Identification of mirror repeats within the egl-1 gene of Caenorhabditis elegans

[1]Mamta Dhankhar

Ph.D. Scholar, Department of Zoology, Baba Mastnath University, Rohtak,Haryana, India

Email: mamtadhankhar48@gmail.com

[2]Deepti Yadav

Ph.D. Scholar, Department of Zoology, Baba Mastnath University, Rohtak, Haryana, India

Email: deeptiyadav304@gmail.com

[3]Vikash Bhardwaj* (corresponding author)

Associate professor, Department of zoology, Baba Mastnath University, Rohtak, Haryana, India
Email: vikashbhardwaj@gmail.com

ABSTRACT

Caenorhabditis elegans (*C. elegans*) is a model organism widely used for biological research. The genome of *C. elegans* is 100MB distributed over six chromosomes and comprises several repetitive sequences element like tandem repeats, direct repeat, palindromic sequences, transposons, and retrotransposons. One type of repeat which is not well studied in *C. elegans* is mirror repeat. The current study aims to determine mirror repeats within the *egl-1* apoptotic gene of *C. elegans*. A simple and swift method was used to extract perfect and imperfect mirror repeats within the *egl-1* gene. We identified 31 mirror repeats within the *egl-1* gene. The identified mirror repeats sequences are not limited to *egl-1* gene of *C. elegans* but are well distributed within the genome of *C. elegans, C. vulgaris, Xenopus tropicalis,* and *Drosophila melanogaster*. The future molecular and biophysical studies will explore the unknown function of these identified mirror repeats.

Keywords: *Caenorhabditis elegans, egl-1* apoptotic gene, BLAST tool, DNA and mirror repeats.

DOI Number: 10.14704/nq.2022.20.9.NQ44135 **Neuro Quantology 2022; 20(9):1196-1203**

1. INTRODUCTION

Caenorhabditis elegans is one of the most widely used model organism in various field of genomic, cell biology, and biological research [1,2]. It is a unique choice for research based on its features such as small size, short life cycle, minimal growth, and nutritional requirements, compact genome, and well-studied development biology [3,4]. The *C. elegans* has a smaller genome size of 100 MB spreaded over six chromosomes, and it was the first multicellular organism whose genome was completely sequenced in 1998 [5,6]. The genome of *C. elegans* annotated with 19,735 protein-coding genes, forming 22,269 unique polypeptide sequences and 1300 noncoding RNA genes, including 630 tRNA, 17snoRNA, and 78 snRNAs genes, etc. [7]. A total of 83% *of C. elegans* proteome have human orthologs and 84% of *C. elegans* proteome have mouse orthologs [8,9]. So, most *C. elegans* protein are evolutionary conserved with respect to humans and mice proteins [10].

A major fraction of genome sequence is repetitive in nature in multicellular organism [11]. Within the genome of *C. elegans*, several types of repeats sequences have been identified, and these are randomly distributed throughout the genome [12,13]. Repetitive DNA is fragments of DNA that repeat itself from hundreds to thousands of times, either in tandem or interspersed manner along the genome [14]. Tandem repeats contain mini and microsatellite DNA, whereas dispersed repeats contain long and short interspersed elements (LINEs and SINEs). Based on the orientation of the repeating unit, some other repeats such as direct repeat; inverted repeat, and

mirror repeats, can be found in any organisms [15]. Direct repeats are repeating unit that is exactly duplicated within the same strand while inverted repeats are repeating unit that is reversed and duplicated on the opposite strand. Mirror repeats are those repeats that possess a center of symmetry and repeating unit is duplicated within the same strand but in opposite orientation. For example, ACGTAATGCA is a mirror repeat in which ACGTA is a repeating unit that is duplicated within the same strand but in reverse order. Mirror repeats have a unique potential to form triplex DNA and other non-canonical B DNA such as H-DNA and associated with many neurological diseases such as Friedreich ataxia, Fragile X syndrome, myotonic dystrophy and amyotrophic lateral sclerosis [16-20]. Apoptosis or programmed cell death is an essential and vital process in *C. elegans*, and the *egl-1* gene is the main apoptotic gene within the *C. elegans* [21]. Future molecular and biophysical study will aid us to find whether these identified mirror repeats have any role in the mechanism of programmed cell death in *C. elegans*.

