Points of View

Syst. Biol. 47(4):696–701, 1998

Conflicting Phylogenetic Patterns Caused by Molecular Mechanisms in Mitochondrial DNA Sequences

RICHARD E. BROUGHTON,[1] GAVIN J. P. NAYLOR,[2] AND THOMAS E. DOWLING

Department of Biology, Arizona State University, Tempe, Arizona 85287-1501, USA

Mutations in DNA are the ultimate source of alternative character states used in molecular systematic studies. If the mutation process is stochastic (Zuckerkandl et al., 1971), homoplasy should tend to be randomly distributed among taxonomic units, whereas variation reflecting historical relationships should be additive (Miyamoto and Cracraft, 1991). Thus, where homoplasy is evident, one should still be able to extract hierarchical patterns from the noise (Farris, 1983). Even with transition or compositional bias, the directionality and phylogenetic distribution of mutations may remain largely random. However, the existence of deterministic mutation processes could cause serious complications in phylogenetic analyses by skewing the distribution of homoplastic characters to suggest spurious hypotheses of relationships. Recent reports have revealed nonrandom patterns of mutation in eukaryotic nuclear (Selker, 1990; Kricker et al., 1992) and bacteriophage (Cunningham et al., 1997) genomes. Here we analyze a peculiar distribution of nucleotide variation in mitochondrial DNA (mtDNA) of a cyprinid fish and suggest that nonrandom homoplasy in molecular data may be more widespread than is currently recognized.

Sequence duplications are a common feature of animal mtDNAs (Rand, 1993), particularly in the control region. Duplicated sequences often form tandem arrays that may vary extensively in copy number among individuals or taxa. Unlike nuclear DNA, copy number changes in mtDNA must be the result of intramolecular processes (e.g., Buroker et al., 1990) because recombination in mtDNA is extremely rare (Moritz et al., 1987). If phylogenetic relationships of individual repeats are examined, two different patterns may be predicted. In one, copy addition events are rare, and homology is maintained among copies in specific positions in the arrays (Fig. 1a). The result is that copies in the same position will be more closely related among arrays than are copies in different positions within individual arrays. In the alternative (Fig. 1b), prodigious addition and deletion of copies will tend to "homogenize" nucleotide variation within arrays, resulting in closer relationships among copies within arrays than between them. Differences in the resulting patterns between these hypotheses are similar to the distinction between paralogous and orthologous duplicated nuclear sequences (Fitch, 1970; Sanderson and Doyle, 1992); however, nucleotide similarity among copies in mtDNA is attributed solely to common ancestry rather than intermolecular exchange.

We have previously reported variation in repeated mtDNA sequences in the North American cyprinid fish *Cyprinella spiloptera* (Broughton and Dowling, 1994, 1997). Restriction site analysis revealed that, among 48 individuals, all possessed between two and four copies of a 260-base pair (bp) tandem repeat located in the control re-

[1]Present address (and address for correspondence): Section of Ecology and Systematics, Corson Hall, Cornell University, Ithaca, New York 14853-2701, USA; E-mail: reb17@cornell.edu

[2]Present address: Department of Zoology and Genetics, Iowa State University, Ames, Iowa 50011, USA.

