Comparative Evolution of the Mitochondrial Cytochrome b Gene and Nuclear β-Fibrinogen Intron 7 in Woodpeckers

Thomas M. Prychitko and William S. Moore
Department of Biological Sciences, Wayne State University

Most molecular phylogenetic studies of vertebrates have been based on DNA sequences of mitochondrial-encoded genes. MitDNA evolves rapidly and is thus particularly useful for resolving relationships among recently evolved groups. However, it has the disadvantage that all of the mitochondrial genes are inherited as a single linkage group so that only one independent gene tree can be inferred regardless of the number of genes sequenced. Introns of nuclear genes are attractive candidates for independent sources of rapidly evolving DNA: they are pervasive, most of their nucleotides appear to be unconstrained by selection, and PCR primers can be designed for sequences in adjacent exons where nucleotide sequences are conserved. We sequenced intron 7 of the β-fibrinogen gene (β-fibint7) for a diversity of woodpeckers and compared the phylogenetic signal and nucleotide substitution properties of this DNA sequence with that of mitochondrially-encoded cytochrome b (cyt b) from a previous study. A few indels (insertions and deletions) were found in the β-fibint7 sequences, but alignment was not difficult, and the indels were phylogenetically informative. The β-fibint7 and cyt b gene trees were nearly identical to each other but differed in significant ways from the traditional woodpecker classification. Cyt b evolves 2.8 times as fast as β-fibint7 (14.0 times as fast at third codon positions). Despite its relatively slow substitution rate, the phylogenetic signal in β-fibint7 is comparable to that in cyt b for woodpeckers, because β-fibint7 has less base composition bias and more uniform nucleotide substitution probabilities. As a consequence, compared with cyt b, β-fibint7 nucleotide sites are expected to enter more distinct character states over the course of evolution and have fewer multiple substitutions and lower levels of homoplasy. Moreover, in contrast to cyt b, in which nearly two thirds of nucleotide sites rarely vary among closely related taxa, virtually all β-fibint7 nucleotide sites appear free of selective constraints, which increases informative sites per unit sequenced. However, the estimated gamma distribution used to model rate variation among sites suggests constraints on some β-fibint7 sites. This study suggests that introns will be useful for phylogenetic studies of recently evolved groups.

Introduction

Molecular phylogenetic studies of recently evolved animal groups have been based overwhelmingly on mitochondrial DNA (mtDNA), because it evolves rapidly (Brown et al. 1982; Miyata et al. 1982; Moritz, Dowling, and Brown 1987; Edwards, Arctander, and Wilson 1991) and because it has a number of properties that make it relatively easy to isolate, amplify, and sequence (Kocher et al. 1989). Aside from Sibley and Ahlquist's (1990) massive study based on DNA-DNA hybridization, very few systematic studies of birds have been based on nuclear genes, and most of those have concerned ancient divergences (Caspers, Wattel, and De Jong 1994; Hedges 1994; Hedges et al. 1995; Caspers et al. 1997; Cooper and Penny 1997; Groth and Barrowclough 1999).

Despite the success of mtDNA studies, there are some disadvantages in mtDNA's use that could only be overcome by sequencing nuclear genes. One problem is that the mitochondrial genome is a single linkage group and a mitochondrial-haplotype tree could misrepresent the species tree because of lineage sorting or past hybridization. Although the mitochondrial haplotype tree has a higher probability of reflecting the species tree than does a nuclear gene tree under most circumstances (Moore 1995, 1997; but see Hoelzer 1997), independent

corroboration of the species tree could only come from a nuclear-gene tree. Another limitation of mitochondrial genes is that the base composition of mitochondrial protein-coding genes is biased, with substantial excess of C and A at third positions of codons (Anderson et al. 1981; Roe et al. 1985; Desjardins and Morais 1990). Codon bias in conjunction with different probabilities of substitutions (e.g., C→T vs. C→A) leads to biased parsimony reconstructions of ancestral sequences (Collins, Winbrmer, and Naylor 1994; Lockhart et al. 1994; Penna and Kocher 1995) and higher levels of homoplasy.

Introns from nuclear genes are obvious places in which to search for rapidly evolving nuclear sequences that might supplement the mitochondrial-haplotype tree. A phylogeny inferred from a nuclear-gene intron would certainly represent a sampling of the species phylogeny that is independent of the mitochondrial-haplotype tree with regard to lineage sorting, and several independent nuclear-gene trees could be sampled by sequencing introns from genes located on distinct chromosomes. Introns have several additional properties that would seemingly make them ideal sources of DNA sequence data for phylogenetic studies. The number and positions of introns in a gene are usually highly conserved through evolution, but the lengths of introns are somewhat variable as a result of indels (insertions and deletions), and the nucleotide sequence is highly variable (Lewin 1997). The rapid rate of nucleotide substitution suggests that these noncoding sequences are for the most part adaptively neutral, although conserved sequences within introns have been reported which function as regulatory elements (Jackson and Hoffmann 1994) or in alternative splicing of exons (Leicht et al. 1995; Clark,
Leicht, and Muse 1996). Nonetheless, most nucleotide positions in introns are free to vary, in contrast, for example, to the protein-coding genes of the mitochondrial genome, in which first and second codon positions are highly conserved. This results in potentially more phylogenetically useful variation per length of intron sequence. Also, implicit in the apparent neutrality of intron sequences is that characters (nucleotide positions) are independent and identically distributed. Character independence is a necessary assumption of phylogenetic inference (Swofford et al. 1996), and distributional identity simplifies analyses, such as maximum-likelihood analyses, based on models of character state change. Thus, it is not surprising that introns have proved useful in analyses in which they have been employed (Koop et al. 1989; Slade, Moritz, and Heideman 1994; Schneider et al. 1996; Prychitko and Moore 1997).

