Uttar Pradesh Journal of Zoology

Volume 45, Issue 17, Page 539-545, 2024; Article no.UPJOZ.3946
ISSN: 0256-971X (P)

Deciphering the Nucleotide Distribution Dynamics of rDNA Internal Transcribed Spacer 1 in Fish Species *Rita rita* Using R Programming

Mohd Imran [a*]

[a] *Section of Genetics, Department of Zoology, Aligarh Muslim University-202002, India.*

Author's contribution

The sole author designed, analysed, interpreted and prepared the manuscript.

Article Information

DOI: https://doi.org/10.56557/upjoz/2024/v45i174398

Open Peer Review History:
This journal follows the Advanced Open Peer Review policy. Identity of the Reviewers, Editor(s) and additional Reviewers, peer review comments, different versions of the manuscript, comments of the editors, etc are available here: https://prh.mbimph.com/review-history/3946

Short Research Article

Received: 16/06/2024
Accepted: 19/08/2024
Published: 28/08/2024

ABSTRACT

The internal transcribed spacer region 1 (ITS1) of *Rita rita* was explored using R programming language with an approach to better interpret its secondary structure in terms of its sequence and structural attributes. It has shown a multi-helical secondary structure composed of loops and self-complementary helical regions. A series of functions from tidyverse core packages and bioseq package in RStudio were used to extract the information from ITS1. The distribution frequency of nucleotides and base pairs were traced on the constituent helices of secondary structure. Moreover, the self-complementary sequence motifs were also extracted and tabulated. Additionally, statistical computing and data visualization using R programming approach has made it easier to represent the sequence data graphically using ggplot by providing direct functions which in turns provided a very effective preliminary characterization of nucleotide composition and dynamics of this non-coding ITS1 region.

**Corresponding author: Email: mimran@myamu.ac.in;*

Cite as: *Imran, Mohd. 2024. "Deciphering the Nucleotide Distribution Dynamics of RDNA Internal Transcribed Spacer 1 in Fish Species Rita Rita Using R Programming". UTTAR PRADESH JOURNAL OF ZOOLOGY 45 (17):539-45. https://doi.org/10.56557/upjoz/2024/v45i174398.*

Keywords: R programming; ITS1; tidyverse; secondary structure; Rita rita.

1. INTRODUCTION

The programming languages has become an integral part of biological research whether it is the data analysis, manipulation or retrieving large datasets from public databases [1-3]. For the same purpose the RStudio which is a working platform based on R programming language has been proved as an efficient platform for bioinformatic analysis without limits, and its capability to keep evolving with the development of new packages as per the target data type and its subsequent analysis [4]. On the same pattern, the present study has targeted the rRNA gene internal transcribed spacer 1 (ITS1) region by incorporating the bioseq package which is comprehensive tool to target the biological sequence including DNA, RNA and Amino acids. With this background, the study aims to do the nucleotide sequence analysis of ITS1 region from the catfish species *Rita rita* with perspective of looking into the nucleotide dynamics along its secondary structure using R programming.