2. MATERIAL AND METHODS

The complete nucleotide sequences of the *egl-1* gene in FASTA format were procured from National Center for Biotechnology information (NCBI) by gene name (*egl*-1 gene and Accession No-003283). The *egl-1* gene sequence was divided into two parts of 500 base pairs and 552 base pairs. These nucleotides represent query sequence. Using the reverse complement tool, a parallel complement of each part was constructed. The complement sequence depicts your subject sequence. Both query and subject sequences were aligned for a similar search using the BLAST tool taking algorithm parameter for word size limit of 7, threshold value as shown in Table 1. The mirror repeats can be identified if the number of nucleotides is reversed in both query and subject [22]. Based on the spacer element, identified mirror repeats were classified into perfect and imperfect mirror repeats. To find out similar repeats within the genome of *C. elegans*, as well as other species like *Caenorhabditis vulgaris, Drosophila melanogaster*, and *Xenopus tropicalis,* megablast tool was used. Algorithm parameters were taken in megablast for short input sequences with a minimum word size of 16 at an expect threshold value of 100.

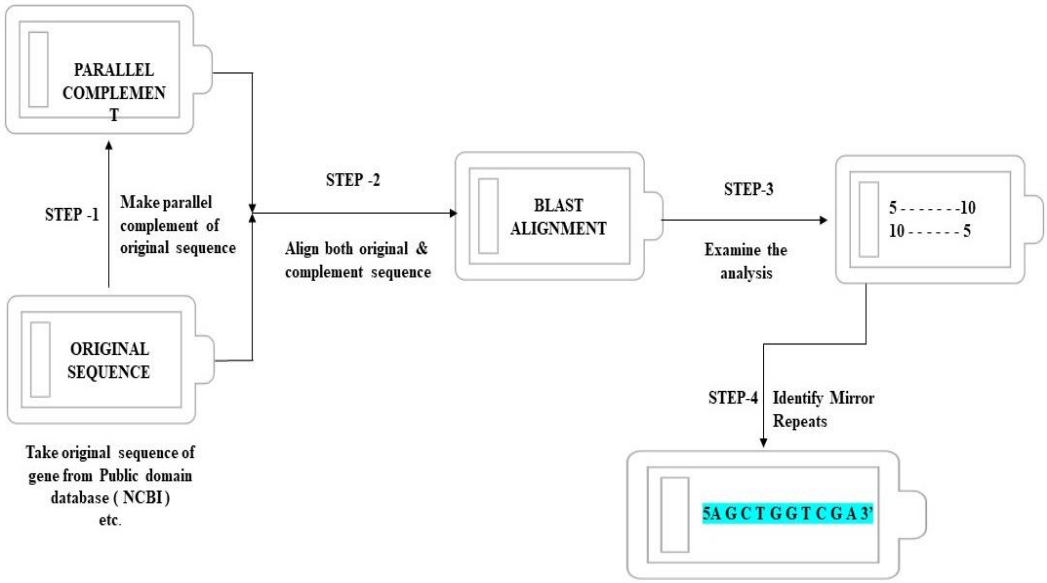

Figure: 1 Diagrammatic Representation of Methodology to Identify Mirror Repeats.

3. RESULT AND DISCUSSION

The *egl-1* gene belongs to BH3 family of proteins which initiate apoptosis in *C. elegans* [23]. More than a dozen of genes are involved in programmed cell death, but *egl-1, ced-9*, *ced-4* and *ced-3* genes are significant genes for apoptosis. During development of *C. elegans*, if mutation or loss of function occur in *egl-1* gene, then all those cells survive that would ordinarily die. The activity of *egl-1* gene is regulated by various transcriptional positive and negative regulator such as TRA-1, HLH-2/HLH-3 and CES-1 [24,25].

In the current study, a simple computational approach was used to find out mirror repeats present within the *egl-1* gene of *C. elegans*. The *egl-1* gene is located on chromosome V in *C. elegans,* and consists of 1052 base pairs with three exon sequences. We identified a total of 31 mirror repeats in the whole gene.

Table: 1 shows the number of identified mirror repeats present in different exons, gene parts, and complete *egl-1* gene of *C. elegans* at different E value varying from 0.05 to 100.

