The Origin and Possible Functional Role of Short Dispersed Repeats

Abstract

Repetitive sequences are common not only in eukaryotic genomes but also in prokaryotic genomes. Bacterial genomes contain many types of repeats including tandem repeats and short dispersed repeats. This research aims to expand our understanding of short dispersed repeats (SDR), a novel type of repetitive sequences in cyanobacterial genomes. SDR are found in the genomes of *Nostoc punctiforme* as well as *Anabaena* PCC 7120 and *Anabaena variabilis*. The comparison of orthologs in which SDR elements are present revealed that SDR insertion was a recent evolutionary event. Evidences that suggest the mobility of these elements were seen, and a few different means of SDR propagation were proposed. The function of SDR is still unknown, but there is evidence that suggests the potential function of SDR in transcriptional processing, which was experimentally tested.

Introduction

Both eukaryotic and prokaryotic genomes are known to have repetitive DNA sequences. However, it was thought that repeats were rare in prokaryotic genomes due to their compact genomes and minimal non-coding regions[1]. The advancement in computational genome analysis has permitted a more thorough examination of DNA sequences. As a result, it is now clear that repeat sequences are widespread in prokaryotic genomes. Short interspersed repeats are widely distributed within and among different bacterial species. One such sequence, REP (Repetitive Extragenic Palindrome), was first identified in *E.coli*. REP sequence is characterized as a 38-nt palindromic sequence capable of forming stem-loop or cruciform[2, 3]. Another dispersed sequence, ERIC (Enterobacterial Repetitive Intergenic Consensus), is preferentially located in non-coding transcribed regions and consists of conserved inverted repeats[3]. The ubiquitous presence of these patterned sequences suggests a function that selects for their persistence and/or a rapid means of propagation.

Several functional roles of these dispersed sequences have been hypothesized and tested, but no role is fully understood. The function of BIMEs (Bacterial Interspersed Mosaic Elements) is by far the most extensively studied. BIMEs are short repetitive sequence with conserved palindromic core called PU (palindromic units). Espéli et al. studied bacterial repeat elements called BIMEs (Bacterial Interspersed Mosaic Elements) for their potential involvement in transcription attenuation and mRNA stabilization. BIMEs are flanked by palindromic units (PUs) and either located on the 3' end of transcriptional units or in between genes that belong to the same operon. Insertion of BIMEs between two genes on artificial operons resulted in the reduction of the level of full-length mRNA and in the accumulation of smaller mRNA that corresponds to the size of the upstream gene[4].

Short dispersed repeats (SDR) are a novel type of dispersed repeats found in *Nostoc punctiforme* genome. During the examination of heptameric tandem repeats in $tRNA^{leu}$ introns by Costa el al., it was discovered that a few of the strains contain a short, non-repetitive sequence within the repeat regions[5]. These inserts vary in size and

sequence, and a few of these sequences have been found in other parts of the genome. These repeated elements were examined in *Nostoc* genome as well as the genome of its closely related cyanobacteria.

We studied the distribution of these elements and examined their origin and propagation. In addition, an experimental approach was taken to investigate the possible function of SDR.

Materials and Methods

Computational Survey of Short Dispersed Repeats

The short insertion sequences from $tRNA^{leu}$ intron were collected and used to look for similar sequences. Three of those (Nos30, Nos37/Nos38, and Nos51) were considered for further study and were designated as reference sequences for SDR1, 2, and 3, respectively. The study of SDR1 revealed that there are other short repeated sequences in the genome as well. Collectively, we used eight reference sequences for this study.

The genomes of *N. punctiforme, Anabaena PCC7120*, and *Anabaena variabilis* were examined to search for similar sequences to each reference sequence. The search was conducted using BioBIKE, a web-based bioinformatics tool that facilitates programming and integrated analysis of biological data[6]. Unlike traditional BLAST (Basic Local Alignment Search Tool), BioBIKE allows searches on the basis of sequence identity, near identity, and pattern. This allows more thorough search of the interested sequences. For this study, we allowed up to three mismatches in the subject sequences. The BioBIKE function SEQUENCE-SIMILAR-TO allows the user to set the number of mismatches to be allowed in the subject. Each returned subject in the initial query was used as the reference sequence for the second turn and on, also allowing up to three mismatches. The returned results were examined by hand to see the patterns and significance.