2. METHODOLOGY

Multiple individuals of *Rita rita* were taken to investigate the internal transcribed spacer region 1 (ITS1) for their different sequence motif and properties using R programming language [4]. The specimens were collected from Aligarh (27°58'41.57"N, 78° 8'17.96"E). Approximately 0.5 ml of the blood extracted from the heart and caudal vein in EDTA coated vial. Extraction of total genomic content, including both nuclear and mitochondrial genome, was performed according to the high salt method [5]. ITS1 sequences were obtained following the methodology using ITS1 specific primers from Imran & Nafees [6]. The sequences were submitted to NCBI (accession no. MT105366-MT105369) The sequence alignment was done using Clustal W [7] which showed all the individual possess the similar ITS1 sequence. The secondary structure reconstructed using RNAfold web server [8] and different attributes of secondary structure were explored using R programming like complementary regions in each helix of secondary structure, GC and AT content, nucleotide frequencies, basepair frequencies including non-canonical basepairs. The R programming based ITS1 sequence analysis was done using the bioseq package [9] along with the tidyverse [10]. The tidyverse consist of several core package which provide diverse functions for data analysis and visualization in R programming language such as dplyr, readr, forcat, stringer, ggplot, tibble, lubridate, tidyr and Purrr. Out of the two vignettes available on Github for the bioseq, the introduction to bioseq was used as a reference guide to execute different function. First of all, RNA vector of the transcribed ITS1 sequence was created using the rna() function for further analysis. Nucleotide sequence of different helices and subhelices of *R.rita* ITS1 secondary structure were extracted by position number function seq_extract_position (), including the self-complementary regions of the ITS1 sequence. All the extracted sequence motifs were simultaneously converted to table format using function as_tibble.bioseq_rna () from tibble package (ver. 3.2.1) [11], and all the individual sequence table were later combined into one using the rbind() function from the R base package to form a single dataset characterizing the different complementary parts of ITS1 sequence. The nucleotide frequencies were calculated separately for each helix with bioseq functions: seq_stat_prop() and seq_stat_gc(), converted into tibble format and combined into single table one by one using the function full_join() from dplyr package (ver. 1.1.4) [12] and exported into csv format using the function fwrite() from *data.table* package (ver.1.15.4) [13] and then after performing required editing further imported back into the R from data directory using function read_csv() from tidyverse package readr (ver. 2.1.5) [14]. The wide format table was converted into long format using the function pivot_longer() from tidyr package (ver. 1.3.1) [15], where the nucleotide frequency values for different helices from different column were placed into one column to make it easier to work with data visualization using ggplot2 (ver.3.5.1) package available with tidyverse [16]. All the tabular data generated in the R were exported using the package data.table with function fwrite() which directly exports the data file into csv format.

3. RESULTS

The ITS sequence of *R. rita* have shown a multi-helical secondary structure compose of loops and self-complementary helical regions by RNAfold web server (Fig. 1). These complementary regions were extracted from the RNA vector of the transcribed rDNA sequence taken from the RNAfold into the R using the extract position function seq_extract_position()

manually one by one and combined into single table (Table 1). Appropriate labels were added to table using table editor function edit().These regions varied in length from a two G-C bps long segment in helix2 up to 19-22 bps complementary nucleotide string in the same helix, and interspersed with the non-complementary loops. The nucleotide frequencies were calculated separately for each helix where the GC composition found to be much higher in each helix than AU composition with overall GC content reaches more than 70% (Fig. 2). The basepair frequencies data including non-canonical basepairs was generated manually in csv format and imported into R for data visualization with ggplot which shows how the four nucleotides vary among themselves and how much GC overmasked this non-coding ITS1 sequence. Consecutively, G-C bps found contributing the maximum to the self-complementary base pairing (Fig. 3).

Table 1. Complementary sequences of the *R.rita* ITS1 secondary structure. The helix position label shows start and end point of the respective sequence in each helix. The complementary sequences are given in pairs. The sequences were extracted in R using the bioseq function seq_extract_position.

S.No.	Helix_Position_start_end	Sequence
1.	helix1_pos_3_13	GGGUAGCGCCC
	helix1_pos_18_27	GGGCGCGCCU
2.	helix2_pos_30_35	UCUACC
	helix2_pos_318_323	GGUGGA
3.	helix2_pos_42_49	CGAGGCUC
	helix2_pos_274_281	GAGCCUCG
4.	helix2_pos_50_65	GGGUUCUCGGGGGGUC
	helix2_pos_70_80	GACCCUCUCCC
5.	helix2_pos_81_89	GGCCUUAGG
	helix2_pos_227_233	CCUGGCC
6.	helix2_pos_91_102	CGCUCGUAACGG
	helix2_pos_218_225	CCGGGGCG
7.	helix2_pos_104_113	CUCCCUGGAU
	helix2_pos_209_215	GUCGGAG
8.	helix2_pos_115_117	CGG
	helix2_pos_205_207	CCG
9.	helix2_pos_120_123	GGGA
	helix2_pos_199_202	UCCC
10.	helix2_pos_127_129	GGA
	helix2_pos_195_197	UCC
11.	helix2_pos_132_134	GGC
	helix2_pos_191_193	GUC
12.	helix2_pos_136_154	CGAGGGUACCUGCUGCCCG
	helix2_pos_168_189	CGGGCGAGGUUCAAAGACCCCG
13.	helix2_pos_158_159	CC
	helix2_pos_163_164	GG
14.	helix2_pos_239_245	UCGUUCC
	helix2_pos_267_273	GGAACGA
15.	helix2_pos_248_251	CACA
	helix2_pos_258_261	UGUG
16.	helix2_pos_302_306	GCUGA
	helix2_pos_313_317	UUAGC
17.	helix3_pos_328_333	UCGGCU
	helix3_pos_335_341	GUGCGUC
18.	helix3_pos_348_356	GAACGCAGC
	helix3_pos_358_363	AGCUGA