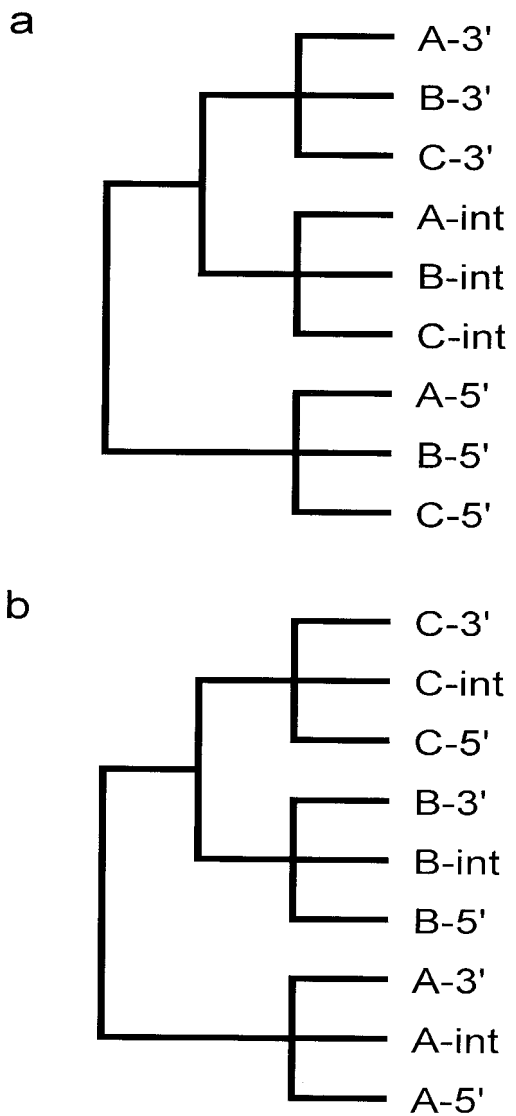

FIGURE 1. Alternative phylogenetic relationships for tandem repeats predicted under different models of repeat evolution. (a) Repeat relationships under a low rate of copy addition/deletion. (b) Repeat relationships under homogenization resulting from a rapid rate of copy addition/deletion. In each panel, letters indicate mitochondrial genomes from different individuals, and specific copies of the repeat are designated based on position in the array: 5′, internal (int), 3′.

gion between the origin of replication and $tRNA^{phe}$ (Broughton and Dowling, 1994). Phylogenetic analysis of nucleotide sequences among repeats from 19 individuals revealed extreme conflict among characters. Variation was distributed such that phylogenetic support was nearly equal for each of the two evolutionary hypotheses outlined above (Broughton and Dowling, 1997). Briefly, one set of characters included an 8-bp deletion and an adjacent A to G transition (collectively referred to hereafter as G/deletion); this was present in all individuals but in only some of the copies (see Fig. 2). In each individual, the G/deletion was present in all copies except those in the 5′ position (5′ = proximal to the origin of replication). Copies in the 5′ position, which did not possess the G/deletion, aligned perfectly with homologous sequences from related minnow species without duplications. Nucleotide variants other than the G/deletion (and characters 1, 2, 6, 7, 8) were less frequent but were present in all copies in the individuals where they occurred (Fig. 2). The frequency and distribution of the two sets of characters resulted in almost no phylogenetic resolution (Broughton and Dowling, 1997).

These nucleotide data are remarkable for two reasons. First, conflicting sets of characters are distributed in highly structured patterns and not at random, as would have been predicted under a stochastic model of mutation. Second, the two conflicting groups of characters each provide clear and consistent support for two essentially opposite hypotheses for the evolution of the repeats. In the absence of recombination, there can be only one history of the repeats and only one phylogeny; thus, one set of characters must have arisen through parallel or convergent mutation. Here we describe new analyses of the nucleotide data aimed at identifying which set of characters reflects the phylogeny of the repeats, and accordingly, the set that does not (i.e., the homoplastic characters). Our approach employed comparison of different measures of character quality, including successive approximations (Farris, 1969), compatibility analysis (Meacham and Estabrook, 1985), and the optimization method of Goloboff (1993).

