The mitochondrial genome of *Paraspadella gotoi* is highly reduced and reveals that chaetognaths are a sister group to protostomes

Kevin G. Helfenbein†‡§, H. Matthew Fourcade‡, Rohit G. Vanjani†, and Jeffrey L. Boore†‡¶

| Department of Biology, University of Michigan, 830 North University Avenue, Ann Arbor, MI 48109; | Department of Energy Joint Genome Institute and Lawrence Berkeley National Laboratory, 2800 Mitchell Drive, Walnut Creek, CA 94598; | Department of Integrative Biology, University of California, Berkeley, CA 94720 |

We report the complete mtDNA sequence from a member of the phylum Chaetognatha (arrow worms). The *Paraspadella gotoi* mtDNA is highly unusual, missing 23 of the genes commonly found in animal mtDNAs, including ATP6, which has otherwise been found universally to be present. Its 14 genes are unusually arranged into two groups, one on each strand. One group is punctuated by numerous noncoding intergenic nucleotides although the other group is tightly packed, having no noncoding nucleotides, leading to speculation that there are two transcription units with differing modes of expression. The phylogenetic position of the Chaetognatha within the Metazoa has long been uncertain, with conflicting or equivocal results from various morphological analyses and rRNA sequence comparisons. Comparisons here of amino acid sequences from mitochondrially encoded proteins give a single most parsimonious tree that supports a position of Chaetognatha as sister to the protostomes studied here. From this analysis, one can more clearly interpret the patterns of evolution of various developmental features, especially regarding the embryological fate of the blastopore.

**Materials and Methods**

**DNA Amplification, Sequencing, and Informatics.** A preparation of *Paraspadella gotoi* total DNA was a gift of Nori Satoh, Kyoto University. Conserved regions within the mitochondrial genes *cox1* and *cox3* were amplified by PCR by using primers designed to match conserved regions. These products were sequenced, and specific primers were designed facing “out” from each gene and were used in long PCR with the enzyme LA-PCR from Takara, resulting in amplifications between a forward *cox1* primer and a reverse *cox3* primer and vice versa (Fig. 1). The long PCR products were sequenced by a combination of primer walking by using an Applied Biosystems 377 sequencer and random shotgun sequencing of sheared PCR products.

The final assembly of all reads was performed with Sequencher (Gene Codes, Ann Arbor, MI). *P. gotoi* protein and rRNA-encoding genes were annotated based on comparisons with homologous genes from other animal species. Searches for tRNA genes were performed by using tRNA scan-SE (7) and GCG (8).

**Sequences.** For phylogenetic analyses, we chose a broad representation of taxa from the available complete mtDNA sequences. We omitted those taxa with the most highly divergent sequences, an approach shown previously to be highly effective for reconstructing metazoan phylogeny using the small nuclear rRNA (9). The following sequences were retrieved from GenBank:

- *Ciona intestinalis* (GenBank accession no. DQ250865)*
- *Haplochlamys guineensis* (Accession no. AY583183)
- *Takifugu rubripes* (Accession no. AY583184)
- *Artemia salina* (Accession no. AY583185)
- *Ctenophora* (Accession no. AY583186)
- *Pisces* (Accession no. AY583187)
- *Mammalia* (Accession no. AY583188)
- *Vertebrata* (Accession no. AY583189)
- *Amphioxus* (Accession no. AY583190)
- *Arthropoda* (Accession no. AY583191)

This paper was submitted directly (Track II) to the PNAS office.

Data deposition: The sequence reported in this paper has been deposited in the GenBank database (accession no. AY619710).

§Present address: American Museum of Natural History, Department of Invertebrate Zoology, Central Park West at 79th Street, New York, NY 10024.

¶To whom correspondence should be addressed. E-mail: jboore@lbl.gov.

