BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net *Hypothesis*

Volume 6(4)

AP-APSE dpol intein: A novel family A DNA polymerase intein domain

Senguttuvan Rajarajan, Kalibulla Syed Ibrahim, Shunmugiah Karutha Pandian*

Bioinformatics Infrastructure Facility, Department of Biotechnology, Alagappa University, Karaikudi-630003, India; Shunmugiah Karutha Pandian - Email: sk_pandian@rediffmail.com; Phone: +91 4565 225215; Fax: +91 4565 225202; *Corresponding author

Received September 28, 2010; Accepted February 03, 2011; Published May 07, 2011

Abstract:
Inteins are "protein introns" that remove themselves from their host proteins through an autocatalytic protein-splicing. After their discovery, inteins have been quickly identified in organisms from all three kingdoms of life - eucarya, bacteria and archaea, but their distribution is sporadic. Here we report the identification and bioinformatics characterization of intein in DNA polymerase A gene of bacteriophage APSE (*Acyrthosiphon pisum* Secondary Endosymbiont bacteriophage) infecting the Aphid secondary endosymbionts of eukaryotic insects such as Acyrthosiphon *pisum, Uroleucon rudbeckiae*. The insertion site of intein within APSE family A DNA polymerase extein was identified to be dpola. Hence we propose this as a unique intein of family A DNA polymerase (dpola insertion site) and only reported intein in podoviridae family.

Background:
Bacteriophage genomes are in a state of constant flux, reflecting a propensity to undergo rampant recombination **[1]**. As such, phages are an important vector for horizontal gene transfer (HGT) within and between bacterial species **[2]**. In addition, integrated temperate phage (prophage) frequently constitutes the single greatest source of genomic variation among closely related bacterial strains **[3]**. The lysogenic bacteriophage APSE infects *Candidatus Hamiltonella defensa* which is a facultative endosymbiont of aphids and other sap-feeding insects. This endosymbiont has established a beneficial association with aphids, increasing survivorship following attack by parasitoid wasps. Different strains of APSE phages have been found in aphids other than *Acyrthosiphon pisum* **[4]**. The Inteins are basically 'protein introns' found inside other proteins which remove themselves from their host proteins post-translationally by autoproteolytic protein-splicing reaction **[5]**. Besides their protein splicing activity, many inteins also have endonucleolytic activity which is believed to mediate the homing of their genes to corresponding unoccupied integration sites. Both these functions benefit the inteins in two ways. One, protein splicing averts deleterious effects caused by their insertion into the protein hosts and next homing disseminates inteins by horizontal gene transfer. Inteins therefore appear to be selfish genetic elements **[6]**. Since their first discovery in 1990 **[7]**, inteins have been identified in a wide variety of organisms, including bacteria, archaea, and unicellular eukaryotes, with sporadic distribution **[8]**. Similarly, they are observed in many unrelated bacterial clades, but appear often limited to several species within each clade. It was suggested that viruses were potential "vectors" of inteins across species and are responsible for the sporadic distribution of inteins **[9]**. Accordingly, inteins have been identified in many bacteriophages and prophages. Here, we describe the first report of DNA polymerase A intein at dpola insertion site in a bacteriophage genome.

Methodology:
Sequence homology searches were carried out using BLAST programs **[10]** against the NCBI Protein database and the New England Biolabs Intein Database **[11]**. Intein sequence motifs were identified by comparing with the other motifs using EMBOSS pairwise alignment **[12]** available at EBI. The Intein sequences and annotation data for the APSE 2, 4 & 5 was deposited in intein database (InBase, http://www.neb.com/neb/inteins.html; Perler, 2002) **[11]**. Multiple sequence alignments were generated with the help of PRALINE **[13]** using the homology extended alignment strategy **[14]**. Three PSI-BLAST iterations were carried out against all non redundant databases with the e value cut off set to 0.01. The alignment file was retrieved for further phylogenetic tree construction using ClustalX version 2.0 **[15]**. The tree thus generated is provided in **Supplementary material**. Conserved sequences in other DNA polymerase A family and APSE were identified using PRALINE program using standard progressive strategy **[16]**.

