

Basal Divergences in Birds and the Phylogenetic Utility of the Nuclear RAG-1 Gene

Jeff G. Groth and George F. Barrowclough

Department of Ornithology, American Museum of Natural History, New York, New York 10024

Received April 30, 1998; revised September 22, 1998

The single-copy RAG-1 gene is found throughout higher vertebrates and consists of a single 3.1-kb exon without intervening introns. A 2.9-kb region of the RAG-1 locus was sequenced for 14 basal taxa of birds plus the crocodylian outgroups *Alligator* and *Gavialis*. Phylogenetic analysis of the sequences supported the hypothesis that the deepest evolutionary split in extant birds separates paleognaths from neognaths. A deep division among neognaths separates the chicken- and duck-like birds (“galloanserines”) from a clade consisting of all other birds (“plethornithines”). The relationships of these three basal clades in Aves were supported by high bootstrap (98 to 100%) and large decay index values (above 14). Additionally, the plethornithine clade is characterized by a 15-bp (five-codon) synapomorphic deletion relative to all other birds. RAG-1 evolves slowly, with a number of properties favoring its phylogenetic utility, including rarity of indels, minimal saturation of transition changes at 3rd positions of codons, nearly constant base composition across taxa, and no asymmetry in directional patterns of reconstructed change. However, RAG-1 does not evolve in a clocklike manner, suggesting that this gene cannot easily be used for estimating ages of ancient lineages. © 1999 Academic Press

INTRODUCTION

The Class Aves has been considered one of the best known groups of organisms because nearly all extant species have been discovered and described. However, owing in part to a fossil record that does not include transitional forms among modern birds (Feduccia, 1996), the higher-level systematics within Aves remains controversial. Over the past century, the many disagreements among systematists (summarized in Sibley and Ahlquist, 1990) have generally not concerned boundaries of major divisions (orders) of birds; rather, most controversy has concerned the relationships among orders and the placement of the root node in the phylogeny of living birds.

The voluminous DNA–DNA hybridization studies by

Sibley and colleagues produced a higher-level phylogenetic classification (Sibley *et al.*, 1988) that included hypotheses for the deepest divisions among birds. Based on a “modified” UPGMA clustering of DNA hybridization distances, the first split in their tree separated the Infraclass Eoaves (ratites, tinamous, plus chicken- and duck-like birds) from the Infraclass Neoaves (all other birds). The chicken- and duck-like birds (Galloanserae) were later moved to the Neoaves in a revised classification (Sibley and Ahlquist, 1990). This revision located the avian root node at a position between ratites plus tinamous and all other birds, reflecting the widely accepted division between paleognaths and neognaths first formalized by Pycraft (1900).

Mitochondrial DNA sequencing studies have not corroborated the paleognath–neognath division as the basal partition among extant birds. Using over 13 kb of mitochondrial DNA sequence, Mindell *et al.* (1997) showed that a crocodylian outgroup attached to the avian tree at a passeriform bird (a group conventionally considered derived within Aves). Härlid *et al.*'s (1997, 1998) analyses of complete mitochondrial cytochrome *b* genes also supported placement of passeriforms as the sister group of other birds. The study of Cooper and Penny (1997), using a combination of nuclear *c-mos* (600 bp) and mitochondrial 12S rRNA (390 bp) genes, did not include a crocodylian and could not unambiguously place the root to Aves. These studies demonstrate the persistence of a longstanding problem in avian systematics years after the advent of both modern molecular techniques and explicit character-based phylogenetic methods.

If mitochondrial sequences are evolving too rapidly for effective studies of the ancient evolution of birds, then it is possible that the major obstacle for molecular phylogenetic analyses at this level is the lack of an inventory of slowly evolving nuclear genes. It is our view that, due to a reduction in the potential problems caused by alignment ambiguities, nuclear exons will prove more amenable to objective analysis than introns. Furthermore, nuclear exons are usually short (less than 500 bp) and surrounded by introns of varying lengths (which may not reliably serve as conserved

sites for PCR primers); therefore, if one could locate exons of more than 1 kb, then many of the logistic difficulties associated with PCR amplification over a diverse range of taxa would be avoided.

In this paper, we document the phylogenetic utility of the recombination activating gene (RAG-1) in ordinal systematics of birds. This 3-kb, single-copy gene is found throughout higher vertebrates and consists of one exon uninterrupted by introns (Schatz *et al.*, 1989; Carlson *et al.*, 1991; Bernstein *et al.*, 1996). We show that the RAG-1 gene is a useful tool for higher-level systematic studies of vertebrates.

MATERIALS AND METHODS

Taxon Sampling

We sampled 16 taxa (acronyms indicate collectors' initials or tissue collections; see under Acknowledgments) representing the putative basal splits among generally recognized major groups of birds and crocodylians. For paleognaths, one ratite (ostrich, *Struthio camelus*; PRS 1636) and one tinamou (*Tinamus guttatus*; PEP 2032) were included. Representing the "Galloanserae" were a duck (*Anas strepera*; PRS 976), a screamer (*Chauna torquata*; PRS 1634), a chicken (red junglefowl, *Gallus gallus*; Zuk laboratory, Univ. of California at Riverside), and a megapode (*Megapodius freycinet*; MKL 67). A hemipode (*Turnix hottentotta*, courtesy of T. Crowe) was included to test the hypothesis (Sibley and Ahlquist, 1990) that hemipodes are basal to all birds not previously mentioned. We also included taxa that have a history of being considered basal in Aves, including a loon (*Gavia immer*; PRS 1000), a penguin (*Spheniscus humboldti*; PRS 1606), a shorebird (*Charadrius vociferus*; AMNH AC36), and a crane (*Grus canadensis*; PRS 1381). In addition, we included a coraciiform roller (*Coracias caudata*; PRS 756), an oscine passeriform sparrow (*Passer montanus*; PRS 697), and a subsocine passeriform flycatcher (*Tyrannus tyrannus*; PRS 1298). An alligator (*Alligator mississippiensis*; AMNH OTC73) and a gharial (*Gavialis gangeticus*, Bronx Zoo) constituted outgroups representing the deepest split in extant crocodylians (see Brochu, 1997).

