Data in brief 29 (2020) 105284

Contents lists available at ScienceDirect

Data in brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Bioinformatic tools for tRNA gene analyses in mitochondrial DNA sequence data

Elena V. Romanova [a, *], Yurij S. Bukin [a, b], Dmitry Yu. Sherbakov [a, b]

[a] Laboratory of Molecular Systematics, Limnological Institute, Irkutsk, Russian Federation
[b] Faculty of Biology and Soil Studies, Irkutsk State University, Irkutsk, Russian Federation

ARTICLE INFO

Article history:
Received 17 January 2020
Received in revised form 4 February 2020
Accepted 6 February 2020
Available online 22 February 2020

Keywords:
tRNA genes
Mitochondrial genomes
Sequence alignment
Genetic distance
R script

ABSTRACT

The data presented here are related to the research article entitled "Hidden cases of tRNA genes duplication and remolding in mitochondrial genomes of amphipods" (Romanova et al., 2020) [1]. Correct tRNA gene sequence annotation in mitochondrial (mt) and nuclear genomes sometimes can be a challenging task because of the differential performances of tRNA annotation/prediction programmes. These programmes may cause false positive or false negative predictions. Moreover, additional difficulties with annotation may be caused by the presence of duplicated tRNA genes and those coding tRNAs with altered identities occurring as due to a mutation in their anticodon sequence (tRNA gene remolding/recruitment).

We developed an R script automating the diagnosis of ancestor tRNA gene coding specificity regardless of anticodon sequence based on genetic distance comparison. Some of the predicted tRNA genes from the mt genomes of amphipods are presented. We also developed an R script for estimation of the best mode of sequence alignment, which was applied to determine the best alignment of tRNA genes in [1], but is also suitable for testing of any nucleotide alignment sets used in phylogenetic inferences.

DOI of original article: https://doi.org/10.1016/j.ympev.2019.106710.
* Corresponding author.
E-mail address: elena_romanova@lin.irk.ru (E.V. Romanova).

https://doi.org/10.1016/j.dib.2020.105284

Specifications Table

Subject	Biochemistry, Genetics and Molecular Biology
Specific subject area	Bioinformatic studies of mitochondrial tRNA gene sequences
Type of data	Code, Table, Figure
How data were acquired	Codes, *In silico* analysis of tRNA gene sequences performed using R scripts.
Data format	Raw, R scripts text, analysed data
Parameters for data collection	tRNA gene sequences from available mitochondrial (mt) genomes of amphipods were predicted using MiTFi [2]
Description of data collection	Custom written R scripts. Analysis of tRNA gene sequences from amphipod mt genomes performed using R scripts.
Data source location	Limnological Institute, Irkutsk, Russia
Data accessibility	The raw data files are provided in the Data in Brief article. The text of codes and the examples of input and output files were deposited in a repository.
Related research article	Romanova, E.V., Bukin, Y.S., Mikhailov, K.V., Logacheva, M.D., Aleoshin, V.V., Sherbakov, D.Yu. 2020. Hidden cases of tRNA genes duplication and remolding in mitochondrial genomes of amphipods. *Molecular Phylogenetics and Evolution*, 144, 106710.

Value of the Data

- The data provide a useful tool for the selection of an optimum mode of alignment of any set of nucleotide sequences that is essential for robust phylogenetic inferences.
- The data present a bioinformatics tool to define true tRNA gene identity regardless of its codon sequence. This will help with performing correct annotation of the tRNA genes in mt genomes and with identifying the true origination of tRNA gene copies that underwent changes to their identity through a mutation in their codon sequence (tRNA gene remolding).
- The data from the pairwise identity analysis estimated for duplicated and some single tRNA genes of interest in mt genomes of amphipods provide additional evidence about the true evolutionary origin of these genes that is useful for understanding their evolutionary dynamics.

1. Data

The data describes the two R scripts: 1) The script for the identification of tRNA gene isoacceptor types based on genetic distance analysis. Additionally, the data from the genetic distance analysis of duplicated and some single tRNA genes from the mt genomes of amphipods are presented in Table S1. 2) The script for choosing the best mode of sequence alignments and the output of its application based on differently aligned groups of tRNA genes from the mt genomes of amphipods are presented in Table S2.

