Prediction of Monophyletic Groups Based on Gene Order and Sequence Similarity in Organelle DNA

Matyas Cserhati

Independent Scholar
Email: csmatyi@protonmail.com

Abstract

Organelle genomics has become its own field of study. Much information can be gleaned from the study of cell organelles. The differences in the genomes of organelles, such as the mitochondrion and the chloroplast are amenable to phylogenetic and cladistic studies. These differences include the genome sequence, GC%, genome length and gene order. The conserved nature of the organelle genomes and the gene inventory of both mitochondrial and chloroplast genomes also make this easier to accomplish. This paper includes a review of existing organelle genome software. These include gene annotation and genome visualization tools as well as organelle gene databases for both mitochondrion and plastid. A new R tool, available on github, called "Organelle DNA Lineages", or ODL, was written to compare and classify organelle genomes based on their genome sequence and gene order. The software was run on the mitochondrial genomes of a set of 51 cephalopod species, delineating ten separate monophyletic groups, including argonauts, nautiluses, octopuses, cuttlefish, and six squid groups. This new tool can help enrich and expand the field of organelle genomics.

Keywords

Organelle, Genome, Mitochondrion, Plastid, Organelle DNA Lineage, Chloroplast

1. Introduction

The DNA inside mitochondria and chloroplasts can very useful in analyzing phylogenetic relationships between species. Several characteristics demonstrate why organelle DNA is amenable to such analyses. But despite its small size, organelle DNA is easy to isolate and sequence. Many species which do not have

their nuclear genomes available have mitochondrial and chloroplast genome sequences instead in the NCBI database. The Organelle Genome Database at NCBI contains 19,320 organelle genome sequences as of July 28, 2021. The distribution of these organelle genomes can be seen in Table 1. This paper will focus on the mitochondrion and the plastid. Second, the mitochondrial DNA (mtDNA) overwhelmingly follows maternal inheritance patterns, avoiding complex biparental inheritance patterns.

Despite its rapid mutation rate, organelle DNA (oDNA) retains a very conserved gene inventory, very similar gene order, genome sequence similarity, and genome size between related species. Gene rearrangements include recombination, inversions, transpositions, inverse transpositions, and tandem duplication genes losses, but are thought to be rare [1] [2]. The insertion of introns and mobile elements are also thought to result in gene order rearrangement [3]. Plastid DNA (plDNA) may expand or contract due to the presence or absence of inverted repeats [4]. In general, genes, rRNAs and tRNAs are colinear with one another between species in the same monophyletic group. This makes the identification and classification of newly sequenced and annotated oDNA relatively easy.

Mitochondrial genes, for example, take part in the production of adenosine triphosphate (ATP), the energy molecule, as well as oxidative phosphorylation (OXPHOS), the tricarboxylic acid cycle (TCA), the β-oxidation of fatty acids, calcium handling, regulating apoptosis, and participation in the cell cycle. Because mtDNA lacks histones, it is more prone to molecular stress. Damaged mtDNA lies behind a number of human diseases [5] [6] [7]. Other illnesses related to mitochondrial dysfunction include skeletal muscle dysfunction, lung injury, acute renal failure, and immune function dysregulation [8].

An example of the classical mtDNA genome is that of human (NC_01290, Figure 1), which is 16,569 bp long and has 37 genes, coding for 13 protein subunits, 14 tRNAs and 2 rRNAs [9]. The 13 protein subunits form the oxidative phosphorylation complexes I, II, IV and V [7]. The mtDNA consists of two strands, the heavy (H) strand and the light (L) strand, which are differentiated based on their GC-content, the H-strand having a higher GC-percentage. This correlates to the sense and antisense terminology. The codon usage of the mtDNA also differs from that of the nuclear genome (see Table 2), and also has a ten-fold higher mutation rate [10]. The thirteen protein subunits coded on the mtDNA all take part in energy metabolism. A list of genes in the mitochondrial genome can be seen in Table 3.

The mtDNA of different groups can also differ based on the presence or absence of specific genes. These genes serve as diagnostic markers of these groups. For example, several algal lineages have a DnaB helicase encoded in the plastid genome [11]. Opisthokonts (animals and fungi) use DNA polymerase γ (Polγ) for the replication of the mtDNA [12]. The number and type of DNA polymerase also varies across eukaryotes. For example, parasite apicomplexans have lost them altogether [11].

Figure 1. Graphical depiction of the human mitogenome, created using the GenomeVx software. The human mitogenome accession number is NC_012920.1.

Table 1. List of different kinds of organelle genome sin the NCBI Organelle Browser as of July 28, 2021.

Type of organelle	Number of genomes in NCBI
Mitochondrium	12,528
Chloroplast	5660
Plastid	1064
Apicoplast	62
Kinetoplast	3
Chromatophore	2
Cyanelle	1

Table 2. Divergent nucleotides in the mtDNA compared to the nuclear genome.

Codon	Universal genetic code	Mitochondrial genetic code
UGA	Stop	Trp
AUA	Ile	Met
AGA	Arg	Stop
AGG	Arg	Stop

Table 3. The 37 genes encoded by the human mitochondrial genome.