Results:
APSE dpol intein:
While searching for inteins in sequence databases with previously identified intein blocks (NCBI BLAST), we identified APSE family A DNA polymerase with some high homology. Hence we proceeded further to analyse for presence of any intein. Sequence homology searches (BLAST against NCBI non redundant protein databases) with the DNA polymerase family A gene sequences of APSE 2, 4 & 5 (retrieved from NCBI protein with accession numbers APSE 2- YP_002308549, APSE 4- ACJ10154.1, APSE 5- ACJ10149.1 respectively) showed a high homology against previously identified DNA polymerase A inteins from salmonella phages (SETP3, SETP5& SETP12), some other phages like *Glycaspis brimblecombei* secondary endosymbiont phage1, Bordetella phage BPP-1 and prokaryotic DNA polymerases like *Escherichia coli, Yersinia pseudotuberculosis*. However APSE 2, 4 & 5 family A DNA polymerase is much larger than APSE 3 by approximately 300 amino acids. Their pairwise alignment with other DNA polymerase family A sequences revealed an extraneous segment. Focusing on this extra segment we identified a 306 amino acid intein in APSE dpolA at position S608 in APSE 2, L388 in APSE 4 & 5. Interestingly this intein segment was not found in dpolA gene of APSE 3, but APSE 1 intein was previously reported in InBase **[11]**. The absence of intein in the APSE-3 DNA polymerase family A is quite surprising and the reason behind that is unknown.

Though salmonella phage consists of homologous DNA polymerase family A intein, the insertion site between APSE (dpola insertion site) and salmonella phage (dpolb insertion site) differs. This makes the APSE DNA polymerase family A intein to be unique in that family. Removing this 306 aa intein region from dpolA sequence of APSE 2, 4 & 5 showed a better optimal alignment with the other counterparts.

```
AP_APSE1_dpol_intein
S|CLAKGTLVLTITGWMPIEIVSQDAYVWDGIEWVRTDGSVFNGNQEVIQAYGVGMTADHQVLTE
KGWKSASQSKRYNRSSCRLPDGYKLPRFRRKEINLESTLHLWTRNNHSSNRITKTKKTRYNCLLR
MPKGTNNIMQKPKARNVKTPRFCCMEQHVSQMYSPFPQSMVKLWWSGNNGLQTLAKKFQQFLGRH
GQDIPTRLIFRSHQQQCRLPPQKLPLGYVASTSSKYSTSTIRANSPRHNEYTGISSPNRDCSKHA
LLSPGKKGKSSTTSGAPKHIAEVYDLINCGPRNRFVIATPDGPLIVHN|C

AP_APSE2_dpol_intein
S|CLAKGTLVLTITGWMPIEIVSQDAYVWDGIEWVRTDGSVFNGNQEVIQAYGVGMTADHQVLTE
KGWKSASQSKRYNRSSCRLPDGYELPRFRRKEINLESTLHLWTRNNHSSNRITKTKKTRYSCLLR
MPKGTNNIMQKPKARNVKTPRFCCMEQHVSQMYSPFPQSMVKLWWSGNNGLQTLAKKFQQFLGRH
GQDIPTRLIFRSHQQQCRLPPQKLPLGYVASTSSKYSTSTIRANSPRHNEYTGISSPNRDCSKHA
LLSPGKKGKSSTTSGAPKHIAEVYDLINCGPRNRFVIATPDGPLIVHN|C

AP_APSE4_dpol_intein
L|CLAKGTLVLTITGWMPIEIVSQDAYVWDGIEWVRTDGSVFNGNQEVIQAYGVGMTADHQVLTE
KGWKSASQSKRYNRSSCRLPDGYELPRFRRKEINLESTLHLWTRNNHSSHRITKTKKTRYSCLLR
MPKGTNNIMQKPKARNVKTPRFCCMEQHVSQMYSPFPQSMVKLWWSGNNGLQTLAKKFQQFLGRH
GQDIPTRLIFRSHQQQCRLPPQKLPLGYVASTSSKYPTSTIRANSPRHNEYTGISSPNRDCSKHA
LLSPGKKGKSSTTSGAPKHIAEVYDLINCGPRNRFVIATPDGPLIVHN|C

AP_APSE5_dpol_intein
L|CLAKGTLVLTITGWMPIEIVSQDAYVWDGIEWVRTDGSVFNGNQEVIQAYGVGMTADHQVLTE
KGWKSASQSKRYNRSSCRLPDGYELPRFRRKEINLESTLHLWTRNNHSSHRITKTKKTRYSCLLR
MPKGTNNIMQKPKARNVKTPRFCCMEQHVSQMYSPFPQSMVKLWWSGNNGLQTLAKKFQQFLGRH
GQDIPTRLIFRSHQQQCRLPPQKLPLGYVASTSSKYPTSTIRANSPRHNEYTGISSPNRDCSKHA
LLSPGKKGKSSTTSGAPKHIAEVYDLINCGPRNRFVIATPDGPLIVHN|C
```

Figure 1: The APSE1, 2, 4 & 5 DNA Polymerase family A inteins. The inteins are named according to the intein nomenclature. The inteins are shown along with their N terminal extein, C terminal extein and its splice junction. Conserved intein sequence motifs are indicated by bold letters.

