MOLECULAR PHYLOGENETICS AND EVOLUTION
Vol. 4, No. 4, December, pp. 420–432, 1995

Evolution of NADH Dehydrogenase Subunit 2 in East African Cichlid Fish

THOMAS D. KOCHER,* JANET A. CONROY,* KENNETH R. MCKAYE,†
JAY R. STAUFFER,‡ AND SAMUEL F. LOCKWOOD*

*Department of Zoology, University of New Hampshire, Durham, New Hampshire 03824; †Appalachian Environmental Laboratory, University of Maryland, Frostburg, Maryland 21532; and ‡School of Forest Resources, Penn State University, University Park, Pennsylvania 16802

Received July 20, 1994; revised February 13, 1995

The complete sequence (1047 bp) of the mitochondrially encoded ND2 gene was obtained from 31 species of cichlid fishes to investigate the evolutionary history of the species flocks of the East African lakes. The observed pattern of nucleotide substitution in this sequence is typical of mitochondrial genes, showing a high transition bias and rapid mutational saturation, especially at codon positions where base frequencies are unequal. The base composition of the third position of codons is heterogeneous among species, suggesting frequent shifts in the pattern of substitution. Phylogenetic analysis of the sequences shows that the mtDNA variation in Lake Malawi cichlids is nested monophyletically within the range of variation shown by Tanganyikan cichlids. The closest Tanganyikan relatives of the Malawi flock are members of the tribe Tropheini. Classifications based on morphology are generally supported by the mtDNA data, with some significant exceptions in the Tropheini and Lamprologini. Because of an apparently rapid radiation of the Tanganyikan lineages, it is difficult to assess the basal topology of the Tanganyikan radiation at this time. Divergences among tribes are consistent with an intralacustrine radiation within the past 10 million years.

INTRODUCTION

The cichlid fishes of the East African Rift Valley lakes have been touted as the premier example of "explosive" adaptive radiation in vertebrates for more than 50 years. Lake Malawi contains over 400 described species of endemic cichlids (Konings, 1990; Greenwood, 1991), more species of fish than any other lake in the world. Other East African lakes show similar levels of diversity (Greenwood, 1974; Fryer and Iles, 1972). Many of the nominal species have been shown to be assemblages of sibling species (e.g., McKaye *et al.*, 1982, 1984; Ribbink *et al.*, 1983), causing some workers to estimate that as many as 1000 species of cichlids may reside within Lake Malawi alone (Lewis *et al.*, 1986). These "species flocks" are an ideal system in which to study the ecological and genetic mechanisms of speciation and morphological evolution in vertebrates (Echelle and Kornfield, 1984).

The geologic history of these lakes is complex. The Rift Valley has been extremely active geologically, especially during the past 2–5 million years. Banister and Clarke (1980) suggested that the current Tanganyikan basins were formed about 2.6 mybp and that Lake Malawi is even more recent (<1 my). Cohen *et al.* (1993), however, have estimated from sedimentation data that the Tanganyikan basins began to form 9–12 mybp. The continuous block faulting of the region makes reconstruction of drainage patterns difficult. Furthermore, because these basins have been closed systems for most of their history, minor climatic fluctuations have had major effects on lake levels. Twenty-five thousand years ago, water levels in lakes Malawi and Tanganyika were 250–500 meters below the current water lines (Scholz and Rosendahl, 1988). Extreme lake level fluctuations have occurred in even more recent times and may have aided microallopatric differentiation of taxa (Owen *et al.*, 1990).

Surprisingly, even though cichlids are a morphologically diverse group, there are few morphological characters which can be used to describe intrafamilial relationships (Stiassny, 1991). The constant repatterning of the cichlid head has been achieved not by the development of new features, but by small changes in relative growth rates among structures (Strauss, 1984). Discrete characters useful for developing phylogenies are few. Dissatisfaction with traditional morphological approaches has prompted the investigation of new kinds of characters, such as reproductive behavior and bower form (McKaye, 1991; McKaye *et al.*, 1993) to resolve relationships.