2. Experimental design, materials, and methods

2.1. R script for estimation of tested tRNA gene sequence resemblance to certain isoacceptor type

For the cases when it is necessary to verify tRNA gene sequence predictions by different programmes or to distinguish between original and remolded/recruited tRNA genes [3,4] we developed an R script that performs identification of tRNA isoacceptor type based on estimation and comparison of genetic distance data using “ape” package [5]:

```
library(ape)
data<-read.table("input.txt",header=TRUE, stringsAsFactors=F, sep="\t")
P_value<-c()
D_mean<-c()
D_mean_test<-c()
Ratio<-c()
for(i in 1:nrow(data))
 {
  pas1<-read.dna(data$first_file[i], format="fasta")
  pas2<-read.dna(data$second_file[i], format="fasta")
  d1<-as.vector(dist.dna(pas1, "raw", pairwise.deletion=T))
  d2<-as.vector(dist.dna(pas2, "raw", pairwise.deletion=T))
  nint<-ceiling(1+log2(length(d1)))
  xmax<-max(c(d1,d2))
  xmin<-min(c(d1,d2))
  int<-(xmax-xmin)/nint
  vint<-rep(0, (nint+1))
  vint[1]<-xmin
  vint[nint+1]<-xmax
  for(i in 2:(nint)) vint[i]<-vint[i-1]+int
  xl<-c(xmin-0.15*xmin, xmax+0.1*xmax)
  h1<-hist(d1, breaks=vint, plot=F)
  h2<-hist(d2, breaks=vint, plot=F)
  tab<-data.frame(h1=h1$counts, h2=h2$counts)
  s<-cbind(h1$counts, h2$counts)
  sh<-rowSums(s)
  s<-s[sh!=0,]
  chisq.test(s, correct=FALSE)$p.value
  P_value<-c(P_value, chisq.test(s, correct=FALSE)$p.value)
   d_mean<-mean(d1)
  D_mean<-c(D_mean, d_mean)
  d<-as.matrix(dist.dna(pas2, "raw", pairwise.deletion=T))
  d_pas<-d[nrow(d), -nrow(d)]
  d_test<-mean(d_pas)
  D_mean_test<-c(D_mean_test, d_test)
  Ratio<-c(Ratio, 1-d_mean/d_test)
}
data<-cbind(data, P_value=P_value, D_mean=D_mean, D_mean_test=D_mean_test, Ratio=Ratio)
write.table(data, file="output.txt",row.names=F, col.names = TRUE, sep = "\t")
```

The distributions of pairwise genetic distances were obtained for tRNA genes of interest placed among each isoacceptor tRNA gene set. The content of amphipods species used to create isoacceptor tRNA gene sets for each test is shown in Ref. [1]. tRNA gene sets of every isoacceptor type consist of two input files: the first one contains aligned tRNA genes of a certain isoacceptor type (d1) and the second one has the tested tRNA gene added to the d1 set, and the sequence of the same species as that being examined is removed (d2). The input file (in the script input.txt) is a tab delimited table looks like this:

No	tRNA	first_file	second_file
1	A	A1.fas	A2.fas
2	C	C1.fas	C2.fas
3	D	D1.fas	D2.fas
4	E	E1.fas	E2.fas
5	F	F1.fas	F2.fas
6	G	G1.fas	G2.fas
7	H	H1.fas	H2.fas
8	I	I1.fas	I2.fas
9	K	K1.fas	K2.fas
10	L1	L1-1.fas	L1-2.fas
11	L2	L2-1.fas	L2-2.fas
12	M	M1.fas	M2.fas
13	N	N1.fas	N2.fas
14	P	P1.fas	P2.fas
15	Q	Q1.fas	Q2.fas
16	R	R1.fas	R2.fas
17	S1	S1-1.fas	S1-2.fas
18	S2	S2-1.fas	S2-2.fas
19	T	T1.fas	T2.fas
20	V	V1.fas	V2.fas
21	W	W1.fas	W2.fas
22	Y	Y1.fas	Y2.fas

The first column is sequence numbers, the second column designates the tRNA gene isoacceptor type, the third column contains the names of aligned fasta files of d1 dataset for tRNA genes of every isoacceptor type (in the script from A1.fas to Y1.fas), and the fourth column contains the names of the aligned fasta files of d2 dataset for tRNA genes of every isoacceptor type (in the script from A2.fas to Y2.fas). The script calculates p-distances matrixes in d1 and d2 groups. These calculated values are used for building of histograms, ranging between the similar minimum and maximum values on the scale and subdivided into bins (number of bins is estimated using Sturges' formula [6]). The script then compares two histograms using the Fisher's chi-square test [7] and writes *p*-values in the fifths column of the output file (in the script output.txt):

