BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis
Volume 7(4)

Comparative Study of different msDNA (multicopy single-stranded DNA) structures and phylogenetic comparison of reverse transcriptases (RTs): evidence for vertical inheritance

Rasel Das[1], Tadashi Shimamoto[2], Sultan Mohammad Zahid Hosen[3], Mohammad Arifuzzaman[1*]

[1]Department of Biochemistry and Biotechnology, University of Science and Technology Chittagong (USTC), Foy's Lake, Pahartali Chittagong – 4202, Bangladesh; [2]Laboratory of Food Microbiology and Hygiene, Graduate School of Biosphere Science, Hiroshima University, Higashi-Hiroshima, Hiroshima 739-8528, Japan; [3]Department of Pharmacy, BGC Trust University, Bangladesh. Mohammad Arifuzzaman - E-mail: larif67@yahoo.com; Phone: +91-880-31-659070-71, Ext: 280; Fax: 880-31-659545; [*]Corresponding author:

Received September 26, 2011; Accepted October 02, 2011; Published October 14, 2011

Abstract:
The multi-copy single-stranded DNA (msDNA) is yielded by the action of reverse transcriptase of retro-element in a wide range of pathogenic bacteria. Upon this phenomenon, it has been shown that msDNA is only produced by Eubacteria because many Eubacteria species contained reverse transcriptase in their special retro-element. We have screened around 111 Archaea at KEGG (Kyoto Encyclopedia of Genes and Genomes) database available at genome net server and observed three *Methanosarcina* species (*M.acetivorans, M.barkeri* and *M.mazei*), which also contained reverse transcriptase in their genome sequences. This observation of reverse transcriptase in Archaea raises questions regarding the origin of this enzyme. The evolutionary relationship between these two domains of life (Eubacteria and Archaea) hinges upon the phenomenon of retrons. Interestingly, the evolutionary trees based on the reverse transcriptases (RTs) and 16S ribosomal RNAs point out that all the Eubacteria RTs were descended from Archaea RTs during their evolutionary times. In addition, we also have shown some significant structural features among the newly identified msDNA-Yf79 in *Yersinia frederiksenii* with other of its related msDNAs (msDNA-St85, msDNA-Vc95, msDNA-Vp96, msDNA-Ec78 and msDNA-Ec83) from pathogenic bacteria. Together the degree of sequence conservation among these msDNAs, the evolutionary trees and the distribution of these *ret* (reverse transcriptase) genes suggest a possible evolutionary scenario. The single common ancestor of the organisms of Eubacteria and Archaea subgroups probably achieved this *ret* gene during their evolution through the vertical descent rather than the horizontal transformations followed by integration into this organism genome by a mechanism related to phage recognition and/or transposition.

Keywords: msDNA, reverse transcriptase, phylogenetic tree.

Background:

Bacterial chromosomes often carry integrated genetic elements (for example transposons, prophages and islands) whose precise functions and contribution to the evolutionary fitness of the host bacterium are still unknown **[1]**. These elements are often associated with the pathogenicity of the organisms, for example the CTXφ prophage, which encodes cholera toxin in *Vibrio cholerae* **[1]**. Retron is also a transposable element **[2]** found in various pathogenic bacteria **[3].** The retron is consisting of three regions- *msr* (encodes RNA part of msDNA), *msd* (encodes DNA part of msDNA) and a *ret* gene for reverse transcriptase (RT) **[4-6]**. This enzyme is responsible for the production of multi-copy single-stranded DNA (msDNA) containing both DNA and RNA covalently linked by a branched rG residue **[4-6]**. Although, reverse transcriptase (RT) was first discovered in virus **[7]**, now this enzyme is frequently found in bacteria which are then prokaryotic reverse transcriptase to be discovered **[4-6]**. The discovery of reverse transcriptase in bacteria raises questions regarding the origin of this enzyme. The reverse transcriptase (RT) consisting of several conserved regions common to all retrons, can thus be used for a comprehensive evolutionary analysis of retrons. To solve a question regarding the origin of retron encoding *ret* (reverse transcriptase) gene, we have compared the retro-element reverse transcriptase (RT) sequences in different bacteria. The phylogenetic trees constructed from this present study provide a framework to appraise possible hypothesis for the origin and evolution of different categories of retrons present in different organisms. We have taken Eubacteria and Archaea for screening purpose to see an evolutionary relationship between these two domains of life on the basis of there *ret* (reverse transcriptase) genes.