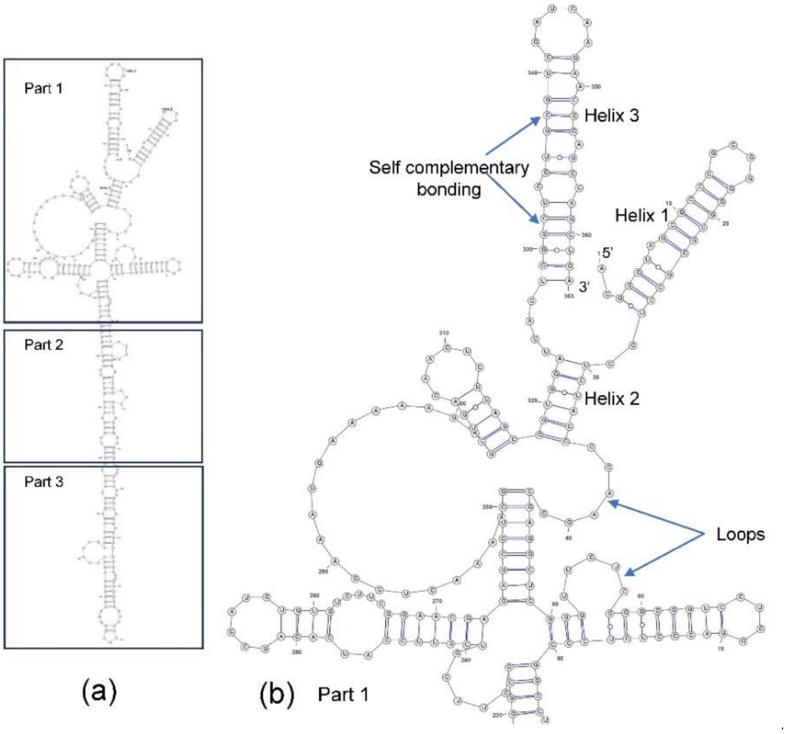

Fig. 1. Secondary structure of *R. rita* ITS1. The structure is shown in magnified view in parts as part 1, 2 and 3. The structure is composed of three helices with multiple branching in helix 2. The helices are interspersed with loop of different size. The structure visualization was done with VARNA software [17].

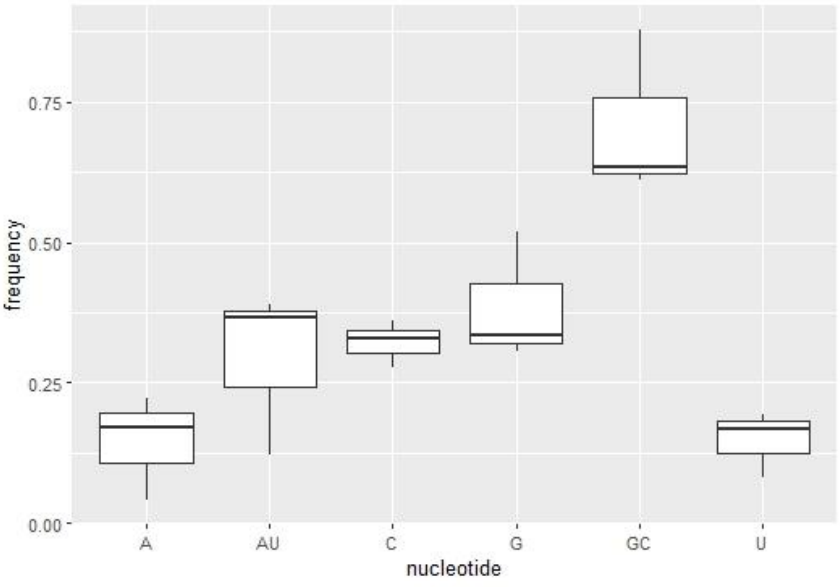

Fig. 2. Boxplot showing the nucleotide composition range of the *R. rita* ITS1 across its three helices. All four nucleotides have been shown separately along with the AU and GC content combined. The data visualization was done in RStudio.