Gene symbol	Function	Strand
RNR1	12S ribosomal RNA	Heavy
RNR2	16S ribosomal RNA	Heavy
ATP6	ATP synthase F0 subunit 6	Heavy
ATP8	ATP synthase F0 subunit 8	Heavy
CYB	cytochrome b	Heavy
CO1	cytochrome c oxidase subunit I	Heavy
CO2	cytochrome c oxidase subunit II	Heavy
CO3	cytochrome c oxidase subunit III	Heavy
ND1	NADH dehydrogenase subunit 1	Heavy
ND2	NADH dehydrogenase subunit 2	Heavy
ND3	NADH dehydrogenase subunit 3	Heavy
ND4	NADH dehydrogenase subunit 4	Heavy
ND4L	NADH dehydrogenase subunit 4L	Heavy
ND5	NADH dehydrogenase subunit 5	Heavy
ND6	NADH dehydrogenase subunit 6	Light
TA	tRNA-Ala	Light
TR	tRNA-Arg	Heavy
TN	tRNA-Asn	Light
TD	tRNA-Asp	Heavy
TC	tRNA-Cys	Light
TE	tRNA-Gln	Light
TQ	tRNA-Glu	Light
TG	tRNA-Gly	Heavy
TH	tRNA-His	Heavy
TI	tRNA-Ile	Heavy
TL1	tRNA-Leu-UUR	Heavy
TL2	tRNA-Leu-CUN	Heavy
TK	tRNA-Lys	Heavy
TM	tRNA-Met	Heavy
TF	tRNA-Phe	Heavy
TP	tRNA-Pro	Light
TS1	tRNA-Ser-UCN	Light
TS2	tRNA-Ser-AGY	Heavy
TT	tRNA-Thr	Heavy
TW	tRNA-Trp	Heavy
TY	tRNA-Tyr	Light
TV	tRNA-Val	Heavy

Besides genes, the presence or absence of certain DNA motifs can help classify species into different groups because of their genome structure. They are abundant, have a simple genomic structure, and are also conserved, and can be used as genetic markers [13]. For example, Xi Wen *et al.* [14] found that 218 simple sequence repeats (SSR) in the plastid genome of *M. grandiflora* correlate with three other Magnoliaceae species. These SSRs range from mononucleotides to hexanucleotides.

In contrast with the mtDNA, the plDNA is much larger than the mtDNA, and contains around 250 genes, including tRNAs and rRNAs, involved in photosynthesis and plastid gene regulation. Xiao-Ming *et al.* [15] found that in a study of 126 plDNA genes in 272 angiosperm species, 106 (84.1%) were found in 245 species (90.1%). Genome sizes range from 10 - 250 kbp in size, and can be present inside the plastid between 100 to 1000 copies Kbp [16]. Some genes in the plastid genome of some species also have introns, such as in *Magnolia grandiflora* [14]. Similar to the mtDNA, the plDNA is inherited from one parent only, except for 20% of angiosperms, where plastids are inherited from both parents. Plastid genes are also double in number and three-fold in silent substitution rates as compared to mitogenomes [17]. Plastid genomes are also conserved in structure and are poor in repeat elements. There are several types of plastids, which fulfil different roles in the plant. The well-known chloroplasts harvest light energy during photosynthesis. Chromoplasts are colored chloroplasts which store different pigments. Leucoplasts are colorless plastids, which store proteins, fat, monoterpenes, tannins, polyphenols, and other secondary metabolites [18].

In this paper, a new resource called the Organelle DNA Lineages (ODL) software will be described and compared to existing organelle DNA resources.

2. Research Materials and Methods

2.1. Description of Software

The Organelle DNA Lineages (ODL) software was written R, version 4.0.3. It is available at https://github.com/csmatyi/odl together with all supplementary files and figures. The script is called odl5.R. As input, it takes a list with two columns: the first column contains the Latin name of the species in the study. The second column is the NCBI accession number of the oDNA sequence. Species and their corresponding accession numbers can be downloaded at the Organelle Genome Browser at https://www.ncbi.nlm.nih.gov/genome/browse#!/organelles/.

The software removes duplicate species entries. Using the "getAnnotations-GenBank" function from the "ape" package, it downloads and stores the annotation for each accession. Using the "read.GenBank" function, it retrieves the genome sequence for each accession. The software adds all of the organelle genome sequences into a DNAStringSet, and then calls the "msa" function to create a multiple sequence alignment. Then, in an all-versus-all pairwise fashion, the script calculates the sequence similarity between all possible pairs of genome

sequences. The sequence similarity values are then stored in a symmetric square sequence similarity matrix.

The gene, tRNA and rRNA elements are also written to an output file, which is compatible with the input format of the CREx software [19]. The output can be directly uploaded to the CREx website at http://pacosy.informatik.uni-leipzig.de/crex. The GenBank records for all sequences are also written to a subdirectory as part of the output. These can be used as input for the GenomeVx visualization software [20].

A gene order distance matrix is calculated for each species pair by comparing the indexes of each gene/tRNA/rRNA between the two species using the following equation:

$$d_{a,b} = \sum_{i=1}^{k} \left| \text{mean}\left(\text{index}_a\left(i\right)\right) - \text{mean}\left(\text{index}_b\left(i\right)\right) \right| + \text{unique}\left(a\right) + \text{unique}\left(b\right) \quad (1)$$

In Equation (1), the difference between the mean index value of element *i* (meaning genes, tRNAs, and rRNAs) between species *a* and *b* is summed up for all common elements between the two species. The sum also includes the unique number of elements pertaining to species *a* and *b*, respectively.