Methodology:

Sequence Retrieval

The amino acid sequences of reverse transcriptase (RT) and the nucleotide sequences of 16S ribosomal RNA (rRNA) such as Archaea: *Methanosarcina* species: (*M.acetivorans, M.barkeri* and *M.mazei*); Eubacteria: *Yersinia* species (*Y.frederiksenii*, *Y.pestis*); Myxobacteria species (*S.cellulosum*, *M.lichenicola*, *S.aurantiaca*, *N.exedens* and *M.xanthus*); *Vibrio* species (*V.cholerae, V.parahaemolyticus* and *V.mimicus*); *Salmonella* Typhimurium; and *Escherichia coli* species (Strains 161,110, RT-Ec73 specific Enterobacteria phage phiR73, ECOR70, ECOB, ECOR35 and ECOR58)- were retrieved from ExPASy proteomics server (http://expasy.org/). In addition, the 16S ribosomal RNA (16S rRNA) nucleotide sequences of Archaea: *Methanosarcina* species (*M.acetivorans, M.barkeri* and *M.mazei*) and Eubacteria: *Yersinia* species (*Y.frederiksenii* and *Y.pestis*); Myxobacteria species (*S.cellulosum*, *M.lichenicola*, *S.aurantiaca*, *N.exedens* and *M.xanthus*); *Vibrio* species (*V.cholerae, V.parahaemolyticus* and *V.mimicus); S.*Typhimurium; and *Escherichia coli* K-12 were retrieved from the KEGG organism database, Japan (http://www.genome.jp/).

Generation of Gene and Phylogenetic Trees

Identical conserved regions of reverse transcriptase (RT) amino acid sequences of Archaea: *Methanosarcina* species: (*M.acetivorans, M.barkeri* and *M.mazei*); Eubacteria: *Yersinia* species (*Y.frederiksenii*, *Y.pestis*); Myxobacteria species (*S.cellulosum*, *M.lichenicola*, *S.aurantiaca*, *N.exedens* and *M.xanthus*); *Vibrio* species (*V.cholerae, V.parahaemolyticus* and *V.mimicus*); *Salmonella* Typhimurium; and *Escherichia coli* species (Strains 161,110, RT-Ec73 specific Enterobacteria phage phiR73, ECOR70, ECOB, ECOR35 and ECOR58) scored in the alignment constructed by using (CLUSTALW) (http://www.ebi.ac.uk/Tools/msa/clustalw2/) **[8]** were used to generate a gene (*ret*) tree. The sequence alignment was performed under default conditions and the gene tree was constructed by the neighbor-joining **[9]** and distance matrix methods. The poorly aligned N-terminals and C-terminals sequence of alignments and also the internal gaps residue were taken off from the alignments to make a precise evolutionary tree by using the Jalview program **[10]**. The phylogenetic tree of 16S ribosomal RNAs of those organisms was also constructed based on their nucleotide sequences by using (CLUSTALW) (http://www.ebi.ac.uk/Tools/msa/clustalw2/) **[8].**