© 2004 by The National Academy of Sciences of the USA
Bank: Asterina pectinifera (Echinodermata) (accession no. NC_001627), Balanoglossus carnosus (Hemichordata) (NC_001887), Daphnia pulex (Arthropoda) (NC_000844), Drosophila yakuba (Arthropoda) (NC_001322), Florometra serratisima (Echinodermata) (NC_001878), Homo sapiens (Chordata) (NC_001907), Katharina tunicata (Mollusca) (NC_001636), Lachesis rubellus (Brachiopoda) (NC_002322), Limulus polyphemus (Arthropoda) (NC_003057), Locusta migratoria (Arthropoda) (NC_001712), Loligo bleekeri (Mollusca) (NC_002507), Lumbricus terrestris (Annelida) (NC_001673), Metridium senile (Cnidaria) (NC_000933), Mustelus manazo (Chordata) (NC_000890), Paracentrotus lividus (Echinodermata) (NC_001572), Phoronis architecta (Phoronida) (AY368231), Pichia canadensis (Fungi) (NC_001762), Platynereis dumerilii (Annelida) (NC_000931), Podospora anserina (Fungi) (NC_001329), Teredotalia transversa (Brachiopoda) (NC_003086), and Terebratulina retusa (Brachiopoda) (NC_000941).

Sequence Alignments and Phylogenetic Analysis. Inferring amino acid sequences of the 11 protein-encoding genes held in common among the 19 ingroup and three outgroup taxa were aligned by using CLUSTALW. Eight of these genes (cob, cox1, cox2, cox3, nad1, nad3, nad4, and nad5) were subjectively judged to be of reliable and unambiguous alignment. Pairwise alignments were done in slow mode with an open gap penalty of 10 and an extend gap penalty of 1 by using the BLOSUM similarity matrix. The multiple alignment used the same parameters as the pairwise plus a “delay divergent sequences” setting of 40%. Ambiguously aligned positions at the termini of each gene alignment were trimmed.

The phylogenetic tree was obtained by parsimony analysis by using a heuristic search with 10,000 random additions in PAUP*. The bootstrap values were calculated by using 500 bootstrap replicates, each with 10 random addition sequence replicates. Bremer support values were calculated by using TREEROT (11), and the statistical significance of this value found for the protostomes plus chaetognath clade was tested by using the Templeton (Wilcoxon’s signed ranks) test (12).

Results and Discussion

Genome Evolution. The P. gotoi mtDNA is remarkably small, consisting of only 11,423 bp. It contains only 14 genes rather than the 37 almost universally found in the mitochondrial genomes of other triploblastic animals. Missing are 21 of the 22 tRNA genes normally present and the protein-encoding genes atp6 and atp8. The only tRNA gene that can be identified by searching for sequences with tRNA-like potential secondary structures is trnM, which is also the case for the mtDNAs of the diploblasts Sarcophyton glaucum and Renilla kolikeri, and nearly the same as for M. senile mtDNA, which has genes for only two tRNAs, trnM and trnW (13-15). These three animals are in the phylum Cnidaria, wherein this gene loss must have occurred independently to their loss in the Chaetognatha. One might expect that the methionyl-tRNA would be the least “expendable” due to its unique role in initiating mitochondrial proteins with formylmethionine (16).

The long-PCR reactions that amplified the mtDNA in two overlapping fragments (Fig. 1) produced strong, singular bands during electrophoretic analysis, bolstering confidence that these products completely represent this circular molecule. However, it is not possible to rule out the alternative that the mtDNA contains a widely separated duplication of either cox1 or cox3 such that sequencing from the 3' end of one duplicate gene copy to the 5' end of the other would yield an apparently complete but truncated version of the mtDNA. The genes seeming to be absent, then, would actually be located between the two duplicated copies in an unstudied portion of the mtDNA. If that were the case, it would compel us to the remarkable coincidence that all of the 21 unfound tRNAs and one of the two unfound protein encoding genes (atp8) comprise the exact set of genes previously found to be missing in any animal mtDNA.

Genes encoding tRNAs capable of carrying the 19 aa other than methionine must be present somewhere in the P. gotoi cell. Because they do not seem to be encoded in this DNA molecule, there remain two possible alternatives. First, the genes could be encoded by a second DNA in the mitochondrion. Cases of second, or multiple, mtDNAs include linear plasmids in fungi (17), and diverse mtDNAs in the nematode Globoderia pallida, each with one or a few genes and much noncoding sequence (18). Neither the fungal plasmids nor any of the sequenced G. pallida mtDNAs contain tRNA genes. Secondly, the genes may be in the nucleus, with their products imported from the cytosol into the mitochondria. There are examples of tRNA import into the mitochondrion from diverse taxonomic groups. Apicomplexans and trypanosomatids, which encode no tRNAs in their mtDNAs (19), and plants, encoding only some tRNAs in their mtDNAs, import multiple tRNAs from the cytosol into the mitochondrion (20). In yeast, tRNA(K) is imported from the cytosol redundantly (i.e., there is a functional, mitochondrially encoded copy) (21), and in marsupials the same tRNA is also imported, although here the mitochondrial copy is a pseudogene (22).