Canonical blocks of intein:

A BLAST search for DNA polymerase family A intein against InBase reported several hits with known inteins. But a high sequence similarity of over 99% was observed with Salmonella phage intein family A polymerase. The specific blocks of those inteins were used to identify the APSE dpolA intein blocks and further splice junctions were also characterized using pairwise alignment **(Figure 1)**. The splice junction amino acids were also identified to be similar to the previously identified inteins and held to the specific intein motifs. There is a difference in the conservation of splice junction amino acid residues when compared to inteinless prokaryotic DNA polymerase family A and which was found to be 2% difference in the base composition (GC %) between the intein and the gene, though these differences are not that significant still these can be a valid reason to say that the inteins are recently acquired **[17]**.

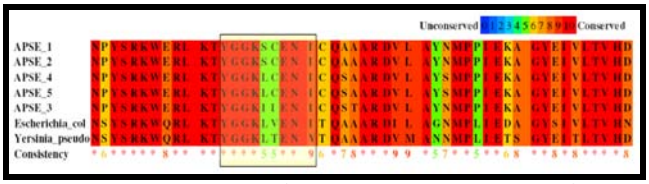


Figure 2: Sequence alignment of APSE phage DNA polymerase family A protein excluding the intein with other prokaryotic DNA polymerase family A shows the presence of conserved region around the insertion site (YGGK*|*ENI). The splice junction amino acids (-1 and +1 extein residues) shows conservation.

Identification of conserved region at insertion site:

Upon removing the intein segment from dpolA protein of APSE 1, 2, 4 & 5 and multiple sequence alignment along with other inteinless dpolA proteins of APSE3, *E. coli and Y. pseudotuberculosis* revealed a highly conserved amino acid domain exactly around the insertion site of the intein (YGGK*|*ENI) **(Figure 2)**.

APSE intein belong to a specific prototype allele:

Perler *et al*. (1997) observed that inteins present in the same location within homologous genes ("intein alleles") tend to be more similar with each other than with inteins in different locations of the same gene or in different genes. This phenomenon appears not only the simple consequence of regular vertical transmission of inteins, but also the result of lateral acquisitions through "homing" at the same site of highly similar genes (i.e. "alleles") by the mechanism involving gene conversion **[8]**. Neither Salmonella phage (SETP3, SETP5 and SETP12) inteins nor APSE (1, 2, 4 & 5) has a homing endonuclease domain or any endonuclease activity. The absence of homing endonuclease domain nullifies the theory of any lateral transmission to this or from this phage. Still it is tempting to say that these DNA phages might act as central reservoir for transmission of the inteins at least to the level of its host as it was already proved that bacteriophages are a central mechanism for HGT in bacteria **[2]**. Moreover, these APSE phages may act as a conduit for ongoing gene exchange among heritable endosymbionts **[4]**. Multiple sequence alignment of APSE intein and its only identified DNA polymerase family A counterpart salmonella phage showed difference in the insertion site of the intein. From that it was identified that APSE 1, 2, 4 & 5 DNA polymerase family A intein insertion site was dpola rather than that of SETP 3, 5 & 12 which belong to dpolb insertion site. This makes the APSE intein unique and the only identified Family A DNA polymerase intein in dpola insertion site **(Figure 3)**.

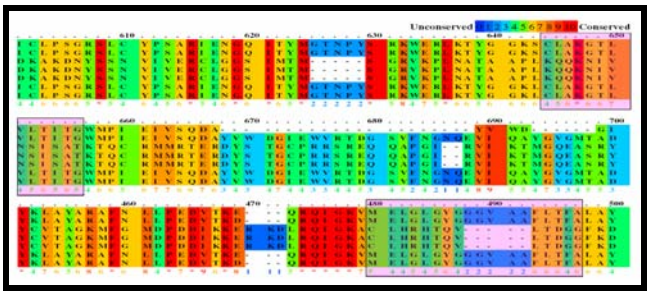


Figure 3: Sequence alignment of APSE 1, 2, 4 & 5 DNA polymerase family A with salmonella phage 3, 12, & 5 DNA polymerase family A shows the difference in insertion site. Block A of both the intein domains are indicated within boxes. The order of the sequences taken for alignment is APSE 1, 2 SETP 3, 12, 5, APSE 4 & 5 from top to bottom respectively.

Shift in base composition:

A shift in the base compositions between intein and extein coding sequences indicates recent acquisition of inteins **[17]**. Analysis of base composition of family A DNA polymerase gene of APSE and its specific intein showed a difference of 2% GC content between them (Gene-45% GC content, intein-47% GC content).