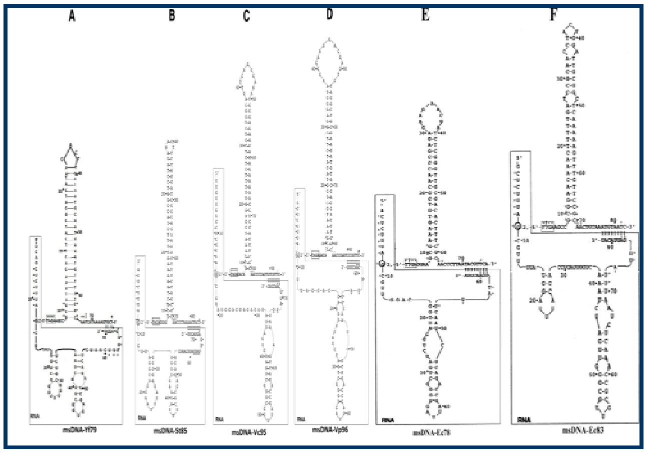

Figure 1: Possible secondary structures of multi-copy single-stranded DNAs (msDNAs) from pathogenic bacteria (*Y.frederiksenii*, *S.typhimurium*, *V.cholerae*, *V.parahaemolyticus* and *Escherichia coli* strains 110 & 161). The branching guanine (G) residue in RNA portion of msDNAs is circled and forming a 2', 5'-phosphodiester bond with DNA portion of msDNAs (A-F). Both the DNA and RNA secondary stem loop structures were suggested on the basis of their sequences. The RNA portion was boxed and the numbers of RNA and DNA were begun from 5' ends. The conserved nucleotides are indicated by stars in all msDNAs. The msDNA-Yf79 (A) is predicted from *Yersinia frederiksenii* **[11],** msDNA-St85 (B) is isolated from *S.typhimurium* **[13]**, msDNA-Vc95 (C) is from *V.cholerae* **[14],** msDNA-Vp96 (D) is from *V.parahaemolyticus* (Shimamoto T, 2003, unpublished data), msDNA-Ec78 (E) is from *E.coli* strain 110 **[12]** and msDNA-Ec83 (F) from *E.coli* strain 161 **[3]**. The conserved sequence for cleavage is indicated by boxes and base substitution pointed out by arrows.

Results and discussion:

Comparative Study of Multicopy Single-Stranded DNA (msDNA) Structures

Recently, we have perceived that a new msDNA-Yf79 exist in *Yersinia frederiksenii* ATCC 33641 contig01029 cell types and compared it's properties to that of St-85, Vc-95 and Vp-96 **[11]**. However, the present study has revealed the close relatedness of this msDNA-Yf79 with msDNA-Ec78 and msDNA-Ec83 from *E.coli* strains 110, 161 respectively **(Figure 2) [12, 3]**. These msDNAs shared a number of highly conserved nucleotides in their DNA-RNA complex sequences **(Figure 1).** The sequence 5'-TAGA-3' box was conserved in msDNA-Yf79 **[11]**, msDNA-St85 **[13]**, msDNA-Vc95 **[14]** and msDNA-Vp96 (Shimamoto T, 2003, unpublished data) **(Figure 1A-1D).** The box 5'-TTGA-3' was conserved in msDNA-Ec78 and msDNA-Ec83 **[12, 3]**. The conserved tetra nucleotides (5'-TTGA-3') would play an important role in the recognition and cleavage of msDNA by a hypothetical enzyme **[15]**. Furthermore, the second nucleotide thymine (T) was substituted by adenine (A) in the 5'-TAGA-3' box of msDNA-Yf79 **[11]**, -St85 **[13]**, -Vc95 **[14]** and -Vp96 (Shimamoto T, 2003, unpublished data) **(Figure 1A-1D).** The third nucleotide guanine (G) in the two boxes of all msDNAs is conserved **(Figure 1A-1E**) and indicates the higher efficiency of cleavage of these msDNAs, because when the guanine (G) at the third position change to cytosine (C), the effort of cleavage is moderately reduced **[15]**. In addition to these boxes, the fifth nucleotide adenine (A) at the 5' end of DNA part of msDNA-Yf79 **[11]** and msDNA-Ec83 **[3]** **(Figure 1A and 1F)** may become a target, because when this adenine (A) was substituted by any pyrimidines, it reduces the overall msDNA accumulation **[15]**. Within conserved nucleotides among all msDNAs **(Figure 1)** the nucleotide cytosine (C) at position 67 of msDNA-Yf79 **[11]** was substituted by thymine (T), and thymine (T) at position 96 of msDNA-Vp96 (Shimamoto T, 2003, unpublished data) was substituted by cytosine (C) **(Figure 1A, 1D).** Furthermore, msDNA-Ec83 **[3]** seems to be mutagenic **(Figure 1F)** because msDNA with any mismatched base pair in their DNA stems could be mutagenic **[16, 17].** However, other remaining msDNAs **(Figure 1A-1E)** contained no such type of mismatched base pairing in their DNA stem structures.

