

# Comparing Molecular Evolution in Two Mitochondrial Protein Coding Genes (Cytochrome *b* and ND2) in the Dabbling Ducks (Tribe: Anatini)

Kevin P. Johnson<sup>\*,1</sup> and Michael D. Sorenson<sup>†,2</sup>

<sup>\*</sup>Department of Ecology, Evolution, and Behavior, and Bell Museum of Natural History, University of Minnesota, St. Paul, Minnesota; and <sup>†</sup>Museum of Zoology, University of Michigan, Ann Arbor, Michigan

Received June 6, 1997; revised October 23, 1997

**Rates of sequence evolution were estimated for the cytochrome *b* (cyt *b*) and NADH dehydrogenase subunit 2 (ND2) genes using a phylogeny of the dabbling ducks (Tribe: Anatini) and outgroups. This speciose group was densely sampled, reducing the impact of undetected homoplasy on rate comparisons. Phylogenies based on sequences of the two gene regions and various weighting schemes differed, but most of the differences involved weakly supported nodes. In addition, partition homogeneity tests show that these differences were not due to statistically significant conflict between the data sets. Cyt *b* and ND2 also showed similar rates and types of both nucleotide and amino acid substitutions. For both genes, substitutions between isoleucine and valine and between alanine and threonine were most common; both of these substitution types are the result of A-G transitions at first positions of codons. Rates of sequence evolution varied substantially and significantly among nucleotide positions, and even within a given codon position (first, second, or third), rates were significantly heterogeneous among sites. Within Anatini, cyt *b* and ND2 show similar levels of variation and homoplasy, and are equally useful for reconstructing the species level phylogeny of this group.** © 1998 Academic Press

**Key Words:** mtDNA coding gene evolution; weighting; cytochrome *b*; ND2; molecular phylogeny; *Anas*.

## INTRODUCTION

Differences in rates of evolution between and within gene regions have important implications for phylogeny reconstruction. These differences have been well documented for several gene regions (Holmquist *et al.*, 1983; Li and Graur, 1991). Certain proteins (e.g., cytochromes, histones) are subject to more biochemical constraints (in terms of protein structure) than others (e.g., ATPase, NADH dehydrogenase), and the pattern

and rate of amino acid substitution may differ between these proteins (Jacobs *et al.*, 1988; Mindell and Thacker 1996). Ideally, differences in rates of evolution within and between gene regions would be reflected in phylogenetic weighting schemes by giving more weight to relatively conserved sites and substitution types which occur less frequently, thus emphasizing sites and changes with a lower probability of homoplasy.

Comparisons of the evolution of mitochondrial gene regions (Jacobs *et al.*, 1988; Li and Graur 1991; Mindell and Thacker, 1996; Russo *et al.*, 1996) suggest considerable variation in rates of change within and between gene regions. Comparisons of widely divergent taxa suggest differences in the rate and mode of evolution of cytochrome *b* (cyt *b*) and NADH dehydrogenase subunit 2 (ND2) (Jacobs *et al.*, 1988; Meyer, 1994; Russo *et al.*, 1996). Cytochromes may have a more functionally constrained protein structure than does NADH dehydrogenase (Meyer, 1994). If there are dramatic differences in rate of evolution, phylogenies based on sequences of the two genes may differ even though they occur on the same linkage group because homoplasy in the rapidly evolving region may cause long branches to attract (Bull *et al.*, 1993). It is important to determine whether or not these differences in constraints are evident at all taxonomic levels and whether these differences cause sequences of the two gene regions to be phylogenetically incongruent.

Many rate variation analyses to date, however, have been made without reference to phylogenetic relationships or include only a few widely divergent taxa where unrecovered homoplasy (especially at third sites) will obscure estimates of the rates of base and amino acid substitutions (Graybeal, 1994; Cummings *et al.*, 1995; Mindell and Thacker, 1996; Russo *et al.*, 1996). Few studies have examined DNA sequence evolution at a finer scale by using many taxa spanning several lineages differing in age, or have directly compared different protein coding genes. Here we compare rates of sequence evolution in two mitochondrial protein coding genes (cyt *b* and ND2) by densely sampling taxa in a single large avian genus and several outgroups. This sampling scheme minimizes unrecovered homoplasy,

<sup>1</sup> Present address: Department of Biology, University of Utah, Salt Lake City, UT.

<sup>2</sup> Present address: Department of Biology, Boston University, Boston, MA.

providing better estimates of the rates and patterns of base substitution.

To determine the nature of DNA sequence variation between *cyt b* and ND2 and whether phylogenies based on sequences of these genes show conflict, we sequenced 1047 bp of *cyt b* and all of ND2 (1041 bp) for the dabbling ducks (Tribe: Anatini) and 11 outgroup taxa in the subfamily Anatinae. The goal of this study was to provide an extensive molecular data set to compare molecular evolution of two coding regions in the mtDNA genome.

We compare *cyt b* and ND2 with respect to rates of transition and transversion substitution at first, second, and third positions of codons and rates of amino acid substitutions. Phylogenies constructed from each gene region are compared and examined for conflicting phylogenetic signal. We discuss these results in relation to levels of homoplasy, utility of these two genes in reconstructing phylogenies, and implications for phylogenetic reconstruction and weighting methodology.

## METHODS

### *DNA Sequencing*

Samples for genetic analysis included blood and feathers taken from live birds in the wild and captivity, feathers from hunter-shot birds, and muscle from museum tissue (Table 1). Genomic DNA was isolated from 0.1 g of muscle tissue, 20  $\mu$ l of whole blood in lysis buffer, or the quill (c. 3 mm) of a single medium-sized (c. 100 mm) feather. Blood and tissue samples were digested in a total volume of 300  $\mu$ l including 0.5 mg/ml proteinase K and 1% sodium dodecyl sulfate (SDS) and were subjected to standard phenol/chloroform extraction and ethanol precipitation. For feathers, the digestion buffer included 10 mg/ml dithiothreitol (DTT), and Centricon-30s (Amicon) were used to purify samples following extraction (Cooper, 1994). For some samples, genomic DNA from both body tissue and feather samples was also isolated with a QIAamp Tissue Kit (QIAGEN) according to the manufacturer's protocol but with the addition of 30  $\mu$ l of 100 mg/ml DTT for feather samples.

The mitochondrial *cyt b* and ND2 genes were amplified in 550- to 700-bp fragments and sequenced with the primers listed in Table 2. PCR amplifications were in 50  $\mu$ l total volume with 1.25 units AmpliTaq DNA Polymerase (Perkin Elmer), the manufacturer's buffer, 2.5 mM MgCl<sub>2</sub>, 0.25 mM each dNTP, and 1  $\mu$ M each primer. PCR products were gel-purified in 1.5% low-melt agarose, excised from the gel under UV light, and purified with a QIAquick Gel Extraction Kit (QIAGEN). Approximately 75 ng of double-stranded PCR product was used in cycle sequencing reactions using fluorescent dye terminators and AmpliTaq FS (Applied Biosystems). Unincorporated dyes were removed from sequencing reaction products with Centri-Sep columns (Princeton Separations). Reaction prod-

ucts were run on an ABI 373 or 377 automated DNA sequencer. Sequences for forward and reverse strands were examined and reconciled using Sequence Navigator (Applied Biosystems) and submitted to GenBank (accession numbers AF059053-AF059174).

Nuclear pseudogenes of mitochondrial origin have been documented in a number of avian taxa and if mistaken for mitochondrial sequences can introduce error in phylogenetic analysis (Quinn, 1992; Arctander, 1995; Sorenson and Fleischer, 1996). In birds, PCR amplification of nuclear copies has usually been associated with total DNA extracts from avian blood samples, the source of DNA for only one of the samples in this study. In addition, nuclear copies of mtDNA sequences usually coamplify with the genuine mtDNA copy and a number of additional criteria can be used to recognize such sequences where they occur (Sorenson and Quinn, 1998). Ambiguities at a small number of positions ( $n = 12$  and 4, respectively) in sequences for *Anas sparsa* and *Anas smithii* (the latter extracted from blood) may be due to the coamplification of nuclear and mtDNA sequences. We found no other evidence for nuclear sequences in our study.

