Dynamic microsatellites in transcribed regions of gastropod mitochondrial 16S rDNA

Markus Pfenninger and Monika Bugert

Abstract: Length variations of repetitive sequences in different AT-rich loop-coding regions of mitochondrial 16S rDNA in two gastropod species were discovered during intraspecific haplotype surveys. Examination of the discrete length variation of the basic repeat unit in a phylogenetic framework led to the conclusion of a microsatellite-like mutational dynamic. The observations suggest that the presence of a repetitive sequence structure alone is sufficient to trigger this dynamic.

Key words: microsatellites, 16S rRNA, mitochondrial genome, repetitive DNA, gastropods.

Résumé : Des variations quant à la longueur de séquences répétitives au sein de diverses régions codantes à boucle AT-riche du gène codant pour l'ADNr 16S mitochondrial ont été découvertes chez deux gastéropodes lors de la comparaison d'haplotypes interspécifiques. Le fait que cette variation de taille, de type discrète, impliquait la perte ou le gain de monomères de la séquence répétitive a été interprété dans un contexte phylogénétique. Il en ressort que cette variation s'apparente à une dynamique mutationnelle analogue à celle des microsatellites. Ces observations suggèrent que la présence d'une structure formée de séquences répétitives suffit, à elle seule, à enclencher ce genre de dynamique.

Mots clés : microsatellites, ARNr 16S, génome mitochondrial, ADN répétitif, gastéropodes.

[Traduit par la Rédaction]

Introduction

Microsatellites are repetitive DNA elements of 1–10 nucleotides repeat unit length (Litt and Luty 1989). They are ubiquitously found in eukaryotic nuclear genomes (Tautz and Renz 1984) but reports are comparatively rare for organelles such as mitochondria (Dobrowolski et al. 1998) and chloroplasts (Weising and Gardner 1999). However, variable monomers have been found in non transcribed regions (Berg et al. 1995).

We report here the presence of variable dinucleotide microsatellite motifs in two AT-rich loop-coding regions of mitochondrial 16S rRNA of two gastropod species, *Ancylus fluviatilis* (Basommatophora, Ancylidae) and *Candidula unifasciata* (Stylommatophora, Helicidae). The observations made in the course of phylogeographic studies of this species indicate that the presence of a repetitive structure alone is sufficient to trigger dynamic length mutations that are comparable to nuclear microsatellite loci.

M. Pfenninger[1] and M. Bugert. Abteilung Ökologie & Evolution, BioCampus, Siesmayerstraße, J.W.Goethe-Universität Frankfurt am Main, 60054 Frankfurt am Main, Germany.

[1]Corresponding author
  (e-mail: Pfenninger@zoology.uni-frankfurt.de).

Material and methods

In *Ancylus fluviatilis*, a freshwater limpet, 93 individuals from 34 populations throughout western Europe, North Africa, and the Canary Islands were surveyed for mitochondrial haplotype variation in a 438–446-bp fragment of the 16S rDNA. In *Candidula unifasciata*, a land snail, 170 individuals from 25 populations throughout the species range in Europe were screened for sequence variation in the same 16S rDNA fragment.

Individuals were crushed with their shells in 10% (w/v) laundry detergent solution for storage at room temperature and tissue digestion following the protocol of Bahl and Pfenninger (1997). Samples were shaken for 24 h at 37°C in the laboratory prior to the phenol–chloroform extraction of total genomic DNA following standard protocol (Sambrook et al. 1989).