In an earlier paper (Prychitko and Moore 1997), we reported PCR primers that amplify intron 7 of the β-fibrinogen gene (β-fibint7) for five woodpecker species representing fairly disparate lineages, and we further reported that the intron sequences were easy to align even though divergent sequences included some indels. We also presented a preliminary parsimony analysis, which suggested that the nucleotide substitution patterns of this nuclear intron were very different from those of a mitochondrial-encoded protein gene (cytochrome b [cyt b]), although their levels of phylogenetic signal were similar.

The purposes of this paper were to report a more comprehensive set of β-fibint7 sequences for woodpeckers (13 species plus an outgroup species) and to compare the nucleotide substitution patterns of this nuclear gene intron with those of the mitochondrial-encoded cyt b gene for nearly identical sets of species. Such a comparison could provide insight as to how substitution patterns affect acquisition and retention of phylogenetic information. The larger number of species included in this study encompasses a broader range of divergences and allows more powerful tests of rate variation both among lineages and among nucleotide sites. In contrast to our previous study, we use a maximum-likelihood approach here for the substitution pattern analysis. This avoids several problems inherent in parsimony analysis (Yang 1994) and allows estimation of additional parameters that are useful in understanding how phylogenetic signal is accumulated by the two genes. We also included the indels in a parsimony analysis to determine their value as a supplement to the phylogenetic signal contained in nucleotide substitutions.

Woodpecker Systematics

The subfamily Picinae (family Picidae: order Piciformes; Peters 1948) comprises the woodpeckers and includes 171 nominal species (Short 1982). There are two additional subfamilies of the Picidae, Picumninae (piculets), which is the sister group of the true woodpeckers, and Jynginia (wrynecks), which is sister to the Picinae-Picumninae clade (Short 1982). Short (1982) recognized 23 woodpecker genera, which he divided into six tribes. Moore and DeFilippis (1997) estimated a phylogeny for the woodpeckers based on the mitochondrial cyt b gene that included seven genera representing four of the six tribes defined by Short. The cyt b phylogeny and its implicit classification differed from that proposed by Short in several salient respects. Because the major objective of this paper was to compare substitution properties between a mitochondrial protein-coding gene and a nuclear gene intron, we will not discuss congruence (or incongruence) of the two DNA-based phylogenies with the morphologically based phylogenies, although we will examine congruence of the two DNA sequence-based phylogenies.

The woodpeckers are a relatively recently evolved group. The oldest bona fide fossil dates to the Lower Pliocene (approximately 5 MYA; Short 1982; Feduccia and Wilson 1967; Cracraft and Morony 1969), and Moore, Smith and Prychitko (1999), based on computer simulations of cyt b and β-fibint7 molecular clocks, estimated that the woodpecker lineage separated from the piculet lineage approximately 15 MYA and that the diversification of the extant woodpeckers began 7–8 MYA. Thus, the comparisons of substitution patterns reported here are based on relatively recent divergences, which should reduce the confounding effects of multiple substitutions.

Materials and Methods

β-fibint7 was amplified and sequenced for two specimens of each of 13 species of woodpeckers plus a piculet, Picumnus aurifrons; the latter is a member of the subfamily Picumninae, which is sister to the Picinae (woodpeckers) and serves as an outgroup species (Short 1982; Moore and DeFilippis 1997). Comparison of the sequences from two specimens of each species served as an assurance that tissue vials were not misidentified or mislabeled and that a contaminant DNA was not amplified. (We sequenced one specimen each of Sphyrapicus varius and Sphyrapicus ruber, closely related hybridizing taxa [Cicero and Johnson 1995].) Species and specimens are tabulated in table 1 along with locale information. The 14 sequences used in the phylogenetic analysis (see below) are deposited in GenBank (accession numbers AF240008–AF240021).

Methods of DNA isolation, PCR amplification, PCR product cleaning, DNA quantification, and sequencing, as well as the sequences of the six primers used in this study, were described by Prychitko and Moore (1997).

Because the quality and amount of PCR product varied among the specimens amplified, the parameters used for PCR amplification of the intron as described by Prychitko and Moore (1997) were adjusted according to the primer combination used and the length of the fragment to be amplified. These adjustments did not require changes in the amount or concentration of reagents but did require modifications of the extension time and annealing temperature parameters of the thermal cycle regimen. Primer combinations, followed by length of fragment amplified, annealing temperature, and exten-
The primary focus of this study is on interspecific, rather than intraspecific, variation. However, because we sequenced two specimens for each species, it is possible to get a crude idea of whether intraspecific allelic var-