Primer Design, DNA Extraction, and Amplification

An alignment of RAG-1 sequences available on GenBank for chicken (M58530; Carlson *et al.*, 1991), human (M29474; Schatz *et al.*, 1989), mouse (M29475; Schatz *et al.*, 1989), and *Xenopus* (L19324; Greenhalgh *et al.*, 1993) was constructed to search for conserved areas in which to position primers. After some sequences were obtained, new primers were developed as needed, and a total of 36 primers (Fig. 1) was used for amplification and sequencing. From three to six fragments were

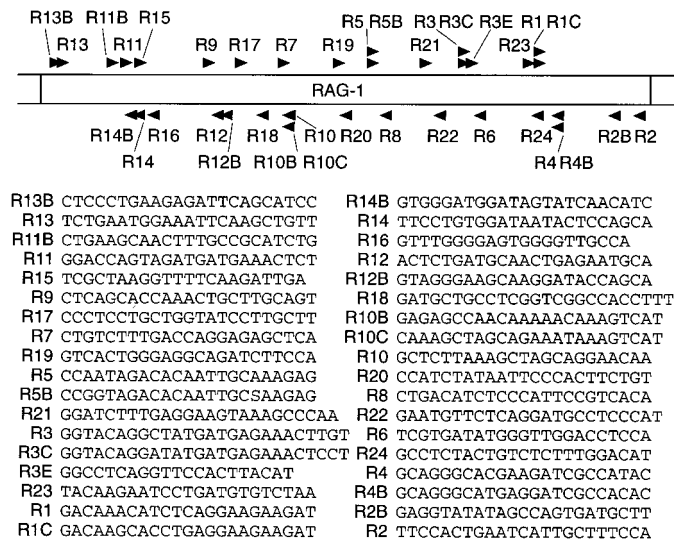


FIG. 1. Names, positions, and sequences of primers used in this study. Sequences are given in the 5' to 3' direction.

amplified independently from the genomic DNA of each taxon, with fragment sizes ranging from 300 to 1200 bp and expected overlap among fragments ranging from 89 to 160 bp.

Total genomic DNAs were extracted using two methods. Blood samples (*Gallus*, *Gavialis*) were extracted by placing 2 μ l whole blood in 300 μ l of 5% Chelex (Bio-Rad) and heating at 100°C for 12 min. All other samples were solid tissue, from which 15–18 mg were digested and extracted using QIAamp tissue kits (Qia-gen) following the manufacturer's instructions, with the exception that the final washed DNA pellet was resuspended in 100 μ l of elution buffer rather than the recommended 400 μ l. DNAs extracted using QIAamp kits were diluted with water by a factor of 20:1 before PCR.

Amplification of target fragments used an Applied Biosystems 9600 thermocycler and Amplitaq Gold kits (Applied Biosystems) following manufacturers' instructions except that total reaction volumes were 20 μ l (instead of 100 μ l) and contained 0.15 μ l enzyme mix (instead of 0.5 μ l). For both the Chelex and diluted QIAamp DNAs, 0.8 μ l was placed in each reaction tube; negative control reactions contained the same amount of water. The two primers in each reaction had final concentrations of 1.0 μ M. The thermocycling procedure (40 total cycles) was a modified hot-start touchdown PCR, with an initial soak at 94°C for 10 min, followed by five cycles of 95°C for 20 s, 61°C for 20 s, 72°C for 1 min. This was followed by two identical 5-cycle phases, with a 2°C reduction in annealing temperature for each phase (59°C, 57°C). The final phase consisted of 25 additional cycles identical to the preceding cycles, but with a 55°C annealing temperature. About one third (5–8 μ l) of each cycled product was run, in the presence

of EtBr, on 2% low-melt TAE agarose gels (NuSieve GTG, FMC). Small portions ("plugs") of desired bands were cut from gels using disposable glass pipets, melted at 73°C for 5 min, diluted with 300 μ l water, and soaked at 73°C for an additional 12 min followed by brief vortexing.

The diluted plug DNAs were used as templates in 30 μ l total volume reamplification PCRs in a hot air thermocycler (Idaho Technology), with 1 U *Taq* (Promega) buffers as suggested by Idaho Technology, and final primer concentrations of 0.6 μ M. These reactions contained either both primers used in the original reactions or one original primer and one internal primer. The thermal profile consisted of 94°C for 3 s, 53°C for 0 s, and 71°C for 22 s. Aliquots (5 μ l) of these reactions were visualized on EtBr-stained agarose gels, and most of the remainder (18 μ l) of each successful reaction was purified using 4 μ l glassmilk and washed using GeneClean (Bio101) kits. Cleaned PCR products were resuspended in 12 μ l water. These products were cycle sequenced (in an ABI 9600 thermocycler) in both directions in 5.5 μ l total volume reactions containing 2.0 μ l cleaned PCR product, 1.0 μ l primer (10 μ M stock), and 2.5 μ l dye-terminator reagent (dRhodamine, Applied Biosystems). Sequenced reactions were cleaned of excess nucleotides using Sephadex columns (Princeton Separations), dried, resuspended in 1.8 μ l formamide-EDTA loading dye, and loaded onto 5% Long Ranger (FMC) gels in an Applied Biosystems 377 automated sequencer.

Data Analysis

Chromatogram output was trimmed of blank ends and primer sequences, assembled as contiguous fragments (contigs) within taxa, and edited using Sequencher software (Gene Codes). The number of separate fragments (600 bp average length) used to construct a contiguous 2.9-kb region ranged from six to eight. All portions of the sequence for each taxon were read at least twice (both directions) and up to four times (overlap zones). Base calling ambiguities on single strands were resolved either by choosing the call on the cleanest strand or using the appropriate standardized IUB ambiguity code if both strands showed the same ambiguity.

Contigs for different taxa were aligned by eye. The final set of aligned contigs was output as a NEXUS file and imported into MacClade (Maddison and Maddison, 1992) for further processing. Phylogenetic analyses were performed using PAUP* 4.0d63 (Swofford, 1998), designating *Alligator* and *Gavialis* as a monophyletic outgroup. Searches for shortest trees were conducted using parsimony, 200 replicate heuristic searches with random taxon addition, TBR branch swapping, gaps coded as missing data, and ambiguity-coded sites as

uncertain. Bootstrap values (Felsenstein, 1985) were estimated using 500 replicate heuristic searches. Decay index values (Bremer, 1994) for each node were obtained by constructing a treefile in MacClade containing an unresolved bush except for the clade of interest (which was also unresolved); this set of treefiles (one file for each clade) was reimported into PAUP*, and each treefile was used as a negative topological constraint (20 full heuristic searches for each treefile) to obtain the shortest trees not containing the clade of interest. A parameter-rich maximum likelihood analysis (in PAUP*) used the general-time reversible (GTR) substitution model with unique probabilities for each of the six possible base transformations, estimation of nucleotide frequencies, estimation of the number of invariable sites (I), and gamma (Γ)-distributed rates at variable sites with estimation of the shape parameter (α); this corresponds to the GTR + I + Γ model of Swofford *et al.* (1996).