Expected Threshold	Exon 1	Exon 2	Exon 3	Part 1	Part 2	Gene
0.5	0	0	3	1	5	4
1	0	2	4	4	5	8
5	0	3	8	9	6	9
10	0	3	11	9	13	14
20	0	3	11	13	18	14
30	0	3	11	13	18	21
50	0	3	11	13	18	22
100	0	3	11	13	18	31

The identified mirror repeats have random distribution within the *egl-1* gene. The identified mirror repeats were classified in two categories perfect and imperfect mirror repeats. Perfect mirror repeats have similar sequences around center of symmetry and while imperfect mirror repeats have mismatches around the central axis. Depending upon the spacer present between them, perfect mirror repeats can be further divided into single, double or multispacer mirror repeats. The frequency of perfect mirror repeats within *egl-1* gene was more than the imperfect mirror repeats. Out of 31, 9 perfect mirror repeats without any spacer, 16 perfect mirror repeats with single spacer and 6 imperfect mirror repeats were identified. Small-sized mirror repeats (7-8 bps) were more commonly seen within *the egl-1* gene. Mirror repeats with their types, position, and length are shown below in Table 2. The maximum length of mirror repeats within the *egl-1* was of 36 nucleotides, and the smallest mirror repeat was of 7 nucleotides. Shorter mirror repeats could not be identified as minimum word size limit of BLAST tool is 7 bp. Mirror repeats that are rich in homopurine and homopyrimidine bases with a size of greater than 10 (>10) can adopt a H-DNA structure [26]. We observed 2 homopyrimidine rich sequences have a size greater than 10 base pairs within the *egl-1* gene.

Table: 2 shows length, types, and position of mirror repeats within the whole *egl*-1 gene.

S. No	Mirror Sequences	Type of mirror repeats	Position	Length
1	TCTAGAAATTATTTATTTTTGATCT	Imperfect mirror	76-100	25
2	ACCT**G**TCCA	Perfect mirror with single spacer	105-113	9
3	CTCCCCC**T**CCCCCTC	Perfect mirror with single spacer (homopyrimidine rich sequences)	125-139	15
4	CCTCC**C**CCTCC	Perfect mirror with single spacer (Homopyrimidine rich sequences)	130-140	11
5	AAA**C**AAA	Perfect mirror with single spacer	145-151	7
6	TTTTT**G**TTTTT	Perfect mirror with single spacer	152-162	11
7	TGTT**T**TTGT	Perfect mirror with single spacer	156-164	9
8	AAA**C**AAA	Perfect mirror with single spacer	216-222	7
9	ACAA**A**AACA	Perfect mirror with single spacer	218-226	9
10	ATT**G**TTA	Perfect mirror with single spacer	279-285	7
11	ATCTTCTA	Perfect mirror	374-381	8
12	GTT**T**TTG	Perfect mirror with single spacer	394-400	7
13	TCAGGACT	Perfect mirror	453-460	8
14	TAGTAA**A**AATGAT	Perfect mirror with single spacer	540-552	13
15	TTTTGAAAGTATT	Imperfect mirror	552-564	13
16	GAAAGTATTTCAAAAAGTTGTCAAAATT-ATGAAAG	Imperfect mirror	556-590	36
17	AAACTTTTCAAA	Perfect mirror	622-633	12
18	AAATTAAA	Perfect mirror	642-649	8
19	AAT**G**TAA	Perfect mirror with single spacer	648-654	7
20	TTT**G**TTT	Perfect mirror with single spacer	657-663	7
21	AGCATCGGCTACGA	Perfect mirror	747-760	14

22	TCGGCTACGAG-ATCGGCT	Imperfect mirror	751-768	19
23	GTCTTCTG	Perfect mirror	853-860	8
24	GACTTTTTCGCTTTTTAAG	Imperfect mirror	861-879	19
25	TTTGTACCATGATT	Imperfect mirror	907-920	14
26	CATAATAC	Perfect mirror	924-931	8
27	AACTCCCCTCAA	Perfect mirror	980-991	12
28	TATTTAT	Perfect mirror with single spacer	1020-1026	7
29	TAATAAT	Perfect mirror with single spacer	1028-1034	7
30	AATAATAA	Perfect mirror	1029-1036	8
31	ATAAATA	Perfect mirror with single spacer	1033-1039	7

1200

Figure 2 represents the location of all identified mirror repeats within the whole *egl-1* gene. The exons are highlighted with different colours, while remaining part represent introns as shown in figure 2. Pink stars indicate mirror repeats present within exons (exon 1, exon 2 and exon 3), and blue stars represent mirror repeats present in introns. Here 17 mirror repeats are present in the intron and 14 are present in exons of *egl-1* gene.

