International Journal of Entomology Research
www.entomologyjournals.com
ISSN: 2455-4758
Received: 17-07-2022, Accepted: 04-08-2022, Published: 20-08-2022
Volume 7, Issue 8, 2022, Page No. 125-131

Missing TRNAs in MITOS gene annotation software: alternative methods to predict and generate secondary structures of tRNA genes within the complete mitogenome of insects

Muzafar Riyaz, Rauf Ahmad Shah, Sivasankaran K*
Division of Taxonomy and Biodiversity, Entomology Research Institute, Loyola College, Chennai, Tamil Nadu, India

Abstract
Apart from playing a significant role in metabolism, apoptosis, illness, and ageing, Mitochondrial (mt) genomes have emerged as a significant source of data for comparative genomics. With a typical insect mtDNA genome size of less than 16 kb, largely preserved gene completion, and a high rate of nucleotide substitution, this small genome makes it the perfect vehicle for a variety of comparative investigations that have advanced our understanding of genome evolution. Insect mitochondrial genomes are now the most widely utilized molecular markers for population genetics, phylogeography, molecular diagnostics, and phylogenetic investigations because of their relative simplicity in amplification and sequencing during the past two decades. The tRNA molecule has a characteristic folded structure with three hairpin loops like a three-leafed clover. One of these hairpin loops has an anticodon sequence, which can detect and decode an mRNA codon. MITOS webserver is a frequently used software for gene annotation. However, during our analysis of complete mitochondrial genome, the *trnH* gene was missing from the annotation table generated in MITOS software for several mitogenomes and there fore secondary structures was not made. To circumvent this problem, we searched for alternative software packages and online webservers that can predict the secondary structures of tRNA genes and enable researchers study and analyze gene studies.

Keywords: tRNA, secondary structure, webservers, softwares, gene studies

Introduction
A certain type of DNA can be found inside the mitochondrion, which is different from the DNA present in the nucleus of a cell. This mitochondrial DNA (mtDNA) is small and circular and has 16,500 base pairs in it (Figure 1). The mtDNA encodes different proteins that are specific to the mitochondria. The insect mitochondrial genomes are commonly circular with about 14–19 kb long which contains 37 genes including 13 protein-coding genes (PCGs): two ATPase genes (atp6 and atp8), three cytochrome c oxidase genes 1–3 (cox1-cox3), one cytochrome B (cob), seven NADH dehydrogenase genes (nad1-6 and nad4L), two ribosomal RNA (rrnL and rrnS) genes and 22 transfer RNA (tRNA) genes and an adenine (A) + thymine (T)-rich region, which comprises the initiation sites for transcription and replication (Sivasankaran *et al*, 2017) [1]. The mitochondrial genome is a compact structure, extremely conserved, maternally inherited, lacking genetic recombination and having a relatively fast evolutionary rate. Apart from playing a significant role in the pathways that are within the mitochondrion for producing energy, mtDNA plays a very crucial role in the ecology and evolution of an insect. Mitochondrial DNA is an excellent molecular marker in analyzing the studies of comparative and evolutionary genomics, molecular evolution, phylogenetics, and population genetics of insects. Mitochondrial genomes have been widely used to address phylogenetic questions in invertebrates, particularly in insects (Riyaz *et al*, 2022) [2].

Gene annotation
The process of extracting the structural and functional information of a protein or gene from a raw data collection using various analysis, comparison, estimate, precision, and other mining approaches is known as genome annotation. Genome annotation is required because genome or DNA sequencing creates sequence information that does not have a functional role (Salzberg, 2019) [3]. The genome must be annotated once it has been sequenced in order to provide more logical information about its structural properties and functional activities. Annotation files include information about the genomic sequence. FASTA, GFF3, and GENBANK are examples of file formats. There are several file formats for representing sequence, structure, pathway information connected to genes and proteins, and internet databases allow you to pick and download a specific file. The genes or proteins that may be recruited by a certain genomic sequence can be predicted using gene annotation algorithms. Functional annotation of these novel genes or proteins can be accomplished by comparing them to well experimentally confirm sequences accessible in databases (Ejigu and Jung, 2020) [4].