Nishida (1991) used allozyme polymorphisms to study relationships within the Tanganyikan flock. He found a number of old (>5 my) lineages among the 12

1055-7903/95 $12.00

tribes of Tanganyikan cichlid defined by Poll (1986). Nishida suggested that Tanganyika represents an evolutionary reservoir from which the Lake Victoria and Lake Malawi flocks arose. Because no representatives of the Malawi or Victoria flocks were included, the relationship of these flocks to Tanganyikan species could not be examined.

Mitochondrial DNA is widely recognized as an important tool for resolving relationships among closely related species. Dorit (1986) was the first to analyze mtDNA to determine relationships among East African cichlids. He showed that there is remarkably little intraspecific variation in the mtDNA sequence of Lake Victoria cichlids and that sufficient interspecific variation exists to reconstruct phylogenies for members of the *Psammochromis–Macropleurodus* lineage. Meyer *et al.* (1990) extended this work by directly sequencing mtDNA from a number of Lake Victoria and Malawi cichlids. These data provided support for an extremely recent, monophyletic origin of each flock. Kocher *et al.* (1993) compared six pairs of morphologically similar taxa from lakes Malawi and Tanganyika and found that the Malawi taxa had a recent, monophyletic origin.

The noncoding region used for these earlier studies has serious deficiencies when applied to deeper divergences. Because of the high rate of substitution, rate heterogeneity among sites, and an unknown pattern of selective constraint (Kocher and Wilson, 1991), it is difficult to properly account for multiple substitutions. Accurate phylogenetic reconstruction, in which branch points are well-calibrated with geologic time, may be easier to accomplish from protein-coding sequences, whose evolutionary properties are more fully understood (Irwin *et al.*, 1991).

The current study focuses on a different mitochondrial gene, NADH dehydrogenase subunit 2 (ND2), because it may exhibit more uniform substitution probabilities among codons. Our goals were first to characterize the patterns and rates of substitution in the ND2 gene and second to use this information to achieve a more accurate reconstruction of the history of East African cichlids.

TABLE 1

Specimens Examined

Lake Malawi	
Buccochromis lepturus	Chembe village, Cape Maclear, Malawi
Champsochromis spilorhynchus	Chembe village, Cape Maclear, Malawi
Lethrinops auritus	Chembe village, Cape Maclear, Malawi
Pseudotropheus zebra	Aquarium trade, U.S.A.
Rhamphochromis sp.	Monkey Bay, Malawi
Lake Tanganyika	
Tribe Bathybatini	
Bathybates sp.	Fish market, Uvira, Zaire
Tribe Cyprichromini	
Paracyprichromis brieni	30 km S Bujumbura, Burundi
Tribe Ectodini	
Callochromis macrops	Marlier's Rocks, Uvira, Zaire
Cardiopharynx schoutedeni	C.R.S.N. guest house, Uvira, Zaire
Opthalmotilapia ventralis	Meriel Shreyen, Bujumbura, Burundi
Xenotilapia flavipinnus	Meriel Shreyen, Bujumbura, Burundi
Xenotilapia sima	Fish market, Uvira, Zaire
Tribe Eretmodini	
Tanganicodus irsacae	Meriel Shreyen, Bujumbura, Burundi
Tribe Lamprologini	
Chalinochromis popeleni	Meriel Shreyen, Bujumbura, Burundi
Julidochromis marlieri	Meriel Shreyen, Bujumbura, Burundi
Lamprologus callipterus	Marlier's Rocks, Uvira, Zaire
Lepidiolamprologus elongatus	30 km S Bujumbura, Burundi
Neolamprologus brichardi	Luhanga, Zaire
Neolamprologus tetracanthus	30 km S Bujumbura, Burundi
Telmatochromis temporalis	Luhanga, Zaire
Tribe Limnochromini	
Gnathochromis pfefferi	1 km S Bemba Springs, Zaire
Limnochromis auritus	Meriel Shreyen, Bujumbura, Burundi
Tribe Perissodini	
Perissodus microlepis (a)	Meriel Shreyen, Bujumbura, Burundi
Perissodus microlepis (b)	Luhanga, Zaire
Tribe Tilapiini	
Boulengerochromis microlepis	Fish market, Uvira, Zaire
Oreochromis niloticus	Fish market, Uvira, Zaire
Tribe Tropheini	
Cyphotilapia frontosa	Meriel Shreyen, Bujumbura, Burundi
Lobochilotes labiatus	Luhanga, Zaire
Petrochromis orthognathus	Meriel Shreyen, Bujumbura, Burundi
Tropheus moorii var. Moba	Meriel Shreyen, Bujumbura, Burundi
Tribe Tylochromini	
Tylochromis polylepis	Fish market, Uvira, Zaire
Central America	
Cichlasoma citrinellum	George Barlow, Berkeley, California

