene is listed

Clade no.	Name of the organism	Division/kingdom	Mean gene G+C content (%)	G+C at 3rd position (%)	Average	
					Mean gene	3rd position
1.	<i>Enterocytozoon bieneusi</i> H348	Fungi	27.49	16.79	39.74	35.21
	<i>Mycoplasma penetrans</i> HF-2	Tenericutes	28.63	17.63	(8.83)	(13.68)
	<i>Chlamydia trachomatis</i> D/UW-3/CX	Chlamydiae	43.83	32.18		
	<i>Caenorhabditis elegans</i>	Animalia	40.05	36.78		
	<i>Escherichia coli</i> UTI89	Proteobacteria	52.45	55.59		
	<i>Coccidioides posadasii</i> C735 delta SOWgp	Fungi	49.13	50.09		
	<i>Saccharomyces cerevisiae</i>	Fungi	38.50	38.59		
2.	<i>Theileria parva</i> strain Muguga	Protista	37.83	34.04		
	<i>Halothermothrix orenii</i> H 168	Firmicutes	35.36	25.45	39.25	33.95
	<i>Dictyoglomus thermophilum</i> H-6-12	Dictyoglomi	34.80	26.47	(4.46)	(7.5)
	<i>Thermobaculum terrenum</i> ATCC BAA-798	Thermobaculum	44.67	41.75		
3.	<i>Ferroplasma acidarmanus</i> fer1	Archaea	38.38	36.62		
	<i>Methylacidiphilum infernorum</i> V4,	Verrucomicrobia	43.06	39.44		
	<i>Fibrobacter succinogenes</i> subsp. <i>succinogenes</i> S85	Fibrobacteres	49.15	51.81	49.01	54.01
	<i>Geobacillus</i> sp. Y412MC61 ctg31	Firmicutes	52.78	64.00	(4.79)	(10.46)
4.	<i>Paenibacillus</i> sp. JDR-2	Firmicutes	51.90	59.95		
	<i>Ruminococcus bromii</i> L2-63	Firmicutes	42.21	40.29		
	<i>Thermanaerovibrio acidaminovorans</i> DSM 6589	Synergistetes	61.39	76.62	63.13	77.49
	<i>Ixodes scapularis</i>	Animalia	58.41	73.65	(3.29)	(6.25)
5.	<i>Rhodococcus jostii</i> RHA1	Actinobacteria	67.57	88.72		
	<i>Candidatus Solibacter usitatus</i> Ellin6076	Acidobacteria	65.64	80.41		
	<i>Meiothermus silvanus</i> DSM 9946	Deinococcus	61.72	72.08		
	<i>Gemmatimonas aurantiaca</i> T-27	Gemmatimonadetes	64.06	73.47		
6.	<i>Gemmata obscuriglobus</i> UQM 2246	Planctomycetes	70.82	75.85	70.52	74.42
	<i>Citricella</i> sp. SE45	Proteobacteria	70.22	73.00	(0.42)	(2.01)
7.	<i>Dehalococcoides ethenogenes</i> 195,	Chloroflexi	51.62	57.40	49.58	54.52
	<i>Desulfurococcus kamchatkensis</i> 1221n	Archaea	44.29	51.90	(4.62)	(2.76)
	<i>Thermococcus kodakarensis</i> KOD1	Archaea	52.84	54.26		
8.	<i>Leptospirillum rubrum</i> sp. Group IV 'UBA BS	Nitrospirae	59.33	64.78	63.78	67.59
	<i>Desulfurivibrio alkaliphilus</i> AHT2	Proteobacteria	62.99	67.24	(5.6)	(3.99)
	<i>Halomicrobium mukohataei</i> DSM 12286,	Archaea	65.70	70.04		
	<i>Pirellula staleyi</i> DSM 6068	Planctomycetes	55.80	60.25		
9.	<i>Roseiflexus castenholzii</i> DSM 13941	Chloroflexi	61.61	68.97		
	<i>Streptosporangium roseum</i> DSM 43021	Actinobacteria	69.01	69.72		
	<i>Catenulispora acidiphila</i> DSM 44928	Actinobacteria	72.06	72.18		
	<i>Aeromonas salmonicida</i> subsp. <i>salmonicida</i> A449	Proteobacteria	60.64	60.99	52.26	58.46
10.	<i>Aliivibrio salmonicida</i> LFI1238	Proteobacteria	40.28	52.65	(7.47)	(4.29)
	<i>Parabacteroides distasonis</i> ATCC 8503	Bacteroidetes	51.89	55.32		
	<i>Chlorobium phaeovibrioides</i> (<i>Prosthecochloris vibrioformis</i>) DSM 265	Chlorobi	55.12	62.86		
	<i>Pelodictyon phaeoclathratiforme</i> BU-1	Chlorobi	53.38	60.50		
9.	<i>Thermoanaerobacterium thermosaccharolyticum</i> DSM 571	Firmicutes	36.93	49.28	44.41	53.31
	<i>Aster yellows</i> witches phytoplasma AYWB	Tenericutes	29.81	39.93	(14.97)	(10.51)
	<i>Leptotrichia buccalis</i> C-1013-b	Fusobacteria	35.28	49.68		
	<i>Eggerthella lenta</i> DSM 2243	Actinobacteria	65.35	66.19		
10.	<i>Akkermansia muciniphila</i> ATCC BAA-835,	Verrucomicrobia	54.68	61.51		
	<i>Sulfurihydrogenibium</i> sp. YO3AOP1	Aquificae	30.00	41.43	35.98	45.79
	<i>Hydrogenivirga</i> sp. 128-5-R1-1	Aquificae	48.93	53.21	(6.90)	(4.71)
	<i>Thermosiphon melanesiensis</i> B1429	Thermotogae	30.87	42.25		
10.	<i>Petrogla mobilis</i> SJ95,	Thermotogae	37.50	44.79		
	<i>Denitrovibrio acetiphilus</i> DSM 12809	Deferribacteres	43.07	50.00		
	<i>Deferribacter desulfuricans</i> SSM1	Deferribacteres	32.39	45.04		