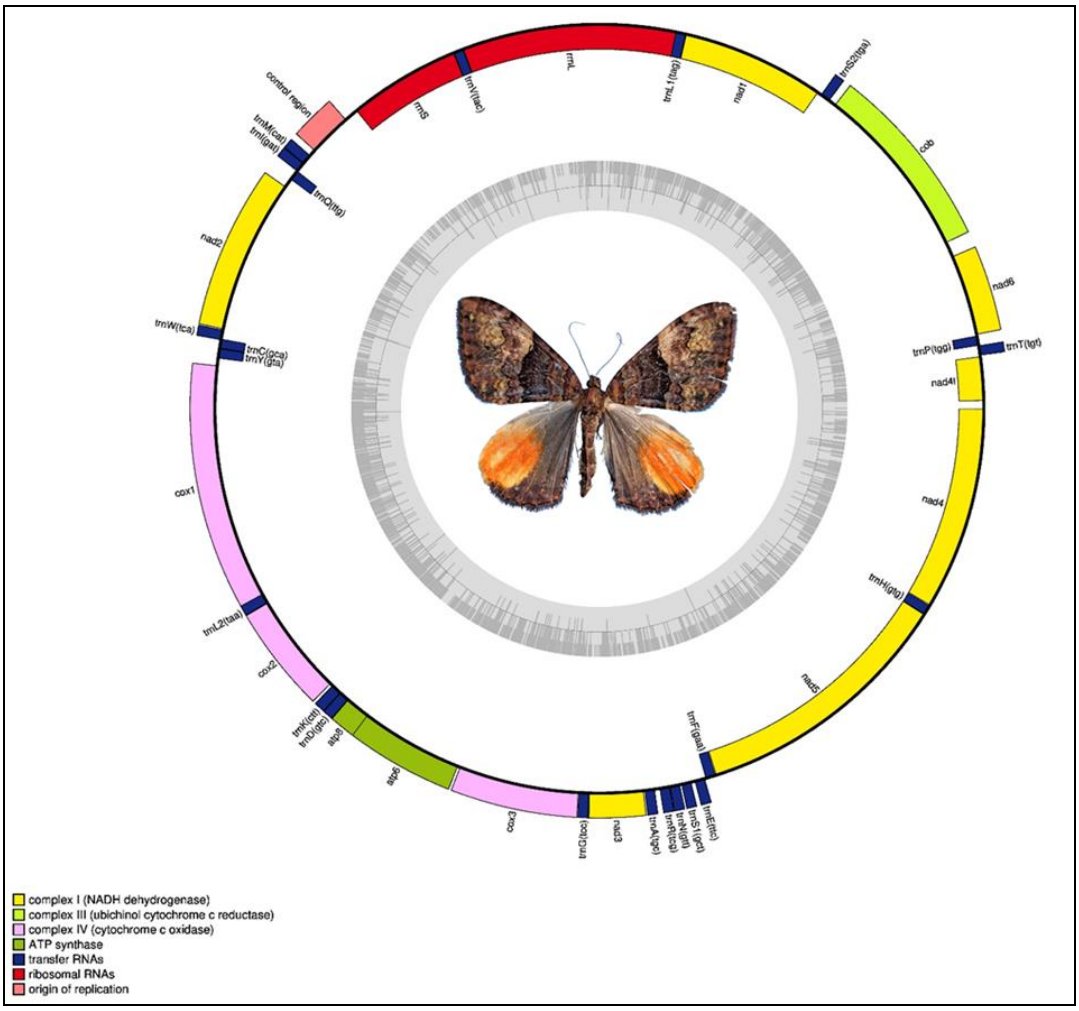

Fig 1: A circular map of complete mitogenome of an anonymous moth species

Annotating genes using MITOS webserver

For a systematic comparative evaluation of genomic sequence data, an accurate and consistent genome annotation is a necessary factor. The study of the mechanisms underlying genomic rearrangements, phylogenetic reconstruction, and the analysis of the impacts of sequence variation have all been instances where it would be particularly true (Figure 2). When automated procedures are used to analyze the progressively larger volumes of data that are becoming accessible, as a result of new sequencing technologies, the necessity for accurate and impartial annotations becomes even more crucial.