MATERIALS AND METHODS

DNA sources. Most fish were collected by scuba divers using monofilament gill nets. Additional wild-caught and F_1 specimens were also studied (Table 1). Species identifications were made immediately upon capture in the field, before body colors were lost. Voucher specimens have been deposited with the Penn State University Fish Museum. Whole fish were placed in several volumes of 100% ethanol for 24–48 h, wrapped in alcohol-moistened cheesecloth, and sealed in a plastic bag for transportation to the laboratory. In the lab, samples were returned to 70% alcohol for long-term storage. DNA was extracted from the dorsal musculature of each fish using a standard proteinase K digestion/phenol–chloroform extraction procedure (Kocher *et al.*, 1989). Two individuals of *Perissodus microlepis* were included as a blind control to test the accuracy of the sequencing reactions and gel scoring in the laboratory.

ND2 sequences. The sequencing strategy is displayed in Fig. 1. The ND2 gene was first amplified using primers directed at the glutamine and asparagine tRNAs. Sequences obtained from the ends of this fragment were then used to design additional primers. Typically, the entire gene was amplified with primers 1 and 6, and the double-stranded product sequenced with the internal primers 2–5 and 7–11. Amplified products were purified by separation on agarose gels, and the

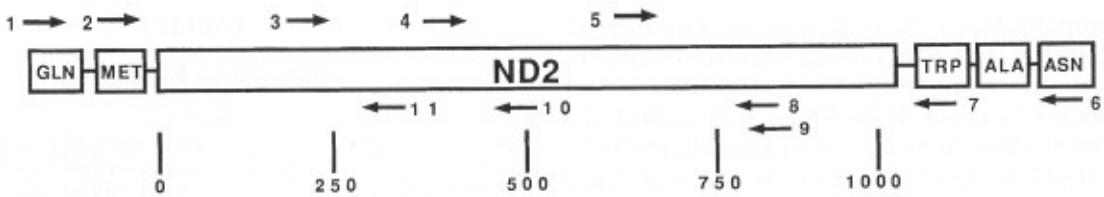

FIG. 1. Amplification and sequencing strategy for the ND2 gene. Bars depict the ND2 and flanking tRNA genes. Arrows indicate the positions of primers used. The primer sequences are 1, GLN 5′-ctacctgaagagatcaaaac-3′; 2, MET 5′-catacccaaacatgttggt-3′; 3, ND2.1 5′-acaggtcaatgagaaattcaacaa-3′; 4, ND2.2A 5′-ctgacaaaaacttgccccctt-3′; 5, ND2.T 5′-acatcctgaacaaaagcccc-3′; 6, ASN 5′-cgcgtttagctgttaactaa-3′; 7, TRP 5′-gagattttcactcccgctta-3′; 8, ND2.4 5′-ttggttagttcttgaaggat-3′; 9, ND2.4A 5′-aagcccttgtttggttagttct-3′; 10, ND2B 5′-tggtttaatccgcctcagcc-3′; 11, ND2A 5′-aacttcggggagtcaagccat-3′.

