G&I Genomics & Informatics

eISSN 2234-0742
Genomics Inform 2016;14(1):29-33
http://dx.doi.org/10.5808/GI.2016.14.1.29

ORIGINAL ARTICLE

Compiling Multicopy Single-Stranded DNA Sequences from Bacterial Genome Sequences

Wonseok Yoo, Dongbin Lim, Sangsoo Kim*

Department of Bioinformatics and Life Science, Soongsil University, Seoul 06978, Korea

A retron is a bacterial retroelement that encodes an RNA gene and a reverse transcriptase (RT). The former, once transcribed, works as a template primer for reverse transcription by the latter. The resulting DNA is covalently linked to the upstream part of the RNA; this chimera is called multicopy single-stranded DNA (msDNA), which is extrachromosomal DNA found in many bacterial species. Based on the conserved features in the eight known msDNA sequences, we developed a detection method and applied it to scan National Center for Biotechnology Information (NCBI) RefSeq bacterial genome sequences. Among 16,844 bacterial sequences possessing a retron-type RT domain, we identified 48 unique types of msDNA. Currently, the biological role of msDNA is not well understood. Our work will be a useful tool in studying the distribution, evolution, and physiological role of msDNA.

Keywords: msDNA, retron, reverse transcriptase, RNA secondary structure

Introduction

A retron is a 2-kb-long bacterial retroelement that carries a promoter and three genes: msr, msd, and ret. The msr/msd part of the transcript is processed by a reverse transcriptase encoded by ret, resulting in an RNA/DNA chimera called multicopy single-stranded DNA (msDNA) [1]. The RNA and DNA parts of msDNA are encoded by msr and msd, respectively. The precursor RNA molecule possesses a palindrome between the 5′ and 3′ ends, forming a double-stranded region between both ends (Fig. 1A). The 5′ part of the palindrome is followed by a G residue, which is the branching point for covalent bonding to the cDNA (Fig. 1B) [2]. The msr portion of the precursor RNA possesses two stem-loop structures, which are recognized by the reverse transcriptase [3]. The msd part of the resulting chimera, shown in black in Fig. 1C, is palindromic, also forming a large hairpin structure.

While msDNA is found in many bacterial species, its biological role is not well understood. Furthermore, msDNA is not well annotated in public sequence databases. In order to facilitate the biological characterization of msDNA, we set out to compile msDNA sequences from known bacterial genomes. Using a few known msDNA sequences, we extracted common sequence and structural features. We screened National Center for Biotechnology Information (NCBI) RefSeq bacterial genome sequences using this rule

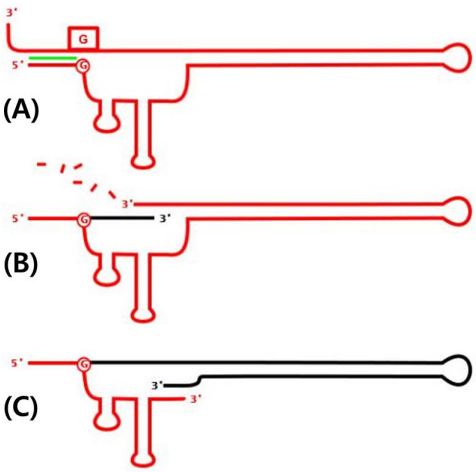

Fig. 1. Schematic diagram of multicopy single-stranded DNA synthesis. (A) A precursor RNA molecule with the characteristic palindrome (green). "G" residues in a circle or box (red) are conserved. (B) Reverse transcription yielding cDNA (black). (C) The final msDNA chimera. Adapted from Wikipedia (https://en.wikipedia.org/wiki/Multicopy_single-stranded_DNA).

Received December 3, 2015; Revised December 23, 2015; Accepted February 23, 2016

***Corresponding author:** Tel: +82-2-820-0457, Fax: +82-2-824-4383, E-mail: sskimb@ssu.ac.kr

to compile potential msDNA sequences. Their authenticity as msDNA needs to be confirmed experimentally.