### *Phylogenetic Analysis*

We constructed maximum parsimony trees from the coding regions of *cyt b* and ND2 both independently and combined using heuristic searches with 20 random addition replicates (PAUP\*; Swofford, 1997). Theoretically, weighting schemes which incorporate the underlying variation in rates among sites should be used in parsimony reconstructions of phylogeny with more slowly evolving sites given more weight because of the decreased chance of homoplasy. To explore the impact of weighting on tree topology, we applied several weightings of transversions over transitions including 1:1, 2:1, 3:1, 5:1, 9.5:1, 15:1, and six-parameter weighting (Williams and Fitch, 1989; see Table 3). In each heuristic search, we initially set the number of random addition replicates to 20; if the shortest trees were found in less than 5 (25%) of the replicates, we repeated the search with 200 random addition replicates.

We performed partition homogeneity tests with PAUP\* using 100 random partitions of the entire data set (Farris *et al.*, 1995; Swofford, 1997) to determine if these two gene regions exhibited significantly different phylogenetic signal. In addition, because substitution rates are known to vary by codon position, the partition homogeneity test using first, second, and third sites as partitions was performed to determine if there was significantly different phylogenetic signal due to rate variation at these different sites (Bull *et al.*, 1993). We also combined first and second sites and performed the partition homogeneity test against third sites, because there were very few variable second sites.

The number of steps at each codon position was calculated over the trees resulting from the 1:1 weight-

TABLE 1

## Samples Used for DNA Sequencing

Scientific name	Common name	Type	Source
<i>Anas acuta</i>	Northern pintail	T	LSU B-20714, La Salle Par., LA
<i>Anas georgica spinicauda</i>	Chilean pintail	F	SHW, c
<i>Anas b. bahamensis</i>	Bahama pintail	F	SHW (CRC-210958), c
<i>Anas b. rubrirostris</i>	White-cheeked pintail	T	BM (42278), nw of Buenos Aires, Argentina
<i>Anas erythrorhyncha</i>	Red-billed pintail	F	SHW, c
<i>Anas capensis</i>	Cape teal	T	LSU B-10357, c
<i>Anas c. crecca</i>	Eurasian green-winged teal	F	SHW, c
<i>Anas carolinensis</i>	American green-winged teal	F	FWS, Solano Co., CA
<i>Anas gibberifrons gracilis</i>	Australian grey teal	F	SHW, c
<i>Anas castanea</i>	Chestnut teal	F	SHW, c
<i>Anas bernieri</i>	Madagascar teal	F	JWPT (B3260), c
<i>Anas a. aucklandica</i>	Auckland Island teal	F	MW (S-48561), Ewing Island, Auckland Islands
<i>Anas a. nesiotis</i>	Campbell island teal	F	MW (S-71044), Dent Island, off Campbell island
<i>Anas chlorotis</i>	Brown teal	F	MW (61024), Clendon Cove, Bay of Islands, Northland, NZ
<i>Anas f. flavirostris</i>	Speckled teal	T	BM (42275), nw of Buenos Aires, Argentina
<i>Anas f. oxyptera</i>	Sharp-winged teal	F	SHW, c
<i>Anas laysanensis</i>	Laysan teal	T	LSU B-10358, c
<i>Anas luzonica</i>	Philippine duck	T	LSU B-19204, c
<i>Anas platyrhynchos 1</i>	Mallard	T	Maryland
<i>Anas platyrhynchos 2</i>	Mallard	T	Maryland
<i>Anas poecilorhyncha</i>	Indian spot-billed duck	F	SHW, c
<i>Anas zonorhyncha</i>	Chinese spot-billed duck	F	SHW, c
<i>Anas diazi</i>	Mexican duck	F	FWS, Santa Cruz Co., AZ
<i>Anas rubripes</i>	American black duck	F	FWS, Montgomery Co., MD
<i>Anas fulvigula</i>	Mottled duck	F	FWS, Terrebonne Co., LA
<i>Anas superciliosa rogersi</i>	Australian black duck	T	MV (762), Victoria, Australia
<i>Anas melleri</i>	Meller's duck	T	JWPT (B3523), c
<i>Anas undulata</i>	Yellow-billed duck	T	JWPT (B3555), c
<i>Anas sparsa</i>	African black duck	T	LSU B-18923, c
<i>Anas americana</i>	American wigeon	F	FWS, Brazoria Co., TX
<i>Anas sibilatrix</i>	Chiloé wigeon	T	LSU B-20764, c
<i>Anas penelope</i>	Eurasian wigeon	T	LSU B-20719, Colusa Co., CA

TABLE 1—Continued

Scientific name	Common name	Type	Source
<i>Anas strepera</i>	Gadwall	F	FWS, Bell Co., TX
<i>Anas falcata</i>	Falcated duck	F	SHW, c
<i>Anas versicolor</i>	Silver teal	T	BM (42276), nw of Buenos Aires, Argentina
<i>Anas puna</i>	Puna teal	F	SHW, c
<i>Anas hottentota</i>	Hottentot teal	F	SHW, c
<i>Anas querquedula</i>	Garganey	F	SHW, c
<i>Anas r. rhynchotis</i>	Australian shoveler	T	MV (2187), Victoria, Australia
<i>Anas smithii</i>	Cape shoveler	B	SHW, c
<i>Anas clypeata</i>	Northern shoveler	F	FWS, Henderson Co., TX
<i>Anas platalea</i>	Red shoveler	T	BM (42273), nw of Buenos Aires, Argentina
<i>Anas c. cyanoptera</i>	Argentine cinnamon teal	T	BM (42277), nw of Buenos Aires, Argentina
<i>Anas c. septentrionalium</i>	Northern cinnamon teal	F	FWS, Fresno Co., CA
<i>Anas discors</i>	Blue-winged teal	F	FWS, Brevard Co., FL
<i>Anas formosa</i>	Baikal teal	T	BM (42241), c
<i>Amazonetta brasiliensis</i>	Brazilian teal	F	SHW, c
<i>Lophonetta specularoides</i>	Crested duck	F	SHW, c
<i>Specularias specularis</i>	Bronzed-winged duck	F	SHW, c
<i>Tachyeres ptereres</i>	Magellanic flightless steamer duck	F	SHW, c
<i>Aix sponsa</i>	Wood duck	F	SHW, c
<i>Asarcornis scutulata</i>	White-winged wood duck	F	SHW (130), c
<i>Aythya americana</i>	Redhead	T	FWS, Laguna Madre, TX
<i>Cairina moschata</i>	Muscovy duck	F	SHW, c
<i>Callonetta leucophrys</i>	Ringed teal	T	CC, MN, c
<i>Chenonetta jubata</i>	Maned goose	F	SHW, c
<i>Cyanochen cyanopterus</i>	Abyssinian blue-winged goose	F	SHW, c
<i>Marmaronetta angustirostris</i>	Marbled teal	F	SHW, c
<i>Pteronetta hartlaubi</i>	Hartlaub's duck	F	SHW, c
<i>Sarkidiornis melanotos</i>	Comb duck	F	SHW, c
<i>Tadorna tadorna</i>	Common shelduck	F	SHW, c

Note. F, feather; T, body tissues; B, blood; c, captivity with no known geographic origin. BM, Bell Museum of Natural History, University of Minnesota; CC, Cedar Creek Natural History Area, University of Minnesota; JWPT, Jersey Wildlife Preservation Trust; LSU, Louisiana State University Museum of Zoology; MV, Museum of Victoria, Australia; MW, Murray Williams, New Zealand; SHW, Sylvan Heights waterfowl, Scotland Neck, North Carolina; UWBM, University of Washington Burke Museum; FWS, U.S. Fish & Wildlife Service.

