Specific mitochondrial ss-tRNAs in phylum Chaetognatha
Eric Faure, Roxane-Marie Barthelemy

To cite this version:
Eric Faure, Roxane-Marie Barthelemy. Specific mitochondrial ss-tRNAs in phylum Chaetognatha. Journal of Entomology and Zoology Studies, 2019, 7, pp.304 - 315. hal-02130653

HAL Id: hal-02130653
https://hal-amu.archives-ouvertes.fr/hal-02130653
Submitted on 16 May 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - ShareAlike 4.0 International License
Specific mitochondrial ss-tRNAs in phylum Chaetognatha

Eric Faure and Roxane-Marie Barthélémy

Abstract

Chaetognaths are marine invertebrate organisms that constitute a small phylum of very abundant animals. Chaetognaths exhibit several morphological, physiological, anatomical and molecular peculiarities. In all known chaetognath mitochondrial (mt) genomes, the number of genes encoding tRNAs is always small. In addition, the primary sequences and secondary structures of chaetognath tRNAs are not conventional. Several metazoan mt-trn genes (encoding tRNAs) exhibit nucleotide triplets corresponding to stop codons (TAG/TAA) and/or start codons (ATG/ATA) at specific conserved positions. The products of these genes that bear one or both types of these codons are known as ss-tRNAs (ss for stop/start). Indirect analyses strongly suggest that in chaetognath mt genomes, some of these start and stop codons could be functional. Moreover, taking these codons into account in the algorithms predicting tRNAs makes it possible to identify and correct several potential annotation errors. Furthermore, a type of ss-trn gene appears to have emerged by duplication in order Aphragmophora.

Keywords: Chaetognatha, mitochondrial, ss-tRNA, start codon, stop codon, overlapping genes

Introduction

Chaetognaths are a small phylum of marine invertebrates with a size range of 2-120 mm. This phylum comprises approximately 130-140 reported species and is subdivided into two orders, Phragmophora and Aphragmophora [1], based on the addition of the transverse musculature (phragma) in the body. Chaetognaths live in various habitats, being found in the open sea, on or near the bottom, from polar to tropical regions, at all depths, in dark submarine caves, in the interstitial milieu and even around hydrothermal vent sites [2]. Chaetognaths were long considered strictly carnivorous but different aspects of their feeding biology have led to the conclusion that they feed primarily on dissolved and fine particulate matter and not on prey [3]. Despite their soft bodies, early Cambrian (~540-520 million years ago) chaetognath fossils with morphologies almost identical to recent forms have been discovered in China, suggesting a Precambrian origin [4]. Their exact phylogenetic position remains controversial [5-7]. Besides, the two longest species of giant viruses identified to date have been found in this taxon [8]. They also exhibit numerous other peculiarities including certain characteristics of their nuclear and mitochondrial (mt) genomes (e.g., [9-13]).

Until 2016, it was thought that chaetognath mitogenomes bore no [14] or only one trn gene (specifying the tRNA-Met) [15,17], rather than the 22 such genes found in most of the other invertebrates that employ the same mt genetic code. However, four studies invalidated the view that all chaetognath mitogenomes include zero or a single trn gene [12,18-20]. Nevertheless, in chaetognaths, most of the mt-trn sequences are non-canonical, and in a given chaetognath mt genome, only approximately half a dozen trn-like sequences have been found [12]. As 22 trn genes are required, the missing tRNAs may be nuclear encoded and imported from the cytosol, or they may be highly post-transcriptionally modified or "bizarre" trn genes that have escaped our detection [21].

In 2004, searching for chaetognath mt-trn genes [16], we incidentally observed that tRNAs bear nucleotide (nt) triplets corresponding to stop or start codons at precise conserved positions. The products of these genes that bear one or two types of these codons are referred to as ss-tRNAs (ss for stop/start) [22]. As most of the studies on this topic focus on DNA sequences, these codons are usually annotated TAR or ATR instead of UAR or AUR (R for purine). While numerous trn genes exhibiting nt triplets corresponding to stop codons (TAG/TAA) at precise conserved positions have been found in all taxa and genomic systems examined to date, relatively high frequencies of start codons (ATG/ATA) occur principally in
fungal/metazoan mt-trn genes \(^{22}\). The last nucleotide of these triplets is the first nucleotide involved in the 5'-D- or 5'-T-stem, so they are referred to as TAR10 and ATR49, respectively. In the chaetognath mt genomes, which are "very constrained", the length of the intergenic regions between two protein-encoding genes is quite often compatible with the average distance found between the TAR10 codon and the ATR49 codon; thus, the main focus of the present study is to analyse the ss-trn sequences found in chaetognath mtDNAs for the first time and to propose plausible scenarios concerning the origin of these genes.

![Fig 1: Typical cloverleaf secondary structure of a metazoan mt-ss-tRNA (A) with a 3D image of an L-shaped tRNA (B). In the 2D structure, standard numbering was applied. The first two nucleotides of the variable region and those of the D-loops and T-loops are represented by circles. The diagonal dashed line indicates the approximate separation between the "top half" and the "cherry-bob"/"bottom half". Nucleotide types are given for UAG10 and AUG49 triplets, the discriminator base (preferentially an A), and the CCA tail at the 3'-end. Short lines connect nucleotides undergoing pairing within stems. Colouring: acceptor-stem; purple; D-arm; red; anticodon-arm; blue with the anticodon in black; T-arm; green; and CCA tail; orange. The yellow segments in descending order of size, represent the variable region (connector 2), connector 1 and nt 26. Figure adapted from Faure and Barthélemy \(^{22}\) and 3D image reproduced with the kind permission of Prof. N.R. Voss (Roosevelt University, Ill., USA)](image)

**2. Materials and Methods**

**2.1 Mitogenomic sequences**

This study analysed complete or almost complete chaetognath mtDNAs available in GenBank (Table 1). For the mt genomes of *Sagitta ferox* and *Pterosagitta draco*, the GenBank staff noted that they were "unable to verify sequence and/or annotation provided by the submitter"; however, these mtDNAs were still included in the analyses even if reannotation was carried out, and possible sequence problems have been discussed.