DNA was recovered by hot phenol extraction (Maniatis *et al.,* 1982) or by digestion with GELase (Epicentre Technologies). Direct cycle sequencing with dye-labeled terminators was performed according to the manufacturer's instructions (Applied Biosystems Inc.). Labeled fragments were analyzed on an ABI 373A automated DNA sequencer. Sequences were edited with the aid of the SeqEd program (Applied Biosystems). Typically, 300 bases could be scored from a single reaction.

Data analysis. The sequences were compiled and translated with the mammalian mitochondrial code using the ESEE multiple sequence alignment editor (Cabot and Beckenbach, 1989). OBSDIS, a modification by Arend Sidow of Felsenstein's DNADIST algorithm (Felsenstein, 1989), was used to count transition and transversion differences among all possible pairs of taxa. Distance trees were constructed using the neighbor-joining algorithm (Saitou and Nei, 1987) as implemented in MEGA (Kumar *et al.,* 1993). Parsimony analysis was conducted with PAUP Ver. 3.0s (Swofford, 1991) on a Macintosh Quadra 950 computer. Programs in the GCG package (Ver. 7; Devereux *et al.,* 1984) were used to study the composition and structure of the gene. Tests for compositional stationarity were performed using the method of Rzhetsky and Nei (1995).

RESULTS

ND2 Sequences

Sequences of 1047 nucleotides for each of the 31 species have been deposited in GenBank (Accession Nos. U07239–U07270). The predicted proteins are all 348 aa long and no indels were observed. The two individuals of *Perissodus* included as a blind control differed at only three sites (all silent), which is within the expected range of intraspecific variation. The fraction of first, second, and third position sites which were variable was 31.5, 14.9, and 90.5%, respectively.

Compositional Bias

Strong bias in base composition can seriously compromise some methods used to correct observed differences to actual divergence (Saccone *et al.,* 1990). In addition, many phylogenetic reconstruction algorithms assume identical base composition in all lineages. Compositional bias has typically been summarized with a single statistic, which obscures patterns in the distribution of nucleotides on the two strands. Deviation from an expectation of equal frequencies of all four bases can take three forms (Perna and Kocher, 1995). The number of GC base pairs may not equal the number of AT pairs (quantified as the percentage GC). Paired bases can exhibit *skew,* which is the nonrandom orientation of paired bases in the helix. For example, among A–T base pairs, most of the A's might be found on one strand. Two kinds of skew are possible, that involving A–T and that involving G–C. The values of GC%, GC, and AT skew vary among positions in codons.

Table 2 lists the base composition at the first, second, and third positions of codons and at the fourfold degenerate sites of the ND2 gene. Base composition varies significantly among codon positions. First positions are the most GC-rich (52.7%), presumably because of selection for a high leucine and alanine content in the protein. First positions exhibit moderate positive AT skew and negative GC skew. Second positions exhibit a low frequency of A and a high frequency of T. Both GC and AT base pairs exhibit strong negative skew. Again, composition probably reflects selective constraints arising from the membrane-spanning structure of this protein, since hydrophobic amino acids are encoded by codons containing C and T at the second position. The bias at first and second positions in these sequences is greater than that reported for mammalian cytochrome b (Irwin *et al.,* 1991). Third positions contain a high proportion of A and T. While AT base pairs at these sites exhibit a slight positive skew (+0.104), GC base pairs are highly skewed (−0.774), reflecting a very low frequency of G on the sense strand. The highest levels of GC skew are found at fourfold degenerate positions. By way of comparison, the carp shows a similar GC skew, but also a much stronger AT skew (+0.73) than the cichlids at the fourfold sites.