### Table 1

<table>
<thead>
<tr>
<th>Species</th>
<th>Locale</th>
<th>Museum*</th>
<th>Catalog No.</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Campephilus haematogaster</em> (Crimson-bellied woodpecker)</td>
<td>Esmeraldes, Ecuador</td>
<td>LSU</td>
<td>11786</td>
</tr>
<tr>
<td>Colaptes atricollis (black-necked flicker)</td>
<td>Darien, Panama</td>
<td>LSU</td>
<td>2188</td>
</tr>
<tr>
<td>Colaptes auratus (northern flicker)</td>
<td>Huanuco, Peru</td>
<td>LSU</td>
<td>3581</td>
</tr>
<tr>
<td>Colaptes rusticolus (Andean flicker)</td>
<td>Kentucky</td>
<td>WSU</td>
<td>8611</td>
</tr>
<tr>
<td>Dryocopus lineatus (lineated woodpecker)</td>
<td>Pasco, Peru</td>
<td>WSU</td>
<td>81144</td>
</tr>
<tr>
<td>Melanerpes carolinus (red-bellied woodpecker)</td>
<td>Texas</td>
<td>LSU</td>
<td>8634</td>
</tr>
<tr>
<td>Picoides villosus (hairy woodpecker)</td>
<td>Kentucky</td>
<td>LSU</td>
<td>8615</td>
</tr>
<tr>
<td>Piculus rubiginosus (golden-olive woodpecker)</td>
<td>Kentucky</td>
<td>LSU</td>
<td>8614</td>
</tr>
<tr>
<td>Piculus auricollis (black-bellied sapsucker)</td>
<td>Alabama</td>
<td>LSU</td>
<td>8625</td>
</tr>
<tr>
<td>Sphyrapicus varius (yellow-bellied sapsucker)</td>
<td>Arizona</td>
<td>WSU</td>
<td>86107</td>
</tr>
<tr>
<td>Sphyrapicus ruber (red-bellied sapsucker)</td>
<td>California</td>
<td>WSU</td>
<td>86144</td>
</tr>
<tr>
<td>Veniliornis callonotus (scarlet-backed woodpecker)</td>
<td>Puno, Peru</td>
<td>LSU</td>
<td>525</td>
</tr>
<tr>
<td>Veniliornis nigricaps (bar-bellied woodpecker)</td>
<td>La Paz, Bolivia</td>
<td>LSU</td>
<td>1274</td>
</tr>
<tr>
<td>Campephilus haematogaster</td>
<td>Lambayeque, Peru</td>
<td>LSU</td>
<td>5162</td>
</tr>
<tr>
<td>Lampropius doliolus</td>
<td>Lambayeque, Peru</td>
<td>LSU</td>
<td>5222</td>
</tr>
<tr>
<td>Lampropius carbonarius</td>
<td>Bolivia</td>
<td>LSU</td>
<td>18254</td>
</tr>
<tr>
<td>Lampropius auriceps</td>
<td>Michigan</td>
<td>WSU</td>
<td>8513</td>
</tr>
<tr>
<td>Lampropius olivaceus</td>
<td>California</td>
<td>WSU</td>
<td>86148</td>
</tr>
<tr>
<td>Lampropius lineatus</td>
<td>Lambayeque, Peru</td>
<td>LSU</td>
<td>5175</td>
</tr>
<tr>
<td>Lampropius lineatus</td>
<td>Lambayeque, Peru</td>
<td>LSU</td>
<td>5178</td>
</tr>
<tr>
<td>Lampropius lineatus</td>
<td>La Paz, Bolivia</td>
<td>LSU</td>
<td>1305</td>
</tr>
<tr>
<td>Lampropius lineatus</td>
<td>La Paz, Bolivia</td>
<td>LSU</td>
<td>8176</td>
</tr>
</tbody>
</table>

* LSU = Louisiana State University Museum of Natural History; WSU = Wayne State University Department of Biological Sciences.

### Results

**DNA Sequences**

The alignment consisting of the best (see below) of the two sequences for each species are presented so that the extent and nature of the indels can be visualized for discussion. The sequences shown represent the upper (sense) strand from the \( \beta \)-fibrinogen 7 intron. (We presented the antisense strand rather than the sense strand in Prychitko and Moore [1997]; that was an error in judgement on our part, and we now adopt the convention of presenting the sense strand.) The length of sequence varied slightly among the 14 species due to indel variation. Two alignment procedures were performed in CLUSTAL W. The input order of sequences for the first alignment was in decreasing order of phylogenetic relationship, as inferred by CLUSTAL W; for the second alignment procedure, the order of sequence input was random. The two resultant alignments are identical, indicating that the alignment is not an artifact of input order. Alignments are available at http://www.molbiolevol.org.

### Intraspecific Variation

Statistical analyses of sequences and phylogenetic analyses were performed with the computer program packages PHYLIB, version 3.57c (Felsenstein 1995), MEGA, version 1.01 (Kumar, Tamura, and Nei 1993), and PAUP* (test version 4.0 d64 written by D. L. Swofford).
Intraspecific variation was slight. We observed only a single nucleotide difference between specimens of 5 of the 14 species, giving a sequence divergence of 0.0012 in each case. In all cases, each specimen appeared to be a homozygote; however, because we sequenced PCR products as opposed to gene clones, it is likely that we would not detect heterozygotes. Although these data are few, they suggest that intraspecific variation in intron sequences would not confound estimation of the species phylogeny.

Because sequences from different specimens of the same species were identical or nearly so in all cases, we selected the “best” sequence from each species for interspecific statistical and phylogenetic analyses. Selection of the best sequence was based on quality of the sequence data (completeness, high signal-to-noise ratio, fewest ambiguous bases, and greatest amount of overlapping sequence among distinctly amplified fragments).

Base Composition and Nucleotide Substitution Patterns

Base composition for $\beta$-fibrin7 is biased, but not to the extreme exhibited by cyt $b$ third positions of codons. Comparing $\beta$-fibrin7 with cyt $b$ third positions, percentage compositions are as follows (mean ± 1 SD): $\beta$-fibrin7—A, 30.9 ± 0.48; T, 33.6 ± 0.39; C, 17.4 ± 0.51; G, 18.1 ± 0.18; cyt $b$—A, 29.8; T, 13.1; C, 53.1; G, 4.0 (Moore and DeFilippis 1997).