Table 3 (continued)

The average value of the G+C content of gene and the G+C content at 3rd position over a give clade is also listed. The standard deviation is shown within parenthesis

codon position of the endonuclease IV gene as well as the mean G+C content of gene makes gene sequence-based phylogenetic tree construction noisy which was also noticed earlier (Kanchan et al. 2015).

We compare gene tree and species tree to examine possible gene duplication and speciation events which could explain some of the observed tree incongruences. In Fig. 4, the gene duplication and speciation events with high bootstrap support value are mentioned by “S” and “D” symbol, respectively. Our analysis predicts a total of ten speciation events and 6 duplication events. The six duplication events followed by six speciation events. So the remaining speciation events without obvious duplication events possibly indicate hidden paralogy. Even we may not possibly rule out horizontal gene transfer (HGT) events during the endonuclease IV evolution as it is observed that endonuclease IV homologs of archaeal and bacterial species have shared same clade (Clade 2, 6 and 7). Similarly, endonuclease IV protein homologs of eukaryotic species in gene tree shared the same clade with bacterial species (clade 1).

Overall 11 distinct clades are identified from endonuclease IV protein-based phylogenetic tree (Fig. 5). Interestingly, the topology of gene-based and protein-based trees are quite different indicating a strong influence of synonymous nucleotide substitution in the gene level phylogenetic tree. All six eukaryotic species (*Caenorhabditis elegans*, *Ixodes scapularis*, *Saccharomyces cerevisiae*, *Coccidioides posadasii*, *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 clades 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 endonuclease IV protein reveals an insertion of ~137 amino acids in *Theileria parva*, ~116 amino acids in *C.elegans*, ~118 amino acids in *Coccidioides posadasii*, at N-terminal region.

All the endonuclease IV proteins are 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 are predicted within N-terminal region of *Coccidioides* and *Saccharomyces cerevisiae*, respectively. TargetP has also predicted that these signal peptides are localized at mitochondria. Another possible event of importance in endonuclease IV

protein family is the transfer of endonuclease IV genes by endosymbiotic transfer from bacterial species to eukaryotes. *E. coli* (Proteobacteria) is placed with eukaryotic species in the same clade suggests a possible endosymbiotic transfer event from bacteria (Proteobacteria) to eukaryotes. Prediction of the signal peptide in fungal species *Saccharomyces* and *Coccidioides* and positioning of these fungal homologs with proteobacterial species *E. coli* in the same clade suggests the endosymbiotic transfer of endonuclease IV genes from proteobacteria to lower eukaryotes. However, endonuclease IV homolog is not present in higher eukaryotes indicating loss of gene during course of the evolution. Pairwise sequence alignment of endonuclease IV of *E. coli* with AP endonuclease 1 (Ape1) of human (exonuclease IV homolog) shows a very poor sequence identity of 3.8% only. These two proteins were also highly dissimilar in their three-dimensional structures. However, structural and biochemical comparisons of these two enzymes showed that APE1 catalytic residues were functionally similar to endonuclease IV active site components (Tsutakawa et al. 2013) which suggest that the endonuclease IV function is possibly carried by an Ape1 in human.