To test the hypothesis of a molecular clock, we performed likelihood ratio tests (Felsenstein, 1981) comparing the difference in log likelihoods ($-\Delta\text{Ln}$) of trees with branch lengths free to vary versus trees with the same branching pattern and terminal taxa aligned such that the molecular clock was enforced. The test statistic (which is $-2\Delta\text{Ln}$) can be approximated using the χ^2 distribution with degrees of freedom equal to the number of taxa minus two (see also Yang *et al.*, 1995).

RESULTS

Indels

The primer R2 (Fig. 1) near the 3' end of RAG-1 was initially used to generate sequence, but failed to produce amplification products for some templates. The final region compared among all taxa contained the sequence between primers R13 and R2B; these sequences have been deposited in GenBank (Accession Nos. AF143724–AF143739). The total number of bases between these primers (not including the primer sites) ranged from 2869 (*Grus*) to 2902 (*Turnix*). A total of eight indels was required for the final alignment of 2932 bp; lengths of all indels were multiples of 3 bases (ranging from 3 to 33 bp), and all occurred no more than 718 bp from the 5' end of the sequenced region. Only one indel was not autapomorphic (i.e., found in a single taxon); this indel (about 150 bp downstream from primer R13) consisted of 15 nucleotides (5 codons) missing in *Grus*, *Gavia*, *Spheniscus*, *Turnix*, *Charadrius*, *Coracias*, *Passer*, and *Tyrannus*. A long stretch of 33 extra bases (corresponding to 11 codons) was found in *Turnix* at a position 80 bp from primer R13. No stop codons were observed in the sequenced region for any taxon.

Heterozygosity

Forty-two sites out of the set of 46,114 compared nucleotides were coded as ambiguous. Each of these sites was given an IUB code corresponding to a two-base ambiguity (no three- or four-base ambiguities were scored). These sites were double checked on the chromatograms corresponding to both forward and reverse directions, and in no case could these polymorphisms be attributed to recognizable sequencing artifacts. Furthermore, the hypothesis that these ambiguities were the result of cross-contamination was rejected because only rare sites contained the ambiguities; more would have been expected given the much higher level of divergence among taxa in this study (see below). Given that these sequences were generated by PCR and not by cloning, it is reasonable to conclude that these polymorphisms correspond to differences between two alleles within individuals that were amplified simultaneously. The total numbers of base differences between the two putative alleles within individuals ranged from zero (five taxa) to nine (0.31%; *Coracias*), with a mean of 2.62 (0.09%). Of the 42 polymorphic sites, 12 were at 1st positions of codons, 3 at 2nd positions, and 27 at 3rd positions; assuming the two alleles within a taxon share a recent common ancestry, 30 polymorphisms represent transitions and 12 represent transversions.

Sequence Divergence

The lowest divergence between taxa, as measured by uncorrected ("p") sequence distance, was 2.49% (*Alligator* vs *Gavialis*). Distances between pairs of bird taxa ranged from 3.06% (*Spheniscus* vs *Charadrius*) to 12.60% (*Tinamus* vs *Gallus*), between birds and crocodylians ranged from 15.36 to 17.65%, and between the new sequences and two previously reported mammalian sequences (*Mus* and *Homo*; see Materials and Methods) ranged from 24.57% (*Gavialis* vs *Homo*) to 28.59% (*Chauna* vs *Mus*). The two mammal sequences were 13.78% distant from each other. More than one-third (36.43%) of the sequenced sites were variable and nearly one-quarter (23.02%) were phylogenetically informative in this dataset (Table 1). Numbers of variable and informative characters at 3rd positions were ap-

TABLE 1

Tree Statistics and Homoplasy Indices at Codon Positions

Position	Number of variable characters	Number of informative characters	Tree length ^a	CI ^a	RI
First	233 (23.8%)	135 (13.8%)	370 (270)	0.738 (0.641)	0.627
Second	155 (15.9%)	90 (9.2%)	262 (188)	0.721 (0.612)	0.637
Third	680 (69.5%)	450 (46.0%)	1276 (1009)	0.678 (0.593)	0.579
Total	1068 (36.4%)	675 (23.0%)	1908 (1467)	0.696 (0.604)	0.594

^a Numbers in parentheses exclude uninformative characters.

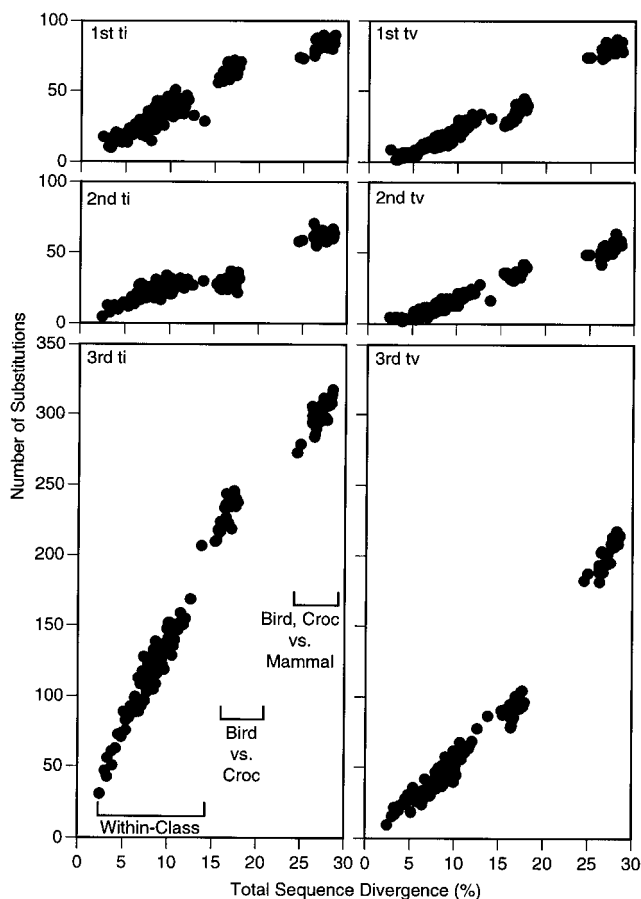


FIG. 2. Plots of numbers of transition (ti) and transversion (tv) substitutions per partition (Y axes) versus total percentage sequence divergence (X axes). Dots represent pairwise comparisons ($n = 153$) among taxa in each partition.

proximately double the number at 1st and 2nd positions combined.