Transition (ts) substitutions are plotted as a function of transversion (tv) substitutions for $\beta$-fibrin7 in figure 1 for all species pair comparisons. Rigorous statistical analysis of these data was not attempted, because many of the data points were not statistically independent because species pairs share varying degrees of common ancestry. The least-squares fitted regression, forced through the origin, is shown in figure 1; the slope of 1.24 is an estimate of the transition:transversion ratio. Although not tested statistically, it is apparent that the transition:transversion ratio does not differ much, if at all, from 1:1. Moreover, the number of transitions as a function of transversions appears linear over the range of nucleotide divergences represented by woodpeckers and the piculet. This is in contrast to the substitution pattern in the mitochondrial cyt $b$ gene for the same set of taxa (Moore and DeFilippis 1997), where the ts/tv ratio is approximately 5.77:1 and the number of transitions plateaus between more divergent pairs of woodpecker sequences as a result of multiple hits. Cicero and Johnson (1995) reported similar saturation of cyt $b$ in a comparison of four sapsucker species with two more distantly related woodpecker species.

Phylogenetic Analysis

We computed the neighbor-joining tree based on the Tajima-Nei distance (Tajima and Nei 1984) for the $\beta$-fibrin7 data set. Like the Jukes-Cantor formula, the Tajima-Nei formula corrects for multiple hits, but the latter also corrects for heterogeneity in nucleotide frequencies, which is apparent in $\beta$-fibrin7. More complex estimators that consider variation in the probabilities of transitions and transversions require estimation of more parameters and hence have larger variances (Kumar, Tamura, and Nei 1993). Because the transition:transversion ratio is not strongly biased if it is biased at all, the Tajima-Nei formula is the estimator that has the lowest variance that adequately considers the complexities apparent in the data. In contrast, the more complex Tamura-Nei distance measure is better suited for phylogenetic analysis of the mitochondrial-encoded protein genes because these genes vary substantially in nucleotide frequencies, and the transition:transversion ratio is strongly biased.

The midpoint-rooted neighbor-joining tree derived from Tajima-Nei distances, based on 886 bp of $\beta$-fibrin7 sequence, is illustrated in figure 2a, juxtaposed against
the midpoint-rooted neighbor-joining tree (fig. 2b) based on 1,047 bp of mitochondrial \textit{cyt b} sequence from Moore and DeFilippis (1997). (The topologies of the two midpoint-rooted gene trees are identical to the respective gene trees rooted by specifying the piculet, \textit{P. aurifrons}, as the outgroup.) Three species in the \textit{\beta-fibint7} tree were not included in the mitochondrial-haplotype tree; these are enclosed in parentheses in figure 2a. For the species shared in common, the topologies of the two gene trees are identical in all respects except one. The clade including \textit{P. villosus} and two species of \textit{Veniliornis} (henceforth called the \textit{Picoidees-Veniliornis} clade) is inferred to be the sister group of a clade including \textit{Campephilus}, \textit{Melanerpes}, and \textit{Sphyrapicus} in the \textit{\beta-fibint7} tree, whereas it is inferred to be the sister group of a clade that includes the genera \textit{Colaptes}, \textit{Piculus}, and \textit{Dryocopus} in the mitochondrial-haplotype tree. The bootstrap proportion is not significant in either case (57% and 68%, respectively).

We found maximum-parsimony trees for two variations of the \textit{\beta-fibint7} data matrix: (1) without indels and (2) with indels. The without-indels analysis found two equally most-parsimonious trees, and the with-indels analysis found one most-parsimonious tree.

For the without-indels parsimony analysis, nucleotide sites where gaps were inserted to achieve alignment were excluded from the analysis (gaps treated as missing, PAUP*). The first of the two most-parsimonious trees differed from the neighbor-joining tree only in that \textit{Colaptes rupicola}, \textit{Piculus rivolii} and the clade of \textit{Colaptes atricollis} and \textit{Piculus rubiginosus} formed an unresolved trichotomy in the most-parsimonious tree, whereas \textit{C. rupicola} and \textit{P. rivolii} are paired as sister species in the neighbor-joining tree, but the bootstrap support was weak (54%). The second most-parsimonious tree differed from the first in one respect: \textit{Sphyrapicus varius} paired with \textit{C. haematogaster} rather than with \textit{Melanerpes carolinus}. The character statistics were identical for both trees: steps = 159; consistency index (CI) = 0.93; homoplasy index (HI) = 0.076; CI excluding uninformative characters = 0.81; HI excluding uninformative characters = 0.19; retention index (RI) = 0.91; rescaled consistency index (RC) = 0.85.

Each nucleotide position at which there is an apparent indel could be considered a distinct character; however, strings of contiguous inserted or deleted nucleotides probably result from single evolutionary events and should be scored as such when computing parsimony scores (e.g., nucleotide sites 845–851). Most of the indels are simple to interpret, involving single bases (e.g., site 94), but others are complex and their interpretation is somewhat subjective (e.g., sites 29–34). Our interpretation of the indels as they pertain to the parsimony analysis is summarized in table 2. For the with-indels parsimony analysis, an indel matrix was added to the nucleotide sequence matrix (FORMAT INTERLEAVE symbols = “01”), and options and commands were implemented such that all nucleotide substitutions plus all indels that were phylogenetically informative were included in the parsimony analysis, with...
all indels considered single characters regardless of their lengths (changed “−−” to “?” at sites 19–22 and 496–503; option: gap = newstate; command: exclude 29–33, 164–166, 439–440, 845–851; add five characters [0/1] to the interleave matrix representing indels 3, 7, 11, 12, and 18). The total number of characters in the without-indels analysis was 890.