Structural evolution of endonuclease IV homologs

Among the homologs of an endonuclease IV protein family, the structures of *E. coli* (PDB ID: 1qum) and *Geobacillus sp.* (PDB ID: 3aal) were solved experimentally. Both the structures belonged to $\alpha_8\beta_8$ TIM barrel fold with main chain RMS deviation of around 1.39 Å (shown in Fig. 6) although the sequence identity between these two structures is only 32.3% which indicates that structural fold plays an important role in the enzymatic activity. The conformations of loops in these two structures mostly deviate from each other. As it is observed in the superimposed structure (Fig. 6), 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 modelled representative sequence of remaining 9 distinct clades. Homology model of nine representative endonuclease IV protein sequences is generated using *E. coli* endonuclease IV crystal structure as the template structure. The structure quality of each model is shown in Table 4. The RMSD values of modelled structures with their template structure are in the range of

Fig. 5 Endonuclease IV protein sequences-based maximum likelihood tree. The evolutionary distances were computed using the JTT as substitution model. Bootstrap support values were presented next to the tree branches for each clade with ≥ 50

0.47–1.63 Å. Among all the modelled structures homolog of *Halothermothrix* (Fig. 7a) shows the closest match with the template structure whereas *Thermococcus* (Fig. 7b) shows the largest deviation from the template crystal structure. Each model is also validated by Ramachandran plot which shows that these protein structures are stereochemically stable as 99% of all amino acids are within favourable regions

of Ramachandran plot. The QMEAN scores and QMEAN Z-scores also indicated that the qualities of the models were reasonably good. A close inspection of modelled structures reveals that though the sequence identity between template and target sequences varies between 25 and 44%, the organization of helices and β -strands within the protein core area is conserved. During the evolution of this protein, insertion/

Fig. 6 Superimposition of the crystal structures of endonuclease IV proteins of *E. coli* (red) and *Geobacillus* (cyan). The main chains of both the structures are shown by ribbon representation. 3D diagrams of these models were generated by Accelrys Discovery studio ViewerPro 5.0

deletion is observed majorly on those loop regions which are structurally away from DNA and Zn binding sites. Thus, shortening or elongation of loops does not possibly affect protein function though it can alter the positions of connecting helix or β -strand (Fig. 7b). Overall, the orientation of the majority of loops differs marginally indicating possible rigidity of this enzyme structure over the long evolutionary period.

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 the conservation of amino acid positions, four consensus sequences have been identified for minor groove motif and other regions of importance. Endonuclease IV protein-based phylogenetic tree reveals possible HGT events from bacteria to archaea. Endosymbiosis events may exist during endonuclease IV evolution. Modelling studies indicate that structural fold of endonuclease IV protein family is

Table 4 Endonuclease IV protein is modelled using *E. coli* Endonuclease IV (PDB ID: 1qum) as a template

Species name (Clade no.)	% Identity with the template	RMSD with template structure (in Å)	Ramachandran plot statistics	Errat 2 score	Qmean score/Z-score
<i>Saccharomyces</i> (Clade 2)	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)
<i>Gemmata</i> (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)
<i>Halothermothrix</i> (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)
<i>Petrotoga</i> (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)
<i>Ruminococcus</i> (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)
<i>Hydrogenivirga</i> (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)
<i>Thermobaculum</i> (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)
<i>Rhodococcus</i> (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)
<i>Thermococcus</i> (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)

Quality of the modelled endonuclease IV homologs is evaluated through Ramachandran plot, Errat server and Qmean server

Fig. 7 Endonuclease IV crystal structure (PDB ID:1qum, Red) is superimposed on **a**. Homology modelled structure of endonuclease IV of *Halothermothrix* (Closest, Blue) **b**. Model structure of endonuclease IV in *Thermococcus* (Farthest, Green). Only backbone of the protein structure is shown

quite rigid and conserved during the long evolutionary period.