Saturation

To test for degree of saturation (i.e., multiple hits), the observed numbers of transitions (ti) and transversions (tv) at the three codon positions were plotted against total sequence divergence (Fig. 2). In such a plot, a lack of vertical increase between ingroup and ingroup-outgroup comparisons would indicate that substitutions have reached saturation. For example, transitions at 3rd positions of codons in the mitochondrial cytochrome *b* gene of vertebrates become saturated at total sequence divergence levels near 10% (e.g., Moritz *et al.*, 1992). In this study of RAG-1, there was no indication of saturation effects in any partition. Even 3rd position transition substitutions within the bird-crocodylian ingroup do not reach the level seen in pairwise comparisons between these taxa and mammals. These observations suggest no reason for eliminating or downweighting any partition prior to phylogenetic analysis.

Phylogenetic Analysis

All of the 200 replicate heuristic searches in PAUP* found the same single shortest tree (Fig. 3). The longest and best-supported internodal branch was that separating crocodylians from birds. Nodes for six other clades achieved bootstrap values at or near 100% and decay indices of at least 15: (1) the paleognaths (tinamou plus ostrich), (2) neognaths (all other birds), (3) duck- and chicken-like birds, (4) neognaths excluding duck- and chicken-like birds, (5) the galliforms (megapode + chicken), and (6) the passeriforms (flycatcher + sparrow). Except for marginal support for a screamer + duck clade, other clades obtained bootstrap values of less than 50%. The same six well-supported clades and the root position in birds were also resolved using a neighbor-joining analysis with Kimura 2-parameter distances, a parsimony analysis of translated amino acid characters (unordered), and a parsimony analysis using only transversions and with the GTR + I + Γ maximum likelihood model; only minor rearrangements at weakly supported nodes were found in these analyses. We also added *Homo* and *Mus* in a separate parsimony analysis using 18 taxa and found that this did not alter the composition or position of major clades in the ingroup.

Because the position of the root is a major focus of this study, we measured the number of additional steps required (in MacClade) to attach crocodylians at alternate sites on the tree. Forcing the paleognaths plus duck- and chicken-like birds to be monophyletic, with the root separating these taxa from the other birds (as in the trees of Sibley *et al.*, 1988; Sibley and Ahlquist, 1990), requires 25 extra steps. Movement of the root to the base of the passeriforms (as in Mindell *et al.*, 1997)

requires 46 extra steps, and attachment of the root to the base of the duck- and chicken-like birds requires 28 extra steps. None of these alternate rootings is favored by the data.

We also performed a quartets analysis (implemented in PAUP*) that corrects for long branches ("evolutionary parsimony" of Lake, 1987; see also Swofford *et al.*, 1996) to evaluate the statistical significance of attaching crocodylians to the paleognath branch. In this test, taxa were separated into four clades: (1) the crocodylians, (2) the paleognaths, (3) the duck- and chicken-like taxa, and (4) the eight other birds. One taxon from each clade was included in 128 possible combinations ($2 \times 2 \times 4 \times 8$) of these four groups. There are three possible arrangements for each quartet. The outcome of interest is a significantly nonzero value (as measured by χ^2) for one arrangement and near-zero support for the other two. Quartets with a crocodylian-paleognath linkage were significantly supported ($P = 0.018$) and other arrangements were not. Additionally, all 128 combinations found this arrangement to have the shortest length using parsimony. Thus, the placement of the root in Fig. 3 is strongly supported by the data.

Homoplasy and Distribution of Character Change

Mapping of character changes on the preferred topology in Fig. 3 indicated only modest differences among codon positions in levels of reconstructed homoplasy (CI and RI values; Table 1). Most parsimony-informative characters were inferred to have changed twice or more at all three codon positions. The average number of steps per informative character was 2.000 at 1st positions, 2.089 at 2nd positions, 2.252 at 3rd positions, and 2.180 for all positions combined.

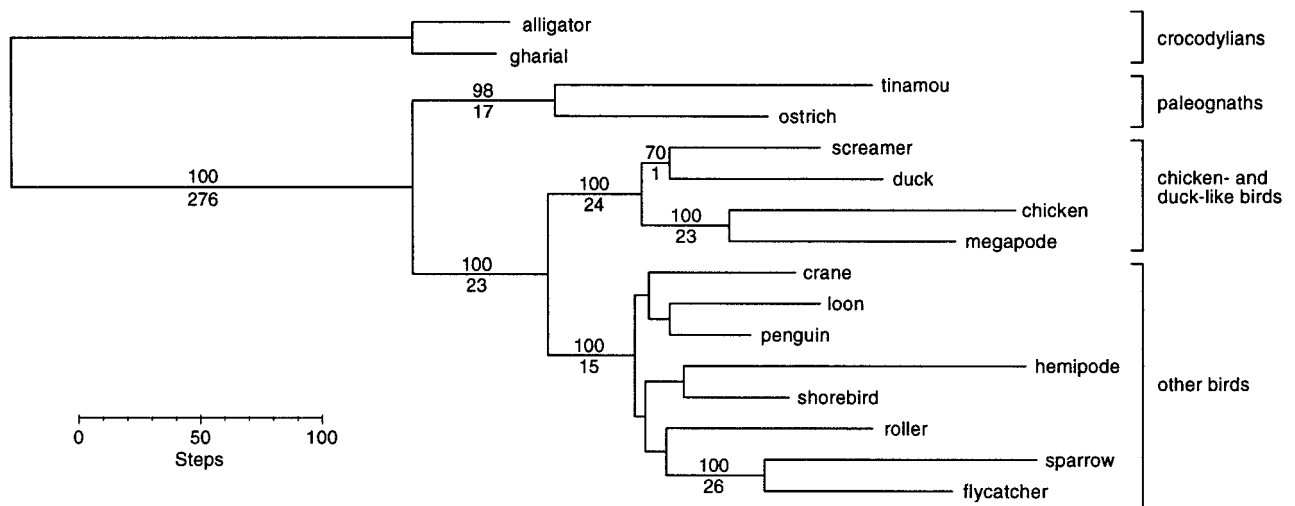


FIG. 3. Single shortest tree (length = 1908 steps) among taxa sequenced for this study. Numbers above internodes are bootstrap percentages and those below are decay index values. Internodes with no values had bootstrap values <50% and decay index values of 1. Branch lengths (scale at lower left) are drawn proportional to average numbers of changes over all reconstructions (using equivocal cycling in MacClade 3.06 [Maddison and Maddison, 1992]).