The atp8 gene encodes subunit 8 of the F0 portion of the enzyme ATP synthase and, in yeast, this protein has been implicated in structure and assembly of the ATPase and for coupling proton transport from the F0 sector to the F1 sector where ATP synthesis occurs (reviewed in ref. 23). In metazoans, Atp8 is the smallest mitochondrially encoded protein, only ~50–65 aa long, and only about a half dozen of these amino acid residues are well conserved across animal mtDNAs. Among the major animal taxa that have been sampled, atp8 has been lost from the mtDNA independently in bivalve mollusks (e.g., ref. 24), scercernentean nematodes (25), and platyhelminths (26). Future mtDNA sequences may more clearly define the limits of this loss within these taxa and, as in the case of this chaetognath, reveal more taxa with this condition. Our attempts to find a copy by BLAST search of atp8 in the nuclear genome of Caenorhabditis...
elegans (one of the secernentean nematodes) have failed. It may be that the ATP synthase of some metazoans can function acceptably without this protein.

Atp6 is directly involved in ATP synthesis (27). Although atp6 is encoded in all other sequenced animal mtDNAs, it is not encoded in some nonmetazoan mtDNAs. This is the case in Chlamydomonas reinhardtii, where atp6 is a nuclear encoded gene (28). A barrier to movement of protein-encoded genes from the mtDNA to the nucleus may be the variation in the genetic code, especially the use of TGA to specify tryptophan in mitochondria instead of as a stop codon as in the "universal" genetic code. It is notable that C. reinhardtii's mitochondrial mRNAs are translated with the standard code, wherein TGA encodes a stop signal, so this would not apply in that case. However, for P. gotoi, TGA encodes tryptophan; in fact, TGA is the preferred tryptophan codon, 65 to 10, over TGG in the chaetognath mitochondrial protein-encoding genes. Thus, in order for atp6 to have transferred to the nucleus, either (i) the gene would not have contained any TGA codons when in the mtDNA, or (ii) the necessary A→G mutations to change all TGA to TGGs took place before the rise of a disabling mutation after transfer to the nucleus. As mentioned above for tRNAs, it is possible that the missing protein genes may be encoded by a second DNA in the mitochondrion rather than by the nuclear genome. Alternatively, in the case of either atp6 or atp8, the genes may be lost; perhaps their function has become dispensable or been subsumed by other proteins.

The organization of the genes in the P. gotoi mtDNA is distinctive among metazoan mtDNAs. Unlike any other studied animal mtDNA, the genes encoded in each strand are grouped together, as shown in Fig. 2. The protein genes in the clockwise orientation are separated by an unusually large number of nucleotides (Figs. 2 and 3) for an animal mtDNA. The protein genes in the opposite transcriptional orientation (shown counterclockwise in Fig. 2) are not separated at all. In fact, although we have annotated these genes such that they end on abbreviated stop codons (see below and Fig. 3), the reading frames of nad8, cob, and nad4 are open such that they could otherwise overlap the downstream gene. It has been shown that overlapping atp8/atp6 and nad4L/nad4 gene pairs form dicistronic mRNAs for some taxa (29), perhaps due to the smaller genes' (atp8 and nad4L) inability to bind solo to the ribosome (30). This result would not be true for cob, which is not particularly small for a mitochondrial encoded protein. These differences suggest that the transcripts of the two strands may be processed differently at one or more steps before translation, or that translation itself is different for the two sets of genes.