The 16S target DNA was amplified by PCR with primers of the sequence 16S1 5′-CGC AGT ACT CTG ACT GTG C-3′ and 16S2 5′-GTC CGG TTT GAA CTC AGA TC-3′. Amplification was performed with Boehringer Taq polymerase in 12.5 µL total reaction volume with standard reaction conditions. Samples were amplified for 10 cycles (92°C for 50 s, 44°C for 50 s, and 72°C for 40 s) and 36 cycles (92°C for 30 s, 48°C for 40 s, and 72°C for 40 s) after initial incubation of 90°C for 2.5 min. A fraction of the PCR product was electrophoresed on a 1.4% agarose gel to verify the presence of a single band of the expected size. The PCR products were then purified using the Promega Wizard®PCR Preps DNA Purification System according to the recommendations of the provider, prior to the sequencing reaction. Both strands of the purified amplification products were cycle-sequenced with the Perkin Elmer Taq DyeDeoxy™ Terminator Cycle Sequencing Kit after the protocol of the supplier and read automatically on the ABI Prism 377® sequencing device (Perkin Elmer). Sequences were deposited in GenBank (Accession Nos.). Sequences were aligned manually after an initial alignment with the Clustal option available in the

Table 1. Sequence variation in AT-rich 16S rDNA loop-coding region in *Ancylus fluviatilis*.

Loop sequence	Alignment	Loop length
Portugal 1-I	tttaga**TTATATATATA**———AATTATT-tctttt	18
Portugal 1-II	tttaga**TTATATATATATATATA**AATTATT-tctttt	24
Portugal 1-III	tttaga**TTATATATATATA**——AATTATT-tctttt	20
Portugal 1-IV	tttaga**TTATATATA**———AATTATT-tctttt	16
Slovenia 1-I	tttagaATATATTTATAC——-AATTATT-tctttt	19
Poland 1-I	tttagaATATATCTATAC——-AATTATT-tctttt	19
Portugal 2-I	tttagaTAAATATTATA———-AATAATT-tctttt	17
Portugal 2-II	tttagaTATATTATA———AACTATT-tctttt	16
Portugal 2-III	tttagaCAGTATTGTA———-AATTATTTtctttt	18
Portugal 2-IV	tttagaTAGTATTGTA———-AATTATTTtctttt	18
Germany 1-I	tttagaTAAATATATCGTA——-AATTATT-tctttt	19
Gomera 1-I	tttagaATATATTATA———-AATTATT-tctttt	17
Madeira 1-I	tttagaATATATTGTA———-AGTTATT-tctttt	17
Ethiopia 1-I	tttagaATATATAATA———-AATTATT-tctttt	17
	******..............*...***.******	

Note: Unambiguous alignment is possible only for the conserved flanking sequence that is given in lower case letters and few positions at the 3′ end of the loop. The unpaired loop sequence is in upper case letters. $[TA]_x$ microsatellite motifs are in bold style. Loop length is the length of the loop, i.e., the non-paired regions. Conserved positions in the alignment are indicated in the last row with asterisks, variable positions with a point. Haplotype designations indicate the sampling location.

computer program SEQUENCENAVIGATOR (Perkin Elmer, Applied Biosystems).

The 95% probability of a parsimonious linkage between haplotyes differing at a certain number of sites was calculated using the PARSPROBE V. 1.1 program provided by David Posada. A parsimonious haplotype network was then constructed using the algorithm given in Templeton et al. (1992) with version 1.0 alpha of the TCS program written by M. Clement, D. Posada, and K.A. Crandall.