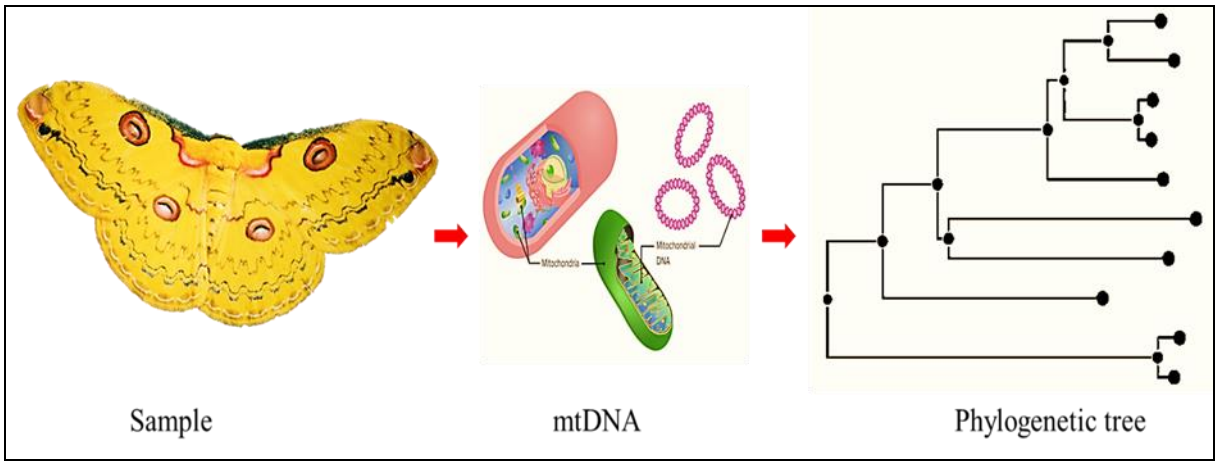

Fig 2: Utilizing mitogenome, a species' phylogenetic tree can be reconstructed

For the de *novo* annotation of metazoan mitochondrial genomes, a completely automated pathway is accessible through the mitochondrial genome annotation Server (MITOS) (http://mitos2.bioinf.uni-leipzig.de/index.py). It does not require a built-in database of well-selected protein models since it employs a revolutionary technique

based on combining BLAST searches with existing annotated protein sequences to find protein-coding genes. For each of the structured RNAs, particular covariance models are used to annotate the tRNAs and rRNAs. Metazoan mitochondrial genomes may be automatically annotated using the web server MITOS (Jühling *et al*, 2012) [5]. The annotation of proteins and non-coding RNAs is made possible by MITOS. The innovative structure-based covariance models reported in (Jühling *et al*. 2012) [5] are used to annotate tRNAs. As opposed to ARWEN and tRNAscan-SE, this method was proven to have an unequalled sensitivity as well as an accuracy that was on par with tRNAscan-SE and greater than ARWEN.

Using MITOS to annotate mitochondrial genomes is a credible yet effective and uncomplicated process, relieving the user of the burden of attempting to pick cutoff values and other similar tasks (Bernt *et al*, 2012) [6]. The only thing that has to be provided is a FASTA file, which should include the genome that needs to be annotated together with the relevant genetic code. There is no need for the user to provide either their name or email address. However, in the case that a user choose to do so, after the task has been finished, a link to the results page will be sent to the address that was provided. After submitting the task, the user will be taken to a page that shows the current state of the work, a link that enables the job to be deleted, and as soon as the job is done, the results will be displayed on the same page. In addition to a graphical summary of the findings, the results are provided in the following file formats: BED, GFF, FASTA, and TBL format (Table 1). In addition, a file that may be used for analysis of genome rearrangement, such as those performed using CREx, is given. This file contains the gene order, which lists the gene names in the order in which they exist on the genome. In addition, the raw data, which includes all of the files that have been produced by MITOS, may be downloaded in the form of a zip package.

So far, we have submitted 36 complete mitochondrial genome sequences and all of them have been given accession numbers by the GENBANK. While the sequences are being uploaded to the MITOS webserver, the tRNA gene was not predicted by the software and was missing from the gene annotation table predicted by the software. Furthermore, we were unable to generate secondary structures of the tRNA gene due to this issue. New software that can generate the cloverleaf secondary srtcutres of tRNA genes and predict the missing genes has been explored to bypass this problem in the present study.