Methods

Prototype msDNA sequences and their features

We used one of the published *Escherichia coli* msDNA sequences (GenBank accession No. U02551) [2] as the query in a Basic Local Alignment Search Tool (BLASTN) search of NCBI GenBank High Throughput Genomic Sequence (HTGS) division. From the BLASTN hits, we manually curated seven other known msDNA sequences (Table 1). By visual inspection of these prototype msDNA sequences, several common features were recognized: (1) the length of palindrome between the 5′ and 3′ ends ranged from 5 bp to 10 bp; (2) the distance between both ends ranged from 140 bp to 200 bp; (3) the residues flanking the palindrome were conserved G residues in both strands; and (4) the distance between msr/msd and ret was less than 2 kb.

We developed a local python script that enforces these rules in the candidate retron sequences.

Gene prediction and protein domain search

From the downloaded DNA sequences, genes were predicted with the GeneMark suite [4]. The input sequences were split into 20-kb batches, and GeneMark was run for each batch with the default settings. The predicted genes were translated into protein sequences, which were then scanned for the Clusters of Orthologous Group (COG) domain using Reverse Position Specific (RPS)-BLAST. COG is a microbial domain database distributed by NCBI. RPS-BLAST is one of the BLAST applications distributed by NCBI and uses position-specific scoring matrices (PSSMs) as the target database [5]. We looked for hits of the accession COG3344 (PSSM ID: 225881), which stands for "Retron-type reverse transcriptase (mobilome: prophages, transposons)."

Palindrome motif search

One of the characteristic features of msDNA is a palindrome formed by its 5′ and 3′ ends. In order to detect such a feature in the potential retron sequences, we used a local installation of the software tool palindrome in the European Molecular Biology Open Software Suite (EMBOSS) bioinformatics software package [6]. It requires several options to be set. The minimum and maximum lengths of a palindrome and the maximum distance between a palindrome pair were set to 5, 10, and 200 bp, respectively, while the number of mismatches between a palindrome pair was set to 1.

Profile search of msDNA

A multiple-sequence alignment of the eight prototype msDNA sequences was generated with ClustalW (Fig. 2). We then built a hidden Markov model (HMM) profile from the multiple-sequence alignment using the software tool HmmerBuild in the HMMER package [7]. The potential retron sequences were screened using HmmerSearch with the HMM profile [7]. The cutoff for a specific hit was determined by leave-one-out crossvalidation as follows: (1) among the eight seed sequences, one was set aside, and the remaining seven were used to build the profile; (2) what was set aside was then evaluated using the profile; (3) this process was repeated for each of the eight seeds, and the minimum score was defined as the cutoff (45.0).

Table 1. Prototype msDNA sequences used for the development of detection rules

GenBank accession No.	Bacterial strain	GenBank definition	Size	Position
U02551.1	*Escherichia coli*	*E. coli* Ec 110 reverse transcriptase gene, complete cds	144	12–155
Z12832.1	*E. coli*	*E. coli* proA and ret genes encoding gamma-glutamylphosphate reductase and reverse transcriptase	166	542–707
CP000462.1	*Aeromonas hydrophila* subsp. hydrophila ATCC 7966	*A. hydrophila* subsp. hydrophila ATCC 7966, complete genome	141	2242058–2242198
CP000749.1	*Marinomonas* sp. MWYL1	*Marinomonas* sp. MWYL1, complete genome	163	2734544–2734706
CP002433.1	*Pantoea* sp. At-9b	*Pantoea* sp. At-9b, complete genome	150	46523–46672
CP007235.1	*Salmonella enterica* subsp. enterica serovar Typhimurium str	*S. enterica* subsp. enterica serovar Typhimurium str. USDA-ARS-USMARC-1899, complete genome	158	4823808–4823965
CP006947.1	*Vibrio cholerae* O1 str. KW3	*V. cholerae* O1 str. KW3 chromosome I, complete sequence	168	534292–534459
JPPS01000005.1	*Yersinia frederiksenii* ATCC 33641	*Y. frederiksenii* ATCC 33641 DJ58.Contig268, whole genome shotgun sequence	153	869165–869317