The single most-parsimonious tree found in the with-indels analysis was identical to the first of the two most-parsimonious trees found in the without-indels analysis and to that of the neighbor-joining tree (except for the unresolved trichotomy noted above). The character statistics for this tree are as follows: steps = 177; CI = 0.92; HI = 0.079; CI excluding uninformative characters = 0.81; HI = 0.19; RI = 0.91; RC = 0.84. The bootstrap proportions for all three of these trees are indicated in figure 2a.

The topology of the maximum-likelihood tree is also identical to the topology illustrated in figure 2a (likelihood score = -2,181,29756). As noted above, the major difference between the β-fibint7 tree and the mitochondrial cyt b tree is where the Picoides-Veniliornis clade joins the tree. To test whether these two estimated topologies are significantly different, we moved this branch from its position as sister group to a Campephilus-Melanerpes-Sphyrapicus clade (fig 2a) to a position such that it was sister to the Colaptes-Picuslus-Dryocopus clade, as in the mitochondrial-haplotype tree (fig. 2b). The likelihood score for this topology was -2,183.48596; the difference in log-likelihoods is not significant (Kishino and Hasegawa 1989, pp. 171–172; Felsenstein 1995).

Discussion
Woodpecker Phylogeny

Congruence of phylogenies based on independent data sets is generally taken as evidence of true evolutionary relationships (Mickey and Johnson 1976; Mickey 1978; Penny and Hendy 1986; Miyamoto and Fitch 1995). As noted above, for the species shared in common, the β-fibint7 tree differed from the cyt b tree in just one respect, the relationship of the Picoides-Veniliornis clade to other genera (fig. 2), and the likelihoods of these alternative topologies do not differ significantly. Because there is no theoretical reason to suspect nor empirical data to suggest that nucleotide substitutions in a nuclear intron and a mitochondrial gene are not completely independent, it is highly likely that the evolutionary history portrayed by the consensus of the two gene trees is correct.

Phylogenetic Properties of β-fibint7 and cyt b

Base Composition

The base composition of neither gene varies much over the diversity of woodpeckers; thus, the evolutionary relationships inferred by the β-fibint7 gene tree and the cyt b gene tree are not likely to be incorrect as a result of convergent evolution of base composition, as explicated by Lockhart et al. (1994). Moreover, it is unlikely that the two gene trees would be as congruent as they were if base composition had biased tree estimation, because the bias was quite different in the two genes.

The greater concern with regard to base composition bias is that it effectively limits the number of character states that nucleotides are likely to enter over the course of evolution. This, in turn, leads to higher homoplasy or, stated another way, lesser accumulation of phylogenetic information by the gene. This will be discussed below in conjunction with the comparison of the nucleotide substitution matrices of the two genes. In this regard, however, the more equitable base composition of β-fibint7 gives it an advantage over cyt b.

Site-Specific Constraints

Because cyt b is a protein-coding gene, the variance in substitution rates between first, second, and third codon positions is great (Irwin, Kocher, and Wilson 1991; Krajewski and King 1995; Moore and DeFilippis 1997). For cyt b, we estimated the frequencies of first-, second-, and third-position substitutions between two pairs of relatively closely related species, C. auratus versus P. rubiginosus and Veniliornis callonotus versus Veniliornis nigriceps. Restricting the comparison to just these pairs of closely related species reduces the amount of data but assures that the data points are independent and that bias resulting from multiple hits is minimized. There are a total of 116 substitutions between species of these two pairs: 14 at position 1, 3 at position 2, and 99 at position 3. Eleven substitutions result in amino acid replacements, leaving 105 that are synonymous (90.5%).

β-fibrinogen intron 7, in contrast, is presumably not a protein-coding sequence, although parts of its sequence could function as regulatory elements, as alternative splice sites, or in some other way such as maintenance of secondary structure of the transcript. For assurance that the sequence is, in fact, not protein-coding,
we numbered the nucleotide sites as first, second, and third positions as though the intron were protein-coding. A total of 143 $\beta$-fibint7 substitutions were observed over the evolution of the 14 specimens (species) included in the phylogenetic analysis. The numbers of substitutions were roughly evenly distributed among the arbitrarily numbered sites: first = 54, second = 50; third = 39. This distribution does not differ significantly from a random distribution ($\chi^2 = 2.531, df = 2, 0.25 < P < 0.50$). In addition, several stop codons occur in the intron, regardless of where the reading frame is initiated. These facts confirm that the intron is not protein-coding.