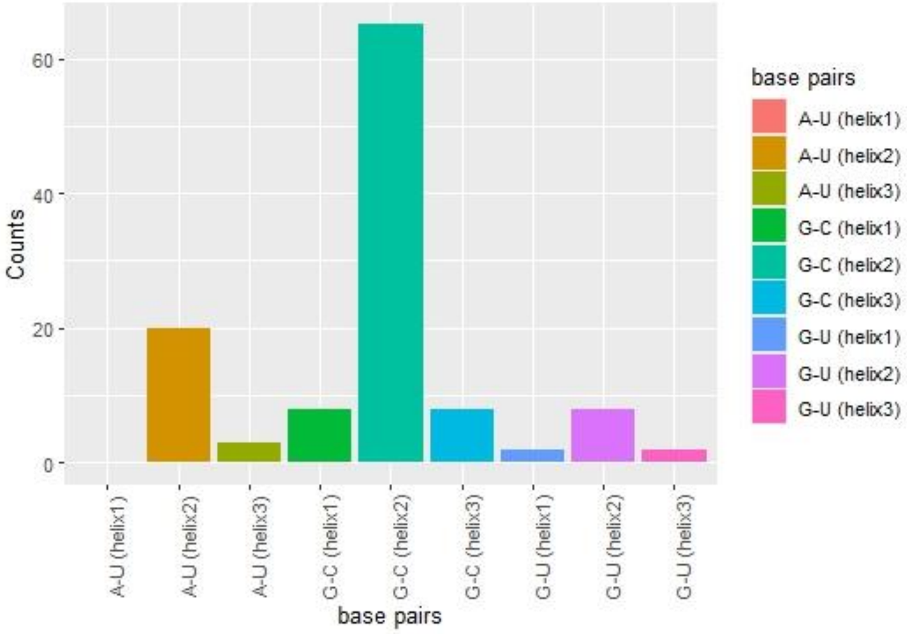

Fig. 3. The barplot showing the composition of the different basepair combinations in helix 1, 2 and 3. The data visualization was done in RStudio.

4. DISCUSSION

The R programming based analysis of ITS1 secondary structure of *R. rita* using *bioseq* package has generated good amount of information about its nucleotide dynamics and self-complementary sequences. The data generated has brought a better understanding about this noncoding segment of vertebrate genome. This has shown about how the four nucleotides A,T/U,G,C are variably distributed and how this distribution is defining the secondary structure where AU is least at the 5’end and hence majority of the self-complementary bonding are the G-C bps with no A-U bps, an indication towards the stability of the helix 1. Likewise, other two helices have also shown a comparative higher G-C base pairing with an overall contribution of around 70% in complete ITS1 sequence. The nucleotide frequencies effect the basepair compositions along the secondary structure which in turn responsible for the self-complementary regions where Guanine contributing the maximum as G-C and G-U bps. The nucleotide frequencies, basepair compositions, and self-complementary regions all together appears providing a shape and definition to the secondary structures of ITS1.

The bioseq R package has made it easier to calculate all the parameters and visualize them for better understanding of the nucleotide dynamics especially when non-canonical basepairs are also a part of the structure. The bioseq package which consist of direct functions to calculate the several parameters of DNA/RNA sequence, along with the basic R functions of manipulation and data representation, has made it easier to extract the desired sequence motifs, frequency and distribution detail of constituent nucleotide and basepairs in the secondary structure.

5. CONCLUSION

The data extracted from ITS1 sequence of *Rita rita* using bioseq has brought further clarity to the structure. The calculation of this nucleotide dynamic is relevant, as it appears a key factor responsible for how a secondary structure looks like especially when this structure plays a crucial role in the maturation of the precursor rRNA molecule, where it provides spatial geometry vital for the binding of snoRNA and RAC protein complex necessary for proper ribosomal maturation, and at this point, using R programming with bioseq package is a quick and precise method to estimate these parameters. Moreover, using programming language like R which consist of diverse packages as per the target analysis makes it easier to interpret the

DNA or RNA sequence data in graphical format for better interpretation.