Although base composition appears relatively homogeneous among the African species (Table 2) statistical tests (Rzhetsky and Nei, 1995) indicate considerable

TABLE 3

Comparison of the Amino Acid Composition of Mitochondrial Genes in a Cichlid, Carp, Human, and Drosophila

	ND2				Cytb		
	Cic	Car	Hum	Dro	Car	Hum	Dro
Alanine	31	39	20	13	30	25	21
Arginine	4	6	4	2	8	7	8
Asparagine	9	8	20	30	17	15	24
Aspartic acid	1	4	0	1	11	11	10
Cysteine	1	1	0	2	4	2	3
Glutamine	15	14	10	7	6	8	8
Glutamic acid	8	5	6	8	6	4	3
Glycine	19	18	13	13	25	24	25
Histidine	7	6	4	3	12	12	11
Isoleucine	25	28	32	34	32	39	45
Leucine	77	66	64	65	60	64	58
Lysine	8	9	12	10	10	9	7
Methionine	11	19	24	27	13	15	11
Phenylalanine	13	9	15	37	30	24	31
Proline	23	19	23	7	21	23	23
Serine	25	23	28	36	22	29	17
Threonine	46	49	43	19	23	30	19
Tryptophan	11	11	11	10	13	12	13
Tyrosine	7	8	10	10	14	17	18
Valine	7	6	8	7	24	10	23
Total	348	349	347	341	381	380	378
Bias[a]	0.35	0.36	0.34	0.39	0.26	0.28	0.27

[a] Bias in amino acid composition is calculated as

$$B = \frac{\text{sum}|f_i - 0.051|}{1.95},$$

where f_i is the frequency of the ith amino acid.

of several mitochondrially encoded peptides which reside in the hydrophobic fragment of complex I. This fragment is thought to wrap around the other two components of the complex, the flavoprotein and iron-protein fragments. The high content of hydrophobic amino acids in the ND2 sequence supports the idea that it is almost entirely embedded in the membrane. Comparison of the cichlid amino acid sequences reveals generally higher levels of variation in the second half of the protein. Several segments (amino acids 19–24, 126–141, 165–178, 249–260) show unusually low levels of variation, which may be indicative of functional constraint. The latter two regions are conserved in the sequence of human ND2 and appear to be located in turns between membrane-spanning helices. Conservation of these regions, together with the maintenance of a transmembrane configuration, seem to be the major constraints on the evolution of this protein.

Although the assumption is rarely tested, most phylogenetic reconstruction methods assume an equal probability of substitution among sites. We estimated the number of potentially variable amino acids by two methods. Using the capture/recapture method (Sidow *et al.,* 1992), a comparison of *Buccochromis* and *Cichlasoma* yields an estimate of 86.7 ± 18.8 variable codons. This might be an underestimate of the actual number of variable codons because functional constraint for hydrophobicity of this protein may reduce the variability of second position sites, violating assumptions of the method. A second estimate was made by counting the number of changes per site in a tree relating four Malawi and seven closely related Tanganyikan taxa. A least-squares fit of the number of substitutions per site to a Poisson distribution yields a similar estimate of about 100 variable codons. Therefore, it seems likely that approximately 25% of the amino acids in the ND2 molecule are free to vary at any particular instant. This value compares to an estimate of 37% for 13 diverse calmodulin sequences and 39% for 4 vertebrate albumin sequences (Sidow *et al.,* 1992).

We also estimated the number of third position sites which are free to undergo a transversion (Uzzell and Corbin, 1971; Holmquist *et al.,* 1983). A larger number of these sites are expected to be variable, since nucleotide substitutions at many sites will not result in amino acid replacement. We counted the number of substitutions at each site over parsimony trees (constructed using all substitutions weighted equally) at three different time depths (Table 4). Among closely related taxa, the Poisson model fits reasonably well, suggesting that all sites are free to vary. Among more distantly related taxa, the Poisson model fits only if an increasingly large invariant class of sites is postulated. Fitting of a negative binomial distribution does not lead to a consistent estimate of the gamma parameter (k), and methods for estimating the standard error of k are not yet available (Tamura and Nei, 1993). Since approximately 215 codons of the ND2 gene are fourfold degenerate and at least some transversions must be possible among the 133 twofold degenerate codons, we estimate that between 250 and 300 third position sites are typically free to undergo a transversion.