**Table 1: List of the chaetognath mt genomes used in this study.**

<table>
<thead>
<tr>
<th>Orders</th>
<th>Species</th>
<th>Accession numbers</th>
<th>Abbreviations</th>
<th>Complete genomes</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aphrag-morphora</td>
<td><strong>Decipisagitta decipiens</strong></td>
<td>AP011546</td>
<td>D.d.</td>
<td>yes</td>
<td>([15])</td>
</tr>
<tr>
<td></td>
<td><strong>Sagitta elegans.1</strong></td>
<td>KP899780</td>
<td>S.el.1</td>
<td>no</td>
<td>([17])</td>
</tr>
<tr>
<td></td>
<td><strong>Sagitta elegans.2 (lineage G)</strong></td>
<td>KP899787</td>
<td>S.el.2</td>
<td>no</td>
<td>([17])</td>
</tr>
<tr>
<td></td>
<td><strong>Sagitta enflata</strong></td>
<td>AP011547</td>
<td>S.en.</td>
<td>yes</td>
<td>([18])</td>
</tr>
<tr>
<td></td>
<td><strong>Sagitta ferox</strong></td>
<td>KT818830</td>
<td>S.f.</td>
<td>yes</td>
<td>([18])</td>
</tr>
<tr>
<td></td>
<td><strong>Sagitta setosa</strong></td>
<td>KP899756</td>
<td>S.s.</td>
<td>no</td>
<td>([13])</td>
</tr>
<tr>
<td></td>
<td><strong>Pterosagitta draco</strong></td>
<td>KU017531</td>
<td>P.d.</td>
<td>yes</td>
<td>([20])</td>
</tr>
<tr>
<td></td>
<td><strong>Zonosagitta nagae.1</strong></td>
<td>AP011545</td>
<td>Z.n.1</td>
<td>yes</td>
<td>([17])</td>
</tr>
<tr>
<td></td>
<td><strong>Zonosagitta nagae.2</strong></td>
<td>KF051939</td>
<td>Z.n.2</td>
<td>yes</td>
<td>([19])</td>
</tr>
<tr>
<td>Phrag-morphora</td>
<td><strong>Paraspadella gotoi</strong></td>
<td>AY619710</td>
<td>P.g.</td>
<td>yes</td>
<td>([15])</td>
</tr>
<tr>
<td></td>
<td><strong>Spadella cephaloptera</strong></td>
<td>AY545549</td>
<td>S.c.</td>
<td>yes</td>
<td>([14])</td>
</tr>
</tbody>
</table>

**2.2 Bio-informatic analyses**

Some of the *trn* genes were detected by the authors reporting the sequences but were not identified as ss-trn genes. Otherwise, previously unidentified tRNAs were detected using tRNAscan-Se 1.21 \(^{23}\) or via alignments of intergenic regions with known *trn*-sequences or BLASTn searches. Multiple sequence alignments were performed with ClustalW software \(^{24}\); the alignments were manually refined at positions where clear misalignments were produced by the algorithm (i.e., they were verified for the conservation of important features and sites, principally in 2D alignments). Analyses of possible sequence homology with non-chaetognath *trn* genes were performed using two databases that include primary sequences and graphical representations of tRNA 2D structures: tRNAdb and mitotRNAdb \(^{25}\).
3. Results
3.1 trnM genes

A

<table>
<thead>
<tr>
<th>Species</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z.n.1/2</td>
<td>ATTATTGTAAGCTA</td>
</tr>
<tr>
<td>D.d.</td>
<td>CTATTGTAAGCTA</td>
</tr>
<tr>
<td>S.en</td>
<td>ATTATTGTAAGCTA</td>
</tr>
<tr>
<td>S.e.1.1</td>
<td>TTTTATATTGTAAGCTA</td>
</tr>
<tr>
<td>S.e.1.2</td>
<td>TTTTATATTGTAAGCTA</td>
</tr>
<tr>
<td>S.f.</td>
<td>ATTATTGTAAGCTA</td>
</tr>
<tr>
<td>S.c.</td>
<td>TTTTATATTGTAAGCTA</td>
</tr>
<tr>
<td>F.g.</td>
<td>TTTTATATTGTAAGCTA</td>
</tr>
</tbody>
</table>

Fig 2: Alignments of charitygnath mt-trnM-like sequences with their secondary (A) and primary structures (B). For abbreviations see Table 1. All the trnM genes are found within the cox3-rrnS gene block except for lineage G of S. elegans (S.e.2), which is located within the nd3-rrnS gene block. The positions of the various genes in their respective mtDNAs are as follows: Z. nagae (for long sequences, 3643-3720 and 5201-5281 for Z.n.1 and Z.n.2, respectively; for short sequences, 3640-3701 and 5201-5262 for Z.n.1 and Z.n.2, respectively), D. decipiens (2892-2970 and 2890-2952 for long and short sequences, respectively), S. enflata (3647-3730 and 3647-3709 for long and short sequences, respectively), S. setosa (4756-4834 and 2379-2437 for long and short sequences, respectively), S. elegans.1 (3395-3456), S. elegans.2 (826-887), S. ferox (6182-6242), S. cephaloptera (11182-11254) and P. gotoi (5157-5216). The TAA1 and ATG49 triplets, and the anticodons (CAT) are indicated in boldface. The pseudo-ATG49 of the D. decipiens sequence and all the nucleotide flanking regions are underlined. The complete and incomplete stop codons which the authors who reported the sequences highlighted in red. Moreover, when the authors noted an incomplete stop codon, the first in-frame complete stop codon is also highlighted in red. In (A), the lines with the "*" and "~" symbols correspond to the secondary structure predictions; in addition, for some sequences, two possible lengths for the same trn type are proposed and the upper and lower lines correspond to the 2D structures of the long and short trn sequences, respectively. In (B), asterisks (*) indicate conserved nucleotides. NB, the presence of trnM genes was not detected by the authors who reported the sequences of S. cephaloptera, S. elegans and S. ferox; similarly, the possible short trnM sequences of S. enflata and D. decipiens are proposed by our group.