Table 1: Annotation table of an anonymous complete mitogenome generated in MITOS webserver with missing *trnH* gene

Name	Start	Stop	Strand	Length	Intergenic nucleotides	Codons	Infos
OH_1-b	4	44	+	41	0		
trnM(cat)	45	112	+	68	0		svg ps
trnI(gat)	113	179	+	67	-3		svg ps
trnQ(ttg)	177	245	-	69	61		svg ps
nad2	307	1320	+	1014	7	ATT/TAA	
trnW(tca)	1328	1396	+	69	-8		svg ps
trnC(gca)	1389	1454	-	66	7		svg ps
trnY(gta)	1462	1527	-	66	15		svg ps
cox1	1543	3081	+	1539	-5	ATG/TAA	
trnL2(taa)	3077	3143	+	67	0		svg ps
cox2	3144	3828	+	685	-3	ATG/T(AA)	
trnK(ctt)	3826	3896	+	71	2		svg ps
trnD(gtc)	3899	3964	+	66	0		svg ps
atp8	3965	4129	+	165	-7	ATT/TAA	
atp6	4123	4800	+	678	-1	ATG/TAA	
cox3	4800	5588	+	789	2	ATG/TAA	
trnG(tcc)	5591	5655	+	65	3		svg ps
nad3	5659	6009	+	351	12	ATT/TAA	
trnA(tgc)	6022	6086	+	65	-1		svg ps
trnR(tcg)	6086	6149	+	64	9		svg ps
trnN(gtt)	6159	6225	+	67	12		svg ps
trnS1(gct)	6238	6303	+	66	2		svg ps
trnE(ttc)	6306	6371	+	66	4		svg ps
trnF(gaa)	6376	6442	-	67	4		svg ps
nad5	6447	8165	-	1719	45	ATT/TAA	
nad4	8211	9536	-	1326	105	ATG/TAA	
nad4l	9642	9929	-	288	27	ATG/TAA	
trnT(tgt)	9957	10020	+	64	0		svg ps
trnP(tgg)	10021	10087	-	67	37		svg ps
nad6	10125	10628	+	504	-7	ATT/TAA	
OH_2	10622	10660	+	39	-12		
cob	10649	11797	+	1149	-1	ATG/TAA	

trnS2(tga)	11797	11862	+	66	23		svg ps
nad1	11886	12824	-	939	1	ATG/TAA	
trnL1(tag)	12826	12894	-	69	3		svg ps
OH_1-a	12898	12972	+	75	-39		
rrnL	12934	14235	-	1302	37		svg ps
trnV(tac)	14273	14339	-	67	-1		svg ps
rrnS	14339	15072	-	734	98		svg ps
OH_0	15171	15567	+	397	3		

TRNA Secondary Structure

Transfer RNA, sometimes referred to as tRNA, is a type of Ribonucleic acid that aids in the synthesis of protein from mRNA. During translation, tRNA serves as an adaptor molecule that connects amino acids and nucleic acids. It transports the amino acid that has to be included in the peptide chain and decodes the mRNA molecule's codon for that amino acid (Khorana, 1995) [7]. The length of tRNAs ranges from 76 to 90 nucleotides. The tRNAs for each amino acid are specific and none of the tRNAs recognize stop codons. Base pairing in at least four locations helps the tRNA molecules fold into a cloverleaf secondary structure and maintain its form. As a result, three loops and four stems or arms are produced.

The cloverleaf structure of tRNA has three distinctive loops. The loop closest to the 5' end in the Figure 3 is known as the dihydrouridine arm (D arm) as it incorporates dihydrouridine bases, which are unique nucleotides found exclusively in tRNA (Holley *et al*, 1965) [8]. The sequence of thymine, pseudouridine, and cytosine in the loop closest to the 3' end is known as the T arm. The anticodon, which complementarily binds to the mRNA codon, is located in the loop at the bottom of the cloverleaf. Since anticodons pair with codons antiparallel to one another, they are transcribed backwards, from 5' to 3'. Through its acceptor stem, a particular tRNA binds to a specific amino acid. The cloverleaf reflects the actual tRNA structure in a two-dimensional structure. Therefore, the cloverleaf is considered a secondary structure. In addition, the cloverleaf continues to fold into a tertiary structure that resembles an ill-defined L-shape. The anticodon is located at one end of the L, while the acceptor stem is located at the other. The anticodon and acceptor stem, the two active ends of tRNA, are just amplified by the L-shaped structure (Cramer *et al*, 1969) [9].