Accumulation of Sequence Difference

Figure 2 displays the accumulation of pairwise sequence difference among cichlid species, including the distantly related *Cichlasoma*. Because no reliable fossil dates are available for the divergence of these taxa, the least saturated component of the data (the number of third position transversion differences) was used to plot the accumulation of sequence difference. Transitions and transversions are plotted separately for each of the three codon positions. The largest third position transversion difference observed among African taxa was about 10%. Comparisons between *Cichlasoma* and the African taxa displayed about 25% third position transversion difference.

The strong transition bias characteristic of animal mtDNA is clearly evident at the third positions. The initial transition/transversion ratio observed at third positions is about 10/1. Saturation does not occur until

TABLE 4

Estimates of the Number of Codons Free to Undergo a Transversion at the Third Position

	Observed changes per site					Poisson	Modified Poisson		Negative binomial	
	0	1	2	3	4	χ^2	Sites	χ^2	K	χ^2
Tree 1	315	31	1	0	0	0.186	—	—	20	0.232
Tree 2	304	39	2	4	0	0.514	250	0.003	2	0.004
Tree 3	285	51	8	2	1	3.908	195	0.080	1	0.022

Note. Taxa for tree 1, Pseudotropheus, Champsochromis, Lethrinops, Rhamphochromis, Tropheus, Gnathochomis, Petrochromis, Lobochilotes, Opthalmotilapia; taxa for tree 2, Pseudotropheus, Champsochromis, Rhampochromis, Tropheus, Gnathochromis, Petrochromis, Lobochilotes, Xenotilapia flavipinnus, Cardiopharynx, Cyphotilapia; taxa for tree 3, Pseudotropheus, Champsochromis, Rhamphochromis, Tropheus, Opthalmotilapia, Xenotilapia, Julidochromis, Lamprologous, Lepidiolamprologus, Oreochromis.

about 25% sequence difference. The pattern of saturation is complex, however (Fig. 2f). Transitions involving A and G saturate at 12–15% difference, whereas those involving C and T saturate at 30–40% difference. These values are each somewhat less than expectations (23 and 48%, respectively) calculated from base frequencies (e.g., $1 - (G/(G + A))^2 - (A/(G + A))^2$), indicating the additional factor of mild selective constraint at these sites.

In contrast to the third positions, there are strong constraints on the first and second positions, derived from selection on protein structure. Substitutional saturation occurs at much lower levels of divergence for these sites. The overall rate of transitional substitution at first and second positions is at least 10-fold lower than for third positions. The saturating values of transitions are about 7.5% for first positions and only 2.5% at the second position. Transversions accumulate linearly at first positions until at least 7.5%, while at second positions saturation occurs at only 2.5% difference.

Previous studies of some of these taxa have focused on a 350-bp segment encompassing the rapidly evolving first half of the control region. In Fig. 3, total sequence difference in the control region is plotted against third position transversion difference for eight taxa for which control region and ND2 sequences are available. The control region saturates at 15–20% difference. This saturation occurs very rapidly—at about 2% third position transversion difference. The control region initially accumulates change as rapidly as third position transitions, but it saturates earlier, at a lower difference value.

Distance Analysis

The first step in the phylogenetic analysis was to discover which taxa were closely related. For this purpose we used the third position transversion divergence values, corrected for multiple hits using a gamma distribution ($\alpha = 20$) (Tamura and Nei, 1993). The topology of the neighbor-joining tree (Fig. 4) is not sensitive to small changes in the number of variable sites assumed.