The first trn genes found in charitygnathas were those specifying methionyl mt-tRNA [15-17]. The sizes of the trn genes may differ slightly from those proposed by the authors who reported the sequences because nucleotide alignment and prediction of secondary structure allow the pairings to be optimized, especially at the 5'- and 3'-termini of these genes (Fig 2). To date, only the trn genes of order Phragmophora sequenced to date exhibit a canonical structure [15, 16], in contrast to those of order Aphantinophora. In the latter order, in the same region of Z. nagae mtDNA (Fig 2A), the authors who reported one sequence assumed that the trn gene size would be long with an extra loop at the level of the variable region (Z. nagae) [17], whereas according to the authors who reported another complete mtDNA sequence of the same species, the trn gene size is shorter (Z. nagae) [19]. The cox3-rrnS gene block has been found in the mtDNAs of all charitygnath species except for S. elegans lineage "G" named by Marletaz et al. [13] (the S. elegans.2 mtDNA belongs to this lineage), and P. draco [20]. In all the charitygnath mtDNAs exhibiting the cox3-rrnS gene block, a putative trnM sequence has been detected between these two genes. The authors who reported the annotation of the S. ferox mt-DNA did not mention the two rrn genes [18], however, a rapid analysis suggested that the rrnS and rrnL genes are located between
coxl and nad6 genes, and cytB and coxl genes, respectively. The trnM gene of S. ferox, which was previously incorrectly identified as a trnH gene by the authors who reported the sequence, is located within the cox3-trnS block, as expected. In the sequence of S. elegans belonging to lineage G, the trnM gene is found within the nad3-trnS gene block (see S.e12 in Fig 2) and the stop codon of the nad3 gene is located considerably upstream of the trnM sequence (Fig 2A). It must be noted that the TAA49 triplets of trnM genes are never used as stop codons. The TAA49 triplets are in-frame for three sequences (D. decipiens, S. elegans, S. setosa) but are preceded by one or two (for S. setosa) in-frame stop codons. For three sequences (S. ferox, P. gotoi and S. cephaloptera), the first in-frame complete stop codons are found at the 3'-termini of the trnM gene or downstream of this gene. However, Blast analysis of chaetognath Cox3 proteins strongly suggested that trnM sequences are not included in the cox3 sequences (data not shown). So, at least for S. cephaloptera, S. ferox and P. gotoi mtDNAs, there is an incomplete stop codon in the corresponding mRNA upstream the trnM gene (or at the very beginning of this sequence); this had been proposed by authors who reported the sequences, either based on those of the trnM gene (15) or intuitively (16). Moreover, in three trnM genes there is an ATG49 triplet, but because the downstream genes encode 12S rRNA, this triplet cannot be used as a start codon. In addition, no trnM-like gene has been found in the P. draco mt genome.

Fig 3: trnL-like sequences of chaetognaths with their 2D structures. In mtDNAs, these sequences are found within those of the cox2-nad1 gene block in S. ferox (3201-3265), D. decipiens (11035-11113), P. draco (3331-3396), S. elegans (352-422), S. enflata (654-724), S. setosa (11015-11084), and Z. nagae 1 and Z. nagae 2 (640-720 and 2201-2081, respectively), and within the nad1-nad3 gene block in P. gotoi (3149-3312) and S. cephaloptera (9070-9148). TAR10, AT49 and the anticodon are presented in boldface, and the anticodon is underlined. The first in-frame stop codon is highlighted in red, whereas the assumed start codon is highlighted in green. For the Z. nagae 1/2 sequence, the letter R represents a purine nucleotide. For P. gotoi, the amino acids of the COOH terminus and of the NH2 terminus of the Nad1 and Nad3 sequences, respectively, are noted. For S. ferox, the NH2 terminus of Nad1 sequence is also noted. Amino acids are shown in single-letter code.

3.2 trnL genes
Within chaetognath mtDNAs, seven putative leucinyl(1) ss-trn genes have been found as well as two "classical" trn-like sequences (Fig 3). All the mtDNAs of the species of order Aphragmophora that harbour the cox2-nad1 gene block have this trn gene (however, this excludes lineage G of S. enflata). The two other ss-trnL genes belong to the S. cephaloptera and P. gotoi mt genomes (order Pbragmophora) and are found between the nad1 and nad3 genes. Three ss-trnL genes exhibit an extra loop between the Ac- and T-stem-loop-stem (D. decipiens, Z. nagae and S. cephaloptera). Only the sequences of P. draco and S. enflata do not contain the TAG10 triplet, but this codon is a possible stop codon only for the two species belonging to order Phragmophora. Concerning the sequences of order Aphragmophora, in five cases, the TAG10 triplet is out of frame, and for S. enflata, there are two stop codons in tandem, with TAG10 in second position. The intergenic region between the cox2 and nad1 genes in species of order Aphragmophora is generally too long to also exhibit an ATR49 start codon, which is the case for S. cephaloptera mtDNA as well. In contrast, a complete ss-tRNA containing both the putative stop codon TAR10 and the start codon ATR49 has been found in P. gotoi mt genome. This tRNA does not exhibit an extra loop, and only a complementary base pair has been found in the D-arm; however, changes could occur post-transcriptionally.
Alignments of the primary sequences of genes specifying tRNA-Leu. The alignments of the sequences of (A), (B) and (C) correspond to sequences from species of order Aphragmophora, order Phragmophora, and both orders Aphragmophora and Phragmophora, respectively. In (A), asterisks indicate conserved nucleotides in all the sequences on the upper line or only those of D. decipiens, S. enflata, S. setosa, Z. nagae.1 and Z. nagae.2 on the lower line, respectively. For other data, see Fig 3.