DISCLAIMER (ARTIFICIAL INTELLIGENCE)

Author(s) hereby declare that NO generative AI technologies such as Large Language Models (ChatGPT, COPILOT, etc) and text-to-image generators have been used during writing or editing of manuscripts.

ACKNOWLEDGEMENT

The Author is grateful to Prof. Mohammad Afzal for his constructive insights into this work. The authors also acknowledge the Chairman Department of Zoology, Aligarh Muslim University, for providing necessary laboratory facilities. The Council of Scientific and Industrial Research (CSIR), New Delhi, India is also gratefully acknowledged for providing funding support (09/112(0603)/2018-EMR-I).

COMPETING INTERESTS

Author has declared that no competing interests exist.

REFERENCES

1. Rani J, Shah AB, Ramachandran S. Pubmed.mineR: An R package with text-mining algorithms to analyse PubMed abstracts. Journal of Biosciences. 2015; 40:671–682.
2. Newman V, Moore B, Sparrow H, Perry E. The Ensembl Genome Browser: Strategies for Accessing Eukaryotic Genome Data. In: Kollmar, M. (eds) Eukaryotic Genomic Databases. Methods in Molecular Biology. Humana Press, New York, NY. 2018; 1757.
3. Gaynor ML, Landis JB, O'Connor TK, Laport RG, Doyle JJ, Soltis DE, Ponciano JM, Soltis PS. nQuack: An R package for predicting ploidal level from sequence data using site-based heterozygosity. Applications in Plant Sciences. 2024;12:e11606
4. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria; 2021. Available:https://www.R-project.org/.
5. Montgomery GW, Sise JA. Extraction of DNA from sheep white blood cells. New Zealand Journal of Agricultural Research. 1990;33:437-441.
6. Imran M. and Nafees S. Exploring unique sequence repeat patterns and secondary structures in rDNA internal transcribed spacers ITS1 and ITS2 for characterization of catfish species. bioRxiv; 2024. Available:https://doi.org/10.1101/2024.01.10.575031
7. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 1994;22:4673–4680.
8. Gruber AR, Lorenz R, Bernhart SH, Neuböck R. Hofacker IL. The Vienna RNA websuite. Nucleic Acids Res. 2008;36: W70-W74.
9. Keck F. Handling biological sequences in R with the bioseq package. Methods Ecol. Evol. 2020;11:1728–1732.
10. Wickham H, Averick M, Bryan J, Chang W, McGowan DL, François R, et al. Welcome to the tidyverse. J Open Source Softw. 2019;4:1686.
11. Müller K, Wickham H. Tibble: Simple data frames. R package version 3.2.1. 2023. Available:https://CRAN.R-project.org/package=tibble.
12. Wickham H, François R, Henry L, Müller K, Vaughan D. dplyr: A Grammar of Data Manipulation. R package version 1.1.4. 2023. Available:https://CRAN.R-project.org/package=dplyr.
13. Barrett T, Dowle M, Srinivasan A, Gorecki J, Chirico M, Hocking T. data.table: Extension of 'data.frame'. R package version 1.15.4. 2024. Available:https://CRAN.R-project.org/package=data.table.
14. Wickham H, Hester J, Bryan J. readr: Read Rectangular Text Data. R package version 2.1.5. 2024; Available:https://CRAN.R-project.org/package=readr.
15. Wickham H, Vaughan D, Girlich M, tidyr: Tidy Messy Data. R package version 1.3.1. 2024. Available:https://CRAN.R-project.org/package=tidyr.
16. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York; 2016.

17. Darty K, Denise A, Ponty Y. VARNA, Interactive drawing and editing of the RNA secondary structure. Bioinformatics. 2009; 25:1974–1975.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of the publisher and/or the editor(s). This publisher and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Peer-review history:
The peer review history for this paper can be accessed here:
https://prh.mbimph.com/review-history/3946