TABLE 2

**Range of Base Composition (%) at Codon Positions
Among 16 Birds and Crocodylians**

Position	A	C	G	T
First	31.3–33.6	17.9–20.8	28.5–29.7	16.9–18.5
Second	34.0–35.5	18.4–19.6	17.3–18.8	25.7–26.9
Third	25.0–28.7	19.8–23.8	21.3–24.4	25.4–28.5
All	30.6–32.4	19.0–20.7	22.7–24.0	22.8–24.3

The RAG-1 genes in this study exhibit a high degree of base compositional homogeneity across taxa, or stationarity (Saccone *et al.*, 1989; Table 2). A test for departure of homogeneity in base composition across these taxa, as implemented in PAUP*, was not significant ($\chi^2 = 30.877$, $df = 45$, $P = 0.946$). Codon positions, however, varied in their degrees of base compositional bias (Fig. 4A). The 1st positions were biased in favor of adenine and guanine and the 2nd positions were characterized by elevated levels of adenine and thymine residues. These skewed compositional frequencies reflect unevenness in translated amino acid frequencies; average frequencies for lysine (AAR; 8.6%), glutamic acid (GAR; 8.1%), and leucine (CTN or TTR; 9.0%) were high. Unlike the common pattern of deficiencies of T and especially of G at 3rd codon positions of vertebrate mitochondrial DNA, 3rd positions in the RAG-1

gene were more even than 1st and 2nd positions in composition (with slight excess of A and T). The overall pattern for all sites was one in which adenine outnumbered the other nucleotides by factors of about 1.2:1 to 1.5:1.

Patterns of reconstructed character change within codon positions are plotted in Fig. 4A. There is a strong correlation between the frequency of a given nucleotide and the frequency with which it was involved in change (e.g., increased frequency of A \leftrightarrow G changes relative to C \leftrightarrow T at 1st positions, where A and G exceed C and T). Bubble diagrams such as those in Fig. 4A for mitochondrial, and sometimes nuclear (e.g., Pritchko and Moore, 1997), sequence data are often highly asymmetrical. This is due to uneven base composition combined with high rates of change, resulting in inferred (as mapped on phylogenetic trees) directionality of character state changes toward rare states (Collins *et al.*, 1994). However, for these RAG-1 data, the overall pattern of inferred change directionality was highly symmetrical, even though base composition was somewhat biased.

The distribution of numbers of inferred nucleotide changes along the RAG-1 gene is shown in Fig. 4B. A pattern of increased frequency of substitutions at the 5' end of the gene is similar to that previously described for amino acid replacements across taxa in different vertebrate classes (Carlson *et al.*, 1991; Bernstein *et al.*, 1996) corresponding to hypothetical functional regions.

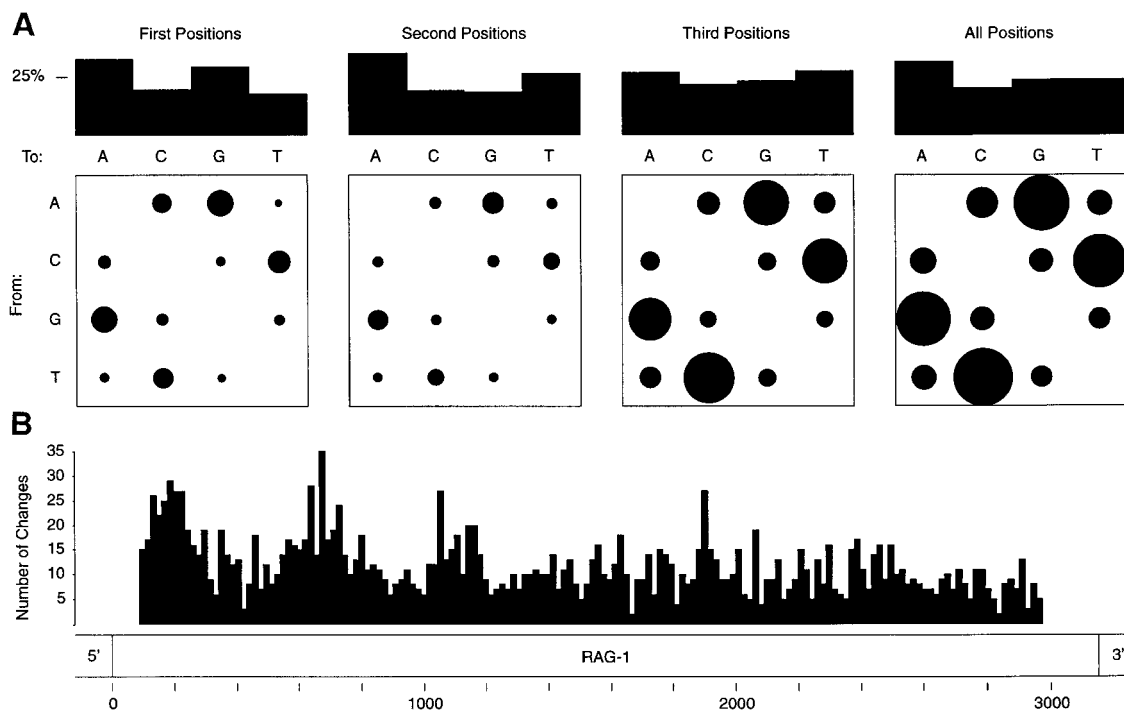


FIG. 4. Patterns of reconstructed character change. (A) Average base frequencies among taxa by codon position (top) and inferred direction of base changes on the shortest tree (equivocal cycling in MacClade [Maddison and Maddison, 1992]); areas of bubbles are drawn proportional to average numbers of changes. (B) Inferred numbers of changes across the sequenced region for the 16 taxa in this study (18-bp window; indels greater than 6 bp removed). Scale at bottom indicates nucleotide position corresponding to the published chicken sequence (Carlson *et al.*, 1991).