**TABLE 2**  
**Primers Used for Amplification and Sequencing**  
**of ND2 and Cyt *b***

Name <sup>a</sup>	Sequence <sup>b</sup>	Ref. <sup>c</sup>
L5219	5'-CCCATACCCCGAAAATGATG-3'	1
L5758	5'-GGCTGAATRGGMCTNAAAYCARAC-3'	1
H5766	5'-GGATGAGAAGGCTAGGATTTTKCG-3'	1
H6313	5'-CTCTTATTTAAGGCTTTGAAGGC-3'	1
L14990	5'-AACATCTCCGCATGATGAAA-3'	2
L15191	5'-ATCTGCATCTACCTACACATCGG-3'	3
H15295	5'-CCTCAGAAKGATATYTGNCCTCAKGG-3'	1
H15298	5'-CCCTCAGAATGATATTTGTCTCA-3'	2
L15517	5'-CACGAATCAGGCTCAAACAACC-3'	1
H15545	5'-GTATGGGTGAAATGGAATTT-3'	5
H15742	5'-TGCTAGTACGCCTCCTAGTTTGTGG- GATTGA-3'	4
H16064	5'-CTTCAGTTTTTGGTTTACAAGACC-3'	1

<sup>a</sup> L and H numbers designate the location of the 3' base in the light or heavy strand, respectively, of the published chicken mtDNA sequence (Desjardins and Morais, 1990).

<sup>b</sup> Degenerate primer positions are as follows: K = G or T; M = A or C; N = A, C, G, or T; R = A or G; Y = T or C.

<sup>c</sup> 1, Sorenson *et al.* (manuscript in preparation). 2, Kocher *et al.* (1989). 3, Lanyon and Hall (1994). 4, Harshman (1996). 5, C. E. McIntosh, personal communication.

ing scheme. We used these values to determine weights for each codon position by ln-transforming the inverse of the relative frequencies. We obtained weights of 3.03, 4.56, and 1.00 for first, second, and third codon positions, respectively. We conducted parsimony analyses using these weights in combination with the 1:1, 5:1, and six-parameter weighting schemes.

We used both bootstrap and decay index analyses to determine the support for various nodes of the phylogenies. Bootstrap analyses (Felsenstein, 1985) with 500 "fast" stepwise addition replicates were performed for each gene for the 1:1, 5:1, and six-parameter weighting schemes. Decay indices (Bremer, 1988) were determined using constraint statements generated by Tree-Rot (Sorenson 1996) for each gene region independently and compared to determine if nodes had similar levels of support for both gene regions.

The numbers of variable sites and phylogenetically informative sites were determined using MEGA (Kumar *et al.*, 1993). In addition, the sequences were translated using MEGA and the numbers of amino acid residues that were variable and phylogenetically informative were determined. Differences between cyt *b* and ND2 in the proportion of variable and phylogenetically informative sites were tested with a *z*-statistic approximation test (Milton and Arnold, 1990).

The numbers of transitions and transversions at first, second, and third codon positions for both coding regions were calculated for all pairwise comparisons of taxa using MEGA (Kumar *et al.*, 1993). The number of

transitions and transversions at first, second, and third codon positions in pairwise comparisons of taxa was plotted against the percentage sequence difference (*p*-distance) for both gene regions. We also plotted the number of pairwise transitions at third sites against the number of transversions at third sites. We used the slope of the initial linear portion of these plots to approximate the native transition to transversion ratio at third sites (Sturmbauer and Meyer, 1992; Funk *et al.*, 1995). In addition, pairwise *p*-distances for ND2 were plotted against those of cyt *b* to compare rates. If rates are equal across the included taxa, these points should fall along a straight line with a slope of one. Regression analyses of these plots were not performed because of nonindependence of the many pairwise comparisons involving a given taxon.

We used the tree resulting from 1:1 weighting of the two genes to reconstruct the numbers of various base and amino acid substitutions using MacClade (Maddison and Maddison, 1992). The inverse of the relative frequencies of these changes (averaging values for forward and reverse changes, e.g., A to G and G to A) was used as a weight in the six-parameter weighting method (Williams and Fitch, 1989) after being ln-transformed and corrected for the triangle inequality using PAUP\* (see Table 3). We used an arbitrary resolution of ambiguous nodes of the equally weighted strict consensus tree to reconstruct these changes. The proportion of substitutions that were transversions was determined for each site for both coding regions and compared between sites and genes using *z*-statistic tests on proportions (Milton and Arnold, 1990).

The frequency distribution of the number of steps per site was determined using MacClade (Maddison and Maddison, 1992). Among site rate variation was tested for each gene region using the test statistic developed by Wakeley (1993). The distribution of steps among sites determined above was used as the data for this test statistic. We performed this test for all sites, for first, second, and third sites independently, for transversions only at first, second, and third sites, and for amino acid changes.

**TABLE 3**  
**Six-Parameter Step Matrix: Weights<sup>a</sup>**

Change from	Change to			
	A	G	C	T
A	0	1	3.2	4.2
G	1	0	4.6	5.6
C	3.2	4.6	0	1.3
T	4.2	5.6	1.3	0

<sup>a</sup> Matrix prior to automatic correction by PAUP\*.

## RESULTS AND DISCUSSION

### *Sequence Variation*

Comparisons of sequence variation provide a first estimate of the pattern of coding sequence evolution. In the *cyt b* data set, 387 of 1047 sites were variable and 297 were phylogenetically informative. There were 47 variable amino acid residues in *cyt b* and 23 of these were phylogenetically informative. In the ND2 data set, 403 of 1041 sites were variable and 326 were phylogenetically informative. There were 52 variable amino acid residues and 34 of these were phylogenetically informative. There were no statistically significant differences ( $p > 0.05$ ) between *cyt b* and ND2 in any of these proportions. Although *cyt b* and ND2 are thought to differ in level of amino acid constraint (Meyer, 1994; Mindell and Thacker, 1996), the lack of strong differences among the Anatid taxa suggests that differences between these two genes do not become apparent until there is a high level of divergence (greater than the maximum divergence in this study of 13% for both genes).

### *Phylogeny Reconstruction and Weighting*

The partition homogeneity test with equal weighting indicated that *cyt b* and ND2 did not represent significantly different partitions of the data ( $p = 0.21$ ), so combining these two gene regions is justified according to the rationale of Bull *et al.* (1993) and Farris *et al.* (1995). That is, differences between the *cyt b* and ND2 phylogenies can be attributed to random sampling error rather than bias (Swofford *et al.*, 1996). Partition homogeneity tests using codon positions as partitions also did not show significant conflict ( $p = 1.0$  for all three partitions;  $p = 0.78$  for first and second sites pooled versus third sites).