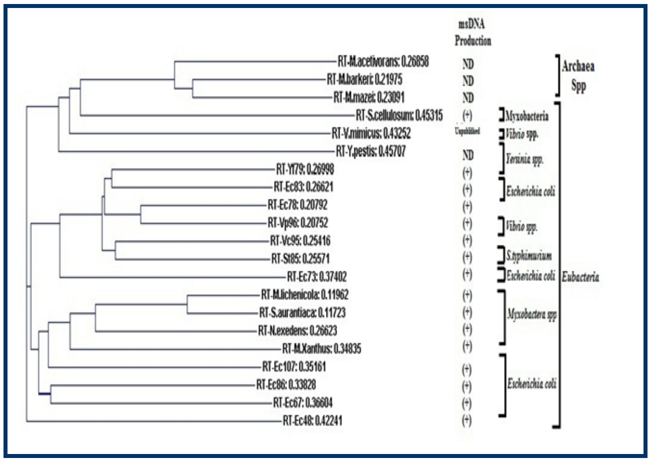

Figure 2: Gene tree among Archaea- *Methanosarcina* species (*M.acetivorans, M.barkeri* and *M.mazei*) and Eubacteria- *Yersinia* species (*Y.frederiksenii* and *Y.pestis*); Myxobacteria species (*S.cellulosum, M.lichenicola, S.aurantiaca, N.exedens,* and *M.xanthus*); *Vibrio* species (*V.cholerae, V.parahaemolyticus* and *V.mimicus*); *S.*Typhimurium; *Escherichia coli* species (Strain 161,110, ECOR70, ECOB, ECOR35 and ECOR58) and msDNA-Ec73 specific RT from Enterobacteria phage phiR73 based on the RT amino acid sequences. Here, ND-indicates that the strains were not tested for msDNA production and the (+) sign indicates the presence of msDNA. The distance between sequences is located just beside each RT. The following ExPASy accession numbers for the RT sequences were used in the phylogenetic construction: (*M.acetivorans*-Q8TMH8, *M.barker*-Q46BR7, *M.mazei*-Q8PTN0); (*Y.frederiksenii* RT-Yf79-C4SUU2, *Y.pestis*-Q7ARB2); (*S.cellulosum*-A9GPU1, *M.lichenicola*- Q50210, *S.aurantiaca*- Q08Y90, *N.exedens*- Q8VRM1, *M.xanthus*-Q1D0F5); (*V.cholerae* RT-Vc95- Q9S1F2, *V.parahaemolyticus*-Q8L0W6, *V.mimicus*- D0HJ73); *S.*Typhimurium- E7UVY4; and *Escherichia coli* species strains 161 (RT-Ec83, Q47526); 110 (RT-Ec78, Q46666); (msDNA-Ec73 specific RT from Enterobacteria phage phiR73, Q7M2A9); ECOR70 (RT-Ec107, Q05804); ECOB (RT-Ec86, P23070); ECOR35 (RT-Ec67, P21325) and ECOR58 (RT-Ec48, P71276).

Evolutionary Relationship of RTs from Pathogenic Bacteria

To explore the evolution of reverse transcriptases (RTs), the phylogenetic analysis was performed among RT amino acid sequences. Result exhibits a fundamental diversity among all the reverse transcriptases (RTs) as RT-Yf79 (*Y. frederiksenii*) **[11]** is closely related to the RT-Ec83 (*E.coli* strain 161) **[3]** rather than to the RT-Ec78 (*E.coli* strain 110) **[12]**, RT-Vp96 (*V. parahaemolyticus*) (Shimamoto T, 2003, unpublished data), RT-Vc95 (*V.cholerae*) **[14]** and RT-St85 (*S.*Typhimurium) **[13]**. RT-Ec78 (*E.coli* strain 110) **[12]** is closely related to RT-Vp96 (*V. parahaemolyticus*) (Shimamoto T, 2003, unpublished data) as well as RT-Vc95 (*V.cholerae*) **[14]** is closely related to the RT-St85 (*S.*Typhimurium) **[13]** **(Figure 2)**. The msDNA-Ec73 specific RT (Enterobacteria phage phiR73) **[18]** is related to RT-Yf79, -Ec83, -Ec78, -Vp96 (Shimamoto T, 2003, unpublished data), -Vc95 and -St85 **[11, 3, 12, 14, 13]**. Although RT-Ec83, -Ec78 and -Ec73 **[3, 12, 18]** were from *Escherichia* species, they spread out from a central point **(Figure 2).** Similarly, RT-Vc95 **[14]** and RT-Vp96 (Shimamoto T, 2003, unpublished data) were from *Vibrio* species, but they are closely related to RT-St85 **[13]** and RT-Ec78 **[12]**, respectively (**Figure 2).** It was also observed that the high similarity and relatedness of all RTs from Archaea species and RT-*S.cellulosum* (Myxobacteria species) **[19]** is originated from Archaea RTs **(Figure 2).** RT (*V. mimicus*) (Shimamoto T, 2003, unpublished data) is closely related to the RT-*S.cellulosum* (Myxobacteria species) **[19]** and RTs from Archaea species **(Figure 2).** In addition, RT (*Y.pestis*) was descended from RTs of *V. mimicus* (Shimamoto T, 2003, unpublished data), *S.cellulosum* **[19]** and Archaea species **(Figure 2).** The remaining RTs (Myxobacteria species) **[19]** are related to RT-Ec107, -Ec86 and -Ec67 **[20, 21]**. Among all RTs (Myxobacteria species) **[19],** especially RTs - *M.lichenicola* and *S.aurantiaca* **[19]** have *shown* higher similarity, as their phylogenetic distance is too low **(Figure 2).** Surprisingly, all the Eubacteria RTs are diverged from the Archaea species **(Figure 2).** Although RT-Ec48 is from *E.coli* species **[22]** this enzyme is distantly related with RTs of Archaea and Eubacteria species **(Figure 2).**