Alignments of the primary sequences of genes specifying tRNA-Leu strongly suggest that at least the sequences of D. decipiens, S. enflata, S. setosa, Z. nagae.1 and Z. nagae.2, which belong to order Aphragmophora, come from a common ancestral sequence (Fig 4A). The two sequences of order Phragmophora also exhibit a relatively high level of homology which each other (Fig 4B). However, in these last two alignments, the homology after the Ac-arm could be an artefact due to the high AT content. Alignments of trnL1 sequences from the two orders exhibit only at the level of TAG10 and the Ac-5′-stem-loop (Fig 4C); however, the high AT content prompts us to remain cautious. In the mtDNAs of two species (P. draco [20] and S. elegans.1), tRNA-Leu.1-like sequences have been found, but there is no TAR10 triplet and sequence alignment with the trnL1 genes of other Aphragmophora reveals a low level of nucleotide conservation. Additionally, analyses using the mitotRNAdb database show that the sequence TAGTAT(TAG10 + 5′-stem)-/TATTAGN(Ac-loop) has only been found in some mt-trnL1 genes (19 of 1374 total mt-tRNA-Leu1 sequences) and these 19 tRNAs belong only to bilaterian animals; however, the database exhibits a strong bias for Metazoa. A total of four, one, nine and five tRNAs have been found among Cephalochordata, Xenacoelomorpha, Nematoda and Mollusca mtDNAs, respectively.

Fig 4: Alignments of the primary sequences of genes specifying tRNA-Leu.1. The alignments of the sequences of (A), (B) and (C) correspond to sequences from species of order Aphragmophora, order Phragmophora, and both orders Aphragmophora and Phragmophora, respectively. In (A), asterisks indicate conserved nucleotides in all the sequences on the upper line or only those of D. decipiens, S. enflata, S. setosa, Z. nagae.1 and Z. nagae.2 on the lower line, respectively. For other data, see Fig 3.

Fig 5: Alignments of the NH3 terminus of Nad3 amino acid sequences of several chaetognath species. For S. enflata and S. setosa, the residues in boldface have been added to the sequences. In the S. enflata sequence, the isocitricine residue, which is underlined, corresponds to a methionine in the original sequence. Asterisks indicate conserved residues, double dots indicate similar residues, and single dots indicate like residues.
### Fig 6: Duplications found in trnL1-like sequences found within the cox2-nd1 gene block of order Aphragmophora. The TAG10 triplets and the anticodons are presented in boldface, and the latter are underlined. The first in-frame stop codons are highlighted in red. The start codons of the nd1 genes are highlighted in green. For other data, see Fig 3. (A) Alignments of duplicated regions. The nucleotides that do not belong to the trn region are underlined. (B) Alignment of the ss-trnL1 gene of *S. elegans* and its flanking sequences with the homologous region found in the *cox1-nd1* gene block of *S. elegans*; the ss-trnL1 sequence is underlined, and the 3' region of the duplicated region is in boldface.

The ss-trnL1-like sequences of mt genomes belonging to order Aphragmophora resulted from a duplication event of a region stretching from approximately TAG10 to the first part of the 3'-Ac-stem (Fig 6A). This duplication involves nearly the entire "cherry-bob" structure. However, in *S. ferox*, the possibility cannot be excluded that the duplication principally involved a region upstream of the *trn* gene. Concerning the two non ss-trn genes, only those belonging to *S. elegans* exhibit a high level of nucleotide conservation. It must be noted that the "G" nucleotide of the anticodons is never found in the duplicated regions. Analysis of Fig 6A suggests that the duplication occurred in the mt genome of the ancestor of suborder Ctenodontina. In the *S. elegans* mtDNA, the cox2-nd1 gene block is not present, and no trnL1-like sequence has been found, but one mt-region exhibits sequence homology to the 3'-part of the trnL1 gene of *S. elegans*.1 (Fig 6B). This homology is found within the *cox1-nd1* gene block of *S. elegans*. The alignment suggests that during gene order rearrangement in lineage G, the 5' region of the *trnL1* gene was lost.

### Fig 7: Alignments of the 5' regions of the ss-trnL1 genes of order Aphragmophora with their 5' flanking sequences. The 5' sequence of the trn-genes is underlined. In *S. ferox*, the 5' region of the duplicated region are in italics. For the *Z. nagae*, 1/2 sequence, the letter Y represents a pyrimidine nucleotide. Asterisks indicate conserved nucleotides for all the sequences and for all the sequences except for those of *D. decipiens* on the upper and the lower lines, respectively. For other data, see Fig 3.

The 5' flanking regions of the ss-trnL1 genes of order Aphragmophora were analysed (Fig 7). For all the sequences, the approximately forty-base pair regions upstream of the TAG10 triplet exhibit a low homology rate with each other, especially compared with the 5' termini of the trn genes (Fig 7). Indels have also been found. The regions upstream of the TAG10 triplet encode the COOH termini of Cox1 proteins (see Fig 3); however, as noted by [12] in chaetognath mt genomes, the intergenic regions between protein-encoding genes may exhibit a higher level of homology than those found in protein-encoding regions. This strongly suggests that, at least in this case, the selection pressure exerted on the trn genes is much higher than that found in regions encoding proteins.