Successive approximations were based on an initial set of 233 most-parsimonious trees, where all characters, including the

```
                                    111111112222
                               1111111111666577888881244
                           12678123456789012967345780508

Raisin 2-5'                TTGCGAAACCCCCTCCCAAAGGAAAAATA
Raisin 2-3'                .....G--------...............
Tiffin 2-5'                ................G...A...G...
Tiffin 2-i                 .....G--------...G...A...G...
Tiffin 2-3'                .....G--------...G...A...G...
Turtle 2-5'                ....................G.A...G...
Turtle 2-i5'               .....G--------.....G.A...G...
Turtle 2-i3'               .....G--------.....G.A...G...
Turtle 2-3'                .....G--------.....G.A...G...
Big 5-5'                   .....................A...GG..
Big 5-i                    ACCGAG--------.......A...GG..
Big 5-3'                   ACCGAG--------.......A...GG..
Big 3-5'                   ..................G..A...G...
Big 3-i                    ..-..G--------....G..A...G...
Big 3-3'                   .....G--------....G..A...G...
Gasconade 1-5'             ..................G.TA...G...
Gasconade 1-i              .....G--------....G.TA...G...
Gasconade 1-3'             .....G--------....G.TA...G...
Susquehanna 2-5'           .....................A...G..G
Susquehanna 2-3'           .....G--------.......A...G..G
French Broad 1-5'          ....................TA.....AG
French Broad 1-i           .....G--------......TA.....AG
French Broad 1-3'          .....G--------......TA.....AG
Stony 6-5'                 ..............ATT...TATTT...G
Stony 6-i                  .....G--------ATT...TATTT...G
Stony 6-3'                 .....G--------ATT...TATTT...G
C. venusta                 ..................T.CA.......
```

FIGURE 2. Phylogenetically informative nucleotides in copies of *C. spiloptera* tandem repeats. Individual fish are identified by river of origin and a numeral distinguishing those from the same collection. Copies within individuals are indicated by physical position (5′, i = internal, 3′). Only individuals with phylogenetically informative variation are included, but all copies from those individuals are present. Numbering refers to position in complete 271-bp alignment (GenBank accessions U73306-U73315), dot = same nucleotide as top sequence, dash = alignment gap.

8-bp deletion, were weighted equally. Although weighting the G/deletion more heavily (e.g., as 8-bp) may alter results, there is no objective criterion for assigning higher weight. A posteriori weights were assigned to each character based on its maximum consistency index (CI; Kluge and Farris, 1969) over all trees and were used in subsequent parsimony searches. Heuristic tree searches were implemented with PAUP 3.1.1 (Swofford, 1993) and using random taxon addition, TBR branch swapping, and ACCTRAN optimization for 10 replicates. The homologous sequence from a related minnow, *Cyprinella venusta*, which lacks the duplication, was used as an outgroup. This procedure was repeated until character weights (and tree topologies) were stable for two consecutive iterations. Final weights (Table 1) were lowest for the G/deletion, which suggests it is more likely to be homoplastic than the other characters. The strict consensus topology of the 18 shortest trees based on these weights (Fig. 3a) requires the G/deletion to be independently gained or lost at least nine times and generally clusters all copies from each individual together.

Compatibility analysis tests differences between the expected and observed number of compatibilities for each character. Observed values are the number of characters in the data matrix with which a particular character is not in conflict. Expected values are the number of compatibilities expected by chance, as estimated by permutation. Characters with compatibilities significantly greater than expected by chance appear to be superior indicators of phylogenetic structure (Penny and Hendy, 1986). Observed and expected character compatibilities were computed with the program Comprob v. 1.1 by C. Meacham. Results (Table 1) indicated that the distributions of all characters except the G/deletion and nucleotide 184 were significantly different from random at $P = 0.05$. Because even characters that appear to be randomly distributed may be informative at some level, complete removal from analysis is not usually desirable. Therefore, we employed the character weighting index (Table 1) described by Penny and Hendy (1986). Weights were calculated as 1 minus the ratio for observed incompatibility/expected incompatibility, with incompatibilities defined as the total possible number of compatible characters (21) minus the respective observed or expected compatibility values. A parsimony search utilizing these weights resulted in the same set of 18 shortest trees as was obtained with successive weighting (Fig. 3a).