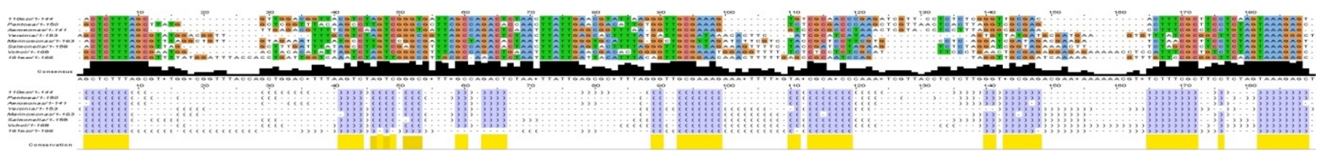

Fig. 2. Multiple-sequence alignment of prototype multicopy single-stranded DNA sequences. The msr/msd portion of each genomic DNA sequence in Table 1 was used to construct the alignment with ClustalW. The corresponding secondary structures, predicted with CentroidFold, are shown in bracket notation. The conserved positions are colored with Jalview.

RNA secondary structure prediction

The precursor RNA molecule of msDNA forms a characteristic secondary structure. We used a local installation of the software tool CentroidFold [8] to predict the secondary structure of the msr/msd portion of the candidate retron. As the palindrome terminates with a G residue in both strands, the 5′ of which is the branching point for covalent bonding to the cDNA (Fig. 1B), sequences without these G residues were removed. The web service version of RNAfold [9] was also used to visually confirm the final prediction set.

Results

Compilation of candidate retron sequences from known genome sequences

Historically, msDNA has been discovered exclusively in bacteria. A keyword search of NCBI RefSeq bacterial genomes with "reverse transcriptase" returned 34,637 hits, as of September 2, 2015. Following the gene prediction of these genomes using GeneMark, a protein domain search for "Retron-type reverse transcriptase (mobilome: prophages, transposons)" using RPS-BLAST resulted in 16,844 candidate retron sequences.

Alternatively, filtering out nonretron-type reverse transcriptase sequences can be also achieved by extracting protein sequences from the RefSeq annotation, followed by the domain search. Instead, we chose to predict genes from the genome sequences using GeneMark, followed by the domain search. Our approach does not rely on annotation information and has the potential to be applied to newly sequenced genomes as an independent tool.

Automatic filtration of candidate msDNA sequences

The candidate retron sequences identified above were filtered by the presence of palindrome motifs. Using the local python script, we searched the 16,844 bacterial sequences for palindrome motifs 5–10-bp-long and separated by less than 200 bp within 2 kb upstream of ret, yielding 7,428,332 such hits. These hits were then screened by the HMM models that were developed based on the prototype msDNA sequences.

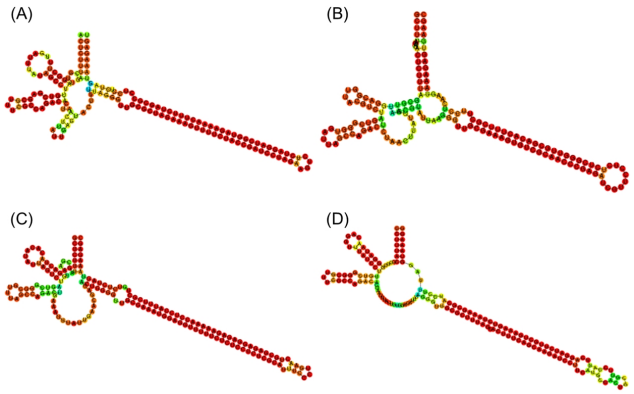

Fig. 3. Representative examples of secondary structures of msr/msd precursors of predicted multicopy single-stranded DNA sequences. (A) NZ_AERV01000006.1 *Salmonella enterica* (J, 86). (B) NZ_ADUL01000064.1 *Escherichia coli* (B, 3). (C) NZ_AELI01000009.1 *Vibrio cholerae* (E, 48). (D) NZ_CBWG010000185.1 *Escherichia coli* (N, 504). The single letters and numbers in parentheses refer to the clustering group and sequence number in Supplementary Table 1.

There were 3,865 hits surpassing the crossvalidation threshold of 45.0. Using a local installation of CentroidFold, the RNA secondary structures were predicted for the remaining hits, and those without the proper double-strand pairing between the 5′ and 3′ ends were removed. Furthermore, those without the conserved flanking G residues were also filtered out. The whole process, which was wrapped in a Linux shell script, resulted in 625 hits.