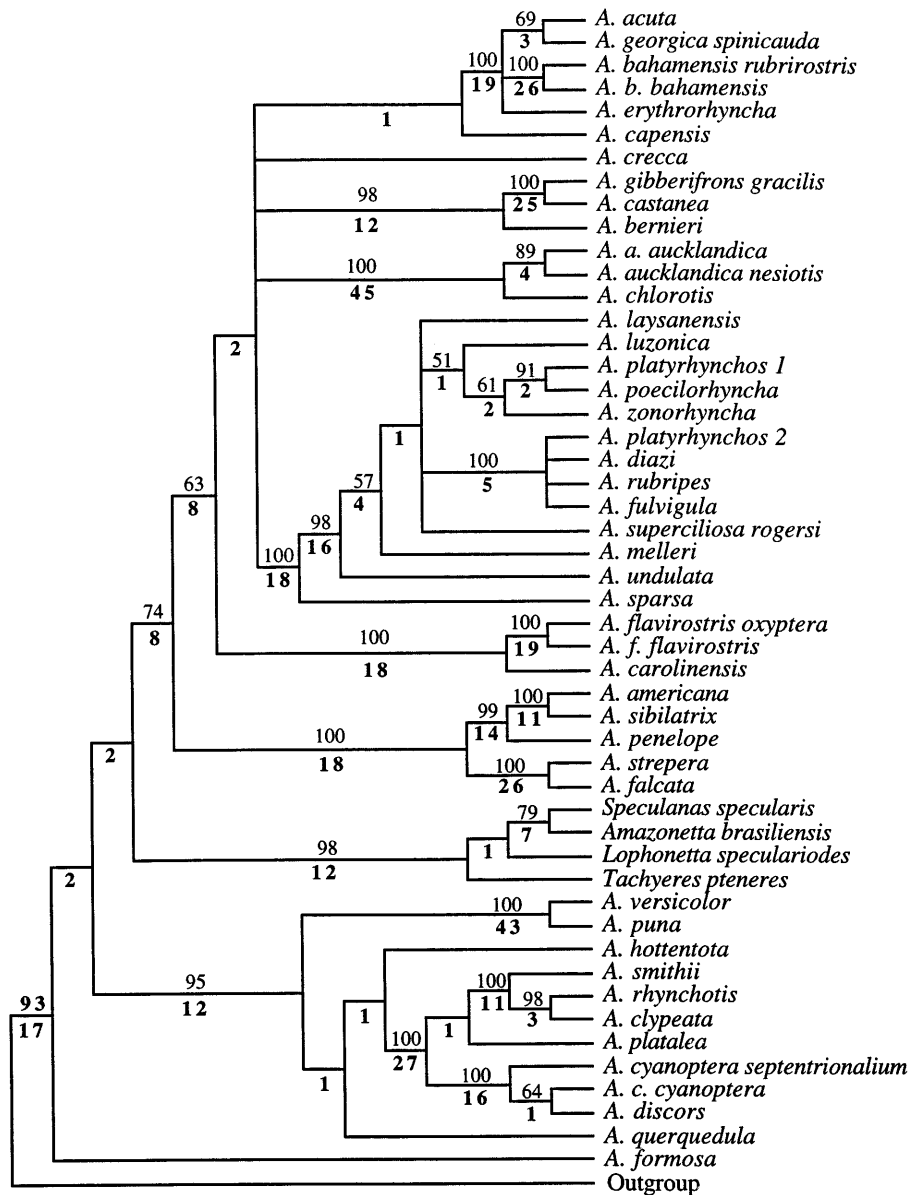
The two gene regions and various weighting schemes resulted in differing topologies, but all analyses yielded topologies that were similar in many respects. The strict consensus tree resulting from the equally weighted (1:1) parsimony analyses is shown in Fig. 1. This tree includes many of the traditional groupings of dabbling ducks. To show the impact of weighting on node stability, the strict consensus of the 1:1, 2:1, 5:1, 9.5:1, 15:1, six-parameter, and codon position weighting schemes is shown in Fig. 2. Nodes that are present in all seven weighting schemes are retained in this tree. Many nodes are insensitive to weighting scheme. In fact, the 3:1, 9.5:1, and 15:1 topologies for the ingroup are identical to the 5:1 topology. In separate analyses of gene regions, differences in tree topology were greater between *cyt b* and ND2 than among weighting schemes. Within a gene, the topology tended to be fairly stable as weighting was increased (42 out of 50 unambiguous nodes retained for ND2 from 1:1 to 5:1 and 35 out of 44 unambiguous nodes retained for *cyt b* from 1:1 to 5:1).

Pairwise symmetric-difference distances [the number of nodes that are present in one tree but not both, the partition metric of Penny and Hendy (1985)] between select pairs of trees are shown in Table 4. Between genes the 5:1 and six-parameter weighting schemes produced the most similar topologies, while 1:1 and site weighting produced more dissimilar trees. This suggests that transversion weighting produces more congruence in phylogenies derived from these different gene regions.

### *Sequence Evolution*

Both genes exhibited similar levels of homoplasy when reconstructed over the combined 1:1 tree. On the 1:1 tree, the rescaled consistency indices (RC) were 0.20 for *cyt b* and 0.21 for ND2. The numbers of transitions and transversions at different codon positions are plotted against the total percentage difference in pairwise comparisons for both gene regions in Figs. 3a–3f. For both genes, third position transitions (Figs. 3a and 3d) accumulate rapidly in a linear manner and then appear to level off very slightly at high divergences (12–13%). Third position transversions accumulate slowly at distances less than approximately 7% but then begin to accumulate more rapidly thereafter. First and second position transitions and transversions accumulate at a much lower rate than at third positions. First position transitions accumulate most rapidly followed by similar rates for first position transversions and second position transitions and a slower rate for second position transversions. The plots of changes at first positions as well as transversions at third positions show no evidence of leveling but continue to increase at higher divergences. It is difficult to determine the general pattern of accumulation at second positions because there is very little accumulation of differences at this position. There is no evidence from this analysis that third position transitions saturate more slowly in ND2 than they do in *cyt b* as suggested by Hackett (1996).

Plots of third site transitions versus third site transversions are shown for *cyt b* and ND2 in Figs. 4a and 4b, respectively. The approximate slope of the linear (unsaturated) portion of the curves from the origin to the horizontal leveling for both gene regions is 15. This suggests that the “native” transition to transversion ratio at third sites is very high (around 15:1) and that transversions should be heavily weighted over transitions, at least at third sites (Sturmbauer and Meyer, 1992; Funk *et al.*, 1995). This provides additional justification for transversion weighting of both gene regions; however, transversion weighting of 3:1 and higher did not change the topology of the ingroup. We prefer weighting to total exclusion of third position transitions because there is still information in third position transitions that is especially useful for resolving relationships between species of low sequence

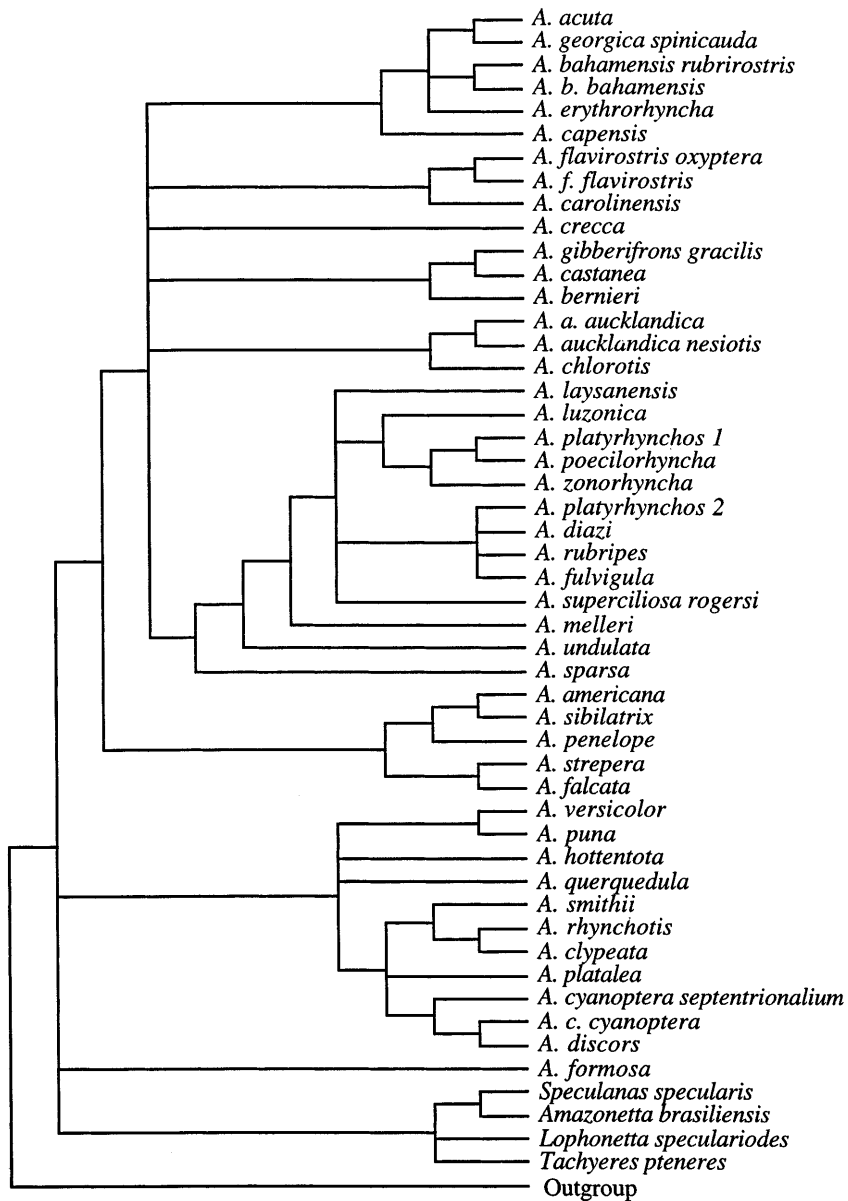


**FIG. 1.** Strict consensus of 18 most parsimonious trees resulting from the combined *cyt b* and ND2 coding regions and 1:1 weighting of transversions and transitions ( $l = 3051$ ). Designated outgroup taxa (used only for rooting the ingroup and collapsed as "Outgroup" in the tree) are *Marmaronetta*, *Pteronetta*, *Cyanochen*, *Aythya*, *Asarcornis*, *Chenonetta*, *Callonetta*, *Tadorna*, *Cairina*, *Aix*, and *Sarkidiornis*. The genus *Anas* is abbreviated by *A.* Numbers above the branches are bootstrap values from 500 "fast" replicates and numbers below the branches (in bold) are decay indices.