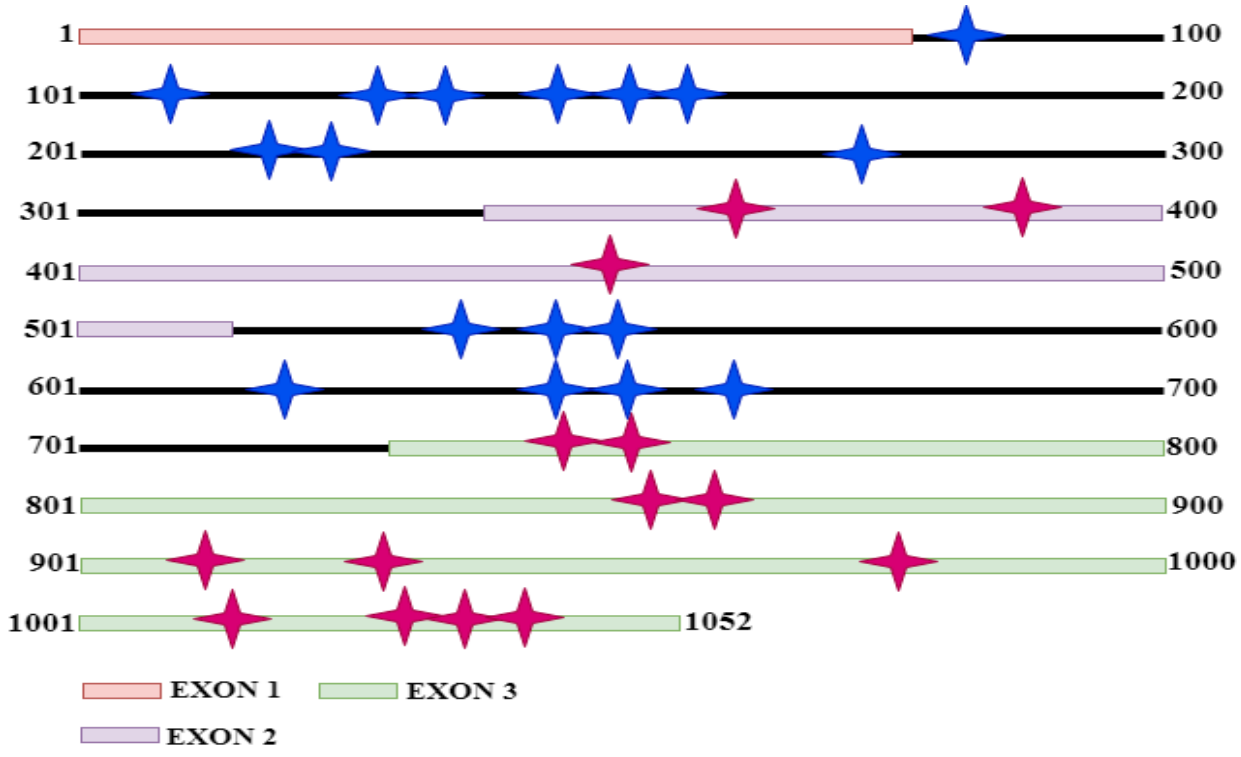

Figure: 2 Diagrammatic representation of position of identified mirror repeats within the *egl-1* gene of the *C. elegans*

Table: 3 shows the identified mirror repeats present in different genera's by using megablast tool

Mirror sequences	*C. elegans*	*C. vulgaris*	*X. tropicalis*	*D. melanogaster*
AAA**C**AAA	-	-	-	-
AAA**C**AAA	-	-	-	-
ATT**G**TTA	-	-	-	-
GTT**T**TTG	-	-	-	-
AAT**G**TAA	-	-	-	-
TTT**G**TTT	-	-	-	-
TAT**T**TAT	-	-	-	-
TAA**T**AAT	-	-	-	-
ATA**A**ATA	-	-	-	-
AAATTAAA	-	+	-	-
ATCTTCTA	-	+	-	-
TCAGGACT	-	+	-	-
GTCTTCTG	-	+	-	-
CATAATAC	-	+	-	-
AATAATAA	-	+	-	-
TGTT**T**TTGT	-	+	-	-
ACCTGTCCA	-	+	-	-
ACAA**A**AACA	-	+	-	-
CCTCCCCCTCC	+	+	+	+
TTTTT**G**TTTTT	+	+	+	+
AAACTTTTCAAA	+	+	+	+
AACTCCCCTCAA	+	+	+	+
TAGTAAAAATGAT	+	+	+	+
TTTTGAAAGTATT	+	+	+	+
AGCATCGGCTACGA	+	-	-	-
TTTGTACCATGATT	+	-	-	-
CTCCCCCTCCCCCTC	+	-	+	+
TCGGCTACGAG-ATCGGCT	+	-	-	-
GACTTTTTCGCTTTTTAAG	+	-	-	-
TCTAGAAATTATT TATTTTTGATCT	+	-	-	-
GAAAGTATTTCAAAAAGTTGT CAAAATT-ATGAAAG	+	-	-	-