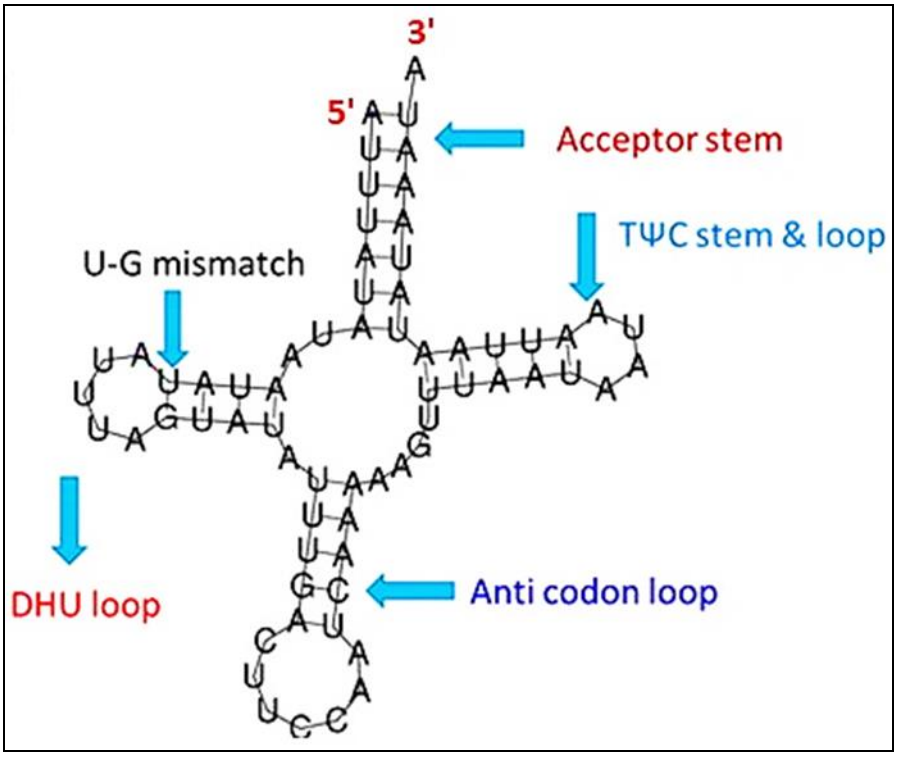

Fig 3: A cloverleaf structure of an anonymous tRNA gene

Alternative approaches for gene annotation and cloverleaf structure

The ability to correctly align sequences is a crucial step in molecular phylogenetic investigations, and so is a firm grasp of RNA's secondary structure. Alignments that take secondary structure into account may produce better phylogenetic trees than those generated by automatic alignment techniques that employ just primary sequences, which may misalign RNA sequences. We need reliable models for the sequences from these species or their relatives before we can utilize secondary structure to align RNA sequences. In the present study, we have taken into account a few online software packages that are designed to make it simpler and more accurate for beginners to annotate genes.

forna server (Webserver)
Forna is a feature-rich, user-friendly, and attractive tool for visualising RNA secondary structure developed by ViennaRNA Web Services (http://rna.tbi.univie.ac.at/). Without installing any software, it enables you to see and change RNA secondary structures online and allows users to download the structures in PNG, SVG, and JSON formats.

Bioinformatics Web Server for RNA
The Bioinformatics webserver is an online collection of various webservers, including CentroidFold, CentroidHomfold, Rtips, Rchange, CapR, Raccess, RintD, and RintW, that enable users to simply paste a homologous sequence and obtain results in a few seconds.

CentroidFold
CentroidFold is one of the most precise techniques for predicting RNA secondary structures based on a generalized centroid estimator. The projected secondary structure is coloured based on the likelihood of base pairing (Hamada *et al*, 2009a) [10].

CentroidHomfold
CentroidHomfold uses automatically gathered homologous sequences of the target to predict RNA secondary structures. LAST is used to extract homologous sequences from the Rfam database. Using homologous sequence information together with the probabilistic consistency transformation for base-pairing probabilities, CentroidHomfold may predict secondary structures for the target sequence more precisely than CentroidFold if homologous sequences are available (Hamada *et al*, 2009b) [11].

RintD
RintD verifies the secondary RNA structure. CentroidFold (McCaskill's inference engine) and RNAfold forecast the secondary structures of the targets (Minimum free energy structure) (Mori *et al*, 2014) [12].