divergence. This information should not be excluded from an analysis. The plot of percentage sequence divergence in cytochrome *b* versus ND2 (Fig. 5) shows a linear trend with a slope very near 1 indicating that the two genes are evolving at roughly the same rate at these levels of divergence.

Analyses of changes at various sites were performed using the unweighted combined tree. The accuracy of these estimates is likely to be higher with this data set than those using a few widely divergent taxa (e.g., Mindell and Thacker, 1996) because homoplasy is more

likely to be recovered by dense sampling of taxa. Across sites and combining *cyt b* with ND2, the relative frequencies of the four bases are A (0.276), C (0.358), G (0.140), and T (0.225). As expected the ratio of transitions to transversions is very high: 9.3 for *cyt b* and 9.5 for ND2 (for results of analysis with 9.5:1 weighting, see above). We suggest that these ratios are lower than the 15:1 estimated using the Sturmbauer and Meyer (1992) method because parsimony reconstruction of changes does not recover all of the homoplasy present at third positions. To illustrate this, we also recon-



**FIG. 2.** Strict consensus of 1:1, 2:1, 5:1, 9.5:1, 15:1, six-parameter, and position-weighted trees from the combined *cyt b* and ND2 coding regions. Tree rooted as in Fig. 1.

structured changes in the outgroup taxa only, the ingroup only, and ingroup clade (*A. falcata*–*A. acuta*). In these analyses, the transition to transversion ratio for *cyt b* increased from 7.5, to 11.5, to 12.5 respectively. Likewise the ratio increased from 7.1, to 12.4, to 13.6 for ND2. This suggests that greater density of taxon sampling leads to higher estimates of the transition/transversion ratio (ts/tv ratio) and approaches the estimates obtained from pairwise distance data. The most frequent transitions are A to G and C to T (combined 64.7% of all changes). This corresponds to the relative frequencies of these bases (A and C being more frequent than G or T). Coding region transver-

sions also tend to follow this pattern with the most frequent pairwise transversions corresponding to a change from the most frequent base to the less frequent base (this is true for all transversions except G to T and T to G which are the rarest type of changes composing only 0.53% of all changes). The most common transversions are C to A and A to C (5.6% of all changes). There is little difference between the two genes in the transition ratio at third sites (see Table 5 for ratios of transitions to transversion across sites and genes). The base composition is even more biased away from G and T at third positions in both genes. When only third positions are considered there is a bias toward changes

TABLE 4

Symmetric-Difference Distances between Cyt *b* and ND2 Trees<sup>a</sup>

Weighting scheme	All taxa	Ingroup only
1:1	48–54	34–38
2:1	44–52	32–38
5:1	40–44	28–32
15:1	46–50	34–38
Six-parameter	40–43	28–31
Site weights	54–55	36–37
Site weights, 5:1	42–46	30–34
Site weights, six-parameter	52–53	36–37

<sup>a</sup> Ranges of pairwise symmetric-difference distances between all-most parsimonious trees found with each weighting scheme.

away from the most frequent base to the less frequent base across all transitions and transversions except for G to C transversions in both cyt *b* and ND2.

As expected there is a strong codon position site bias in the number of changes reconstructed. For cyt *b*, there are 148, 34, and 1347 changes at first, second, and third codon positions, respectively. For ND2 there are 196, 34, and 1285 changes at first, second, and third codon positions, respectively. The ts/tv ratio at third sites does not differ significantly between cyt *b* and ND2 ( $P = 0.50$ ). At first sites ND2 has a slightly higher ts/tv ratio than cyt *b* ( $P = 0.08$ ). Within cyt *b* there is a significantly higher ts/tv ratio at third sites than at first sites ( $P = 0.02$ ). Within ND2, however, there is a higher ts/tv ratio at first sites although this is not significant ( $P = 0.64$ ). In general, the relative ratios of transitions and transversions occurring in cyt *b* and ND2 are extremely similar in our study taxa and lend further support to the conclusion that these genes have very similar modes of evolution at low divergences.

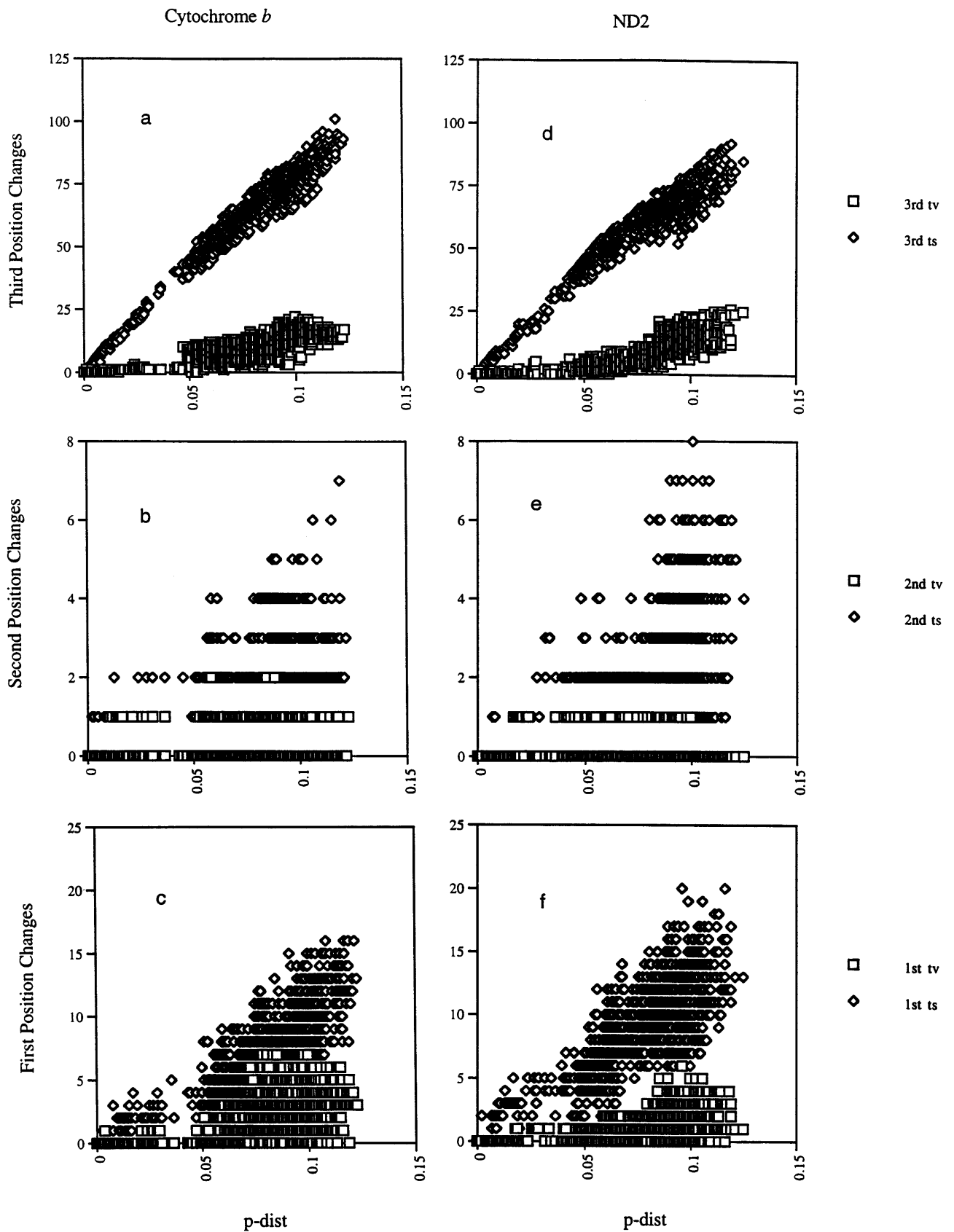
Relative to the number of changes in DNA nucleotide sequences, there is relatively little change in amino acid residues. Cyt *b* exhibits 121 changes and ND2 shows 166. The most common amino acid substitutions for both genes are between isoleucine and valine, accounting for 29.5% of substitutions in cyt *b* and 24.6% in ND2. The next most frequent substitution is between alanine and threonine (13.2% in cyt *b* and 19.9% in ND2). Isoleucine and valine are biochemically and functionally similar; however, alanine and threonine are not (Conn *et al.*, 1987). Both of these amino acid substitutions involve A–G transitions at first sites. At first sites, G is much more common than at third positions (21.7% at first sites vs 9.2% at third sites in cyt *b* and ND2). The similarity of amino acid substitution between cyt *b* and ND2 is very striking given the functional differences between these two genes. This suggests that the most common amino acid changes may be due to similarity in base substitution bias and not differences in amino acid constraints. Large differ-