Investigation of the Possible Functional Role of SDR

pVCU303 was constructed by introducing KpnI and ClaI recognition sequences in the luxAB intergenic region of pRL559. pRL559 is a shuttle vector which bears glnA promoter (P_{glnA}) upstream of luxAB gene (figure 1).

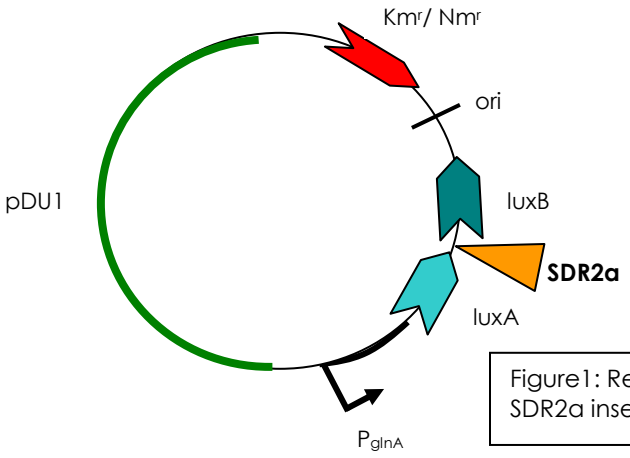

Figure1: Representation of a vector with SDR2a inserted in luxAB intergenic region

Various insertion sequences were constructed by annealing two synthesized oligomers. The double-stranded DNA was inserted into pVCU303 at KpnI and ClaI, creating four different plasmids (table 1).

Table 1: The list of insertion sequences in each construct. Each 24-nt sequence is flanked by 7-nt sequence, **AACTCTT**

Construct	Insertion Sequence	Description
pVCU 304	AACTCTT**GTACAGACGCGATTAATCGCGTCT**AACTCTT	SDR2a sequence flanked by 7-mer
pVCU 305	AACTCTT**AGACGCGATTAATCGCGTCTGTAC**AACTCTT	SDR2a in reverse orientation flanked by 7-mer
pVCU 306	AACTCTT**TTCCGTGCTAATGCACGGAAGCAT**AACTCTT	random 24-nt palindrome flanked by 7-mer
pVCU 307	AACTCTT**GGTCTAACGTTACCAAGTGCCATG**AACTCTT	random non-palindromic 24-mer flanked by 7-mer

Table 2: Bacterial strains and plasmids

Strain or Plasmid	Relavant Characteristics
Anabaena PCC 7120	Wild type
pRL559	Shuttle vector with glnA promoter upstream of luxAB gene; Km^R
pVCU303	pRL559 with ClaI and KpnI recognition sites in the luxAB intergenic region; Km^R
pVCU304	SDR2a sequence with flanking heptamers inserted paralell to luxAB gene in pVCU303 at ClaI and KpnI; Km^R
pVCU305	SDR2a sequence with flanking heptamers inserted antiparalell to luxAB gene in pVCU303 at ClaI and KpnI; Km^R
pVCU306	Random palindromic 24-mer with flanking heptamers inserted paralell to luxAB gene in pVCU303 at ClaI and KpnI; Km^R
pVCU307	Random non-palindromic 24-mer with flanking heptamers inserted paralell to luxAB gene in pVCU303 at ClaI and KpnI
pRL443	RP4-derived conjugative plasmid; Ap^R, Tc^R
pRL528	Helper plasmid; Cm^R, Sm^R

The plasmids bearing different insertion sequences were transferred to Anabaena PCC 7120 by spot mating method[7]. The strains are grown in the presence of nitrogen and total RNA will be extracted for Real-Time PCR analysis. The gene expression of luxA and luxB genes will be measured separately.

Results and Discussion

Distribution and structure of SDR

We have learned that SDR elements are distributed throughout *Nostoc punctiforme* genome as well as the genomes of two *Anabaena* species. These elements were categorized into eight groups by sequence and pattern similarities. For example, SDR1 is characterized by the presence of 10-nt core sequence (**GAGCG.AG.CGA**) and is

flanked by 7-nt tandem repeats. SDR2a and SDR2b have 10-nt palindromic repeats and often flanked by 7-nt tandem repeats. Among the eight SDR families, SDR1, 4, and 5 are the most abundant. The distribution and the location of SDR are summarized in the table below[8].