The Ectodini are one of several major clades easily distinguished in this tree. A second cluster contains *Gnathochromis,* elements of the Tropheini, together with the Malawi taxa. The relationship of these two clusters to the Limnochromini, Perissodini, Cyprichromini, and *Cyphotilapia* is not well defined. Deeper in the tree, there is a well-defined clade containing lamprologine taxa. *Tanganicodus* diverges at about the same point. The deepest branches of the tree lead to *Boulengerochromis, Bathybates, Oreochromis,* and *Tylochromis. Cichlasoma* is too distantly related to be a useful outgroup in this analysis.

A striking feature of this tree is the apparent variation in rate among lineages. In particular, several mtDNAs of the tribe Ectodini appear to be evolving at a very high rate. These differences in rate may be related to the rapidly evolving base composition of this group.

Parsimony Analysis of Tribal Clusters

A parimony analysis was used to further examine three clusters (Lamprologini, Ectodini, Tropheini/Haplochromini) identified in the distance analysis. In each case, all characters were weighted equally, and the reliability of the most parsimonious tree was estimated with 2000 bootstrap samples.

The phylogeny of the lamprologines is resolved well by these data. Bootstrap analysis shows significant support for nearly all segments of the tree (Fig. 5a). *Lepidiolamprologus elongatus* is the clear sister group to the other lamprologine taxa studied here. *Neolamprologus* is clearly not monophyletic. *N. tetracanthus* is a sister to *Lamprologus callipterus,* while *N. brichardi* clusters with *Julidochromis marlieri* and *Telmatochromis temporalis.* This result confirms the assessment of others (Brichard, 1989; Poll, 1986) that *Neolamprologus* is not a natural assemblage.

Parsimony analysis also corroborates previous views of Ectodine phylogeny (Sturmbauer and Meyer, 1993).