Manual curation and clustering of msDNA sequences

The final set of 625 candidate sequences that were identified by an automatic process was highly redundant. The exact copies were pruned, giving rising to a set of 88 unique sequence types. As shown in Fig. 1A, the true msDNA should possess two hairpins in msr and a long hairpin in the msd region. In order to confirm this complex overall topology of msDNA, we examined the RNA secondary structure plots of the 88 candidates visually. In fact, we removed 40 sequences that did not form the appropriate overall topology. See Fig. 3 for a few representative surviving examples. Among the original 625 sequences, 525 can be mapped to these 48

unique sequence types (Supplementary Table 1).

In order to cluster the remaining 48 sequence types further, allowing small divergences within clusters, multiple-sequence alignment was performed with Multiple Alignment using Fast Fourier Transform (MAFFT) [10]. From the resulting dendrogram, we recognized 20 groups by splitting the tree at a branch length of between 0.12 and 0.20 (Fig. 4). The species distribution of the 525 msDNA sequences over the 20 groups is given in Table 2 and Supplementary Table 1.

Table 2. Distribution of 525 msDNA sequences in each group

Species	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
Acinetobacter baumannii	-	-	-	-	-	3	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Enterobacter cloacae	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Enterovibrio calviensis	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	-
Escherichia coli	-	40	-	-	-	-	-	-	-	-	-	-	1	10	-	-	2	-	2	1
Klebsiella pneumoniae	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-
Pectobacterium wasabiae	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	-	-
Photobacterium iliopiscarium	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Pluralibacter gergoviae	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	-	-	-
Pseudoalteromonas	-	-	-	1	-	-	-	-	2	-	-	1	-	-	-	-	-	-	-	-
Salmonella enterica	-	-	-	-	-	-	-	-	-	414	-	-	-	-	-	-	-	-	2	-
Shigella boydii	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	-	-	-	-	-
Vibrio cholerae	-	-	-	-	31	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Vibrio crassostreae	-	-	-	-	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-
Vibrio parahaemolyticus	-	-	1	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-
Vibrio sinaloensis	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Vibrio	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	-	-	-	-
Xenorhabdus bovienii	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-
Yersinia frederiksenii	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

msDNA, multicopy single-stranded DNA.