---

**Journal of Entomology and Zoology Studies**

A

**D. decipiens** (5'-part) 11036-5A[AGAAATATCCATATCTGATATTTTAATCGTTATTTAAATTTGATGATT---11072
(3'-part) 11072-<http:[//RNA-box]

**F. dracon** (5'-part) 3331-AAGGC[AGAATCTCAGATAC-ATATCTTAGTATTC-TACATGCN-403
(3'-part) 3331-AAGGC[AGAATCTCAGATAC-ATATCTTAGTATTC-TACATGCN-403

**S. elegans** 1 (5'-part) 352-NTAAAAGTCGCCTATGATTATTTTAGATTATGGTCT-302
(3'-part) 352-NTAAAAGTCGCCTATGATTATTTTAGATTATGGTCT-302

**S. enflata** (5'-part) 656-TTCCTAGATATTTTAATCGTTATTTAAATTTGATGATT--656
(3'-part) 656-TTCCTAGATATTTTAATCGTTATTTAAATTTGATGATT--656

**S. ferox** (5'-part) 3164-TTATCTCTATATCTGATATTTTAATCGTTATTTAAATTTGATGATT--3207
(3'-part) 3206---TATAATGACTTATTTTAATCGTTATTTAAATTTGATGATT---

**S. setosa** (5'-part) 1104-AAGAATCTCAGATAC-ATATCTTAGTATTC-TACATGCN-1105
(3'-part) 1105-AGAATCTCAGATAC-ATATCTTAGTATTC-TACATGCN-1105

**Z. nagae**, 1/2 (5'-part) 628/2/323--TATCTTAGTATTC-TACATGCN-676/2/323
(3'-part) 676/2/323--TATCTTAGTATTC-TACATGCN-676/2/323

**S. el.1** 390-AAGAATCTCAGATAC-ATATCTTAGTATTC-TACATGCN-390
**S. el.2** 1163-AAGAATCTCAGATAC-ATATCTTAGTATTC-TACATGCN-1163

**S. el.1** 404-ACATCTTACATATTTTAATCGTTATTTAAATTTGATGATT--404
**S. el.2** 1174-CACATCTTACATATTTTAATCGTTATTTAAATTTGATGATT--1174

---

**Z. n. 1/2** 582/2/123-GAGAATCTCAGATAC-ATATCTTAGTATTTTAATCGTTATTTAAATTTGATGATT---582/2/123

**S. en.** 568-AAGAATCTCAGATAC-ATATCTTAGTATTTTAATCGTTATTTAAATTTGATGATT---

**S. el.** 1095-AGAATCTCAGATAC-ATATCTTAGTATTTTAATCGTTATTTAAATTTGATGATT-1095

**S. el.** 1095-AGAATCTCAGATAC-ATATCTTAGTATTTTAATCGTTATTTAAATTTGATGATT-1095

**D. d.** 1058-AGAATCTCAGATAC-ATATCTTAGTATTTTAATCGTTATTTAAATTTGATGATT-1058

---

**Z. n. 1/2** 622/3/223-5ATAC-ATATCTTACATATCTTACATATTTTAATCGTTATTTAAATTTGATGATT---622/3/223

**S. en.** 628-ACATCTTACATATTTTAATCGTTATTTAAATTTGATGATT---

**S. el.** 1095-AGAATCTCAGATAC-ATATCTTAGTATTTTAATCGTTATTTAAATTTGATGATT-1095

**D. d.** 11016-ACATCTTACATATTTTAATCGTTATTTAAATTTGATGATT-11016

---

**Z. n. 1/2** 622/3/223-AGAATCTCAGATAC-ATATCTTAGTATTTTAATCGTTATTTAAATTTGATGATT---622/3/223

**S. en.** 628-ACATCTTACATATTTTAATCGTTATTTAAATTTGATGATT---

**S. el.** 1095-AGAATCTCAGATAC-ATATCTTAGTATTTTAATCGTTATTTAAATTTGATGATT-1095

**D. d.** 11016-ACATCTTACATATTTTAATCGTTATTTAAATTTGATGATT-11016

---

**Notes:** The S. elegans 2 sequence is not included due to the lack of information.
S. cephaloptera

For one chaetognath species (S. cephaloptera) a large collection of expressed sequence tags (ESTs) has been assembled by Marlétaz et al. [9] and submitted to the EMBL website. Blast analyses using all the trn genes analysed in this study as the query against the S. cephaloptera EST database only revealed homology of the trnL1 gene or its flanking regions (Fig 8). Two ESTs contain the 3’ portion of the nadl mRNAs and polyadenylation begins one or two nucleotide(s) upstream of the trnL1 sequence (CU557716) or after the first nt of this trn gene (CU563975). In the latter case, the ‘T’ nt at position 568 could be an artefact, as suggested by analyses of the beginning portions of other polyadenylated sequences.

Even though the analysis only involves two EST sequences, this suggests that the nadl mRNAs could exhibit an incomplete stop codon immediately before the 5' terminus of the trnL1 gene. The third EST (CU694059) is bicistronic and contains a large part of the nadl gene and the trnL1 sequence up to the beginning of the 5'-T-stem. As the 3'-T-stem and the 3'-Acc-stem are missing, the putative truncated tRNA would require post-transcriptional maturation to become functional.

3.3. trnG genes

A

P.g. CoxI    MDFYLSMNLSQN
P.g. TAAGTTTAAACG    ATGTGGATATTGGAAGAATATATATTAA
D.d. CoxI    MFWSAS
D.d. GATCTATGTATGACTTCTTTCTTGAATATTACGCTAGTCT
Z.n.1/2 CoxI    MELNQLNANASFM
Z.n.1/2 GATCTATGTATGACTTCTTTCTTGAATATTACGCTAGTCT

B

P.g.达人--TAA-CTGTTA--GOTAAGAATATATATTAA--
D.d. GATCTATGTATGACTTCTTTCTTGAATATTACGCTAGTCT
Z.n.1/2 GATCTATGTATGACTTCTTTCTTGAATATTACGCTAGTCT

C

S. elegans.1
S. setosa.1
S. faxox
S. nagae.1
S. nagae.2
P. gotoi
S. cephaloptera
S. enflata
D. decipiens.
S. elegans.1
S. setosa.1
S. faxox
S. nagae.1
S. nagae.2
P. gotoi
D. decipiens.