Acknowledgements The authors acknowledge the financial support from the University Grant Commission, Government of India in the form of Basic Science Research fellowship.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215(3):403–410
- Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25(17):3389–3402
- Banner DW, Bloomer AC, Petsko GA, Phillips DC, Pogson CI, Wilson IA, Corran PH, Furth AJ, Milman JD, Offord RE, Priddle JD, Waley SG (1975) Structure of chicken muscle triose phosphate isomerase determined crystallographically at 2.5-angstrom resolution using amino acid sequence data. *Nature* 255(5510):609–614
- Brochier C, Philippe H, Moreira D (2000) The evolutionary history of ribosomal protein RpS14: horizontal gene transfer at the heart of the ribosome. *Trends Genet* 16(12):529–533
- Carstens BC, Knowles LL (2007) Estimating species phylogeny from gene-tree probabilities despite incomplete lineage sorting: an example from *Melanoplus* Grasshoppers. *Syst Biol* 56(3):400–411
- Chan E, Weiss B (1987) Endonuclease IV of *Escherichia coli* is induced by paraquat. *Proc Natl Acad Sci USA* 84(10):3189–3193
- Chatterjee N, Walker GC (2017) Mechanisms of DNA damage, repair, and mutagenesis. *Environ Mol Mutagen* 58(5):235–263
- 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 Res* 37(Database issue):D141–D145
- Crooks GE, Hon G, Chandonia JM, Brenner SE (2004) WebLogo: a sequence logo generator. *Genome Res* 14(6):1188–1190
- Cunningham RP, Saporito SM, Spitzer SG, Weiss B (1986) Endonuclease IV (nfo) mutant of *Escherichia coli*. *J Bacteriol* 168(3):1120–1127
- Daley JM, Zakaria C, Ramotar D (2010) The endonuclease IV family of apurinic/apyrimidinic endonucleases. *Mutat Res* 705:217–227
- Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39(4):783–791
- Garrity GM, Holt JG (2001) The roadmap to the manual. In: Boone DR, Castenholz RW, Garrity GM (eds) *Bergey's manual of systematic bacteriology. The Archaea and the deeply branching and phototropic bacteria*, vol 1, 2nd edn. Springer, New York, pp 119–166
- 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
- Ide H, Tedzuka K, Shimizu H, Kimura Y, Purmal AA, Wallace SS, Kow YW (1994) Alpha-deoxyadenosine, a major anoxic radiolysis product of adenine in DNA, is a substrate for *Escherichia coli* endonuclease IV. *Biochemistry* 33(25):7842–7847
- Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* 8(3):275–282
- Kanchan S, Mehrotra R, Chowdhury S (2015) In silico analysis of the endonuclease III protein family identifies key residues and processes during evolution. *J Mol Evol* 81(1–2):54–67
- Marchler-Bauer A, Bryant SH (2004) CD-Search: protein domain annotations on the fly. *Nucleic Acids Res* 32(Web Server issue):W327–W331
- 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(2):291–293
- Mol CD, Hosfield DJ, Tainer JA (2000) Abasic site recognition by two apurinic/apyrimidinic endonuclease families in DNA

- base excision repair: the 3' ends justify the means. *Mutat Res* 460(3–4):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 Res* 35(Web Server issue):W645–W648
- Nei M, Rooney AP (2005) Concerted and birth-and-death evolution of multigene families. *Annu Rev Genet* 39:121–152
- Puigbo P, Bravo IG, Garcia-Vallve S (2008) CAIcal: a combined set of tools to assess codon usage adaptation. *Biol Direct* 3:38
- Puri RV, Singh N, Gupta RK, Tyagi AK (2013) Endonuclease IV is the major apurinic/apyrimidinic endonuclease in *Mycobacterium tuberculosis* and is important for protection against oxidative damage. *PLoS ONE* 8(8):e71535
- 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. *Mol Cell Biol* 11(9):4537–4544
- Ramotar D, Vadnais J, Masson JY, Tremblay S (1998) Schizosaccharomyces pombe apn1 encodes a homologue of the *Escherichia coli* endonuclease IV family of DNA repair proteins. *Biochim Biophys Acta* 1396(1):15–20
- Rice P, Longden I, Bleasby A (2000) EMBOSS: the European molecular biology open software suite. *Trends Genet* 16(6):276–277
- Richardson CC, Kornberg A (1964) A deoxyribonucleic acid phosphatase-exonuclease from *Escherichia coli*. I. Purification of the enzyme and characterization of the phosphatase activity. *J Biol Chem* 239:242–250
- Tamura K, Nei M (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol* 10(3):512–526
- Tamura K, Nei M, Kumar S (2004) Prospects for inferring very large phylogenies by using the neighbour-joining method. *Proc Natl Acad Sci USA* 101(30):11030–11035
- 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. *Mol Biol Evol* 28(10):2731–2739
- 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. *J Biol Chem* 288(12):8445–8455
- 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
- Zmasek CM, Eddy SR (2001) A simple algorithm to infer gene duplication and speciation events on a gene tree. *Bioinformatics* 17(9):821–828