Goloboff's (1993) method searches for trees that imply the highest "weights" for all characters. Character quality is assessed by fitting a function that relates the fit of a character on a tree to its homoplasy: $k + 1/(s + k + 1 - m)$, where k is a constant describing the concavity of the fit/homoplasy relationship, s is the number of steps required for a character to fit a particular tree, and m is the minimum possible number of steps for that character on any tree. The Goloboff rou-

TABLE 1. Results of successive weighting with CI, compatibility analysis, and Goloboff optimization (weights converted to scale of 1,000).

Nucleotide position[a]	CI weights[b]	Compatibility *P* values[c]	Compatibility weights[d]	G fit[e]
1	1,000	0.0299	1,000	1,000
2	1,000	0.0280	1,000	1,000
6	1,000	0.0265	1,000	1,000
7	1,000	0.0277	1,000	1,000
8	1,000	0.0289	1,000	1,000
11	100	0.6515	0	250
12	100	0.6497	0	250
60	1,000	0.0174	757	1,000
61	1,000	0.0180	756	1,000
62	1,000	0.0201	756	1,000
159	1,000	0.0318	759	1,000
176	1,000	0.0021	760	1,000
177	1,000	0.0067	801	1,000
183	500	0.0022	666	600
184	1,000	0.2305	638	1,000
185	1,000	0.0195	758	1,000
187	1,000	0.0194	756	1,000
188	1,000	0.0201	756	1,000
210	1,000	0.0003	207	500
225	1,000	0.0217	757	1,000
240	1,000	0.0195	758	1,000
248	333	0.0006	721	500

[a]From Figure 2; character 12 is the 8-bp deletion.
[b]Stabilized weights after three iterations.
[c]Probability of observed compatibilities due to chance.
[d]Weights = 1 – (observed incompatibility /expected incompatibility); negative values were changed to zero.
[e]Character fit based on Goloboff optimization.

tine, implemented in PAUP* (D. Swofford, pers. comm.) with a *k* value of 2, recovered 33 best trees. The strict consensus of these (Fig. 3b) was less well resolved than in the other methods, but again, copies were grouped by individual rather than by position in the array. The Goloboff "fit" values for the characters on the consensus tree (Table 1) also indicate the G/deletion to be a poor indicator of phylogenetic pattern.

All three approaches for assessing character reliability consistently downweighted the G/deletion relative to the other characters. This implies that the G/deletion is due to multiple independent (parallel) origins, a result that is counter to the notion that insertion and deletion changes tend to be conservative. Parallel character-state changes are not particularly unusual, but parallel changes in such a nonrandom pattern imply a common mechanism. The resulting trees each indicate that copies within individuals are more closely related to each other than to copies in other individuals. This suggests either that copies within individuals are subject to a homogenization process, or that the duplications have arisen independently in each lineage (Broughton and Dowling, 1997). Both of these hypotheses are difficult to reconcile with the distribution of the G/deletion. In the homogenization scenario, one must explain how nucleotide differences are distributed to all copies without the concomitant homogenization of the G/deletion. Alternatively, the multiple origins model requires parallel formation of the G/deletion in some, but not all, copies in each lineage.

It is possible that the G/deletion is generated as a byproduct of the mechanism that generates new copies of the repeat. Variation of copy number in mtDNA repeat ar-