ences in amino acid substitution patterns may become apparent only when the time since divergence is large (Russo *et al.*, 1996).

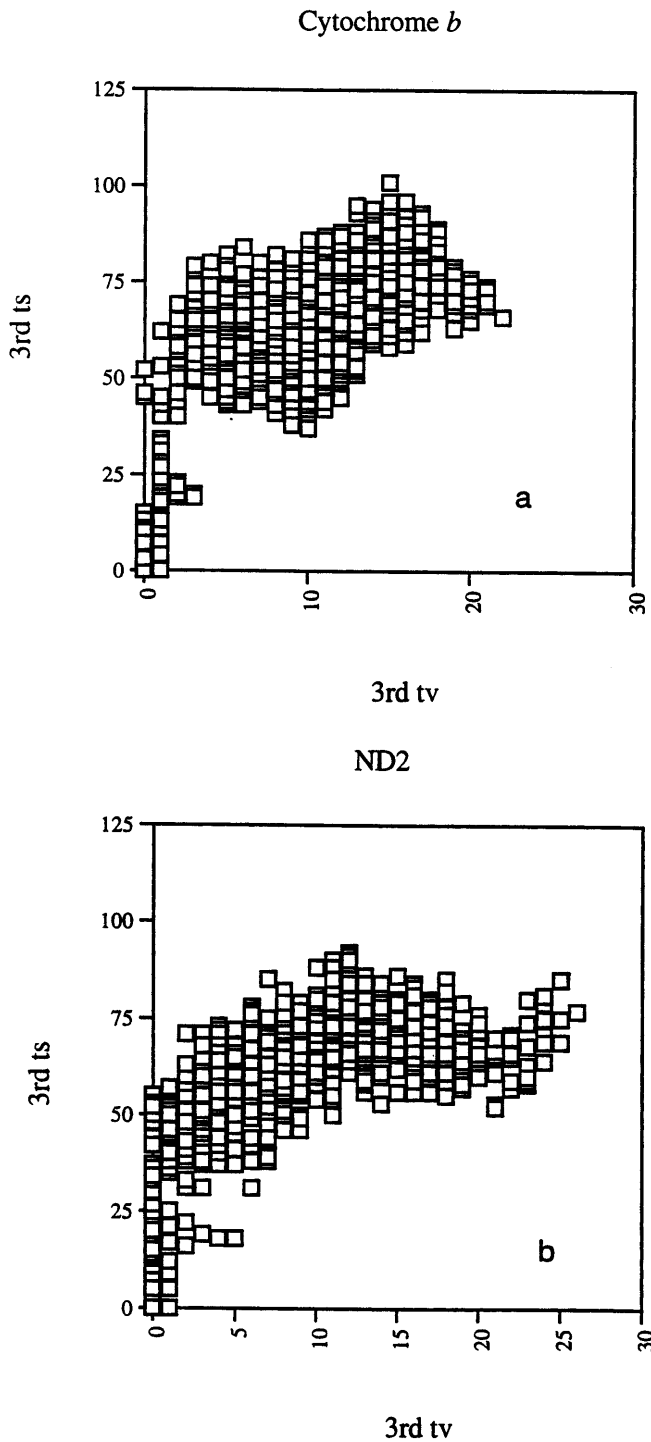
## Tree Support

Bootstrap analyses (Felsenstein, 1985) of the combined data (see Fig. 1) revealed that nodes with low support tended to have low support regardless of weighting scheme. These nodes were also the nodes that tended to vary between weighting schemes. In addition, 18 out of 26 nodes that had higher than 75% bootstrap support in the equally weighted cyt *b* tree also appeared at the 75% level in the equally weighted ND2 tree (not shown). In no case was there a node supported in more than 75% of bootstrap replicates for either gene that had an alternative arrangement supported by more than 75% of bootstrap replicates in the other gene region. This suggests that resolution provided by both of these gene regions is very similar, and that phylogenetic conflict between these two genes is centered around weakly supported nodes. Nonsignificant results in partition homogeneity tests also suggest that differences between gene region topologies are due to sampling error and not real conflict. There are seven lineages (*A. acuta* through *A. carolinensis* in Fig. 1) whose relationships are not resolved by bootstrap analyses in any of the weighting schemes (with one exception: *A. capensis* falls with low support at the base of the *A. acuta* through *A. erythrorhyncha* clade with transversion weighting). Resolution of the relationships among these well supported groups was highly dependent on weighting scheme. In contrast, 50% bootstrap topologies were in general very similar among weighting schemes suggesting that for many nodes there is strong support irrespective of weighting scheme and gene used. There are only 5 nodes that are resolved in the strict consensus of seven weighting schemes (Fig. 2) that are not also resolved in greater than 50% of bootstrap replicates with equal (1:1) weighting (Fig. 1).

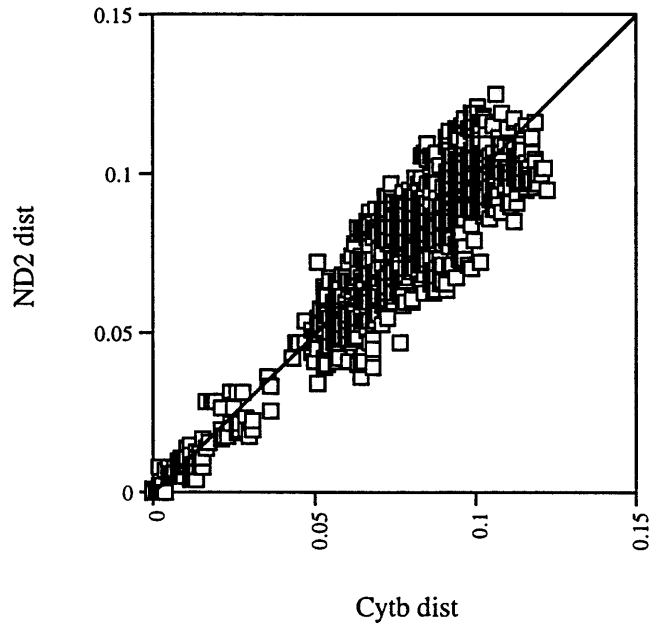
We also examined the congruence between gene regions using decay indices (Bremer, 1988). We calculated decay indices for nodes that were present in either the 1:1 cyt *b* or ND2 trees (see Fig. 6). Nodes that were present in both trees tended to have a high decay index in both trees, while nodes that were present in only one tree tended to have a low decay index for the tree in which they were present. In addition, for many of these nodes, the number of additional steps necessary to obtain the node in the tree derived from the other gene region was generally very small (shown as a “negative decay index” in Fig. 6). These results again suggest that nodes with high support using one gene region also tend to have high support using the other gene region. In contrast, weakly supported nodes in trees generated from one gene region often did not appear in trees derived from the other gene region.