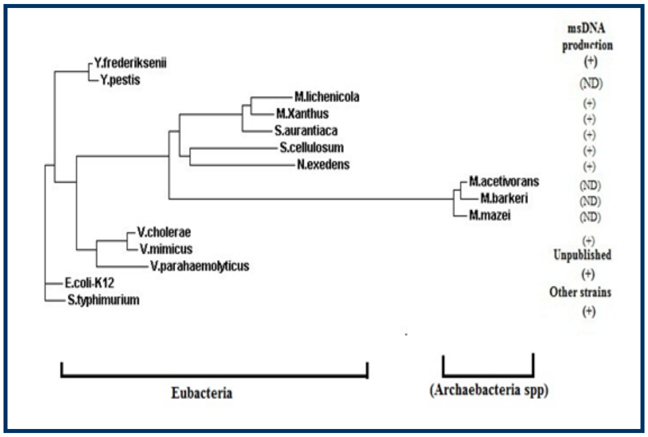

Figure 3: Phylogeny of the two domains of life. The tree among Archaea: *Methanosarcina* species (*M.acetivorans, M.barkeri* and *M.mazei*) and Eubacteria: *Yersinia* species (*Y.frederiksenii* and *Y.pestis*); Myxobacteria species (*S.cellulosum, M.lichenicola, S.aurantiaca, N.exedens* and *M.xanthus*); *Vibrio* species (*V.cholerae, V.parahaemolyticus* and *V.mimicus*); *S.*Typhimurium; and *Escherichia coli* K-12 based on the 16S ribosomal RNA nucleotide sequences. Here, ND-indicates that the strains were not tested for msDNA production and the (+) sign indicates the presence of msDNA. The following Genomenet accession numbers for the 16S rRNA sequences were used in the phylogenetic construction: (*M.acetivorans*-1472788, *M.barker*-3625539 and

M.mazei-2820544); (*Y.frederiksenii*-NR_027544.1 and *Y.pestis*-1172856); (*S.cellulosum*-5807545, *M.lichenicola*- DQ491069.1, *S.aurantiaca*- 9873957, *N.exedens*- AB084253.1 (GenBank accession number), *M.xanthus*-4107927); (*V.cholerae* 2614447, *V.parahaemolyticus*-1187490 and *V.mimicus*-NZ_ADAJ01000005.1); *S*.Typhimurium- 1251767; and *Escherichia coli* K-12 (944897).