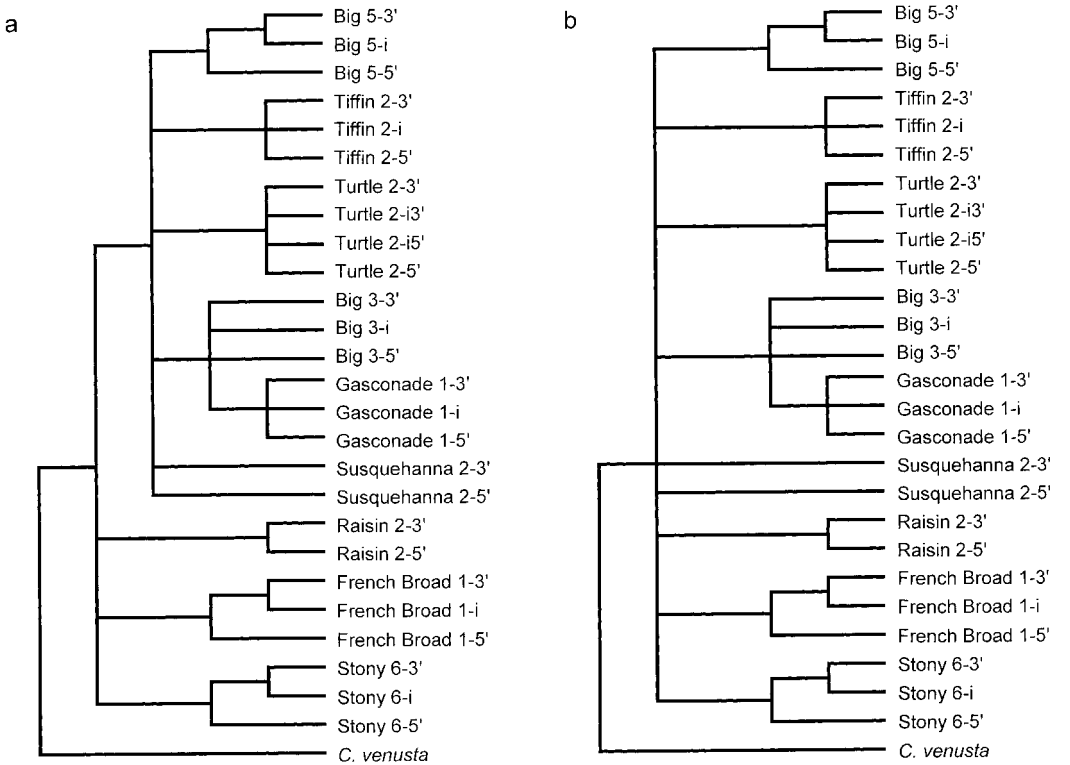

FIGURE 3. Phylogenetic trees generated from weighted characters. (a) Strict consensus of 18 trees recovered with CI successive approximations or compatibility weighting. (b) Strict consensus of 33 trees recovered with Goloboff optimization.

rays is thought to result from slip-strand mispairing during replication (Buroker et al., 1990; Broughton and Dowling, 1994), and the strong association of secondary structures with repeat arrays (Stanton et al., 1994) suggests they play a role in this process. Details of the molecular mechanism(s) are not clear but the G/deletion could conceivably result from replication errors involving secondary structures on the DNA strand that serves as the template for copy duplication. Although it is possible that recombination (or gene conversion) could account for the patterns of variation in these repeated sequences, this hypothesis seems doubtful, as the only evidence for recombination in vertebrate mtDNA is that implied by differences in gene order among higher taxonomic groups (Brown, 1985). Moreover, to obtain the present results via recombination would require (and result in) extensive length and nucleotide heteroplasmy which has not been observed in *C. spiloptera.*

The unusual character distribution we have described appears to be the result of a deterministic mechanism of DNA mutation that has not been previously recognized. Although a minority of molecular systematic studies involve repeated sequences, our results raise concerns about deterministic mutation mechanisms as more general phenomena. We suggest that such parallel mutations could occur in any system where specific secondary structures have the potential to interfere with DNA replication. Thus parallelisms resulting from specific mutational mechanisms might be more common than is currently realized, and this emphasizes

the necessity of understanding the underlying mechanisms of evolutionary change for characters used in molecular phylogenetic analysis.

Acknowledgments

We thank Joe Bielawski, John Gold, Rick Harrison, Rosemary Knapp, Doug McElroy, Tom Turner, and Rob Wood for helpful discussions and/or comments on the manuscript. Support was provided by the Department of Zoology and Graduate Student Association at Arizona State University, Sigma Xi, and the National Science Foundation (DEB-9220683).

References

Broughton, R. E., and T. E. Dowling. 1994. Length variation in mitochondrial DNA of the minnow *Cyprinella spiloptera*. Genetics 138:179–190.