The pattern in birds and crocodylians also shows three hypervariable zones in the first 1.2 kb of sequence; however, the overall degree of peakedness across the RAG-1 gene is less than that across cytochrome *b* (e.g., Irwin *et al.*, 1991).

DISCUSSION

Basal Divergence in Modern Aves

This is the first DNA sequencing study of a slowly evolving nuclear gene in birds that employs crocodylians as an outgroup, includes a reasonable sample of taxa among putative basal lineages, and uses more than 1 kb of sequence. Phylogenetic analysis of the RAG-1 gene shows, with strong indices of nodal support, that birds are divisible into three major basal clades. According to the RAG-1 data, paleognathous birds are the sister group of other extant birds, and neognaths are deeply divided into two clades. The paleognath–neognath division has long been recognized on morphological grounds (e.g., Pycraft, 1900) as the primary division in extant birds, and support for a galliform–anseriform clade has been growing (e.g., Ho *et al.*, 1976; Cracraft, 1981; Sibley *et al.*, 1988; Caspers *et al.*, 1997; Livezey, 1997; however, see Ericson [1996] for an alternative view). Thus, our results are not startling (e.g., Sheldon and Bledsoe, 1993), but are the first with robust indicators of nodal support.

The evidence provided by RAG-1 for a clade of neognaths excluding the duck- and chicken-like birds goes beyond high indices of nodal support; an apparently homologous 15-bp deletion (which was not used in tree building or estimating nodal support) was found in all eight taxa in this clade. Sequences of additional avian orders (unpubl. data) all show the same deletion. Parsimony and maximum likelihood analyses of RAG-1 sequence do not support a basal position of *Turnix* in this clade; the long branch to *Turnix* suggests that its basal placement in a dendrogram (e.g., UPGMA, Sibley and Ahlquist [1990]) might arise as an artifact of distance-based clustering procedures.

Our study found strong nodal support both at the base and near some terminals (megapode + chicken and flycatcher + sparrow) of the tree. However, some intermediate internodes were short and received poor support indices. There is no evidence that these short internodes are artifacts resulting from a set of data compromised by homoplasy, because consistency and retention indices for the data are high (Table 1). Nor can this apparent lack of resolution be due to scarcity of characters, because all taxa in this study differed by at least 88 nucleotide sites. Therefore, RAG-1 should provide opportunity for phylogenetic resolution among more terminal groups of birds, such as families or genera. The short internodes in our tree may simply reflect an ancient history of relatively rapid diversification among many avian orders. Additional sequences are needed to investigate this possibility.

These results corroborate immunological distance results (Ho *et al.*, 1976; Prager and Wilson, 1980) for a basal position of paleognaths (using crocodylian outgroups) in avian phylogeny, with monophyly of both a chicken–duck clade and another clade for all other birds. The DNA–DNA hybridization results of Sibley and Ahlquist (1990) also supported three major clades in birds, but their study did not root the avian tree with an outgroup. Although their dendrograms show a sister relationship between the paleognaths and their Parvclass Galloanserae (which suggests nonmonophyly of neognaths), they preferred (p. 255, based on a discussion of the limits of resolution in DNA–DNA hybridization analysis) an alternative rooting with paleognaths as the sister group to all other birds, which is the same as the RAG-1 tree. It is not possible to determine whether the nuclear *c-mos* data (600 bp) of Cooper and Penny (1997) also support the same arrangement because their analysis included neither an outgroup nor some relevant basal taxa (e.g., tinamou and anseriforms). Caspers *et al.* (1997), with approximately 600 bp of nuclear α -crystallin genes, recovered a topology similar to ours but used only five avian taxa.

Our three-clade pattern would be consistent with an unrooted mtDNA sequence network based on weighted data published by Mindell *et al.* (1997; five avian taxa, 13 kb). However, if the RAG-1 results are correct, then their rooting (with *Alligator*) of the avian tree at a passeriform is an anomaly, perhaps caused by homoplasy and excessively long branch lengths associated with sparse taxon sampling. Their unweighted analysis of the same data placed their single paleognath (*Rhea*) as terminal in the tree, which is a result not corroborated by any other study. Although it is difficult to evaluate their results further (the data have not yet been accessioned in public databases), we suggest that rapid evolution of mitochondrial sequences may accrue too much homoplasy for effective analysis at this level.

The deep and strongly supported three-clade basal subdivision of birds deserves recognition in classification and nomenclature. Based on morphological information, Cracraft (1981) recognized nine major clades (divisions) in birds, including divisions for both paleognath and chicken–duck clades; however, these nine divisions were not placed within a hierarchical scheme. Although Cracraft (1988) recognized the monophyly of the non-chicken–duck neognaths based on biochemical work of others, he found no anatomical synapomorphies supporting this group. This clade, containing neognathous birds excluding the chicken–duck clade, is not named in any currently used classification system. The classification of Sibley *et al.* (1988) grouped the chicken–duck clade (Galloanserae) with paleognaths (in their Parvclass Eoaves) and named a Parvclass Neoaves, which corresponds to our clade of interest. However, their more recent use of Neoaves (Sibley and Ahlquist, 1990; Sibley and Monroe, 1990) included the Galloanserae and refers to all neognaths; they provided

no name or hierarchical level for the clade containing the other avian orders. Therefore, there is currently no commonly used or appropriate name for this third clade. We propose the following classification that reflects the phylogeny resolved in this study:

Class Aves: birds

Subclass Neornithes

Infraclass Palaeognathae: ratites and tinamous

Infraclass Neognathae: typical birds

Cohort Galloanserae: gallinaceous birds, screamers, ducks, and allies

Cohort Plethornithae: all other extant birds

This system preserves the paleognath–neognath (Pycraft, 1900) dichotomy in common use among systematists and in modern literature (e.g., Härlid *et al.*, 1997, 1998; Mindell *et al.*, 1997). We suggest the use of “plethornithines” or “plethornithine birds” to reflect the plethora of birds in this large clade. The level of cohort for the galloanserine and plethornithine clades allows the superordinal level to be used for clades of avian orders that might become evident in future studies. We also suggest that the Subclass Neornithes be applied only to the descendants of the common ancestor of all extant birds, following the recommendations of Cracraft (1988).