Figure 2 depicts a simple example of calculating the gene order distance between three hypothetical species. For example, the distance between species 1 and species 3 is 3 + 1 + 2.5 + 3 = 9.5. The distance in the index value for gene A is 4 − 1 = 3. For gene B it is 3 − 2 = 1. For gene C it is (4 + 3)/2 − 1 = 2.5. For gene D it is 5 − 2 = 3.

It is possible that a certain gene/rRNA/tRNA is present in multiple copies. In this instance, the average index value is taken for that particular gene/rRNA/tRNA. The distance value between two species may be greater than 1, since we are

Example
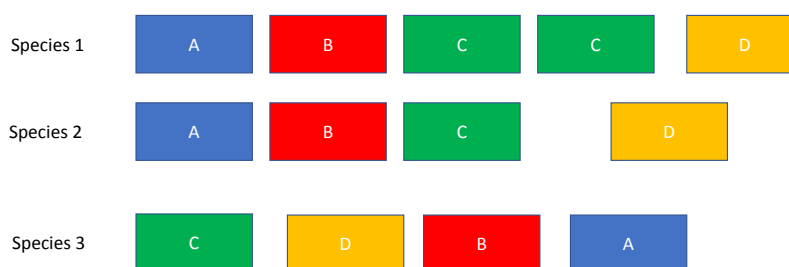
Distance(species 1, species 2) = 0 + 0 + 0.5 + 1 = 1.5 => 0.158 (normalized)
Distance(species 2, species 3) = 3 + 1 + 2 + 2 = 9 => 0.947
Distance(species 1, species 3) = 3 + 1 + 2.5 + 3 = 9.5 => 1.000

Figure 2. Schematic diagram depicting how gene order similarity is calculated between species. In a given study, the gene order distance between all possible species pairs is calculated. Species within the same monophyletic group have a similar gene order (species 1 vs. species 2). Species from two different groups have different gene orders (species 1 vs. species 3, and species 2 vs. species 3). The distances are calculated according to Equation (1), and then normalized to get a number between 0 and 1.

looking at indexes. The gene order distance matrix is then normalized by dividing each element in the distance matrix by the maximum value of the matrix, and then subtracted from 1 to derive a gene order similarity matrix.

The genome sequence similarity matrix and the gene order similarity matrix are then visualized on a heatmap. Furthermore, these two matrixes are also combined into a "combined" matrix, giving equal weight to both the genome sequence similarity matrix and the gene order similarity matrix. The "heatmap" function is called, using the "ward.D2" method to perform species clustering. A Silhouette plot is also created by the software, which depict the average silhouette width for each possible number of clusters, showing the optimal number of clusters with a dashed line, where the average silhouette width is highest. The mean silhouette width shows how close the points in one cluster are to points in another cluster.

Clustering is performed using the "ward.D2" method on the sequence similarity matrix. Several statistics measures are written to an output file, such as the Hopkins clustering statistic, which denotes how well the matrix forms clusters. Other parameters are also recorded for each of the clusters (which correspond to monophyletic groups), including the number of species in the cluster, the average oDNA length ± one standard deviation, the minimum, mean, and maximum similarity value, as well as the standard error of the mean (SEM), as well as the p-value, which calculates statistical significance between similarity values between species within the cluster and between species within the cluster and outside the cluster. This p-value is a good measure of discontinuity between monophyletic groups.

The main output of the software is a gene order map, showing the order of genes/rRNAs/tRNAs along the organelle genome per species per cluster. A color-coded legend helps the user identify a given gene/rRNA/tRNA along the organelle genome of a given species.

2.2. Other Software

GenomeVx was used to create the image of the human reference mitogenome in **Figure 1**.

2.3. Used Sequences

The mitogenomes for 51 cephalopod species were downloaded from the Organelle Genome Browser, specified in the previous subsection. If a given species occurred more than once, only one of its accession number was randomly selected.

3. Results and Discussion

3.1. Review of Existing Organelle Software and Databases

There are a lot of different databases and programs which are tailored for storing and visualizing genome and synteny. Some of these programs include software

which are amenable for the visualization of organelle genomes and gene order. In the following, some of these software and databases will be reviewed. A list of the most commonly used mitogenome and plastid genome annotation tools are listed in Table 4.

GenBank and RefSeq are the standard databases used for storing sequence data, including organelle genomes. Refseq offers curated and non-redundant data for users. Several databases exist, which are built upon GenBank and RefSeq, which store data derived from these databases and which have improved annotation and data quality. Such databases include OGRe [21], MetAMIGA [22], MitoZOA [23], Mamit-tRNA [24], MamMiBase [25], and tRNAdb [26].

CREx is a software which calculates the hypothetical most parsimonious gene order rearrangement between two organelle genomes, and then visualizes the output. Possible rearrangements include transpositions, reverse transpositions, reversals and tandem duplication random losses (TDRLs) [19].

Similar to CREx, EqualTDRL is a software which compares two input genomes (*i.e.* two mt genomes), calculates different series of tandem duplication random losses (TDRL), which hypothetically transforms the gene order of one genome into the other. It creates a diagonally symmetric map showing all the possible series of TDRLs between the two genomes [27].