Using megablast tool, available at NCBI, we further searched identified mirror repeat within the genome of *C. elegans* and other species like *Caenorhabditis vulgaris, Drosophila melanogaster*, and *Xenopus tropicalis.* Here, **+** sign shows presence of mirror and **–** sign shows absence of mirror repeats. megablast tool was unable to detect small size mirror repeats. Very small size mirror repeats like 7-8 base pairs are not well distributed within the genome of *C. elegans* and the genome of other organisms like *Drosophila melanogaster* and *Xenopus tropicalis* and similarly, very large size mirror repeats are also not well distributed as shown in table 3 given below. But mirror repeats size of 11-13 base pairs are well distributed within the genome of other species like *C. vulgaris*, *Xenopus tropicalis* and *Drosophila melanogaster*, as shown in table 3. This result clearly indicates the universal distribution of mirror repeats sequences of size 11-13 base pairs in all genera's like *Caenorhabditis vulgaris, Xenopus tropicalis*, and *Drosophila melanogaster*. These sequences were conserved during the course of evolution. Therefore, they must be performing something important role within the genome of living beings.

In this research, we have identified mirror repeats of different length and types within *egl-1* gene of *C. elegans*. Some identified mirror repeats that are rich in homopurine and homopyrimidine bases that have potential to form non canonical B-DNA structure such as triplex DNA. Various biological processes such as replication pausing, inhibit transcription, repair, recombination regulation of gene expression and genetic instability which leads to human disorder like cancer are influenced by triplex DNA [27-30]. These identified mirror repeats whether play any significant role in gene function and molecular metabolism within *C. elegans* is still unknown. Nonetheless, to understand the mechanism, our future target is to identify the proteins that may bind to mirror DNA repeats.

4. CONCLUSION

The current studies determined mirror repeats within *the egl-1* apoptotic gene in *C. elegans*. A total of 31 mirror repeats were determined by using the BLAST tool. It will be highly important to discover the function of mirror DNA in the transcription and translation process occurring in a cell at the molecular level. Up till now, the exact role of mirror repeats has not been elucidated. Future studies may be required to determine the precise function of mirror repeats. So, this study of identification and assessment of mirror repeats in apoptotic gene *egl-1* in *C. elegans* will continue to contribute in a crucial way to our present knowledge of apoptosis or programmed cell death.

5. FUNDING

This research received no external funding.

6. CONFLICT OF INTEREST

None disclosed.

REFERENCES

1. Meneely PM, Dahlberg CL, & Rose JK. Working with worms: *Caenorhabditis elegans* as a model organism. Current Protocols Essential Laboratory Techniques, 2019 19(1), e35.
2. White CV, & Herman MA. Transcriptomic, functional, and network analyses reveal novel genes involved in the interaction between *Caenorhabditis elegans* and *Stenotrophomonas maltophilia*. Frontiers in cellular and infection microbiology, 2018, 266.
3. Frézal L, & Félix MA. The natural history of model organisms: *C. elegans* outside the Petri dish. elife, 2015,4, e05849.
4. Leung MC, Williams PL, Benedetto A, Au C, Helmcke KJ, Aschner M, & Meyer JN. *Caenorhabditis elegans*: an emerging model in biomedical and environmental toxicology, . Toxicological sciences, 2008,106(1), 5-28.
5. Howe KL. A new reference genome sequence for *Caenorhabditis elegans*?. Lab animal, 2019, 48(9), 267-268.
6. Tejeda-Benitez L, & Olivero-Verbel J. *Caenorhabditis elegans*, a biological model for research in toxicology. Reviews of Environmental Contamination and Toxicology Volume, 2016, 237, 1-35.
7. Hillier LW, Coulson A, Murray JI, Bao Z, Sulston JE, & Waterston RH. Genomics in *C. elegans*: so many genes, such a little worm. Genome research, 2005,15(12), 1651-1660.
8. Kim Y, Park Y, Hwang J, & Kwack K. Comparative genomic analysis of the human and nematode *Caenorhabditis elegans* uncovers potential reproductive genes and disease associations in