Rates and Dates

Considerable recent attention has been paid to the timing of ordinal diversification in birds (e.g., Hedges *et al.*, 1996; Cooper and Penny, 1997; Cooper and Fortey, 1998; Rambaut and Bromham, 1998). Effective dating of the origin of lineages relies upon (1) well-supported estimates of phylogeny, (2) molecular data and distances that behave in a clocklike manner, and (3) reliable fossil calibration. Casual inspection of branch lengths on the tree in Fig. 3 suggests substantial variation in relative rate of evolution in RAG-1 across taxa. We tested the hypothesis of clocklike evolution in RAG-1 by using the likelihood ratio test (Felsenstein, 1981). Using the GTR + I + Γ model, the topology in Fig. 3 has a $-\ln$ likelihood of 13335.54084; the score is increased by 103.6728 ($-\Delta\ln$) when the molecular clock is enforced, resulting in rejection of the clock model ($-2\Delta\ln = 207.3456$, $df = 14$, $P < 0.0001$). Similarly, we removed crocodylians and tested for clocklike behavior among birds only ($-2\Delta\ln = 205.7780$, $df = 12$, $P < 0.0001$), and removed paleognaths to create a tree for neognaths only, $-2\Delta\ln = 189.2090$ ($df = 10$, $P < 0.0001$). Clearly, there is no overall molecular clock for RAG-1 for these birds.

If one were to ignore the above results and assume that a molecular clock exists, there are several dates from the fossil record that yield a wide range of potential rate calibrations. For example, *Alligator* and *Gavialis* have been separated for approximately 80 my (see Brochu, 1997) and show a RAG-1 distance of 2.5%; this suggests a rate of 0.03%/my. Birds and crocodylians can be traced to the early archosaurs (~250 mya);

they have RAG-1 pairwise distances averaging 16%, indicating a calibration of ~0.065%/my. Finally, the earliest passeriform fossils may be as old as 50 my (see review in Feduccia, 1996); the distance between *Passer* and *Tyrannus* (which represents the basal oscine/suboscine passerine division) was 6.5%, producing a rate of ~0.13%/my. Even when genetic distances are corrected for multiple substitutions using HKY distances (Hasegawa *et al.*, 1985), or HKY distances with gamma correction (Yang, 1996), the range of rate estimates still varies by a factor of four to five. Consequently, we are not sanguine about trying to associate a clock to these data. Nevertheless, any of the above calibrations would place the divergences among the three major clades of modern birds (with pairwise RAG-1 distances around 10%) before the Cretaceous–Tertiary (K–T) boundary at 65 mya. However, given the above problems with the clock, combined with identification disputes and lack of detailed phylogenetic analysis for most avian fossils (Feduccia, 1996), RAG-1 date calibrations cannot rule out a hypothesis that all living birds trace to only three divergent ancestors (one paleognath and two neognath) at the K–T boundary.

Utility of the RAG-1 Gene

The RAG-1 gene has many properties desirable for molecular phylogenetic analyses. Indels are rare, and the few that exist did not cause alignment or homology problems. Lack of absolute constraint against indels does, however, allow for the possibility that some might prove useful as markers for higher-level taxa. Allelic variation within individuals (heterozygosity) is low and will not compromise higher-level phylogenetic analysis. Graphical analysis shows that the most rapidly evolving data partitions (3rd position transitions) have not reached saturation levels, even for comparisons involving the deepest divergences (e.g., birds versus crocodylians); consequently, downweighting or eliminating partitions is not necessary. It is known that heterogeneity in base composition among taxa might negatively influence phylogenetic inference (e.g., Saccone *et al.*, 1989; Lockhart, 1994), but the base composition in RAG-1 is highly stationary. Finally, directionality of inferred substitutions was symmetrical, suggesting that estimation of the character states of nucleotide positions at internal nodes will not be biased (Collins *et al.*, 1994). Perhaps as a consequence of these features, RAG-1 provided a robust (in terms of bootstrap and Bremer support) phylogeny for deep divergences in birds for which previous studies using mitochondrial and nuclear genes had been less successful.

ACKNOWLEDGMENTS

We are grateful to Gavin Naylor for suggesting RAG-1 as a useful gene. David Kizirian competently assisted with the lab work. David Swofford kindly made available a test version of the PAUP* program. Carole Griffiths, Brad Livezey, Gavin Naylor, and three anonymous

reviewers gave useful comments on drafts of the manuscript. For donations and assistance with tissue samples, we thank George Amato, Tim Crowe, Mary K. LeCroy, Patricia Escalante-Pliego, Robert E. Jones, Paul R. Sweet, Sea World of San Diego, and the Bronx Zoo. This research was supported by the L. J. Sanford Trust and the Lewis B. and Dorothy Cullman Program for Molecular Systematic Studies, a joint initiative of the New York Botanical Garden and the American Museum of Natural History.