GenomeVx is a web-based software which allows the user to create editable, colorful, publication-ready images of circular genomes, such as that of mitochondrial and plastid genomes as well as large plasmids. As input, it takes raw feature positions or GenBank files, which the user can upload to the GenomeVx website [20].

Table 4. Existing software amenable for comparative depiction of organelle genomes and gene order.

Software	Description	Website
CGAP	Plastid genome collection, visualization, content comparison, and annotation tool	http://www.herbbol.org:8080/chloroplast
CpGAVAS/ CpGAVAS2	Integrated web server for annotating, visualizing, analyzing and submitting plastid genomes to GenBank	http://www.herbalgenomics.org/cpgavas; http://47.96.249.172:16019/analyzer/home
CpGDB	Database storing chloroplast genomes, gene sequences and annotation	http://gndu.ac.in/CpGDB/
CREx/CREx2	Visualizes most parsimonious gene order rearrangements between two species	http://pacosy.informatik.uni-leipzig.de/crex; http://pacosy.informatik.uni-leipzig.de/271-0-CREx2.html
DOGMA	Metazoan mitogenome gene annotation server for both mitogenomes and plastid genomes	https://dogma.ccbb.utexas.edu/
GeSeq	Fast, high-quality HMM-based organelle genome annotator, mainly for plastid genomes	https://chlorobox.mpimp-golm.mpg.de/geseq.html
EqualTDRL	Visualizes hypothetical series of gene order rearrangements between two species	http://pacosy.informatik.uni-leipzig.de/269-0-EqualTdrl.html
GenomeVx	Gene order visualization on small chromosomes and large plastids	http://wolfe.ucd.ie/genomevx/
Mitofish/ Mitoannotator	Annotation database and pipeline for fish mitogenomes	http://mitofish.aori.u-tokyo.ac.jp/
MITOS	Metazoan mitogenome gene annotation server	http://mitos.bioinf.uni-leipzig.de/index.py

Mitofish is a database of annotations and re-annotations for mitochondrial genome for a large number of fish species. Its companion program, Mitoannotator is an extremely rapid, high quality annotation pipeline for fish mitogenomes [28].

The MITOchondrial genome annotation Server (MITOS) is a well-known, high-quality mitochondrial genome annotation pipeline. MITOS uses standardized gene names and gene boundary designations. It uses BLAST to compare existing mitochondrial proteins to the raw genome sequence to annotate the location of these genes. It also has a web server where users can upload their raw genome sequences. After selecting a translation code, the server will calculate and present results for the user. These results include tabular summary of annotated genetic elements and visualization of genes within the uploaded genome sequence. The results can then be downloaded by the user [29] [30]. DOGMA is another web service, which uses BLAST against internal databases to detect protein-coding genes and also rRNAs, and use tRNAscan-SE to discover tRNAs, however, according to Guyeux *et al.* [31] it is outdated. DOGMA and MITOS both use metazoan databases data [29] [32].

For plastid genomes, the Chloroplast Genome Database (CpGDB) stores plastid genome and individual genes sequences, and also annotation records for 3823 species [33]. Plastid genome annotation tools include CpGAVAS (Chloroplast Genome Annotation, Visualization, Analysis, and GenBank Submission) and CpGAVAS2 [34] [35], CGAP (Chloroplast Genome Annotation Platform) [36], and GeSeq [37].

The ChloroMitoSSRDB database is an open-source repository of perfect and imperfect microsatellites, two to six nucleotides long. Information include the position of repeats, size, motif and length polymorphisms. The repeat sequences are hyperlinked to annotated gene regions at NCBI [38].

The current software differs from all the previous methods in that it clusters the organelle genomes based on both gene order and sequence similarity. It then depicts a linear organelle genome map for each cluster and species in the study showing the position of each gene. The software also produces accompanying statistics files and also heatmaps showing species relationships based on gene order and sequence similarity. Other software calculates genome distance by the number of rearrangements needed to transform the organelle genome of one species into another [39]. The present software does this by calculating the difference between the index of the order of each gene.

3.2. Analysis of Cephalopod Mitogenomes

Since the main goal of this paper is to present a new bioinformatics tool, usage of the ODL software will be showcased here in the mitochondrial mapping of cephalopods. Cephalopods are a class of species in the phylum Mollusca (mollusks). They have two subclasses, Nautiloidea (nautiluses) and Coleoidea, which is made up of two superorders, Decapodiformes (squids and cuttlefish), and Octopodiformes (octopuses and argonauts). The mitogenomes of 51 species as well as the outlier

Danio rerio were analyzed to discover putative monophyletic groups.

After running the software, ten putative monophyletic groups were discovered, besides *Danio rerio*, the outlier species. A list of species and accession numbers can be found in Supplementary File 1. The statistics for each cluster can be seen in Table 5. The matrixes, clusters, and statistics for the mitochondrial gene order, sequence similarity, and combined analyses can be found in Supplementary Files 2 - 4. Here results for only the combined matrix are reported. The mitogenome map can be seen in Figure 3, and the corresponding combined heatmap can be seen in Figure 4. The Hopkins clustering statistic for the combined matrix is 0.81, which denotes good clustering. Supplementary Figures 1-3 show the Silhouette plot for clustering based on gene order similarity, mitogenome sequence similarity, and combining both methods.