Selective constraints generally should produce a nonrandom distribution of substitutions. We tested whether substitutions were randomly distributed along the length of the intron by analyzing the number of substitutions falling in successive blocks of 10 nt among the 14 specimens. For unconstrained (neutral) sequences, the number of substitutions occurring among blocks should be Poisson-distributed, and, accordingly, the mean should be equal to the variance. The mean for the combined 886 nt (88 blocks of 10; 1 block of 6) is 1.61 substitutions per block, and the variance is 1.56. A chi-square goodness-of-fit test did not indicate significant deviation from a Poisson distribution ($\chi^2 = 1.479, df = 5, 0.90 < P < 0.95$). $\beta$-fibint7 genes of all species in this study have the GT and AG doublet splice sites that mark the beginning and end of the intron (Breathnach, Mandel, and Chambron 1977; Mount 1982), but there is no apparent conservation of additional sequence within the intron. However, our tests were not very powerful statistically, and it is probable that we would have failed to detect short conserved sequences; that our statistical test did not detect the GT and AG splice sites is a case in point. A more powerful test could be based on comparison among more distantly related species, where more substitutions would be expected and conservation would be more apparent.

The lack of detectable selective constraints on $\beta$-fibint7 is an advantage in that all nucleotides are free to vary, as opposed to just the synonymous sites in cyt $b$, which comprise only third positions plus a small fraction of first positions. Of course, some amino acid substitutions in cyt $b$ may be neutral or weakly selected. Nonetheless, it is apparent that a preponderance approaching two thirds of nucleotide sites in cyt $b$ will rarely be phylogenetically informative among recently evolved species, whereas virtually all $\beta$-fibint7 nucleotide sites are potentially phylogenetically informative.

Rate Variation Among Lineages

Moore and DeFilippis (1997) tested the molecular clock for the cyt $b$ data set by studying the index of dispersion ($R$, the ratio of the mean of branch lengths to the variance) of pairs of species diverging from a common ancestor. If substitutions are neutral, the number of substitutions will follow a Poisson distribution and be clocklike, and the dispersion indices should be approximately 1 (Gillespie 1989). The average dispersion index for cyt $b$ was 0.90, which indicates no rate variation.

Similarly, Prychitko and Moore (1997) reported clocklike evolution of $\beta$-fibint7 based on only five species of woodpeckers. The latter study was based on the ratio of likelihoods without and with the constraint of a molecular clock (DNAML/DNAMLK programs in PHYLIP; Felsenstein 1995). However, the same test applied to the 14 $\beta$-fibint7 sequences used in figure 3a resulted in rejection of the null hypothesis, i.e., there is significant rate variation ($\chi^2 = 30.74, df = 12, P < 0.005$). Inspection of figure 2a suggests that the branches leading to Piculus rivolii, Campephilus, and M. carolinus are exceptionally short. Indeed, if these three species are eliminated from the analysis, the null hypothesis of rate constancy is accepted ($\chi^2 = 8.9, df = 9, P > 0.25$). Dropping various combinations of these three species from the analysis is useful in identifying where rate variation might exist. Dropping $P$. rivolii alone results in an insignificant or marginally significant test ($\chi^2 = 18.76, df = 11, 0.05 < P < 0.10$), whereas dropping $M$. carolinus and $C$. haematogaster but retaining $P$. rivolii still results in rejection of the rate constancy hypothesis ($\chi^2 = 20.66, df = 10, 0.01 < P < 0.025$). This suggests that the apparent rate variation is most probably associated with the $P$. rivolii branch, which effectively has a length of 0 according to the phylogenetic inference (fig. 2a), and that $M$. carolinus and $C$. haematogaster also contribute to deviation from molecular-clock expectations, but to lesser extents than $P$. rivolii.

Although this analysis suggests variation in the rate of $\beta$-fibint7 evolution for woodpeckers, a cautionary note should be made. Specifically, if the DNAML-DNAMLK test is based on an incorrect topology, it is possible, indeed likely, that the null hypothesis will be rejected even though evolution of the molecule is clocklike. In our analysis, the most deviant branch length ($P$. rivolii) is associated with the most weakly supported relationship in the $\beta$-fibint7 tree. Similarly, $C$. haematogaster and $M$. carolinus are paired with other taxa in relatively weakly supported relationships. Thus, it is
possible that rejection of the molecular-clock hypothesis resulted from basing the test on an incorrect topology rather than from rate variation. Additional sequence from other β-fibrinogen introns could result in a more strongly supported tree and, hence, less uncertainty about the test of the molecular-clock hypothesis.

**Relative Substitution Rates in β-fibint7 and cyt b**

The substitution rate for a nucleotide sequence is important in determining the span of evolutionary time over which the sequence will be phylogenetically informative. Mitochondrial DNA evolves rapidly relative to nuclear DNA in vertebrates (Brown, George, and Wilson 1979; Brown et al. 1982; Miyata et al. 1982) and apparently in *Drosophila* as well (Moriyama and Powell 1997). Estimation of substitution rates can be confounded by several factors. Genetic distances between all species in a phylogeny are not independent estimates, because the species are related to varying degrees through common ancestry. This statistical dependence resulting from phylogeny can be controlled to the extent that the topology is correctly resolved by comparing the lengths of identical branches for the two genes. Another potentially confounding factor is saturation of the sequence with multiple hits. If, for example, a branch is very old, it is possible that both the *cyt b* and the *β-fibint7* sequences are saturated, in which case the branch lengths will be similar because the more slowly evolving sequence has caught up with the saturated rapidly evolving sequence, which would lead to the false inference that they evolve at similar rates. Thus, comparison of the lengths of young branches in well-resolved gene trees would provide the ideal circumstances.