**FIG. 3.** Plots of pairwise transition and transversion distances against total percentage sequence divergence by gene and codon position (a–f). Coding regions of *cyt b* and ND2 are shown separately. Abbreviations: ts, transitions; tv, transversions; p-dist, percentage pairwise difference. The number of pairwise substitutions is shown on vertical axes.



**FIG. 4.** Plots of third position transitions against third position transversions for *cyt b* and ND2. The approximate slope of the relationship for both genes was derived from point with less than five transversion differences. Abbreviations: ts, transitions; tv, transversions.



**FIG. 5.** Plot of pairwise percentage divergence between taxa for *cyt b* and ND2. A line of equal rates (slope of 1) is shown for comparison.

*Codon Position Weighting and within-Genes Rate Heterogeneity*

If one is interested in providing the best estimate of phylogeny for weakly supported nodes, then a tree based on a weighting scheme that models the process of molecular evolution most closely would be desired. However, a reasonably accurate model may be extremely difficult to obtain. For most genes, transitions occur more frequently than transversions and third positions evolve more rapidly than first and second positions in protein coding genes (Li and Graur, 1991). Codon position weights of 3.03, 4.56, and 1.0 for first, second, and third sites, respectively, were used to determine the effect of site position weighting on tree topology. Shortest trees ( $I = 4003.46$ ) were found in 31

**TABLE 5**

**Transition:Transversion Ratios over Combined Unweighted Tree<sup>a</sup>**

Position	Transitions:Transversions	
	Cytochrome <i>b</i>	ND2
All sites coding	9.36	9.50
First sites	5.73	10.40
Second sites	10.33	33.00
Third sites	9.99	9.13

<sup>a</sup> First sites had a marginally significant higher ts/tv ratio ( $P = 0.08$ ) in ND2 than in *cyt b*, but ts/tv ratio at third sites did not differ ( $P = 0.50$ ).

of 200 random addition replicates and the consensus of these three trees (not shown) is very similar to that of trees generated by alternative weighting schemes (Fig. 2). While site position weighting in combination with other weighting schemes (5:1 and six-parameter) did not drastically alter tree topologies, there was less congruence between gene regions when weighting by codon position was used (see Table 4).

While the logic behind transversion and codon position weighting is clear, simple uniform weighting schemes fail to reflect variation in the rate and mode of evolution among individual nucleotide positions. For both *cyt b* and ND2 there is considerable rate variation among nucleotide positions [using test developed by Wakeley (1993); Table 5,  $P < 0.01$  for both gene regions]. This is not unexpected because third sites are known to have higher rates of substitution than first or second sites. We also found that within first, second, or third codon positions there is significant rate variation ( $P < 0.01$ ). This suggests that positional effects within the protein may also play an important role in determining the rate of base substitution as has been documented for *cyt b* in mammals (Irwin *et al.*, 1991) and birds (Griffiths, 1997). This rate heterogeneity is significant for third positions even when only transversions are considered (Table 6). The average number of transversions at first and second codon positions was extremely low and significant rate heterogeneity was not

TABLE 6

Results of Rate Homogeneity Test (Wakeley, 1993)					
Region	Sites	Mean no. substitutions	$f$ statistic	Mean no. tvs.	$f$ statistic
ND2	All coding	1.46	87.2**	0.13	3.6**
ND2	First	0.56	30.1**	0.05	0.21
ND2	Second	0.10	5.5**	0.003	0.00
ND2	Third	3.70	28.2**	0.35	3.9**
ND2	Amino acids	0.48	30.8**		
<i>Cyt b</i>	All coding	1.46	100.4**	0.14	4.9**
<i>Cyt b</i>	First	0.42	19.1**	0.06	1.6
<i>Cyt b</i>	Second	0.10	9.0**	0.009	-0.001
<i>Cyt b</i>	Third	3.82	39.9**	0.34	4.9**
<i>Cyt b</i>	Amino acids	0.35	24.4**		

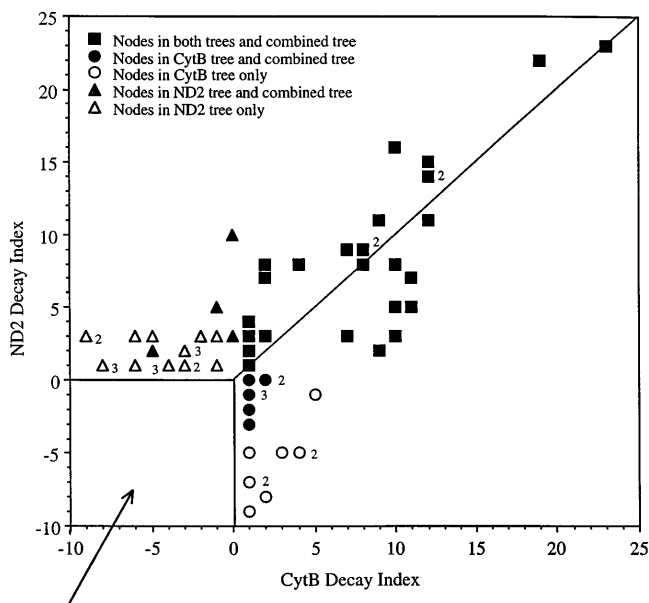
\*  $P < 0.01$ .

\*\*  $P < 0.001$ .

detected, although the low sample size of transversions at these positions reduces the power of the test to detect significant rate heterogeneity (Wakeley, 1993).

These results pose a difficult problem for attempts to determine a reasonable weighting scheme based on rate variation. There is no way to determine a priori which sites are evolving at a faster rate. Only methods that weight sites individually according to their rate of evolution such as iterative reweighting (Farris 1969) and parsimony with implied weights (Goloboff, 1993) have the potential to fully reflect rate variation among sites. While Goloboff's method addresses the problem of circularity inherent in iterative reweighting, neither method (as currently implemented) allows for differential weighting of transitions and transversions. In addition, it is possible that rate and mode of evolution at an individual site may change through time as certain base or amino acid substitutions compensate for others. Thus, it seems that no weighting scheme will be able to account fully for variation in both rate and transition/transversion bias among sites. The general problem as we see it is the difficulty of evaluating the phylogenetic "error" introduced by the necessary approximation inherent in a simplified weighting scheme. For this reason, we use a variety of weighting schemes in a heuristic manner only to explore the sensitivity (or lack thereof) of inferred phylogenetic relationships to weighting.

Our fine-scale comparison of two mitochondrial protein coding genes (*cyt b* and ND2) suggests that they evolve at very similar rates when the time of divergence has been small. The two most common types of amino acid substitutions were shared between the two genes, suggesting that a large part of sequence evolution is governed by base composition bias and transition/transversion bias. While functional constraints on amino acid sequence may play a large role in preventing the fixation of many mutations at first and second positions, differences in constraints between these two



Values here would indicate the number of additional steps required in each of the datasets for nodes present in neither tree

**FIG. 6.** Plot of decay indices for 1:1 ND2 tree versus decay indices for 1:1 *cyt b* tree. Negative values are the number of additional steps needed to obtain a node found in one gene region with sequence data from the other. The lower left portion of the plot is blank (both negative values) because this is the number of additional steps needed to obtain nodes not found in analyses of either gene region. Numbers indicate multiple points in the same position.

genes are apparently not large enough to be reflected in sequence substitution patterns at this taxonomic level. Similarity between genes in base substitution pattern initially suggests that a realistic weighting scheme could be developed but substantial variation in rates among sites makes uniform weighting schemes problematic. We anticipate that there is still a great deal to learn about how coding gene DNA sequences evolve and how patterns of sequence evolution might inform phylogenetic analyses.