Fig 9: Analyses of chaetognath trnG genes. (A). In the mtDNAs of D. decipiens (10402-10468), P. gotoi (1535-1593), and Z. nagae.1 and Z. nagae.2 (1521-1600 and 11419-11459/1-35 respectively); these sequences are at the level of the regions flanking the cox1 and cox2 genes. ATG49 and the anticodon (TTG) are indicated in bold. The putative start codons indicated by the authors who reported the sequences are highlighted in green. For other data, see Fig 3. In the P. gotoi sequence, ATG49 is in frame with the putative start codon ATA. Moreover, the first complete stop codon of the cox1 gene of D. decipiens, a TAA triplet, is highlighted in red. Alignments are presented using secondary structure predictions, and the Cox2 amino acid sequences are indicated. (B). Alignment of primary sequences. (C). Alignment of secondary structure predictions.
Three glycinyl ss-trn genes have been found in chaetognath mt genomes (Fig 9), but a high level of nucleotide conservation has been found only between the two sequences of order Aphragmophora (Fig 9B). Analyses of the ss-trn genes of D. decipiens and P. gotoi suggest that the ATG49 triplets present in these sequences might be the physiological start codons of the cox2 genes (Fig 9). For D. decipiens, the authors who reported the sequences assumed that the start codon was ATG49, whereas for P. gotoi, they hypothesized that a triplet ATA, located upstream of the anticodon, was the translation initiation codon \(^{103}\) (Fig 9A); for Z. nagae trnG genes, according to the authors who reported the sequences, the start codon is downstream from the trn gene, but ATG49 is in frame with the cox2 gene. Blast analyses strongly suggest that the three ATG49 triplets are the physiological start codon for the cox2 gene (Fig 9C). Indeed, the alignment of the regions around the first amino acid residues of chaetognath Cox2 proteins shows that, except for the P. gotoi and P. draco proteins, all the other sequences subjected to Blast searches are shorter by at least six residues (Fig 9C). Additionally, whereas for the cox1 gene of P. gotoi, the stop codon (TAAT) is located upstream of the trn gene, for D. decipiens, the first complete stop codon of the cox1 gene is the TAAT triplet located at the 5' terminus of the 3'-Ac-stem, suggesting overlap between the two protein-encoding sequences. However, in this last hypothesis, Blast analyses revealed that the COOH-terminal portion of the D. decipiens protein, which is abnormally long, does not exhibit any sequence that homologous (or analogous) to other Cox1 protein sequences (data not shown). An incomplete stop codon located upstream of the trnG gene is presumably completed by polyadenylation. In addition, relatively strong sequence alignment is found between the two trnG genes of order Aphragmophora (Fig 9B); whereas, the level of nucleotide sequence identity between the P. gotoi trnG gene and the two sequences of species of order Aphragmophora is approximately 42%. Moreover, the level of sequence identity is lower in the overlapping region (trnG/cox2) than in the non-overlapping area, 17% versus 58%, highlighting against a high degree of sequence conservation between the first residues of the various Cox2 proteins (see Fig 9C). Besides, there could be annotation errors at the beginning of the S. elegans Cox2 proteins, e.g., for S. elegans1, the start codon begins at nt 13364 and not at nt 1, and the CAG codon is probably not an alternative initiation codon. The Cox2 sequence of S. elegans2 has not been included in the analysis because its NH2 terminus differs strongly from those of the other sequences.

![Fig 10: Alignments of trnT and Cox1 sequences of chaetognaths. In the Z. nagae1 (9849-9914), Z. nagae2 (11410-11458/1-16) S. enflata (11020-11081) and P. draco (937-1021) mtDNAs, the trnT-like genes are at the level of the regions flanking the 16S rrm genes and cox1 genes. The ATG49 triplets, the TAG10 triplet of S. enflata and the anticodons are in bold. The start codons given by the authors who reported the sequences are highlighted in green. In the Z. nagae sequences, the ATG49 triplet, which is the first possible start codon, is underlined. (A) Alignment of trnT-like genes with their 2D structures. For the P. draco sequence, the two parts of the duplicated regions are underlined or highlighted in yellow. The NH2 regions of the Cox1 proteins are also reported. (B) Alignment of the primary sequences of the trnT-like genes. Asterisks indicate conserved nucleotides for Z. nagae and S. enflata sequences on the upper line and for the four sequences on the lower line, respectively. (C) Alignment of the NH2 regions of the Cox1 proteins of Aphragmophora specimens. All the sequences correspond to those found in GenBank, except for the four first residues (in bold) of the Z. nagae sequences.](image-url)
In chaetognath mt genomes, four threoninyl ss-trn genes have been found (Fig 10A). All of these sequences have been found among specimens of order Aphragmophora and contain the ATG49 triplet. According to the authors who reported the sequences, this triplet serves as the start codon for the S. enf
tata and P. draco cox1 genes, whereas for the Z. nagae sequences, the initiation codon is an in-frame ATA codon located downstream from the ATG49 triplet [17]. According to Wei et al. [20], the trnT-like sequence of P. draco starts at nt-934 and ends at nt-1021, but our reanalysis shows that the 3’ terminus of the trn-gene of P. draco has been duplicated and that this trn gene is shorter and does not include the duplicated region (Fig 10A). Alignment of the primary sequences of the trnT-like genes revealed that the sequences of Z. nagae and S. enf
tata which are closely related species exhibit a high level of homology (> 89%) (Fig 10B). Alignment of the four sequences shows that the 5’-Acc-stem and the D-loop of the P. draco sequence are very different from those of the two other species. Additionally, alignment of the NH2 regions of the Cox1 proteins of chaetognath specimens strongly suggests that the physiological start codon would be the ATG49 triplet; however, four amino acid residues downstream of the methionine encoded by ATG49 triplets, there is often a methionine residue (for 7 of 12 sequences) suggesting that there may be an alternative initiation codon (Fig 10C).