Our data do not achieve this ideal but are sufficiently close to warrant making an estimate of the ratio of *cyt b* to *β-fibint7* substitution over the evolutionary history of woodpeckers. Specifically, we estimated the relative rate of *cyt b* substitution to *β-fibint7* substitution for all *cyt b* codon sites and also for just the third positions of *cyt b* codons. Most third-position substitutions are synonymous, although some third-position transversions are nonsynonymous. In the comparison of first-, second-, and third-position substitutions between two pairs of closely related species (*C. auratus* vs. *P. rubiginosus* and *V. callonotus* vs. *V. nigriceps*) described above, 98 of 99 third-position substitutions were synonymous; thus, third-position substitutions provide a reasonably good estimate of the synonymous substitution rate for *cyt b*. In turn, the rate ratio of synonymous *cyt b* to *β-fibint7* substitutions should be a reasonably good estimate of the rate of neutral substitution for mitochondrial versus nuclear genes.

We estimated these rate ratios by simple linear regression of the lengths of terminal branches in the *β-fibint7* tree on those of identical branches in the *cyt b* tree. For this analysis, we dropped the outgroup, *P. aurifrons*, from the data sets, because it is estimated to have diverged from woodpeckers approximately 15 MYA, and *cyt b* saturates much sooner than that (Moore, Smith, and Prychitko 1999). Taxa that did not occur in both data sets were also excluded. Maximum-likelihood branch lengths were used in this regression analysis and were obtained as follows: a topology was created identical to figure 2a except the short internode (heavy branch) representing the common ancestor of the *Picoides-Veniliornis* clade, *Campephilus*, *Melanerpes*, and *Sphyrapicus* was collapsed to form a basal trichotomy (*C. rupicola* ... *D. lineatus*), (*V. callonotus* ... *P. willouis*), (*C. haematogaster* ... *S. varius*). Forming this trichotomy allowed comparison of *β-fibint7* and *cyt b* in a consensus tree; i.e., the *β-fibint7* and *cyt b* trees differed only with regard to this short, basal internode and the heavy internode in the *cyt b* tree (fig. 2b); collapsing both internodes made the two gene trees identical. This topology was input as a user tree into DNAML (PHYLIP), and the likelihood branch lengths were calculated for both the *β-fibint7* and the *cyt b* data sets.

The regression of *cyt b* third-position branch lengths on *β-fibint7* branch lengths was significant (*P* < 0.007), and the slope of 14.03 (95% confidence interval: 5.40–22.67) estimates the ratio of third-position *cyt b* substitutions to *β-fibint7* substitutions. When the *cyt b* branch lengths were based on all codon positions, the slope was 2.79 (*P* < 0.03; 95% confidence interval: 0.32–5.26). Because a large majority of all of the substitutions occur at third positions, including the first and second positions in the analyses does little more than rescale the *cyt b* branch lengths. In the foregoing analyses, we excluded *Melanerpes* and *Campephilus* also, reasoning that these two taxa may have evolved at a slower rate for *β-fibint7* than the eight other species or that their true relationships may be misrepresented in the tree.

**Patterns of Nucleotide Substitution**

The relative frequencies of specific substitutions (A→G, C→T, etc.) differ between the fibrinogen intron and mitochondrial-encoded *cyt b* gene in ways that affect the acquisition and retention of phylogenetic signal by each molecule. In our preliminary paper, we estimated the 12 substitution frequencies by a parsimony method (Prychitko and Moore 1997). However, parsimony is known to give biased estimates of these frequencies when base composition is biased, as is the case with *cyt b* and, to a lesser extent, with *β-fibint7* (Collins, Wimberger, and Naylor 1994; Perna and Kocher 1995; Adachi and Hasegawa 1996).

Here, we modeled nucleotide evolution as a reversible, stationary Markov process and estimated “rate ratios” by maximum likelihood (Yang 1994) as implemented in PAUP*. The topology and branch lengths are parameters of the model, as are the rate ratios (Yang, Goldman, and Friday 1995); the process variables that are assumed to be stationary (in equilibrium) are the frequencies of the four nucleotides (*πA*, *πC*, *πG*, *πT*) in the DNA sequence. The Markov process is driven by the *4 × 4* Q-matrix with elements *Qij* = *Rijπj*; the elements of the K-matrix are the rate ratios and are proportional to the conditional probabilities that a nucleo-
tide, \( i \), will be substituted by \( j \) in a small increment of time given that it is \( i \) at the beginning of the time increment \( (i, j = A, C, G, T) \). To achieve mathematical tractability, the reversible model assumes \( \pi_iQ_{ij} = \pi_jQ_{ji} \). This assumption is likely violated by both \( \beta\text{-fbin}7 \) and \( \text{cyt b} \) evolution, but Yang (1994) showed that the reversible model is a good approximation of the unrestricted model (i.e., \( \pi_iQ_{ij} \neq \pi_jQ_{ji} \)). Moreover, our earlier parsimony analysis (Prychitko and Moore 1997) indicated that reverse substitution frequencies are similar to forward substitution frequencies for both genes, as was the case for \( \text{cyt b} \) in the sapsucker study of Cicero and Johnson (1995).

We included all sites in both our \( \text{cyt b} \) analysis (first-, second-, and third-codon positions) and our \( \beta\text{-fbin}7 \) analysis and used the gamma correction in both cases to correct for rate variation among sites (three rate categories; Yang 1996). Because we wanted to compare the rate ratios of \( \beta\text{-fbin}7 \) with those of \( \text{cyt b} \), we transformed the \( \beta\text{-fbin}7 \) rate ratios to the same scale as those of \( \text{cyt b} \) by multiplying the \( \beta\text{-fbin}7 \) values by 2.79, which is the multiple by which the \( \text{cyt b} \) substitution rate exceeds that of \( \beta\text{-fbin}7 \).