Table 3: The number and distribution of each SDR family

				In	Intergenic[5]		
Family	**Length[1]**	**Total[2]**	**Copy[3] ≥ 3**	**genes[4]**	**P**	**C**	**D**
SDR1	24	227	104	23*	49	13	19
SDR2a	24	135	54	11*	27*	13*	3*
SDR2b	25	86	54	0*	27*	24*	3*
SDR3	25?	34	13	1*	7	1	4
SDR4	21	344	145	21*	91*	13*	20*
SDR5	21	188	56	15*	18	8	15
SDR6	26?	71	45	1*	31	7	6
SDR7	28?	46	9	0*	3*	5*	1*
SDR8	24?	36	18	0*	13	3	2
Total counts		1099	498	72	266	87	73
(percentage)			(45%)	(14%)	(53%)	(17%)	(15%)
Fraction of genome in given context				81%	11%	2.6%	5.6%

* signifies P < 2% per chi square test done between in genes vs intergenic and a second test between P vs C vs D
Counts for IPCD are confined to SDR elements with 3 or more instances

Origin of SDR

The hypothesis that SDR insertion was a recent event in evolution was examined by comparing orthologs between three closely related species. There are several instances where SDR insertion was found inside of protein-coding genes. Upon comparing the orthologs, we found that there are few cases where the SDR insertion was unique to *Nostoc*. This finding supports that SDR insertion occurred after *Nostoc* diverged from *Anabaena* species (figure 1).

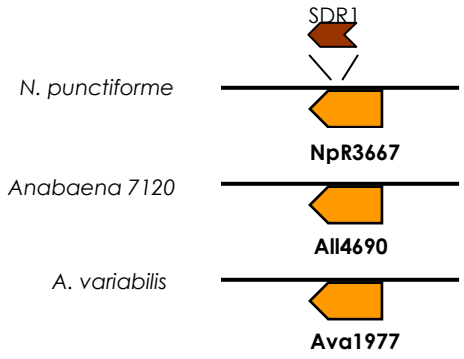

Figure 1: An example of ortholog comparison between three closely related cyanobacteria. In this case, SDR1 is only present in the gene of *Nostoc*.

Mobility of SDR and the mechanisms of insertion

The examination of SDR occurrences in conserved regions shows that SDR had arisen by insertion. There is no mechanism that plausibly explains the insertion of SDR so far. However, we observed several characteristics that indicate how SDR might propagate.

First, there are instances where several different SDR elements are located in close proximity to each other, forming "composites". These composite SDR are often composed of SDR1, 4, and 6. Further analysis revealed that these elements are interrupting an SDR1 sequence (figure 2). This suggests that one SDR1 element was introduced to the genome, which then became the target site for other SDR elements to be inserted.

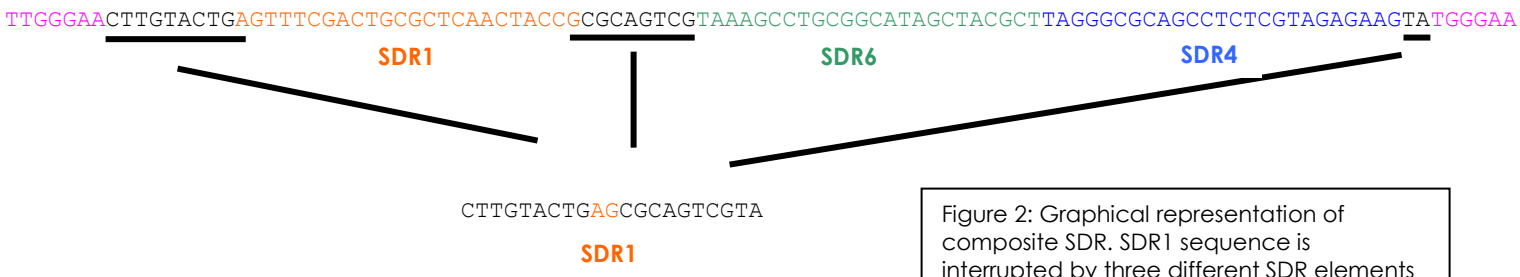

Figure 2: Graphical representation of composite SDR. SDR1 sequence is interrupted by three different SDR elements

Another evidence of mobility was seen where SDR5 was found inside of genes. In one example, SDR5 was present in a gene of *Nostoc* but absent in its orthologous genes. A closer look of the region surrounding SDR5 showed that a few nucleotides on both ends of SDR5 were shared between the orthologs. It appears that SDR5 is inserted by interrupting the 8-nt sequence "GCGATCGC". This 8-nt sequence is known as HIP1 sequence which has been found to be overrepresented in many cyanobacterial genomes[9]. It may by possible that HIP1 sequence is a target site for SDR insertion.