Conclusion:

We have characterized a new bacteriophage intein found in the prokaryotic-type putative DNA polymerase family A of APSE by bioinformatics methods. The conservation of the active site motifs for splicing as well as its insertion at a catalytically important site of the PolA sequence suggests that the intein is most likely to be functional. This intein was found to be absent in APSE 3. Conserved region was identified at the insertion site of the intein and the inteinless prokaryotic DNA polymerase family A proteins. The canonical blocks of inteins and their splice junctions were identified and characterized by bioinformatics method. The insertion site of intein within APSE family A DNA polymerase extein was identified to be dpola. Hence we propose this as a unique intein of family A DNA polymerase (dpola insertion site) and also the only reported intein in podoviridae family till date.

Acknowledgements:

This research was supported by the Bioinformatics Infrastructure Facility grant funded by the Department of Biotechnology (DBT), Government of India (GOI) (Grant No. BT/BI/25/001/2006). Financial support provided to Senguttuvan Rajarajan by the DBT, GOI (Grant No. BT/HRD/TTC/04/2008) in the form of studentship is thankfully acknowledged.

References:

[1] Juhala RJ *et al. J Mol Biol.* 2000 **299**: 27 [PMID: 10860721]
[2] Canchaya C *et al. Curr Opin Microbiol.* 2003 **6**: 417 [PMID: 12941415]
[3] Canchaya C *et al. Microbiol Mol Biol Rev.* 2003 **67**: 238 [PMID: 12794192]
[4] Degnan PH & Moran NA. *Appl Environ Microbiol.* 2008 **74**: 6782 [PMID: 18791000]
[5] Perler FB. *Cell* 1998 **92**: 1 [PMID: 9489693]
[6] Belfort M *et al. J Bacteriol.* 1995 **177**: 3897 [PMID: 7608058]
[7] Hirata R *et al. J Biol Chem.* 1990 **265**: 6726 [PMID: 2139027]
[8] Perler FB *et al. Nucleic Acids Res.*, 1997 **25**: 1087 [PMID: 9092614]
[9] Pietrokovski S. *Curr Biol.* 1998 8: R634 [PMID: 9740808]

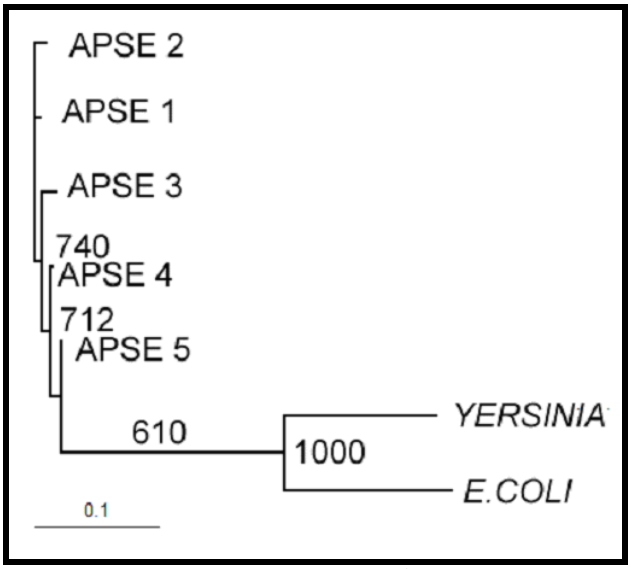
[10] Altschul SF *et al. Nucleic Acids Res.* 1997 **25**: 3389 [PMID: 9254694]
[11] Perler FB. *Nucleic Acids Res.* 2002 **30**: 383 [PMID: 11752343]
[12] Pearson WR & Lipman DJ. *Proc Natl Acad Sci USA* 1988 **85**: 2444 [PMID: 3162770]
[13] Simossis & Heringa J. *Nucleic Acids Res.* 2005 **33**: W289 [PMID: 15980472]
[14] Simossis VA *et al. Nucleic Acid Res.* 2005 **33**: 816 [PMID: 15699183]
[15] Larkin MA *et al. Bioinformatics* 2007 **23**: 2947 [PMID: 17846036]
[16] Heringa J. *Comput Chem.* 1999 **15**: 341 [PMID: 10404624]
[17] Liu XQ & Hu Z. *Proc Natl Acad Sci U S A.* 1997 **94**: 7851 [PMID: 9223276]

Edited by N Gautham
Citation: Rajarajan *et al.* Bioinformation 6(4): 149-152 (2011)

BIOINFORMATION

open access

Supplementary material:



Supplementary data 1: Phylogenetic analysis of APSE phage DNA polymerase family A protein excluding the intein with other prokaryotic DNA polymerase family A.