The results of these two analyses are presented in figure 3. There are two striking contrasts in substitution patterns between \( \beta\text{-fbin}7 \) and \( \text{cyt b} \): (1) the overall rate difference (reflected in smaller-diameter circles for \( \beta\text{-fbin}7 \) and \( \beta\text{-fbin}7 \)) and (2) greater uniformity of rates among categories for \( \beta\text{-fbin}7 \). The consequences of these differences are that \( \text{cyt b} \) saturates quickly and manifests higher levels of homoplasy (HI excluding uninformative characters \( = 0.45 \)) because of its rapid rate of evolution and because a given nucleotide is likely to enter fewer distinct character states over a given number of character state changes. \( \beta\text{-fbin}7 \), in contrast, evolves more slowly and manifests lower levels of homoplasy (HI excluding uninformative characters \( = 0.19 \)), because a given site is likely to enter more distinct character states. For example, in \( \text{cyt b} \), an A will most likely be substituted by a G, but that G is much more likely to revert to an A than to “advance” to a C or a T; in \( \beta\text{-fbin}7 \), an A will most likely be substituted by a G but then has higher relative probabilities of substitution by a C or T. These patterns are consistent with the higher transition: transversion ratio of \( \text{cyt b} \) relative to \( \beta\text{-fbin}7 \).

Finally, comparison of the maximum-likelihood-estimated shape parameter, \( \alpha \), provides an additional contrast in the substitution patterns of \( \beta\text{-fbin}7 \) and \( \text{cyt b} \). As noted above, substitution rates can vary substantially among sites in a DNA sequence; this is obvious in protein-coding sequences where substitution rates vary among first-, second-, and third-codon positions. If the substitution rate were the same among all nucleotide sites, then the number of substitutions falling in blocks of, say, 10 sites would have a Poisson distribution, but if the rates differed among sites, then the distribution would be that of a negative binomial, whose variance is greater than that of a Poisson distribution. One interpretation of the negative binomial distribution is that it is a Poisson distribution where the Poisson parameter (its mean) is a random variable with a gamma distribution.

To take into account rate variation among sites, Yang (1993, 1994) developed a substitution model based on the negative binomial distribution using an approximation of the gamma distribution. He simplified the maximum-likelihood estimation by equating the two parameters that specify a gamma distribution \( (\alpha = \beta) \), which implies that the mean is 1), thus leaving the single shape parameter, \( \alpha \), to be estimated. Substantial rate variation among sites is reflected in smaller values of \( \alpha \), whereas large values are indicative of low rate variation among sites. The negative binomial distribution converges to a Poisson distribution as \( \alpha \to \infty \).

We estimated \( \alpha \) by maximum likelihood for \( \beta\text{-fbin}7 \), \( \text{cyt b} \) (all sites), and \( \text{cyt b} \) third positions (PAUP* 4.0b2, options: empirical nucleotide frequencies, six substitution types, three rate categories, R-matrix estimated, general time-reversible model, topologies as in fig. 2a and b, respectively). The estimated values of \( \alpha \) are 0.791, 0.213, and 3.741, respectively. The low \( \alpha \) value estimated from all \( \text{cyt b} \) sites is as expected because of the rate variation between first, second, and third codon positions. The value for third positions of \( \text{cyt b} \) is higher, as expected, because most of these substitutions are synonymous (see above). The intermediate value for \( \beta\text{-fbin}7 \) suggests that there are relatively uniform rates among sites but that there is more variation than for third positions of \( \text{cyt b} \). This suggests that some sites in the intron may actually be constrained by natural selection but that we were unable to detect this by the weak statistical tests we applied (Poisson goodness-of-fit test). Failure to detect constraints would be probable if the conserved sequences were short.

Conclusions

Although the nuclear-encoded fibrinogen intron evolves more slowly than the mitochondrial-encoded protein-coding gene, it contains at least as much information as the latter about the phylogeny of a comparatively young clade, the woodpeckers. For seven nodes held in common between the \( \beta\text{-fbin}7 \) and \( \text{cyt b} \) trees (fig. 2), bootstrap support is strongest in the \( \beta\text{-fbin}7 \) tree in five cases, equal in one, and less in one. Thus, introns appear to have as much potential as mtDNA for resolving relationships among homeotherms as young as 2–5 Myr old (the approximate ages of splits among younger, resolved lineages) and, at the other extreme, should not saturate significantly until approximately 60 Myr of divergence, which is later than that expected for \( \text{cyt b} \) (Moore, Smith, and Prychitko 1999). Although the phylogenetic information embodied in the two genes is similar, it results from very distinct patterns of substitution: \( \text{cyt b} \) evolves rapidly, but because of base composition bias and substantial inequalities among nucleotide substitution probabilities, homoplasy (parallel and reversed substitutions) is high. The opposite is true for \( \beta\text{-fbin}7 \); this gene evolves slowly, but homoplasy is low because of more equitable base composition and nucleotide substitution probabilities. In a sense, \( \text{cyt b} \) rushes through evolution but soon forgets its history,
while β-fibin7 plods along, clearly remembering its history.

Acknowledgments

We thank the Louisiana State University Museum of Natural History for its generous donation of frozen tissue samples. We thank two anonymous reviewers for their valuable comments on the manuscript. This research was supported in part by grants from the National Science Foundation (grant numbers DEB-9316452 and DEB-9726512).

LITERATURE CITED


Stephen Palumbi, reviewing editor

Accepted March 27, 2000