Over the length of the whole fragment, 42 haplotypes were identified in *Ancylus fluviatilis* by directly sequencing both strands of the PCR products. An AT-rich region showed extensive sequence variation. Comparison with other molluscan 16S secondary structures and their highly conserved positions (Lydeard et al. 2000) showed that this region forms a loop when transcribed into RNA. The loop corresponds to the region between helices G3 and G3′ after the nomenclature used for LSU RNA secondary structure by De Rijk et al. (2000). Considering this loop-coding region alone, 14 different sequences were identified (Table 1). Homology of base positions can not be determined unambiguously within this region for all haplotypes. Except for an AT bias in nucleotide composition, base changes and indels appear to be random. The length varied from 16–19 bp, except in a clade of closely related haplotypes, originating from the same Portuguese population. These haplotypes contained four alleles of a perfect TA dinucleotide microsatellite motif varying in length between 16 and 24 bp. The most parsimonious cladogram was reconstructed following the algorithm given in Templeton et al. (1992). To obtain sufficient resolution, the phylogenetic reconstruction was based on the sequence differences of the entire sequence of the fragment, including the region in question. Figure 1A shows the phylogenetic relations of these haplotypes. The topology of such a cladogram can be used to infer the direction of haplotype evolution (Castelloe and Templeton 1994). Haplotypes at the tip of a cladogram have a higher probability of being evolutionarily younger than haplotypes situated at interior nodes. We can therefore infer that the interior haplotype at the root of this clade contains a $[TA]_5$ repeat (Fig. 1A). At least two independent evolutionary lines with variation in loop length could be observed. In one of them, a deletion of two bases resulted in a $[TA]_4$ repeat. Extensions to $[TA]_6$ and $[TA]_8$ alleles were also found in a different lineage. Even if the assumption of an interior

haplotype bearing a $[TA]_5$ repeat being the root of this clade is not justified, logically only one of the tip haplotypes can be the root of the cladogram, so that we have in any case at least two length variations having arisen independently.

In *Candidula unifasciata*, a different loop-coding region between helices G9 and G9′, further to the 3′ end of the fragment, harboured an $[AT]_3$ motif in a large majority of the specimens. However, in eight individuals sampled from neighbouring French populations, the homologous region was eight base pairs long. Seven sequences contained an $[AT]_4$ motif and one contained a $GT[AT]_3$ sequence. A cladogram for this species was derived using the same principles as described above. Examination of the cladogram shown in Fig. 1B revealed that haplotypes bearing the $[AT]_3$ motif are ancestral to the $[AT]_4$-carrying haplotypes, which in turn gave rise to the $GT[AT]_3$-containing sequence.

Results and discussion

Taking all observations into consideration, we put forward the hypothesis that the length variations of AT-rich loops in both species are due to one or more insertions of the basic dinucleotide repeat motif by a mechanism described typically for nuclear microsatellite loci, and not by sequential insertions or deletions of single bases. Support for this view arises from the observation that only haplotypes with length variations in multiples of the basic dinucleotide repeat motif were found. In *A. fluviatilis*, sequential indels of individual bases would require a minimum of eight independent mutation events of the right bases in the correct order. This explanation seems to be a much less probable explanation for the observed length variation than indels of microsatellite repeat units, which requires five independent mutational events at maximum. Assuming a microsatellite-like length dynamic of these sequences, it is not possible to deduce whether insertions of one and consecutively two repeat units took place to create the haplotypes with the $[TA]_6$ and $[TA]_8$ repeats in a single evolutionary line, or whether two independent insertions of one and three repeat units occurred, respectively, because of a lack of mutations outside the microsatellite

sequence. Both alternatives would require two mutational events.

Our observations are similar to those described by Messier et al. (1996) who found two independently arising and expanding microsatellite motifs in nuclear η-globin pseudogenes of primates. While the repeat extensions at this primate nuclear locus took place in a phylogenetic time frame of several million years, the mutational events in mtDNA reported here occurred on a population level, where the timescale to the most recent common ancestor of a closely related clade is usually magnitudes shorter.

From the proposed extension mechanisms for microsatellite loci, unequal crossing-over during recombination (Jarman and Wells 1989) is unlikely here, since mitochondrial DNA is usually non-recombining. No evidence for mitochondrial recombination was found in our data. Therefore, slipped-strand mispairing (Levinson and Gutman 1987) remains in this case the most plausible of the proposed mechanisms for length variation of the microsatellite motif. Length variations of nuclear structural RNAs due to slippage of short repetitive elements have been described before (Hancock and Dover 1990). However, the length variations described in these two gastropod species seem not to fit into their concept of compensatory slippage, in which mutations at one point of the sequence are compensated by slippage events at another point to maintain the secondary structure mainly of stems. No sequence regions coding for stems were affected in our case and there is no indication of interactions with positions outside of the sequenced fragment (Lydeard et al. 2000).