Genetic Diversity of msDNA Producing Strains

The phylogenetic tree based on 16SrRNA sequences of the bacteria such as Archaea: *Methanosarcina* species (*M.acetivorans, M.barkeri* and *M.mazei*) and Eubacteria: *Yersinia* species (*Y.frederiksenii* and *Y.pestis*); Myxobacteria species (*S.cellulosum*, *M.lichenicola*, *S.aurantiaca*, *N.exedens* and *M.xanthus*); *Vibrio* species (*V.cholerae, V.parahaemolyticus* and *V.mimicus*); *S*.Typhimurium; and *Escherichia coli* K-12 form a phylogenetically related clusters or subgroups **(Figure 3).** The phylogeny tree revealed that four of the five major phylogenetic groups produce msDNAs perhaps acquired this *ret* gene during their evolution through vertical descended rather than horizontal transformation. This observation is consistent with the hypothesis of Rice and Lampson (1995) **[19]**.

Perspectives

This study manifests that, the *ret* genes commonly perceived in Eubacteria are unique compared with *ret* genes found in Archaea. Eubacteria *ret* genes of diverse types are probably widespread and might be descended from the characteristics of the Archaea world. This hypothesis is supported by the prevalence of distribution of *ret* genes, among a wide variety of organisms as documented here and in previous reports **[11, 23]**. Xiong and Eickbush (1990) also have shown the origin and evolution of retro-elements in different organisms based upon reverse transcriptase sequences **[23].** Though Archaea had been grouped with bacteria as prokaryotes (archaebacteria), they have an independent evolutionary history. With respect to the observation of *ret* genes in Archaea, the observation suggests that the RT enzyme played a role in the evolutionary emergence of Eubacteria from Archaea (or vice versa). Furthermore, reverse transcriptase (RT) is also frequently found in retroviruses like Human Immunodeficiency Virus (HIV) **[7].** Now these findings raise questions as to where *ret* gene originates? Why are they so diverse along with retron elements in these organisms? This statement is parallel to the popular chicken versus egg puzzles theory.

References:

[1] Hassan F *et al. Nature* 2010 **467**: 982 [PMID: 20944629].
[2] Hsu MY *et al. Proc Natl. Acad Sci.* USA 1990 **87**: 9454 [PMID: 1701261].
[3] Lim D. *Mol. Microbiol.* 1992 **6**: 3531 [PMID: 1282191].
[4] Lampson BC *et al. Cell* 1989 **56**: 701 [PMID: 2465091].
[5] Inouye S *et al. Cell* 1989 **56**: 709 [PMID: 2465092].
[6] Lim D & Maas W, *Cell* 1989 **56**: 891 [PMID: 2466573].
[7] Temin HM & Mizutani S, *Nature* 1970 **226**:1211 PMID: 4316301].
[8] Larkin MA *et al. Bioinformatics* 2007 **23:** 2947 [PMID: 17846036].
[9] Saitou N & Nei M, *Mol.Biol.Evo.* 1987 **4**: 406 [PMID: 3447015].
[10] Waterhouse AM *et al. Bioinformatics* 2009 **25:** 1189 [PMID: 19151095].
[11] Das R *et al. J Pathogen.* 2011 (Article in press).
[12] Lima TMO & Lim D, *Plasmid* 1997 **38**: 25 [PMID: 9281493].
[13] Ahmed AM & Shimamoto T, *FEMS Microbiol. Let.* 2003 **224**: 291 [PMID: 12892895].
[14] Shimamoto T *et al. Mol Microbiol.* 1999 **33**: 631 [PMID: 10564503].
[15] Kim K *et al. J Bacteriol.* 1997 **179**: 6518 [PMID: 9335306].
[16] Maas WK *et al. Mol Microbiol.* 1994 **14**: 437 [PMID: 788522].
[17] Mao JR *et al. FEMS Microbiol Lett.* 1996 **144**: 109 [PMID: 8870259].
[18] Sun J *et al. J Bacteriol.* 1991 **173**: 4171 [PMID: 1712012].
[19] Rice A Scott & Lampson BC, *J Bacteriol.* 1995 **177**: 37 [PMID: 7798147].
[20] Herzer PJ. *J Bacteriol.* 1996 **178**: 4438 [PMID: 8755870].
[21] Lim D. *Mol Microbiol.* 1991 **5**: 1863 [PMID: 1722556].
[22] Mao JR *et al. J Bacteriol.* 1997 **179**: 7865 [PMID: 9401048].
[23] Xiong Y & Eickbush TH, *EMBO J* 1990 **9**: 3353 [PMID: 1698615].

Edited by P Kangueane
Citation: Das *et al.* Bioinformation 7(4): 176-179 (2011)