No	tRNA	first_file	second_file	P_value	D_mean	D_mean_test	Ratio
1	A	A1.fas	A2.fas	0.9922	0.1954	0.1764	-0.1080
2	C	C1.fas	C2.fas	0.1689	0.2301	0.5185	0.5562
3	D	D1.fas	D2.fas	0.0500	0.2300	0.6145	0.6257
4	E	E1.fas	E2.fas	0.0064	0.1066	0.5939	0.8205
5	F	F1.fas	F2.fas	0.0360	0.2585	0.5885	0.5607
6	G	G1.fas	G2.fas	0.0348	0.1227	0.5207	0.7644
7	H	H1.fas	H2.fas	0.1155	0.2462	0.5040	0.5114
8	I	I1.fas	I2.fas	0.0527	0.2113	0.5811	0.6364
9	K	K1.fas	K2.fas	0.0064	0.1353	0.5830	0.7679
10	L1	L1-1.fas	L1-2.fas	0.1035	0.2828	0.6588	0.5708
11	L2	L2-1.fas	L2-2.fas	0.0532	0.1690	0.4918	0.6564
12	M	M1.fas	M2.fas	0.0316	0.1010	0.5616	0.8201
13	N	N1.fas	N2.fas	0.0231	0.1613	0.6070	0.7342
14	P	P1.fas	P2.fas	0.0174	0.1049	0.5614	0.8132
15	Q	Q1.fas	Q2.fas	0.0700	0.1216	0.5305	0.7707
16	R	R1.fas	R2.fas	0.0173	0.1668	0.6667	0.7497
17	S1	S1-1.fas	S1-2.fas	0.0709	0.1424	0.4941	0.7118
18	S2	S2-1.fas	S2-2.fas	0.0490	0.2780	0.5787	0.5197
19	T	T1.fas	T2.fas	0.1078	0.1553	0.3751	0.5860
20	V	V1.fas	V2.fas	0.0302	0.1895	0.5718	0.6687
21	W	W1.fas	W2.fas	0.0535	0.1535	0.5595	0.7257
22	Y	Y1.fas	Y2.fas	0.0823	0.1653	0.5612	0.7054

The script also calculates the mean distances (m1) for d1 groups and writes the values in the sixth column of the output file, and the mean distances (m2) between the sequence under examination has the latest position between the alignment and the rest of the sequences in d2, and the values are written in the seventh column of the output file. The ratio between m1 and m2 is calculated using formula 1-m1/m2, and the values are written in the eighth column of the output file. The maximum *p*-value and minimum value of the ratio between genetic distances indicate the type of the progenitor tRNA gene of the gene under study. For correct analysis, the mean p-distance of the aligned nucleotide sequences should not exceed 0.75.

2.2. R script for choosing the best sequence alignment

To perform phylogenetic analysis based on tRNA gene sequences, alignments considering different features of the secondary structures are often used [8–10]. The choice of the features is determined by the predicted structures of the tRNA molecules coded by the DNA fragments aligned. Numerous algorithms of the alignment of DNA fragments are also available. They perform differently in cases of short and hypervariable sequences such as tRNA genes. This causes the variation of topologies of trees inferred.

To determine the best mode of alignment among the several different alternatives, we developed an R script, which identifies the best alignment based on its minimum BIC value:

```
library("phangorn")
data<-read.table("input.txt",header=TRUE, stringsAsFactors=F, sep="\t")
BIC<-c()
AIC<-c()
logLik<-c()
for(i in 1:nrow(data))
 {
  dat<-read.dna(data$alignment[i], format="fasta")
  dat<-as.phyDat(dat)
  tre<-read.tree(data$tree[i])
  fit<-pml(tree=tre, data=dat, k=4)
  fit.opt<-optim.pml(fit, optEdge=F, optGamma=data$G[i], optInv=data$I[i], model=data$model[i])
  BIC<-c(BIC, BIC(fit.opt))
  AIC<-c(AIC, AIC(fit.opt))
  logLik<-c(logLik, fit.opt$logLik)
  print(paste("el=> ", i, "\n", sep=""))
 }
rez<-cbind(data, logLik=logLik, AIC=AIC, BIC=BIC)
rez<-rez[order(rez$BIC),]
dBIC<-c()
for(i in 1:nrow(rez)) dBIC<-c(dBIC, rez$BIC[i]-rez$BIC[1])
rez<-cbind(rez, dBIC=dBIC)
write.table(rez, file="output.txt",row.names=F, col.names = TRUE, sep = "\t")
```

The dataset of every version of the alignment tested should contain an alignment file, a deduced substitution model, and the phylogenetic tree. The script utilizes the “phangorn” package [11]. The input file (in the script input.txt) is a tab delimited table looks like:

tree	alignment	model	G	I	how to align
Baik_clust_14	Baik_clust_14.fas	TrN	T	T	Baik_clustalw_14
Baik_mafft_14	Baik_mafft_14.fas	TPM3	T	F	Baik_mafft_14
Baik_tcoffee_14	Baik_tcoffee_14.fas	TPM3u	T	F	Baik_220_tcoffee_14
Baik_all_clust	Baik_all_clust.fas	HKY	T	F	Baik_final_clustalw_220
Baik_all_tcoffee	Baik_all_tcoffee.fas	HKY	T	F	Baik_final_tcoffee_220
Baik_all_mafft	Baik_all_mafft.fas	HKY	T	F	Baik_220_mafft_trex