It has been proposed that the tRNA secondary structures
form in the primary transcript act as endonucleolytic targets, with their excision liberating both the tRNAs themselves as well as the intervening messages (6). Furthermore, it has been proposed that, in lieu of a tRNA, a tRNA-like structure in a primary transcript can serve this function (29). Such structures have been noted for some mtDNAs (31), but in many cases there are genes abutting without any obvious, intervening potential secondary structures, and that is the case with the P. gotoi mitochondrial genome. We have annotated some genes to end on a "T" in the first codon position, inferring that these are modified into complete TAA stop codons by polyadenylation of the mRNA after transcript cleavage. This process seems to be common in mitochondrial systems (6), and for P. gotoi there seems to be no reasonable alternative, because some genes would have to overlap extensively to reach the first in-frame stop codon. However, it is not obvious what might be acting to signal cleavage of the polycistron at these precise locations. The large number of these cases in the P. gotoi mtDNA serves to highlight this issue as one of the major unsolved problems of mitochondrial molecular biology.

**Phylogenetics.** Our analysis recovered a single most parsimonious tree of 16,267 steps (Fig. 4) that robustly supports the Chaetognatha, represented here by P. gotoi, as the sister group to the protostomes included in this study. This result corroborates that of Nielsen *et al.* (2) although their analysis placed the phoronids and brachiopods as deuterostomes, an untenable position given the results of independent molecular data sets (32, 33). The shortest alternative tree that does not include a monophyletic group of protostomes and chaetognaths requires 32 extra steps and is rejected by the Templeton test (12) in favor of the shortest tree uniting the groups (\(n = 116, z = -2.9711, P = 0.003\)). Little variation was noted in any of the branch lengths in this tree as reconstructed by parsimony criteria. There are some embryological features recently noted (34) that also support the placement of chaetognaths as or near protostomes rather than as deuterostomes. Chaetognaths have a tetrahedral four-cell embryo and an arrangement of blastomeres at the four-cell stage corresponding to future body axes, as is found in some protostomes (i.e., Mollusca, Annelida, and some crustaceans).

Thus, morphological features that have been inferred to place the Chaetognatha within the deuterostomes, e.g., secondary mouth formation, mesoderm derived directly from the archenteron (1), and having a trimeric coelom [although this has been recently reevaluated in chaetognaths (35)] have either evolved independently in these groups or are retained from a more ancient lineage (i.e., sympleisiomorphies). If the latter is the case, then one might infer that having the mouth arise from the blastopore during embryological development is a shared and derived condition (i.e., a synapomorphy) of the protostomes, with mouth formation by means of a secondary opening in the embryo being ancestral for triploblastic animals, and being retained by chaetognaths and deuterostomes, as has been suggested by Peterson and Eernisse (36).

In their study (36), chaetognaths were placed either as the sister taxon to the clade of molting animals referred to as the Ecdysozoa (9) (using morphological characters), or within the Ecdysozoa (using small nuclear rRNA data or a combination of these data). The Bremer support values for these hypotheses of relationship were low, however, with the small nuclear rRNA sequence results described as “unstable,” possibly due to extreme GC bias. The chaetognath sequences reported here show no unusual base compositional bias among animal mtDNA sequences. We specifically tested their hypothesis of a Chaetognatha plus Ecdysozoa clade against the most parsimonious tree found, which increased the tree length by 27 steps; thus, our data are inconsistent with this view. (Ecdysozoa is represented only by arthropods in this study.)
A recent Hox gene survey in the chaetognath Spadella cephaloptera (37) led the authors to speculate that chaetognaths diverged from the triploblast lineage before the deuterostome/protostome split. We would suggest, however, that such a study illustrates the difficulty of using paralogous genes to reconstruct the evolutionary history of species; indeed, there is no actual phylogenetic analysis included therein. That the few morphological characters that ally the chaetognaths to the protostomes (see ref. 2) and the various morphological features that are unique to the chaetognaths open the door to these speculations (see ref. 2) and the various morphological features that are unique to the chaetognaths open the door to these speculations is undeniable. However, only strongly supported phylogenetic analyses, rather than narrative ad hoc hypotheses, will enable us to understand the relationships of the animal phyla. We look forward to continued tests of the phylogenetic position of the chaetognaths, and new tests of other poorly studied animal phyla, as more complete animal mitochondrial genome sequences become available.

We thank Nori Satoh of Kyoto University for the gift of P. gotoi DNA and J. R. Macey for help with the Templeton test. This work was supported by National Science Foundation Grant DEB-9807100 and was partly performed under the auspices of the U.S. Department of Energy, Office of Biological and Environmental Research, by the University of California, Lawrence Berkeley National Laboratory, under Contract No. DE-AC03-76SF00098.