4. Discussion

4.1 Chaetognath trnM genes

In some species, only one or two trn genes have been found among their mtDNAs. When only one gene is present, it specifies tRNA-Met (to date, this situation is only observed in some Anthozoa, a class of Cnidaria), whereas when only two are found, one specifies tRNA-Met and the other tRNA-Trp (this occurs in cnidarians and a lineage of demosponges: Keratosa) [27]. This observation supports the inference of special roles of these two tRNAs in animal mitochondria, in translation initiation and the recognition of the standard stop codon UGA as tryptophan. Indeed, in most animal mt genetic codes, the two major deviations from the standard genetic code are the reassignment of the UUA codon from isoleucine to both initial and internal methionine residues (trnM(CAU)) and reassignment of the UGA codon from the termination codon to tryptophan (trnW(UCA)) (see e.g., [16, 27]). This situation suggests that these reassignments would have appeared very early in the evolution of, at least, metazoans and would have resulted from the optimization of the mt genome [22]. A previous analysis showed that a trnW gene exists in the mtDNAs of almost all examined chaetognaths [12], and the present study showed that only the P. draco mt genome does not appear to exhibit the trnM gene. Interestingly, the chaetognath trnM genes seem to correspond to the initiating tRNA (tRNA-I-Met) which may be charged with formyl-methionine to be used for protein initiation. As found in numerous taxa, including chaetognaths, the AUA and AUG codons are used in both initiator and internal positions; if only one tRNA-Met is specified mitochondrially, it can either play the two translational roles, or it might be charged only with formyl-methionine, whereas another tRNA-Met from the nucleus only functions in elongation. All the chaetognath trnM genes identified to date are partial or complete ss-trn genes, but the TAA10 and ATG49 triplets are never used as stop or start codons respectively. Concerning the ATG49 triplets found in the D. decipiens, P gotoi and S. cephaloptera mt genomes, as the downstream genes specifying 12S rRNA, they cannot play a role in translation. In addition, in chaetognath trnM genes, TAA10 triplets are never used as stop codons; when they are in-frame, there is always an upstream complete stop codon. Analyses of chaetognath trnM genes suggest that the use of TAR10 and ATR10 as stop or start codons, respectively, in other ss-trn genes could partially represent an exaptation. This concept suggests that traits that evolved for one purpose have been co-opted for their current use [20]. However, in the ss-trn genes these two types of triplets continue to perform their primitive function, which is essential for maintaining the spatial conformation of tRNAs [22], nevertheless, in some cases, they can also play a crucial role in translation. The presence of ss-trn genes in bacterial and nuclear genomes that are relatively distant from neighbouring protein-encoding genes also argues in favour of this hypothesis.

Furthermore, the S. elegans mt genomes have been divided into eight lineages [13]. In all the lineages except G, the trnM genes are within the cox3-rnS gene block, whereas in the last lineage, the trn gene is within the nd3-rnS gene block (see S.el2 in Fig 2 and Table 2), and the stop codon of the nad3 gene is located considerably upstream of the trnM sequence. Despite the difference in the upstream protein-encoding genes, the sequences of the trn genes are very similar; e.g., the two analysed trnM genes of S. elegans exhibit a high level of nucleotide identity (92%), and the percentage is always 100% at the level of the stems (Fig 2). This finding suggests that, these ss-trn genes at least appeared independent of any interactions with protein-encoding genes; indeed, when these interactions existed, they would have been selected later. The high degree of conservation of the trnM genes of different lineages of S. elegans suggests that strong selection pressure is exerted on this region suggesting that the specified tRNA would be functional.

Despite the fact that in animal mtDNAs, the level of similarity between trn genes, even those of the same type, is relatively low, a relatively large number of trnM genes exhibit sequence similarity at the level of their cherry-bob motif. As these similarities have been found principally in protostomian mt genomes (mostly chaetognaths and insects), this situation could represent evolutionary convergence, or the motif may already have been present in the common ancestor of chaetognaths and insects, considering the assumed age of phylum Chaetognatha, this motif would extend back to the mitochondria of the first protostomians. The latter hypothesis suggests that this motif would have disappeared from the mt genomes of the majority of protostomial taxa. According to the current knowledge, it is difficult to reject one of these two assumptions rather than the other. Additionally, for most of the chaetognath trnM genes, 2D structure predictions show non-canonical spatial arrangements after the Ac-5’-stem suggesting possible post-transcriptional modifications.
4.2 Chaetognath ss-trn genes not specifying tRNA-Met
Within chaetognath mtDNAs, seven putative leucinyl(1), four glycyl- and four threonyl-ss-trn genes have been found. The trnT genes have only been detected in some species of order Aphragmophora. Analyses of the COOH and NH2 termini of proteins encoded by overlapping genes strongly suggest that most of these ss-tRNAs exhibit TAG10 or ATG49 triplets, which are probably used as physiological stop or start codons respectively. Moreover, ss-tRNA-Leu1 of P. gotoi could be a complete true ss-tRNA in which the two types of triplet play a role at the translation level. The predicted 2D structures suggest that these tRNAs do not adopt canonical structures but improvements could occur during post-transcriptional maturation, which can also include the removal of extra loops. Only trnM and trnL1 genes exhibit a relatively high level of nucleotide identity with trn genes of the same type belonging to non-chaothognath species but for the trnL1 genes, this could be due in part to a high AT content [12]. In contrast to the trnL1 genes, the trnG genes, which have also been found in species belonging to the two chaetognath orders, show low inter order-level nucleotide identity. It seems that the trn genes present among the current chaetognath mtDNAs arose via a type of DIY mechanism ("molecular tinkering"), partially independent of phylogenetic relationships. Generally, in nucleotidic regions where two or more genes overlap, selection pressure is such that the sequences are highly conserved, and the reading frame is said to be locked. Paradoxically, in the ss-trnL1, ss-trnG and ss-trnT genes, the level of nucleotide identity is low in overlapping regions. This finding suggests that ss-trn genes, at least in chaetognaths, play a role in the precision of the initiation or termination of transcription (or in transcript maturation) and especially in the translation of mt-mtRNAs but do not act at the level of sequence conservation of protein-encoding genes. Moreover, as suggested previously [22], the genomic organization of ss-trn genes could allow regulation of the synthesis of products of the mt genomes of chaetognaths.