tree	alignment	model	G	I	how.to.align	logLik	AIC	BIC	dBIC
Baik_all_mafft	Baik_all_mafft.fas	HKY	TRUE	FALSE	Baik_220_mafft_trex	-6477.777	13839.555	14913.976	0.000
Baik_all_clust	Baik_all_clust.fas	HKY	TRUE	FALSE	Baik_final_clustalw_220	-6555.615	13995.230	14969.704	55.728
Baik_clust_14	Baik_clust_14.fas	TrN	TRUE	TRUE	Baik_clustalw_14	-6504.009	13896.019	15185.961	271.985
Baik_all_tcoffee	Baik_all_tcoffee.fas	HKY	TRUE	FALSE	Baik_final_tcoffee_220	-6727.613	14339.225	15402.995	489.019
Baik_mafft_14	Baik_mafft_14.fas	TPM3	TRUE	FALSE	Baik_mafft_14	-6755.529	14391.058	15519.371	605.395
Baik_tcoffee_14	Baik_tcoffee_14.fas	TPM3u	TRUE	FALSE	Baik_220_tcoffee_14	-7216.934	15319.867	16465.008	1551.032

The first column contains names of tree files in Newick format, the second column contains names of aligned files in fasta format, the third column contains substitution model notation designated as in Posada (2008) [12], the forth column contains labels T or F (true of false) for the parameter of the gamma-shaped distribution of rates across sites that the model used, the fifth column contains labels T or F (true of false) for the proportion of invariable sites of the model used, and the sixth column contains any user's comments about alignments. The output file (in the script output.txt) is also a tab delimited table in which the columns with estimated mean likelihood, AIC, BIC, and delta BIC values for every alignment dataset are added:

The dataset lines in the output file are sorted from the minimum delta BIC value at the top to the maximum value at the bottom. The current R script is suitable for assessment of the best alignment modes of any sets of aligned nucleotide sequences.

CRediT authorship contribution statement

Elena V. Romanova: Investigation, Validation, Writing - original draft, Visualization. Yurij S. Bukin: Software, Formal analysis, Dmitry Yu. Sherbakov: Conceptualization, Methodology, Resources, Writing – review & editing.

Acknowledgments

The work was supported by the governmentally funded project 0345–2019–0004 (AAAA-A16-116122110060-9).

Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.dib.2020.105284.

References

[1] E.V. Romanova, Y.S. Bukin, K.V. Mikhailov, M.D. Logacheva, V.V. Aleoshin, D.Yu. Sherbakov, Hidden cases of tRNA genes duplication and remolding in mitochondrial genomes of amphipods, Mol. Phylogenet. Evol. 144 (2020) 106710.
[2] F. Jühling, J. Pütz, M. Bernt, A. Donath, M. Middendorf, C. Florentz, P.F. Stadler, 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 Res. 40 (7) (2012) 2833–2845.
[3] P. Cantatore, M.N. Gadaleta, M. Roberti, C. Saccone, A.C. Wilson, Duplication and remoulding of tRNA genes during the evolutionary rearrangement of mitochondrial genomes, Nature 329 (6142) (1987) 853–855.
[4] M.E. Saks, J.R. Sampson, J. Abelson, Evolution of a transfer RNA gene through a point mutation in the anticodon, Science 279 (5357) (1998) 1665–1670.
[5] E. Paradis, K. Schliep, Ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R, Bioinformatics 35 (3) (2018) 526–528.
[6] H.A. Sturges, The choice of a class interval, J. Am. Stat. Assoc. 21 (153) (1926) 65–66.
[7] A. Agresti, An Introduction to Categorical Data Analysis, John Wiley and Sons, New York, 2007, p. 38.
[8] O.V. Popova, K.V. Mikhailov, M.A. Nikitin, M.D. Logacheva, A.A. Penin, M.S. Muntyan, O.S. Kedrova, N.B. Petrov, Y.V. Panchin, V.V. Aleoshin, Mitochondrial genomes of Kinorhyncha: *trnM* duplication and new gene orders within animals, PloS One 11 (10) (2016) e0165072.
[9] A.H. Sahyoun, M. Hölzer, F. Jühling, C. Höner zu Siederdissen, M. Al-Arab, K. Tout, M. Marz, M. Middendorf, P.F. Stadler, M. Bernt, Towards a comprehensive picture of alloacceptor tRNA remolding in metazoan mitochondrial genomes, Nucleic Acids Res. 43 (16) (2015) 8044–8056.
[10] X. Wang, D.V. Lavrov, Gene recruitment – a common mechanism in the evolution of transfer RNA gene families, Gene 475 (1) (2011) 22–29.
[11] K.P. Schliep, Phangorn: phylogenetic analysis in R, Bioinformatics 27 (4) (2011) 592–593.
[12] D. Posada, jModelTest: phylogenetic model averaging, Mol. Biol. Evol. 25 (7) (2008) 1253–1256.