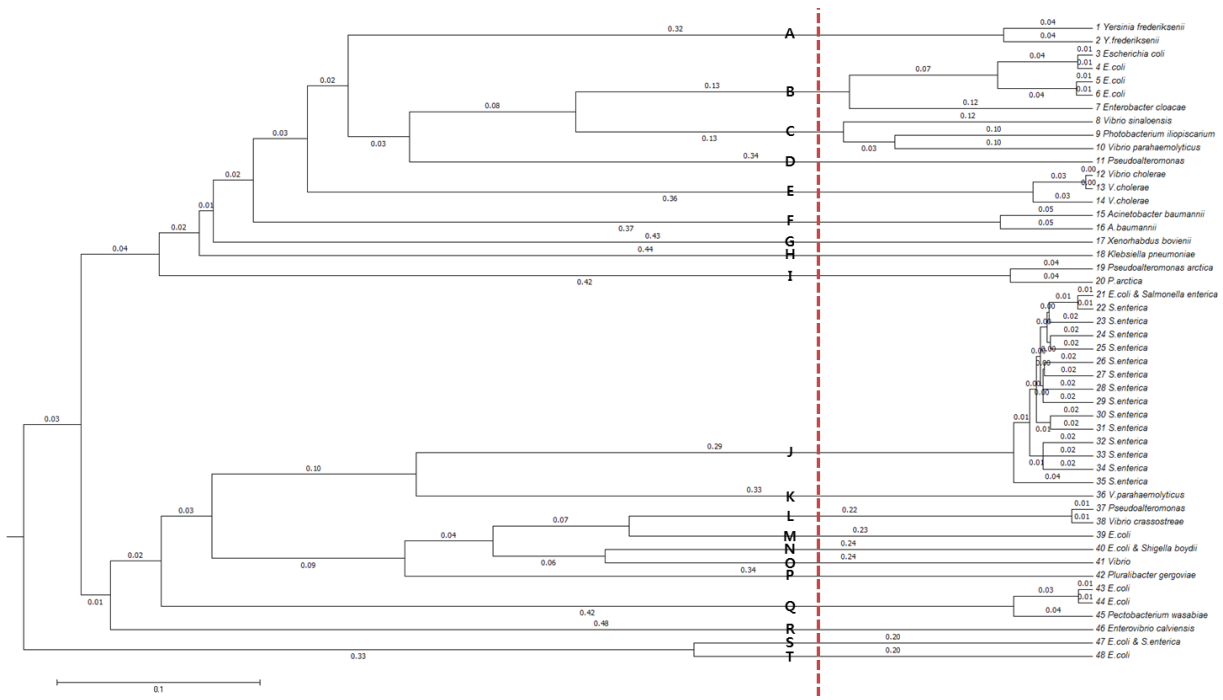

Fig. 4. Dendrogram of the final 48 multicopy single-stranded DNA sequence types. The plot was generated with a multiple-sequence alignment program, Multiple Alignment using Fast Fourier Transform (MAFFT). Distinct clusters are marked with symbols, A–T.

Discussion

We developed a rule-based protocol to detect msDNA in a given genomic DNA sequence. The rule is based on the presence of a 5–10-bp-long palindrome motif in the precursor transcript that is less than 2 kb upstream of a retron-type reverse transcriptase gene. The rule is augmented by enforcing a sequence similarity with known msDNA sequences. This step is implemented with an HMM profile search method. While this was a powerful filtration step, reducing the hits by about 1/20, it would inevitably miss some true msDNA sequences that are distantly related to the prototypes given in Table 1. As more distinct msDNA sequences are discovered, the panel of the seed sequences must be updated.

The list of potential candidate msDNA sequences was filtered by conformation to the known topology of the RNA secondary structure. The RNA secondary structure prediction programs usually output the result in terms of Vienna dot bracket notation [9]. As the manipulation of the string is not straightforward, some manual curation was involved. In the future, we will explore the possibility of automatic implementation, which is critical for web-based service of the prediction tool. To our knowledge, this is the first large-scale annotation of msDNA in all publically available bacterial genome sequences. As our work is computational in nature, the list we compiled can include some false positives. Its authenticity should be validated experimentally. Nevertheless, our list of msDNA sequences and their species distribution profile constitute a useful resource for the future study of msDNA, as its biological role is still elusive.

Supplementary materials

Supplementary data including one table can be found with this article online at http://www.genominfo.org/src/sm/gni-14-29-s001.pdf.

Acknowledgments

This work is based the Bachelor's degree thesis of WY at Soongsil University, Seoul, Korea. We acknowledge the financial support from the National Research Foundation of Korea (NRF-2012M3A9D1054705 and 2010-0023759).

References

1. Inouye S, Herzer PJ, Inouye M. Two independent retrons with highly diverse reverse transcriptases in *Myxococcus xanthus*. *Proc Natl Acad Sci U S A* 1990;87:942-945.
2. Lima TM, Lim D. A novel retron that produces RNA-less msDNA in *Escherichia coli* using reverse transcriptase. *Plasmid* 1997;38:25-33.
3. Inouye S, Hsu MY, Xu A, Inouye M. Highly specific recognition of primer RNA structures for 2'-OH priming reaction by bacterial reverse transcriptases. *J Biol Chem* 1999;274:31236-31244.
4. Borodovsky M, McIninch J. GeneMark: parallel gene recognition for both DNA strands. *Comput Chem* 1993;17:123-133.
5. Tao Tao. 3.5 RPS BLAST. Bethesda: National Center for Biotechnology Information, 2006. Accessed 2015 Jun 6. Available from: http://www.ncbi.nlm.nih.gov/staff/tao/URLAPI/wwwblast/node20.html.
6. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* 2000;16:276-277.
7. Durbin R, Eddy S, Krogh A, Mitchison G. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge: Cambridge University Press, 1998.
8. Sato K, Hamada M, Asai K, Mituyama T. CENTROIDFOLD: a web server for RNA secondary structure prediction. *Nucleic Acids Res* 2009;37:W277-W280.
9. Lorenz R, Bernhart SH, Höner Zu Siederdissen C, Tafer H, Flamm C, *et al*. ViennaRNA Package 2.0. *Algorithms Mol Biol* 2011;6:26.
10. Katoh K, Toh H. Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinform* 2008;9:286-298.