In chaetognath mitogenomes, the rnl gene could constitute a type of trn gene nursery [15], and the annotation of S. ferox mtDNA reinforces this hypothesis [18]; however, in chaetognath rnl genes no sequence exhibiting strong homology with the four types of ss-trn genes analysed in this study has been found within the sequences specifying rRNAs. Among the S. cephaloptera EST collection, only three ESTs exhibit sequence homology to one of the trn genes (trnL1) analysed in this article or to its upstream region. Two ESTs suggest that there is probably an incomplete stop codon just before the beginning of the trn gene. This situation could argue in favour of functionality of this trn gene. No EST that ends after the 3′ terminus of the trn sequence has been found, and the polyadenylated region of an EST starts at the beginning of the putative 5′-T-stem. The three following hypotheses can be proposed: 1/ this site of polyadenylation is artefactual; 2/ the 3′ portion of the trnL1 sequence is shorter; or 3/ as has already been shown, incomplete cloverleaf structures may also be repaired post-transcriptionally [29] and, in some cases, mt-trn genes that have incomplete 3′-ends are completed post-transcriptionally by polyadenylation [30]. However, the polyadenylation of mt RNAs could be ambiguous; indeed, this process can stabilize them or may be a signal of the degradation of the RNAs [31], thus, further experiments are required to reach a conclusion.

Several models have been proposed to explain the origin of tRNA molecules (see reviews [32-34]), which is beyond the scope of this work, but the two major hypotheses can be briefly proposed. Many authors have assumed that the current tRNA cloverleaf structure arose through direct duplication of an ancestral RNA hairpin (e.g.,) [32], whereas according to the "two halves" hypothesis [35], tRNAs are composed of two independent structural and functional domains: the "top half", containing the acceptor-stem and the T-arm, and the "bottom half", equivalent to the cherry-bob structure, which contains the D- and anticodon-arms (Fig 1). Several analyses have strongly suggested that the "bottom half", which would have appeared much later, was integrated into the pre-existing top half structure [36]. Analyses of trnL1-like sequences of species of order Aphragmophora revealed that they are a result of duplication of the cherry-bob region, which constitutes a sort of syncretism between the hypotheses of Di Giulio [32] and of Maizels and Weiner [35]. Additionally, in the living world, mechanisms exist for avoiding the possibility that in biological systems, a molecule that is similar but not strictly identical (at the sequence and therefore structure levels) to another that is functional could interfere with and make the latter molecule non functional in a given system. Under a hypothesis that we refer to as the "double funnel" hypothesis, the first funnel, in which the larger hole is at the top, symbolizes the fact that within multigene families, all members retain almost identical sequences over very long periods of time. This suggests the presence of active mechanisms causing sequences that begin to diverge to become almost all identical again. Gene conversion is one of the mechanisms in this process of "concerted evolution" [37]. In contrast, to avoid the fact that the equivalence principle does not apply between molecules that are relatively closely related at the sequence level, the opposite mechanism corresponding to the second funnel for which the pipe is at the top quickly accentuates differences. The natural mutation rate allows the induction of variations between sequences but, it is probably too slow when the situation is critical, e.g., after allopolyploidization events (see,
e.g.,) [9-11]. Several examples of the "double funnel hypothesis" can also be found in the tRNA world, which could imply the occurrence of duplication of trn genes followed by substitutions (e.g.,) [18, 39] and the existence of ss-trn genes that can combine several specificities related to the type of tRNA and to the overlap of protein-encoding genes in their 5'- and 3'-regions, which confers upon them singular characteristics.

What might have been the event(s) exerting such selection pressure that all the original mt-trn genes would have disappeared? The numerous rearrangements that have occurred in chaetognath mtDNAs would have been detrimental to trn genes because selection pressure would have been exerted preferentially at the level of protein-encoding genes. However, some observations strongly suggest that codon reassignments might be adaptive. Mt-codon reassignments might prevent viral infections as most viruses follow the standard genetic code [40, 41]. Approximately half of known alternative genetic codes are mt genetic codes and three of them have been found in fungi [42]. Viruses infecting mitochondria are relatively rare, but members of genus Mitovirus, family Narnaviridae (which are all capsidless), are ubiquitously detected in filamentous fungi; these viruses exhibit a positive single-stranded RNA genome encoding only an RNA-dependent RNA polymerase [43] and share their mitochondrial host's codon reassignment [44]. Moreover, endogenized mitovirus elements are widespread in land plant genomes [45]; thus, chaetognath nuclear genomes could still contain traces of past infections. Although very hypothetical at present, the ss-trn genes could also play a protective role in viral mt-infections by a completely unknown mechanism.

5. Conclusion
In addition to the loss of many trn genes and those encoding ATP6 and ATP8, the chaetognath mt genomes are characterized by an accelerated mt-mutation rate, which could be related to the extreme size reduction of these genomes (from c.11,000 to 14,000 bp for current sizes) and their propensity for structural rearrangements [17]. We assumed that in chaetognaths, all of the original mt-trn genes were lost during evolution, and that tRNAs allowing the translation of mt-mRNAs have been superseded by nuclear-specified tRNAs. Actually, since mt-tRNAs are perhaps more efficient than their cytosolic equivalents, we are witnessing the neosynthesis of new specific mt-trn genes via what seems, at least partly, to consist of a large-scale DIY mechanism. The fact that the pool of trn genes differs between phylogenetically closely related species and that tRNA-like sequences seem to be in statu nascendi favours this hypothesis. The observation that some of the new trn genes are ss-trn sequences strongly suggests the use of UAR10 or AUR49 as stop or start codons, respectively, as an exaptation. DNA alignments corresponding to COOH- or NH2-terminal protein regions, in which TAR10 and ATR49 are not always the physiological stop and start codons, reinforce this last hypothesis. The high level of nucleotide conservation in non-overlapping regions of ss-trn genes strongly suggests that selection pressure occurs preferentially in these areas, but this suggestion contrasts with known data, which deserves special attention in subsequent works. In the future, research on specific mt-amino-acyl tRNA synthetases could be very informative; direct sequencing of tRNAs and in vitro translation experiments are also needed.

6. References
17. Miyamoto H, Machida RJ, Nishida S. Complete