## ACKNOWLEDGMENTS

We thank Mike Lubbock, H. Glynn Young, Marc Woodin, Woody Martin, Murray Williams, Les Christidis, Kevin McCracken, Shannon Hackett, and John Klicka for their assistance in obtaining tissue and feather specimens. Robert Zink, David Mindell, and Robert Fleischer were generous in allowing us to use laboratory space and equipment to conduct the DNA sequencing. We thank S. Lanyon, F. McKinney, K. Omland, J. Weckstein, U. Wittmann, and R. Zink for reviewing earlier versions of the manuscript. David Swofford kindly gave us permission to publish results from a prerelease version of PAUP\*. This work was supported financially by a National Science Foundation Doctoral Dissertation Improvement Grant, Frank M. Chapman Memorial Grants, and Dayton and Wilkie Funds for Natural History Grants to K.J.

## REFERENCES

- Arctander, P. (1995). Comparison of a mitochondrial gene and a corresponding nuclear pseudogene. *Proc. R. Soc. Lond. B* **262**: 13–19.
- Bremer, K. (1988). The limits of amino-acid sequence data in angiosperm phylogenetic reconstruction. *Evolution* **42**: 795–803.
- Bull, J. J., Huelsenbeck, J. P., Cunningham, C. W., Swofford, D. L., and Waddell, P. J. (1993). Partitioning and combining data in phylogenetic analysis. *Syst. Biol.* **42**: 384–397.
- Conn, E. E., Stumpf, P. K., Bruening, G., and Doi, R. H. (1987). "Outlines of Biochemistry," 5th ed., Wiley, New York.
- Cooper, A. (1994). DNA from museum specimens. In "Ancient DNA: Recovery and Analysis of Genetic Material from Paleontological, Archaeological, Museum, Medical, and Forensic Specimens" (B. Herrmann and S. Herrmann, Eds.), pp. 149–165, Springer-Verlag, New York.
- Cumming, M. P., Otto, S. P., and Wakeley, J. (1995). Sampling properties of DNA sequence data in phylogenetic analysis. *Mol. Biol. Evol.* **12**: 814–822.
- Desjardins, P., and Morais, R. (1990). Sequence and gene organization of the chicken mitochondrial genome. A novel gene order in higher vertebrates. *J. Mol. Biol.* **212**: 599–634.
- Farris, J. S. (1969). A successive approximations approach to character weighting. *Syst. Zool.* **18**: 374–385.
- Farris, J. S., Källersjö, M., Kluge, A. G., and Bult, C. (1995). Testing significance of incongruence. *Cladistics* **10**: 315–319.
- Felsenstein, J. (1985). Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* **39**: 783–791.
- Funk, D. J., Futuyma, D. J., Orti, G., and Meyer, A. (1995). Mitochondrial DNA sequences and multiple data sets: A phylogenetic study of phytophagous beetles (Chrysomelidae: *Ophraella*). *Mol. Biol. Evol.* **12**: 627–640.
- Goloboff, P. A. (1993). Estimating character weights during tree search. *Cladistics* **9**: 83–91.
- Graybeal, A. (1994). Evaluating the phylogenetic utility of genes: A search for genes informative about deep divergences among vertebrates. *Syst. Biol.* **43**: 174–193.
- Griffiths, C. S. (1997). Correlation of functional domains and rates of nucleotide substitution in cytochrome *b*. *Mol. Phylogenet. Evol.* **7**: 352–365.
- Hackett, S. J. (1996). Molecular phylogenetics and biogeography of tanagers in the genus *Ramphocelus* (Aves). *Mol. Phylogenet. Evol.* **5**: 368–382.
- Harshman, J. (1996). "Phylogeny, evolutionary rates, and ducks." Unpublished thesis. University of Chicago.
- Holmquist, R., Goodman, M., Conroy, T., and Czelusniak, J. (1983). The spatial distribution of fixed mutations within genes coding for proteins. *J. Mol. Evol.* **19**: 437–448.
- Irwin, D., Kocher, T., and Wilson, A. (1991). Evolution of the cytochrome *b* gene of mammals. *J. Mol. Evol.* **32**: 128–144.
- Jacobs, H., Elliott, D., Math, V., and Farquharson, A. (1988). Nucleotide sequence and gene organization of sea urchin mitochondrial DNA. *J. Mol. Biol.* **202**: 185–217.
- Kocher, T. D., Thomas, W. K., Meyer, A., Edwards, S. V., Pääbo, S., Villablanca, F. X., and Wilson, A. C. (1989). Dynamics of mitochondrial DNA evolution in animals: Amplification and sequencing with conserved primers. *Proc. Natl. Acad. Sci. USA* **86**: 6196–6200.
- Kumar, S., Tamura, K., and Nei, M. (1993). "MEGA: Molecular Evolutionary Genetics Analysis, Version 1.02." Pennsylvania State University, University Park, PA.
- Lanyon, S. M., and Hall, J. G. (1994). Reexamination of barbet monophyly using mitochondrial DNA sequence data. *The Auk* **111**: 389–397.
- Li, W.-H., and Graur, D. (1991). "Fundamentals of Molecular Evolution," Sinauer, Sunderland, MA.
- Maddison, W. P., and Maddison, D. R. (1992). "MacClade: Analysis of Phylogeny and Character Evolution," Version 3, Sinauer, Sunderland, MA.
- Meyer, A. (1994). Shortcomings of the cytochrome *b* gene as a molecular marker. *TREE* **9**: 278–280.
- Milton, J. S., and Arnold, J. C. (1990). "Introduction to Probability and Statistics," 2nd ed., McGraw-Hill, New York.
- Mindell, D. P., and Thacker, C. E. (1996). Rates of molecular evolution: phylogenetic issues and applications. *Annu. Rev. Ecol. Syst.* **27**: 279–303.
- Quinn, T. W. (1992). The genetic legacy of Mother Goose Q phylogeographic patterns of Lesser Snow Goose *Chen caerulescens* maternal lineages. *Mol. Ecol.* **1**: 105–117.
- Penny, D., and Hendy, M. D. (1985). The use of tree comparison metrics. *Syst. Zool.* **34**: 75–82.
- Russo, C. M., Takezaki, N., and Nei, M. (1996). Efficiencies of different gene and different tree-building methods in recovering a known vertebrate phylogeny. *Mol. Biol. Evol.* **13**: 525–536.
- Sorenson, M. D. (1996). "TreeRot." University of Michigan, Ann Arbor.
- Sorenson, M. D., Ast, J. C., Dimcheff, D. E., Yuri, T., and Mindell, D. P. Primers for a PCR-based approach to complete mitochondrial genome sequencing in birds and other vertebrates. [Manuscript in preparation]
- Sorenson, M. D., and Fleischer, R. C. (1996). Multiple independent transpositions of mitochondrial DNA control region sequences to the nucleus. *Proc. Natl. Acad. Sci. USA* **93**: 15239–15243.
- Sorenson, M. D., and Quinn, T. W. (1998). Numts: A challenge for avian systematics and population biology. *Auk* **115**: 214–221.
- Sturmbauer, C., and Meyer, A. (1992). Genetic divergence, speciation, and morphological stasis in a lineage of African cichlid fishes. *Nature* **358**: 578–581.

- Swofford, D. L. (1997). "PAUP\*: Phylogenetic Analysis Using Parsimony," Version 4.0, Sinauer, Sunderland, MA.
- Swofford, D. L., Olsen, G. J., Waddell, P. J., and Hillis, D. M. (1996). Phylogenetic inference. *In* "Molecular Systematics," 2nd ed." (D. M. Hillis, C. Moritz, and B. K. Maple, Eds.), pp. 407–514, Sinauer, Sunderland, MA.
- Wakeley, J. (1993). Substitution rate variation among sites in hyper-variable region 1 of human mitochondrial DNA. *J. Mol. Evol.* **37**: 613–623.
- Williams, P., and Fitch, W. (1989). Finding the minimal change in a given tree. *In* "The Hierarchy of Life" (B. Fernholm, K. Bremer, and H. Jornvall, Eds.), Elsevier, Amsterdam.