Our findings suggest that the ancestral $[TA]_5$-repeat in *A. fluviatilis* initially arose by random sequence variation of an AT-rich loop-coding motif. Once the microsatellite motif was established, it was inherently predisposed to length variation. From comparison with the highly variable length of this loop in other molluscan species (Lydeard et al. 2000) it seems probable that the length variation has no functional significance and is selectively neutral or almost neutral. In *C. unifasciata*, the sequence region containing the $[AT]_3$ repeat appears to be less variable in length, maybe because a shorter repeat is less prone to length variation (Pupko and Graur 1999).

The extension of randomly emerged microsatellite motifs in loop-coding regions of structural RNAs could be a possible general mechanism for their length variation. However, perfect microsatellite motifs of different lengths have not yet been reported in surveys of intra specific sequence variation of mitochondrial rRNAs. This may change in the future due to the expanding possibility of sequencing the large number of individuals that seems to be necessary to make apparent such phenomena. Another explanation might be the low selective pressure on unpaired regions of structural RNAs. The point mutation rate and length variation is relatively high in loops (Clark et al. 1984), thus possibly destroying perfect repeats in relatively short evolutionary time, as can be seen in *C. unifasciata*. In a data set where the $[AT]_4$ intermediate haplotype would be missing, either due to inadequate sampling or drift, we would have deduced a G and T insertion instead of a single microsatellite extension mutation. Our findings thus stress the notion that appropriate precautions concerning alignment and interpretation should be applied in

Fig. 1. (A) Most parsimonious cladogram of haplotypes derived from 16S rDNA sequences of *Ancylus fluviatilis*. (B) Most parsimonious cladogram of haplotypes derived from 16S rDNA sequences of *Candidula unifasciata*. Haplotypes are given as circles. Diameter of circles is proportional to relative abundance of haplotypes. Within circles, microsatellite repeat sequence and length contained in respective haplotype is specified. Lines between circles designate the most parsimonious connections between haplotypes without the assumption of a microsatellite-like mutational dynamic. Dashes show the number of base differences between haplotypes. Dashes indicating mutations that change the microsatellite repeat sequence or length are bold. Arrows pointing away from haplotypes indicate connections to other haplotypes not shown here.
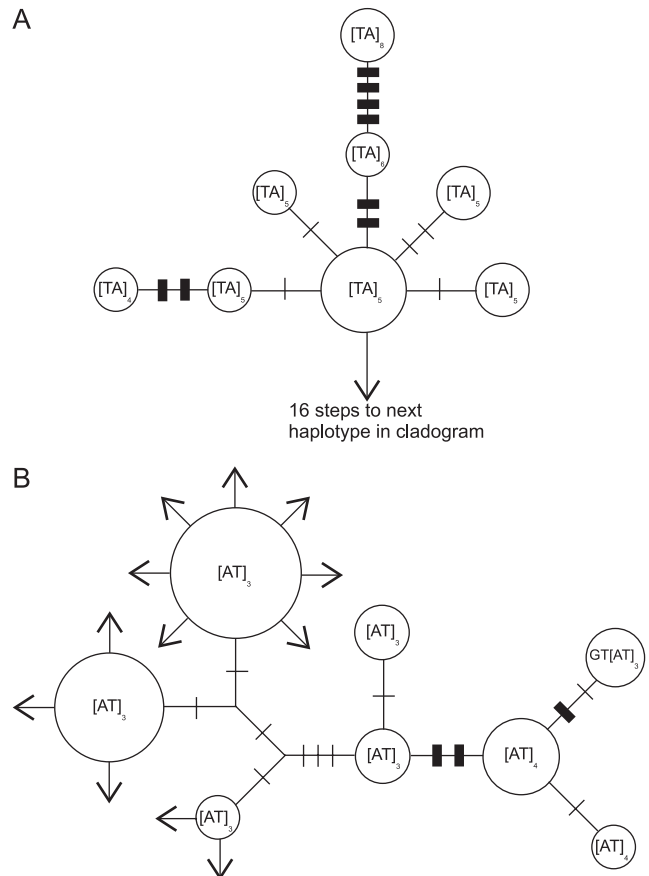
drawing evolutionary inferences from sequence differences in structural RNA molecules (Hancock and Dover 1990; Hancock and Vogler 1998).