RintW
RintW, an efficient programme that determines the base pairing probability distributions based on the Hamming distance from a canonical secondary structure, does the decomposition calculation (Figure 4). This methodology breaks down the base pairing probability matrix to find important alternative secondary structures in an RNA sequence (Hagio *et al*, 2018) [13].

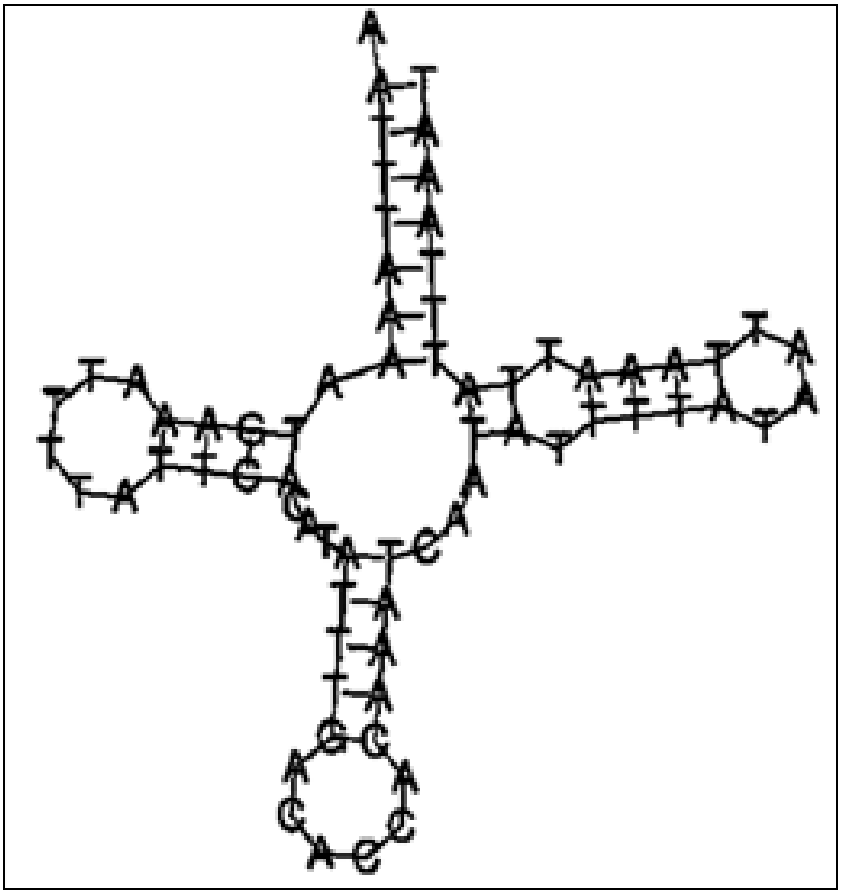

Fig 4: Secondary and alternative structure of anonymous *trnH* sequence generated in RintW using Bioinformatics Web Server for RNA

RNAstructure

RNAstructure is a software tool for predicting and analysing RNA secondary structure (http://rna.urmc. rochester.edu/RNAstructureWeb/). This contribution outlines a new set of web servers that will be used to deliver the functionality. The online server provides RNA secondary structure prediction, including free energy reduction, structure prediction with the highest predicted accuracy, and pseudoknot prediction. Prediction of bimolecular secondary structure is also offered. Furthermore, the server may forecast secondary structures that are preserved in two or more homologs (Mathews, 2014) [14].

MitoZ (Software package)

Even though mitochondrial tRNAs (mt-tRNAs) often have the well-known cloverleaf shape, they exhibit low levels of primary sequence conservation and structural divergence across various lineages. Raw data pre-treatment, de novo assembly, candidate mitochondrial sequences search, mitogenome annotation, and visualization are just a few of the components that MitoZ can provide users. In the event that users just desire a portion of the whole workflow, each module may operate independently (Meng *et al*, 2019) [15].

MitoZ employs MiTFi, a technique based on covariance models (CM), to annotate mt-tRNAs. Using the software Infernal, CM is often built using structurally annotated multiple sequence alignments and includes both the sequence and secondary structure. By default, MitoZ generates tRNA annotation results with an e-value 0.001 using the MiTFi parameters '-cores 1-evalue 0.001-onlycutoff-code 2/5(representing Chordate/Arthropod)'.