Figure 3. Gene order map of 51 cephalopod species and *Danio rerio* as an outlier. There are ten cephalopod clusters, predicted by combining the sequence similarity and gene order methods. Each cluster contains several species, showing the genome length and also the order of the genes. A color legend is in the upper right.

Figure 4. Heat map showing the relationships between the individual cephalopod monophyletic groups based on combining clustering based on gene order and clustering based on sequence similarity. Both clustering schemes were given a 50% weight. Each pixel in the heatmap shows how similar a pair of species is. The heat map is symmetrical, mirrored across the diagonal going from bottom left to top right. Darker, redder colors denote species with high similarity to one another, while brighter, yellow colors denote species which are less similar to one another. The heat map shows the same ten clusters shown in Figure 3.

Table 5. Clustering statistics for putative predicted cephalopod clusters with more than one member based on mt genome sequence similarity and gene order similarity.

cluster	no. species	Mean DNA length ± sd	Min	mean	max	SEM	p-value
1. Nautilus	3	16,027.667 ± 296.598	0.801	0.813	0.834	1.1E-02	0.000421
2. Argonauta	2	15,665.500 ± 656.902	0.75	0.75	0.75	NA	1.42E-68
3. Octopus	17	15,794.824 ± 196.267	0.756	0.819	0.903	7E-03	1.33E-211
4. Squids I	4	20,303.75 ± 34.798	0.629	0.68	0.725	0.023	0.000539
6. Squids II	2	20,564.500 ± 405.172	0.692	0.692	0.692	NA	3.97E-28
7. Squids III	9	17,244.889 ± 255.089	0.729	0.821	0.899	1.6E-02	1.46E-31
8. Squids IV	2	20,200.500 ± 152.028	0.66	0.66	0.66	NA	3.22E-40
10. Sepia	10	16,197.500 ± 26.488	0.801	0.846	0.918	9E-03	9.05E-66

3.3. The Ten Cephalopod Clusters

Argonauta and Nautilus form two separate clusters. Nautilus also has a significantly large non-coding region between the tRNAs for glutamine (Q) and threonine (T) [40]. Nautilus also has a mean GC% of 40% ± 0.26%, whereas for Argonauta this

value is 22.9% ± 0.05%.

Octopodidae with 17 species is the largest and most significant monophyletic group, with a unique morphology of eight legs. These species come from the genera *Amphioctopus*, *Callistoctopus*, *Cistopus*, *Hapalochlaena*, and *Octopus*. The 17 octopus species in this study have a very conserved gene order and genome length between 15,479 bp and 16,084 bp. The octopus genomes have a mean GC% of 24.5% ± 0.88%. The enigmatic *Vampyroteuthis infernalis* [41] is classified as a member of this group.

Ten species of cuttlefish from the genera *Sepia* and *Sepiella* also form a monophyletic group. Their gene order is very conserved, but very different from all other cephalopod groups. The genome length is also very conserved, between 16,163 to 16,244 bp. The GC% is 24.6% ± 1.47%. Takumiya *et al.* [42] also found significant differences between coleoid cephalopods based on mtDNA, separating cuttlefish from all other groups.

Following are several squid groups, including four monophyletic groups and two species which don't belong anywhere else. The first of the four groups is Loliginidae (pencil squids), with species from the genera *Doryteuthis*, *Heterololigo*, *Loliolus*, *Sepioteuthis*, and *Uroteuthis*. The species *Sepioteuthis lessoniana* is an outlier due to the translocation of three tRNA sequences (for isoleucine, valine and tryptophan) and l-rrna dn s-rrna in its mt genome. Yokobori *et al.* [43] also found that this squid group is monophyletic.

Four species, *Architeuthis dux*, *Dosidicus gigas*, *Stenoteithis oualaniensis*, and *Todarodes pacificus* form a separate group. These four species have a mean GC% of 29.8% ± 1.4% and a genome length from 20,254 to 20,331 bp. Xu, *et al.* [44] also separate them from all other squids.

Bathyteuthis abyssicola (the deepsea squid) belongs to its own family, Bathyteuthidae, in the suborder Oegopsina. It possibly even belongs to its own order, according to Uribe and Zardoya [45]. Kawashima *et al.* [46] found that the mitogenome structure of *Semmirossia patagonica* (Patagonian bobtail squid), from the family Sepiolidae is unique among decapods. For example, like *Nautilus* species, the ATP6 and ATP8 genes are not adjacent to one another. *Ommastrephes bartrami* (the neon flying squid) and *Watasenia scintillans* (the firefly squid) and *Chiroteuthis picteti* and *Ilex argentinus* cluster together into the same putative groups.

4. Conclusion

Studying organelle genomes is a new and expanding field of genomics research. Several tools have been devised to determine phylogeny based on gene order. Several tools and databases exist to annotate and visualize organelle genomes and to store organelle sequences. The present software, Organelle DNA Lineages was designed to cluster species into monophyletic groups based on genome sequence similarity and gene order and visualize the results. The software was run on a set of 51 cephalopod species and found ten clusters. This software can help expand the field of organelle genomics.