REFERENCES

- Bernstein, R. M., Schluter, S. F., Bernstein, H., and Marchalonis, J. J. (1996). Primordial emergence of the recombination activating gene 1 (RAG1): Sequence of the complete shark gene indicates homology to microbial integrases. *Proc. Natl. Acad. Sci. USA* **93**: 9454–9459.
- Bremer, K. (1994). Branch support and tree stability. *Cladistics* **10**: 295–304.
- Brochu, C. A. (1997). Morphology, fossils, divergence timing, and the phylogenetic relationships of *Gavialis*. *Syst. Biol.* **46**: 479–522.
- Carlson, L. M., Oettinger, M. A., Schatz, D. G., Masteller, E. L., Hurley, E. A., McCormack, W. T., Baltimore, D., and Thompson, C. B. (1991). Selective expression of RAG-2 in chicken B cells undergoing immunoglobulin gene conversion. *Cell* **64**: 201–208.
- Caspers, G.-J., Uit de Weerd, D., Wattel, J., and de Jong, W. W. (1997). α -Crystallin sequences support a galliform–anseriform clade. *Mol. Phylogenet. Evol.* **7**: 185–188.
- Collins, T. M., Wimberger, P. H., and Naylor, G. P. (1994). Compositional bias, character state bias, and character state reconstruction using parsimony. *Syst. Biol.* **43**: 482–496.
- Cooper, A., and Penny, D. (1997). Mass survival of birds across the Cretaceous–Tertiary boundary: Molecular evidence. *Science* **275**: 1109–1113.
- Cooper, A., and Fortey, R. (1998). Evolutionary explosions and the phylogenetic fuse. *Trends Ecol. Evol.* **13**: 151–156.
- Cracraft, J. (1981). Toward a phylogenetic classification of Recent birds of the world (Class Aves). *Auk* **98**: 681–714.
- Cracraft, J. (1988). The major clades of birds. In “The Phylogeny and Classification of the Tetrapods” (M. J. Benton, Ed.), Vol. 1, pp. 339–361. Clarendon, Oxford.
- Ericson, P. G. P. (1996). The skeletal evidence for a sister-group relationship of anseriform and galliform birds—A critical evaluation. *J. Avian Biol.* **27**: 195–202.
- Feduccia, A. (1996). “The Origin and Evolution of Birds,” Yale Univ. Press, New Haven, CT.
- Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.* **17**: 368–376.
- Felsenstein, J. (1985). Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* **39**: 783–791.
- Greenhalgh, P., Olesen, C. E., and Steiner, L. A. (1993). Characterization and expression of the recombination activating genes (RAG-1 and RAG-2) in *Xenopus laevis*. *J. Immunol.* **151**: 3100–3110.
- Härlid, A., Janke, A., and Arnason, U. (1997). The mtDNA sequence of the ostrich and the divergence between palaeognathous and neognathous birds. *Mol. Biol. Evol.* **14**: 754–761.
- Härlid, A., Janke, A., and Arnason, U. (1998). The complete mitochondrial genome of *Rhea americana* and early avian divergences. *J. Mol. Evol.* **46**: 669–679.
- Hasegawa, M., Kishino, H., and Yano, T. (1985). Dating the human–ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22**: 160–174.
- Hedges, S. B., Parker, P. H., Sibley, C. G., and Kumar, S. (1996). Continental breakup and the ordinal diversification of birds and mammals. *Nature* **381**: 226–229.
- Ho, C. Y.-K., Prager, E. M., Wilson, A. C., Osuga, D. T., and Feeney, R. E. (1976). Penguin evolution: Protein comparisons demonstrate phylogenetic relationships to flying birds. *J. Mol. Evol.* **8**: 271–282.
- Irwin, D. M., Kocher, T. D., and Wilson, A. C. (1991). Evolution of the cytochrome *b* gene in mammals. *J. Mol. Evol.* **32**: 128–144.
- Janke, A., and Arnason, U. (1997). The complete mitochondrial genome of *Alligator mississippiensis* and the separation between recent archosauria (birds and crocodiles). *Mol. Biol. Evol.* **14**: 1266–1272.
- Lake, J. A. (1987). Rate-independent technique for analysis of nucleic acid sequences: Evolutionary parsimony. *Mol. Biol. Evol.* **4**: 167–191.
- Livezey, B. C. (1997). A phylogenetic analysis of basal Anseriformes, the fossil *Presbyornis*, and the interordinal relationships of waterfowl. *Zool. J. Linn. Soc.* **121**: 361–428.
- Lockhart, P. J., Howe, C. J., Bryant, D. A., Beanland, T. J., and Larkum, A. W. D. (1992). Substitutional bias confounds inference of cyanelle origins from sequence data. *J. Mol. Evol.* **34**: 153–162.
- Maddison, W. P., and Maddison, D. R. (1992). “MacClade, Version 3: Analysis of Phylogeny and Character Evolution,” Sinauer, Sunderland, MA.
- Mindell, D. P., Sorenson, M. D., Huddleston, C. J., Miranda, H. C., Jr., Knight, A., Sawchuk, S. J., and Yuri, T. (1997). Phylogenetic relationships among and within select avian orders based on mitochondrial DNA. In “Avian Molecular Evolution and Systematics” (D. P. Mindell, Ed.), pp. 213–247. Academic Press, San Diego.
- Moritz, C., Schneider, C. J., and Wake, D. B. (1992). Evolutionary relationships within the *Ensatina eschscholtzii* complex confirm the ring species interpretation. *Syst. Biol.* **41**: 273–291.
- Prager, E. M., and Wilson, A. C. (1980). Phylogenetic relationships and rates of evolution in birds. *Proc. XVII Int. Ornithol. Congr.* pp. 1209–1214.
- Prychitko, T. M., and Moore, W. S. (1997). The utility of DNA sequences of an intron from the β -fibrinogen gene in phylogenetic analysis of woodpeckers (Aves: Picidae). *Mol. Phylogenet. Evol.* **8**: 193–204.
- Pycraft, W. P. (1900). The morphology and phylogeny of the Palaeognathae (Ratitae and Crypturi) and the Neognathae (Carinatae). *Trans. Zool. Soc. London* **15**: 149–290.
- Rambout, A., and Bromham, L. (1998). Estimating divergence dates from molecular sequences. *Mol. Biol. Evol.* **15**: 442–448.
- Saccone, C., Pesole, G., and Preparata, G. (1989). DNA microenvironments and the molecular clock. *J. Mol. Evol.* **29**: 407–411.
- Schatz, D. G., Oettinger, M. A., and Baltimore, D. (1989). The V(D)J recombination activating gene, RAG-1. *Cell* **59**: 1035–1048.
- Sheldon, F. H., and Bledsoe, A. H. (1993). Avian molecular systematics, 1970s to 1990s. *Annu. Rev. Ecol. Syst.* **24**: 243–278.
- Sibley, C. G., Ahlquist, J. E., and Monroe, B. L., Jr. (1988). A classification of the living birds of the world based on DNA–DNA hybridization studies. *Auk* **105**: 409–423.
- Sibley, C. G., and Ahlquist, J. E. (1990). “Phylogeny and Classification of Birds,” Yale Univ. Press, New Haven, CT.
- Sibley, C. G., and Monroe, B. L., Jr. (1990). “Distribution and Taxonomy of Birds of the World,” Yale Univ. Press, New Haven, CT.
- Swofford, D. L. (1998). “PAUP*: Phylogenetic analysis using parsimony (*and other methods), 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” (D. M. Hillis, C. Moritz, and B. K. Mable, Eds.), 2nd ed., pp. 407–514. Sinauer, Sunderland, MA.
- Yang, Z. (1996). Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol. Evol.* **11**: 367–372.
- Yang, Z., Goldman, N., and Friday, A. (1995). Maximum likelihood trees from DNA sequences: A peculiar statistical estimation problem. *Syst. Biol.* **44**: 384–399.