Acknowledgements

References

Berg, T., Moum, T., and Johansen, S. 1995. Variable numbers of simple tandem repeats make birds of the order Ciconiiformes heteroplasmic in their mitochondrial genomes. Curr. Genet. **27**: 257–262.

Castelloe, J., and Templeton, A.R. 1994. Root probabilities for intraspecific gene trees und neutral coalescent theory. Mol. Phylogenet. Evol. **3**: 102–113.

Clark, C.G., Tague, B.W., Ware, V.C., and Gerbi, S.A. 1984. *Xenopus laevis* 28S ribosomal RNA: A secondary structure model and its evolutionary and functional implications. Nucleic Acids Res. **12**: 6197–6220.

De Rijk, P., Wuyts, J., van de Peer, Y., Winkelmans, T., and de Wachter, R. 2000. The European Large Subunit Ribosomal RNA database. Nucleic Acids Res. **28**: 177–178.

Dobrowolski, M.P., Tommerup, I.C., and O'Brien, P.A. 1998. Microsatellites in mitochondrial genome of *Phytophora cinnamomi* failed to provide highly polymorphic markers for population genetics. FEMS Microbiol. Lett. **163**: 243–248.

Hancock, M.J., and Dover, G.A. 1990. 'Compensatory slippage' in the evolution of ribosomal RNA genes. Nucleic Acids Res. **18**: 5949–5954.

Hancock, M.J., and Vogler, A.P. 1998. Modelling the secondary structures of slippage-prone hypervariable RNA regions: The example of the tiger beetle 18S rRNA variable region V4. Nucleic Acids Res. **26**: 1689–1699.

Jarman, A.P., and Wells, R.A. 1989. Hypervariable minisatellites: Recombinators or innocent bystanders? Trends Genet. **5**: 367–371.

Levinson, G., and Gutman, G.A. 1991. Slipped-strand mispairing: A major mechanism for DNA sequence evolution. Mol. Evol. Biol. **4**: 203–221.

Litt, M., and Luty, J.A. 1989. A hypervariable microsatellite revealed by *in vitro* amplification of a dinucleotide repeat within the cardiac muscle actin gene. Am. J. Hum. Genet. **44**: 397–401.

Lydeard, C., Holznagel, W.E., Schnare, M.N., and Gutell, R.R. 2000. Evolutionary Patterns and Phylogenetic Content of the Molluscan Mitochondrial LSU rDNA Sequences and rRNA Secondary Structure. Mol. Phylogenet. Evol. **15**: 83–102.

Messier, W., Li, S.H., and Stewart, C.B. 1996. The Birth of Microsatellites. Nature, **381**: 483.

Pupko, T., and Graur, D. 1999. Evolution of microsatellites in the yeast *Saccharomyces cerevisiae*: Role of length and number of repeated units. J. Mol. Evol. **48**: 313–316.

Tautz, D., and Renz, M. 1984. Simple sequences are ubiquitous repetitive components of eukaryotic genomes. Nucleic Acids Res. **12**: 4127–4138.

Templeton, A.R., Crandall, K.A., and Sing, C.F. 1992. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. Genetics, **132**: 619–633.

Weising, K., and Gardner, R.C. 1999. A set of conserved PCR primers for the analysis of simple sequence repeat polymorphisms in chloroplast genomes of dicotyledonous angiosperms. Genome, **42**: 9–19.