Acknowledgements

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Boore, J.L., Collins, T.M., Stanton, D., Daehler, L.L. and Brown, W.M. (1995) Deducing the Pattern of Arthropod Phylogeny from Mitochondrial DNA Rearrangements. *Nature*, **376**, 163-165. https://doi.org/10.1038/376163a0

[2] Boore, J.L., Lavrov, D.V. and Brown, W.M. (1998) Gene Translocation Links Insects and Crustaceans. *Nature*, **392**, 667-668. https://doi.org/10.1038/33577

[3] Aguileta, G., de Vienne, D.M., Ross, O.N., Hood, M.E., Giraud, T., Petit, E. and Gabaldón, T. (2014) High Variability of Mitochondrial Gene Order among Fungi. *Genome Biology and Evolution*, **6**, 451-465. https://doi.org/10.1093/gbe/evu028

[4] Ravi, V., Khurana, J.P., Tyagi, A.K. and Khurana, P. (2008) An Update on Chloroplast Genomes. *Plant Systematics and Evolution*, **271**, 101-122. https://doi.org/10.1007/s00606-007-0608-0

[5] Friedman, J.R. and Nunnari, J. (2014) Mitochondrial Form and Function. *Nature*, **505**, 335-343. https://doi.org/10.1038/nature12985

[6] Suomalainen, A. and Battersby, B.J. (2018) Mitochondrial Diseases: The Contribution of Organelle Stress Responses to Pathology. *Nature reviews. Molecular Cell Biology*, **19**, 77-92. https://doi.org/10.1038/nrm.2017.66

[7] Yan, C., Duanmu, X., Zeng, L., Liu, B. and Song, Z. (2019) Mitochondrial DNA: Distribution, Mutations, and Elimination. *Cells*, **8**, 379. https://doi.org/10.3390/cells8040379

[8] Supinski, G.S., Schroder, E.A. and Callahan, L.A. (2020) Mitochondria and Critical Illness. *Chest*, **157**, 310-322.

[9] Anderson, S., Bankier, A.T., Barrell, B.G., de Bruijn, M.H., Coulson, A.R., Drouin, J., Eperon, I.C., Nierlich, D.P., Roe, B.A., Sanger, F., Schreier, P.H., Smith, A.J., Staden, R. and Young, I.G. (1981) Sequence and Organization of the Human Mitochondrial Genome. *Nature*, **290**, 457-465. https://doi.org/10.1038/290457a0

[10] Goodman, S.R. (2007) Medical Cell Biology. Third Edition, Academic Press, Cambridge.

[11] Hirakawa, Y. and Watanabe, A. (2019) Organellar DNA Polymerases in Complex Plastid-Bearing Algae. *Biomolecules*, **9**, 140. https://doi.org/10.3390/biom9040140

[12] Moriyama, T., Terasawa, K. and Sato, N. (2011) Conservation of POPs, the Plant Organellar DNA Polymerases, in Eukaryotes. *Protist*, **162**, 177-187. https://doi.org/10.1016/j.protis.2010.06.001

[13] Jiao, Y., Jia, H.M., Li, X.W., Chai, M.L., Jia, H.J., Chen, Z., Wang, G.Y., Chai, C.Y., van de Weg, E. and Gao, Z.S. (2012) Development of Simple Sequence Repeat (SSR) Markers from a Genome Survey of Chinese Bayberry (*Myrica rubra*). *BMC Genomics*, **13**, 201. https://doi.org/10.1186/1471-2164-13-201

[14] Li, X., Gao, H., Wang, Y., Song, J., Henry, R., Wu, H., Hu, Z., Yao, H., Luo, H., Luo, K., Pan, H. and Chen, S. (2013). Complete Chloroplast Genome Sequence of *Mag-*

nolia grandiflora and Comparative Analysis with Related Species. *Science China. Life Sciences*, **56**, 189-198.

[15] Zheng, X.-M., Wang, J.R., Feng, L., Liu, S., Pang, H.B., Qi, L., Li, J., Sun, Y., Qiao, W.H., Zhang, L.F., Cheng, Y.L. and Yang, Q.W. (2017) Inferring the Evolutionary Mechanism of the Chloroplast Genome Size by Comparing Whole-Chloroplast Genome Sequences in Seed Plants. *Scientific Reports*, **7**, Article No. 1555.

[16] Wakasugi, T., Tsudzuki, T. and Sugiura, M. (2001) The Genomics of Land Plant Chloroplasts: Gene Content and Alteration of Genomic Information by RNA Editing. *Photosynthesis Research*, **70**, 107-118.
https://doi.org/10.1023/A:1013892009589

[17] Wolfe, K.H., Li, W.H. and Sharp, P.M. (1987) Rates of Nucleotide Substitution Vary Greatly among Plant Mitochondrial, Chloroplast, and Nuclear DNAs. *Proceedings of the National Academy of Sciences of the United States of America*, **84**, 9054-9058.
https://doi.org/10.1073/pnas.84.24.9054

[18] Wise, R.R. (2006) The Diversity of Plastid Form and Function. Advances in Photosynthesis and Respiration Book 23. Springer, Berlin, 3-26.