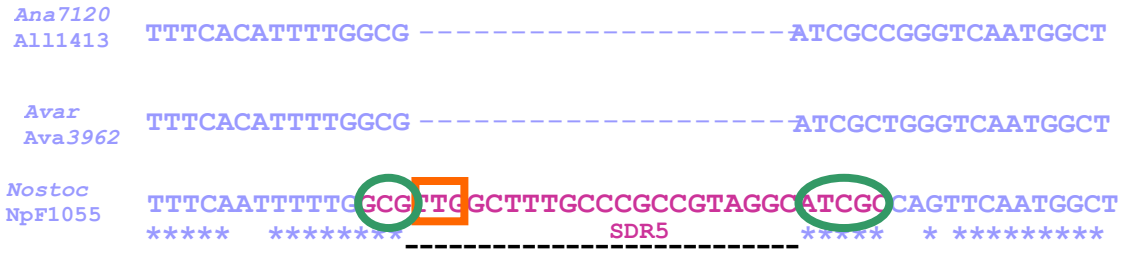

Figure 3: Comparison between orthologs where clean SDR5 insertion is seen

.

Possible Function of SDR

There is nothing that indicates the function of SDR. However, we noticed that SDR2 sequences are located preferentially downstream of coding genes. Table 3 shows that 70% of SDR2 sequences are located between two genes in either parallel or convergent orientation. In addition, SDR2 has a conserved 10-nt core inverted repeat that can adapt stem-loop structure. These characteristics suggest that SDR2 may have a role in transcriptional processing or termination.

The hypothesis was tested by inserting SDR2 sequence between two genes and measuring the level of transcription. The construction of vectors that bear the insertion sequences was successful. The plasmids were introduced to *Anabaena* cells by conjugation.

Total RNA will be extracted from the wild type and the synthesized strains and the level of luxA and luxB transcripts will be measured separately by Real Time PCR. If the presence of SDR affects transcription, decrease in the level of downstream (luxB) transcript will be expected.

Reference

1. Rocha E.P.C, Danchin A, and Viari A (1999). Functional and evolutionary roles of long repeats in prokaryotes. Res. Microbiology. 150: 725-733
2. Stern MJ, Ferro-Luzzi Ames G et al. (1984). Repetitive Extragenic Palindromic Sequences: A major component of the bacterial genomes. Cell. 37: 1015-1026
3. Lupski JR and Weinstock GM (1992). Short, interspersed repetitive DNA sequences in prokaryotic genomes. Journal of bacteriology. 174: 4525-4529.
4. Espéli O, Moulin L, and Boccard F (2001). Transcription attenuation associated with bacterial repetitive extragenic BIME elements. Journal of Molecular Biology. 314:375-386.
5. Costa JL, Paulsrud P, and Lindblad P (2002). The cyanobacterial $tRNA^{leu}$ (UAA) intron: Evolutionary paterns in a genetic marker. Mol. Biol. Evol. 19: 850-857.
6. Massar JP, Travers M, Elhai J, and Shrager J (2005). BioLingua: a programmable knowledge environment for biologists. Bioinformatics. 21: 199-207.
7. Elhai J, Wolk CP (1988). Conjugal transfer of cyanobacteria. Methods in Enzymology167: 747-754.
8. Elhai J, Kato M, Costa JL, Cousins S, and Lindblad P. (2006) Very short mobile repeated elements in the genome of the cyanobacterium *Nostoc punctiforme* ATCC29133. Unpublished.
9. Robinson NJ, Robinson PJ, Gupta A, Bleasby AJ, Whitton BA, and Morby AP (1995). Singular over-representation of an octameric palindrome, HIP1, in DNA from many cyanobacteria. Nucleic Acids Research. 23: 729-735.