Conclusion

RNA is a crucial biomolecule that functions as a catalyst, a controller of post-transcriptional alteration and gene regulation, a therapeutic target, and a pharmaceutical. In general, its organisation is hierarchical. The sequence of nucleotides, which is a covalent bond structure, makes up the primary structure. The set of canonical base pairs is referred to as the secondary structure, while the set of extra connections and the three-dimensional structure are referred to as the tertiary structure. In evolutionary and ecological research, the mitochondrial genome (mitogenome) plays a significant role in investigating the phylogeny of organisms. With the advent of High Throughput Sequencing (HTS) technology, using numerous genes on the mitogenome or the complete mitogenome to examine the phylogeny and biodiversity of target groups has become a routine practice in reconstructing the phylogenetic trees. These new web servers carry on this tradition by making the software available to more users. In earlier studies, the algorithms' accuracy was thoroughly validated. The web servers will be updated to keep RNA structure accessible to the community when existing algorithms are upgraded and new algorithms are developed.

Acknowledgment

The authors wishes to thank Entomology Research Institute, Loyola College Chennai for extended support and guidance.

References

1. Sivasankaran K, Mathew P, Anand S, Ceasar SA, Mariapackiam S, Ignacimuthu S. Complete mitochondrial genome sequence of fruit-piercing moth Eudocima phalonia (Linnaeus, 1763) (Lepidoptera: Noctuoidea). Genomics data.,2017:1(14):66-81.
2. Riyaz M, Shah RA, Savarimuthu I, Kuppusamy S. Comparative mitochondrial genome analysis of Eudocima salaminia (Cramer, 1777) (Lepidoptera: Noctuoidea), novel gene rearrangement and phylogenetic relationship within the superfamily Noctuoidea. Molecular Biology Reports.,2021:48(5):4449-63.
3. Salzberg SL. Next-generation genome annotation: we still struggle to get it right. Genome biology.,2019:20(1):1-3.
4. Ejigu GF, Jung J. Review on the computational genome annotation of sequences obtained by next-generation sequencing. Biology.,2020:18;9(9):295.
5. Jühling F, Pütz J, Bernt M, Donath A, Middendorf M, Florentz C, *et al*. Improved systematic tRNA gene annotation allows new insights into the evolution of mitochondrial tRNA structures and into the mechanisms of mitochondrial genome rearrangements. Nucleic acids research.,2012:140 (7):2833-45.
6. Bernt M, Donath A, Jühling F, Externbrink F, Florentz C, Fritzsch G, *et al*. MITOS: improved de novo metazoan mitochondrial genome annotation. Molecular phylogenetics and evolution.,2013:1;69(2):313-9.
7. Khorana HG. Transfer RNA: discovery, early work, and total synthesis of a tRNA gene. In: Söll D, Rajbhandary UL (eds) tRNA: Structure, Biosynthesis, and Function. American Society Microbiology Press, Washington, 1995, 5-16.
8. Holley RW, Apgar J, Everett GA, Madison JT, Marquisee M, Merrill SH, *et al*. Structure of a ribonucleic acid. Science.,1965:147(3664):1462-5.
9. Cramer F, Erdmann VA, Von Der Haar F, Schlimme E. Structure and reactivity of tRNA. Journal of cellular physiology.,1969:74(S1):163-78.
10. Hamada M, Kiryu H, Sato K, Mituyama T, Asai K. Prediction of RNA secondary structure using generalized centroid estimators. Bioinformatics.,2009a:25(4):465-73.
11. Hamada M, Sato K, Kiryu H, Mituyama T, Asai K. Predictions of RNA secondary structure by combining homologous sequence information. Bioinformatics.,2009b:25(12):i330-8.

12. Mori R, Hamada M, Asai K. Efficient calculation of exact probability distributions of integer features on RNA secondary structures. BMC genomics.,2014:(10):1-10.
13. Hagio T, Sakuraba S, Iwakiri J, Mori R, Asai K. Capturing alternative secondary structures of RNA by decomposition of base-pairing probabilities. BMC bioinformatics.,2018:(1):85-95.
14. Mathews DH. RNA secondary structure analysis using RNAstructure. Current protocols in bioinformatics.,2006:13(1):12-6.
15. Meng G, Li Y, Yang C, Liu S. MitoZ: a toolkit for animal mitochondrial genome assembly, annotation and visualization. Nucleic acids research.,2019:47(11):e63-1-7.