[19] Bernt, M., Merkle, D., Ramsch, K., Fritzsch, G., Perseke, M., Bernhard, D., Schlegel, M., Stadler, P.F. and Middendorf, M. (2007) CREx: Inferring Genomic Rearrangements Based on Common Intervals. *Bioinformatics* (*Oxford, England*), **23**, 2957-2958.
https://doi.org/10.1093/bioinformatics/btm468

[20] Conant, G.C. and Wolfe, K.H. (2008) GenomeVx: Simple Web-Based Creation of Editable Circular Chromosome Maps. *Bioinformatics* (*Oxford, England*), **24**, 861-862.
https://doi.org/10.1093/bioinformatics/btm598

[21] Jameson, D., Gibson, A.P., Hudelot, C. and Higgs, P.G. (2003) OGRe: A Relational Database for Comparative Analysis of Mitochondrial Genomes. *Nucleic Acids Research*, **31**, 202-206. https://doi.org/10.1093/nar/gkg077

[22] Feijao, P.C., Neiva, L.S., de Azeredo-Espin, A.M. and Lessinger, A.C. (2006) AMiGA: The Arthropodan Mitochondrial Genomes Accessible Database. *Bioinformatics*, **22**, 902-903. https://doi.org/10.1093/bioinformatics/btl021

[23] Lupi, R., de Meo, P.D., Picardi, E., D'Antonio, M., Paoletti, D., Castrignanò, T., Pesole, G. and Gissi, C. (2010) MitoZoa: A Curated Mitochondrial Genome Database of Metazoans for Comparative Genomics Studies. *Mitochondrion*, **10**, 192-199.
https://doi.org/10.1016/j.mito.2010.01.004

[24] Pütz, J., Dupuis, B., Sissler, M. and Florentz, C. (2007) Mamit-tRNA, a Database of Mammalian Mitochondrial tRNA Primary and Secondary Structures, *RNA* (*New York, N.Y.*), **13**, 1184-1190.

[25] de Vasconcelos, A.T., Guimarães, A.C., Castelletti, C.H., Caruso, C.S., Ribeiro, C., Yokaichiya, F., Armôa, G.R., Pereira, G., da Silva, I.T., Schrago, C.G., Fernandes, A.L., da Silveira, A.R., Carneiro, A.G., Carvalho, B.M., Viana, C.J., Gramkow, D., Lima, F.J., Corrêa, L.G., Mudado, M., Nehab-Hess, P., *et al.* (2005) MamMiBase: A Mitochondrial Genome Database for Mammalian Phylogenetic Studies. *Bioinformatics* (*Oxford, England*), **21**, 2566-2567.
https://doi.org/10.1093/bioinformatics/bti326

[26] Bernt, M., Braband, A., Middendorf, M., Misof, B., Rota-Stabelli, O. and Stadler, P.F. (2013) Bioinformatics Methods for the Comparative Analysis of Metazoan Mitochondrial Genome Sequences. *Molecular Phylogenetics and Evolution*, **69**, 320-327.
https://doi.org/10.1016/j.ympev.2012.09.019

[27] Hartmann, T., Bernt, M. and Middendorf, M. (2018) EqualTDRL: Illustrating Equivalent Tandem Duplication Random Loss Rearrangements. *BMC Bioinformatics*, **19**,

192. https://doi.org/10.1186/s12859-018-2170-x

[28] Iwasaki, W., Fukunaga, T., Isagozawa, R., Yamada, K., Maeda, Y., Satoh, T.P., Sado, T., Mabuchi, K., Takeshima, H., Miya, M. and Nishida, M. (2013) MitoFish and MitoAnnotator: A Mitochondrial Genome Database of Fish with an Accurate and Automatic Annotation Pipeline. *Molecular Biology and Evolution*, **30**, 2531-2540. https://doi.org/10.1093/molbev/mst141

[29] Bernt, M., Donath, A., Jühling, F., Externbrink, F., Florentz, C., Fritzsch, G., Pütz, J., Middendorf, M. and Stadler, P.F. (2013) MITOS: Improved De Novo Metazoan Mitochondrial Genome Annotation. *Molecular Phylogenetics and Evolution*, **69**, 313-319. https://doi.org/10.1016/j.ympev.2012.08.023

[30] Donath, A., Jühling, F., Al-Arab, M., Bernhart, S.H., Reinhardt, F., Stadler, P.F., Middendorf, M. and Bernt, M. (2019) Improved Annotation of Protein-Coding Genes Boundaries in Metazoan Mitochondrial Genomes. *Nucleic Acids Research*, **47**, 10543-10552. https://doi.org/10.1093/nar/gkz833

[31] Guyeux, C., Charr, J.C., Tran, H., Furtado, A., Henry, R.J., Crouzillat, D., Guyot, R. and Hamon, P. (2019) Evaluation of Chloroplast Genome Annotation Tools and Application to Analysis of the Evolution of Coffee Species. *PLoS ONE*, **14**, e0216347.