humans. Physiological Genomics, 2018 *50*(11), 1002-1014.
9. Henricson A, Sonnhammer E L, Baillie DL, & Gomes A V. Functional characterization in *Caenorhabditis elegans* of transmembrane worm-human orthologs. BMC genomics,2004, *5*(1), 1-18.
10. Lai C H, Chou CY, Ch'ang LY, Liu CS, & Lin WC. Identification of novel human genes evolutionarily conserved in *Caenorhabditis elegans* by comparative proteomics. Genome research, 2000 10(5), 703-713.
11. *C. elegans* Sequencing Consortium*. Genome sequence of the nematode *C. elegans*: a platform for investigating biology,1998. Science, 282(5396), 2012-2018.
12. Pilgrim D. CeRep25B forms chromosome-specific minisatellite arrays in *Caenorhabditis elegans*. Genome research,1998 8(11), 1192-1201.
13. Berks M. The *C. elegans* genome sequencing project. *C. elegans* Genome Mapping and Sequencing Consortium. Genome research,1995, *5*(2), 99-104.
14. Biscotti MA, Olmo E, & Heslop-Harrison J P. Repetitive DNA in eukaryotic genomes, 2015.
15. Gurusaran M, Ravella D, & Sekar K. RepEx: repeat extractor for biological sequences. Genomics,2013, 102(4), 403-408.
16. Paulson H. Repeat expansion diseases. Handbook of clinical neurology, 2018,147, 105-123.
17. Shah K A, & Mirkin S M. The hidden side of unstable DNA repeats: Mutagenesis at a distance. DNA repair, 2015,32, 106-112.
18. Georgakopoulos-Soares I, Victorino J, Parada GE, Agarwal V, Zhao J, Wong H Y, & Ahituv N. High-throughput characterization of the role of non-B DNA motifs on promoter function. Cell genomics, 2022;*2*(4), 100111.
19. Guiblet WM, Cremona MA, Harris RS, Chen D, Eckert KA, Chiaromonte F, et al. Non-B DNA: a major contributor to small- and large-scale variation in nucleotide substitution frequencies across the genome. Nucleic Acids Research, 2021;49(3):1497–1516.
20. Mirkin SM. DNA topology fundamentals. Encyclopedia of Life Sciences, 2001,111.
21. Arvanitis M, Li D, Lee K, & Mylonakis E. Apoptosis in *C. elegans*: lessons for cancer and immunity. Frontiers in cellular and infection microbiology,2013, 3, 67.
22. Vikash B, Swapni G, Sitaram M, & Kulbhushan S. FPCB: a simple and swift strategy for mirror repeat identification. arXiv preprint arXiv:1312.3869, 2013.
23. Cohn JA. Characterizing transcriptional and morphological consequences of long-term activity in a tonic *C. elegans* sensory neuron (Doctoral dissertation),2019.
24. Conradt B, Wu Y C, & Xue D. Programmed cell death during *Caenorhabditis elegans* development. Genetics,2016, 203(4), 1533-1562.
25. Conradt B, & Xue D. Programmed cell death. In WormBook: The Online Review of *C. elegans* Biology [Internet]. WormBook, 2005.
26. Schroth G P, & Ho PS. Occurrence of potential cruciform and H-DNA forming sequences in genomic DNA, 1995 Nucleic acids research, 23(11), 1977-1983.
27. Brazda V, Fojta M, & Bowater R P. Structures and stability of simple DNA repeats from bacteria. Biochemical Journal,2020, 477(2), 325-339.
28. Bacolla A, Wang G, & Vasquez KM. New perspectives on DNA and RNA triplexes as effectors of biological activity. PLoS genetics,2015, 11(12), e1005696.
29. Wang G, & Vasquez KM. Naturally occurring H-DNA-forming sequences are mutagenic in mammalian cells. Proceedings of the National Academy of Sciences,2004, 101(37), 13448-13453.
30. Jain A, Wang G, & Vasquez KM. DNA triple helices: biological consequences and therapeutic potential. Biochimie, 2008, 90(8), 1117-1130.