Broughton, R. E., and T. E. Dowling. 1997. Evolutionary dynamics of tandem repeats in the mitochondrial DNA control region of the minnow *Cyprinella spiloptera*. Mol. Biol. Evol. 14:1187–1196.

Brown, W. M. 1985. The mitochondrial genome of animals. Pages 95–130 *in* Molecular evolutionary genetics (R. Macintyre, ed.). Plenum, New York.

Buroker, N. E., J. R. Brown, T. A. Gilbert, P. J. O'Hara, A. T. Beckenbach, W. K. Thomas, and M. J. Smith. 1990. Length heteroplasmy of sturgeon mitochondrial DNA; an illegitimate elongation model. Genetics 124:157–163.

Cunningham, C. W., K. Jeng, J. Husti, M. Badgett, I. J. Molineux, D. M. Hillis, and J. J. Bull. 1997. Parallel molecular evolution of deletions and nonsense mutation in bacteriophage T7. Mol. Biol. Evol. 14:113–116.

Farris, J. S. 1969. A successive approximations approach to character weighting. Syst. Zool. 18:374–385.

Farris, J. S. 1983. The logical basis of phylogenetic analysis. Pages 7–36 *in* Advances in cladistics, vol. 2 (N. Platnick and V. Funk, eds.). Columbia Univ. Press, New York.

Fitch, W. M. 1970. Distinguishing homologous from analogous proteins. Syst. Zool. 19:99–113.

Goloboff, P. A. 1993. Estimating character weights during tree search. Cladistics 9:83–91.

Kluge, A. G., and J. S. Farris. 1969. Quantitative phyletics and the evolution of anurans. Syst. Zool. 18:1–32.

Kricker, M. C., J. W. Drake, and M. Radman. 1992. Duplication-targeted DNA methylation and mutagenesis in the evolution of eukaryotic chromosomes. Proc. Natl. Acad. Sci. USA 89:1075–1079.

Meacham, C. A., and G. F. Estabrook. 1985. Compatibility methods in systematics. Annu. Rev. Ecol. Syst. 16:431–446.

Miyamoto, M. M., and J. Cracraft. 1991. Phylogenetic inference, DNA sequence analysis, and the future of molecular systematics. Pages 3–17 *in* Phylogenetic analysis of DNA sequences (M. Miyamoto and J. Cracraft, eds.). Oxford Univ. Press, New York.

Moritz, C., T. E. Dowling, and W. M. Brown. 1987. Evolution of animal mitochondrial DNA: Relevance for population biology and systematics. Annu. Rev. Ecol. Syst. 18:269–292.

Penny, D., and M. Hendy. 1986. Estimating the reliability of evolutionary trees. Mol. Biol. Evol. 3:403–417.

Rand, D. M. 1993. Endotherms, ectotherms, and mitochondrial genome-size variation. J. Mol. Evol. 37:281–295.

Sanderson, M. J., and J. J. Doyle. 1992. Reconstruction of organismal and gene phylogenies from data on multigene families: Concerted evolution, homoplasy, and confidence. Syst. Biol. 41:4–17.

Selker, E. U. 1990. Premeiotic instability of repeated sequences in *Neurospora crassa*. Annu. Rev. Genet. 24:570–613.

Stanton, D. J., L. L. Dahler, C. Moritz, and W. M. Brown. 1994. Sequences with the potential to form stem-and-loop structures are associated with coding-region duplications in animal mitochondrial DNA. Genetics 137:233–241.

Swofford, D. L. 1993. PAUP: Phylogenetic analysis using parsimony, version 3.1.1. Illinois Natural History Survey, Champaign.

Zuckerkandl, E., J. Derancourt, and H. Vogel. 1971. Mutational trends and random processes in the evolution of informational macromolecules. J. Mol. Biol. 59:473–490.

Received 11 November 1997; accepted 18 March 1998
Associate Editor: D. Cannatella