[32] Wyman, S.K., Jansen, R.K. and Boore, J.L. (2004) Automatic Annotation of Organellar Genomes with DOGMA. *Bioinformatics* (*Oxford, England*), **20**, 3252-3255. https://doi.org/10.1093/bioinformatics/bth352

[33] Singh, B.P., Kumar, A., Kaur, H., Singh, H. and Nagpal, A.K. (2020) CpGDB: A Comprehensive Database of Chloroplast Genomes. *Bioinformation*, **16**, 171-175. https://doi.org/10.6026/97320630016171

[34] Liu C., Shi L., Zhu Y., Chen H., Zhang J., Lin X. and Guan, X. (2012) CpGAVAS, an Integrated Web Server for the Annotation, Visualization, Analysis, and GenBank Submission of Completely Sequenced Chloroplast Genome Sequences. *BMC Genomics*, **13**, 715. https://doi.org/10.1186/1471-2164-13-715

[35] Shi, L., Chen, H., Jiang, M., Wang, L., Wu, X., Huang, L. and Liu, C. (2019) CPGAVAS2, an Integrated Plastome Sequence Annotator and Analyzer. *Nucleic Acids Research*, **47**, W65-W73. https://doi.org/10.1093/nar/gkz345

[36] Cheng, J., Zeng, X., Ren, G. and Liu, Z. (2013) CGAP: A New Comprehensive Platform for the Comparative Analysis of Chloroplast Genomes. *BMC Bioinformatics*, **14**, 95. https://doi.org/10.1186/1471-2105-14-95

[37] Tillich, M., Lehwark, P., Pellizzer, T., Ulbricht-Jones, E.S., Fischer, A., Bock, R. and Greiner, S. (2017) GeSeq—Versatile and Accurate Annotation of Organelle Genomes. *Nucleic Acids Research*, **45**, W6-W11.

[38] Sablok, G., Mudunuri, S.B., Patnana, S., Popova, M., Fares, M.A. and Porta, N.L. (2013) ChloroMitoSSRDB: Open Source Repository of Perfect and Imperfect Repeats in Organelle Genomes for Evolutionary Genomics. *DNA Research*, **20**, 127-133. https://doi.org/10.1093/dnares/dss038

[39] Bernt, M. and Middendorf, M. (2011) A Method for Computing an Inventory of Metazoan Mitochondrial Gene Order Rearrangements. *BMC Bioinformatics*, **12**, S6.

[40] Boore, J.L. (2006) The Complete Sequence of the Mitochondrial Genome of *Nautilus macromphalus* (Mollusca: Cephalopoda). *BMC Genomics*, **7**, 182. https://doi.org/10.1186/1471-2164-7-182

[41] Robison, B.H., Reisenbichler, K.R., Hunt, J.C. and Haddock, S.H. (2003) Light Production by the Arm Tips of the Deep-Sea Cephalopod *Vampyroteuthis infernalis*. *The Biological Bulletin*, **205**, 102-109. https://doi.org/10.2307/1543231

[42] Takumiya, M., Kobayashi, M., Tsuneki, K. and Furuya, H. (2005) Phylogenetic Relationships among Major Species of Japanese Coleoid Cephalopods (Mollusca: Cephalopoda) Using Three Mitochondrial DNA Sequences. *Zoological Science*, **22**, 147-155. https://doi.org/10.2108/zsj.22.147

[43] Yokobori, S., Lindsay, D.J., Yoshida, M., Tsuchiya, K., Yamagishi, A., Maruyama, T. and Oshima, T. (2007) Mitochondrial Genome Structure and Evolution in the Living Fossil Vampire Squid, *Vampyroteuthis infernalis*, and Extant Cephalopods. *Molecular Phylogenetics and Evolution*, **44**, 898-910. https://doi.org/10.1016/j.ympev.2007.05.009

[44] Xu, L., Wang, X. and Du, F. (2020) The Complete Mitochondrial Genome of Loliginid Squid (*Uroteuthis chinensis*) from Minnan-Taiwan Bank Fishing Ground. *Mitochondrial DNA. Part B, Resources*, **5**, 428-429. https://doi.org/10.1080/23802359.2019.1703599

[45] Uribe, J.E. and Zardoya, R. (2017) Revisiting the Phylogeny of Cephalopoda Using Complete Mitochondrial Genomes. *Journal of Molluscan Studies*, **83**, 133-144. https://doi.org/10.1093/mollus/eyw052

[46] Kawashima, Y., Nishihara, H., Akasaki, T., Nikaido, M., Tsuchiya, K., Segawa, S. and Okada, N. (2013) The Complete Mitochondrial Genomes of Deep-Sea Squid (*Bathyteuthis abyssicola*), Bob-Tail Squid (*Semirossia patagonica*) and Four Giant Cuttlefish (*Sepia apama*, *S. latimanus*, *S. lycidas* and *S. pharaonis*), and Their Application to the Phylogenetic Analysis of Decapodiformes. *Molecular Phylogenetics and Evolution*, **69**, 980-993.

Abbreviations and Acronyms

ATP: adenosine triphosphate

BLAST: Basic Local Alignment Software Tool

CpGAVAS: Chloroplast Genome Annotation, Visualization, Analysis, and GenBank Submission

CpGDB: Chloroplast Genome Database

CREx: Common interval Rearrangement Explorer

DOGMA: Dual Organellar GenoMe Annotator

GC%: GC content

MITOS: MITOchondrial genome annotation Server

Msa: multiple sequence alignment

mt: mitochondrial

mtDNA: mitochondrial DNA

NCBI: National Center for Biotechnology Information

oDNA: organelle DNA

ODL: Organelle DNA Lineages

OGRe: Overlap Graph-based Read ClustEring

OXPHOS: oxidative phosphorylation

Pl: plastid

plDNA: plastid DNA

rRNA: ribosomal RNA

SEM: standard error of the mean

SSR: simple sequence repeat

TCA: tricarboxylic acid

TDRL: tandem duplication random loss

tRNA: transfer RNA

tRNAdb: transfer RNA database