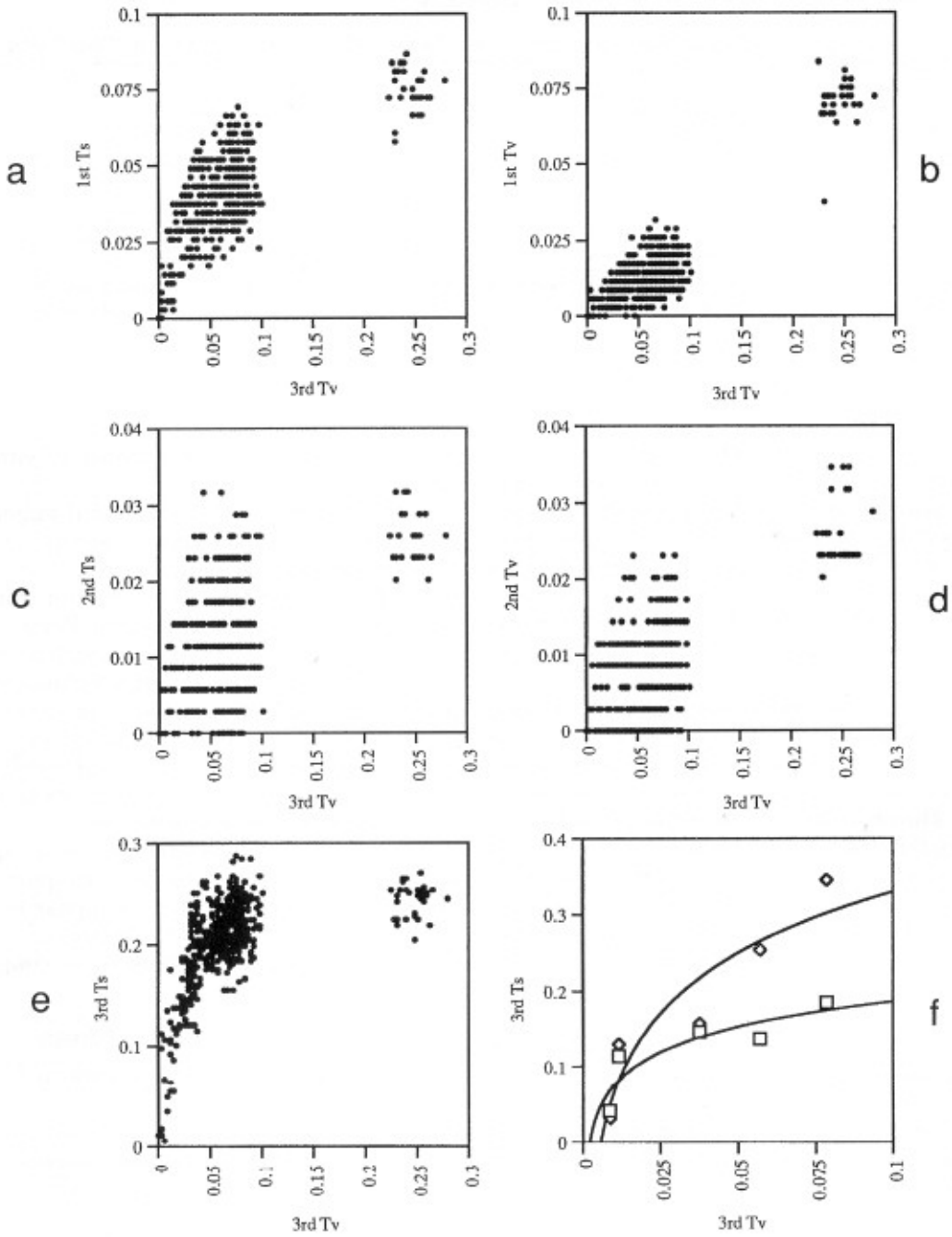

FIG. 2. Accumulation of sequence difference at first, second, and third positions of codons in the ND2 gene. In each panel, the accumulation of difference is plotted against the third position tranversion difference. In f, the accumulation of transitions involving A and G (squares) is plotted separately from those involving C and T (diamonds).

Specifically, it supports the idea that *Cardiopharynx* and *Opthalmotilapia* are sister taxa and that *Callochromis* is the outgroup to the other four taxa (Fig. 5b). It should be noted, however, that this arrangement is at odds with the distance tree (Fig. 4). Our data do not support the idea that the Cyprochromini are the sister group to the Ectodines (Sturmbauer and Meyer, 1993; Meyer, 1993). *Paracyprichromis* falls outside the clade containing the Ectodines, *Limnochromis,* and *Tropheus* in 69% of the bootstrap samples.

The Tropheini/Haplochromine clade was examined using *Xenotilapia* as the outgroup (Fig. 5c). There is a clear division between the Tanganyikan and the Malawi members of this clade. The Malawi taxa cluster

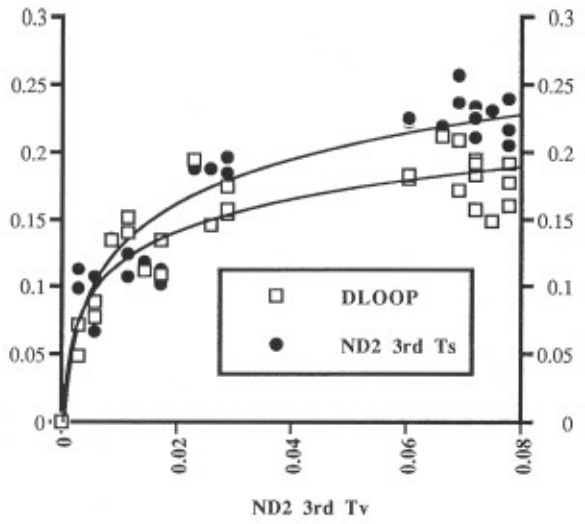

FIG. 3. Comparison of sequence difference in the control region and ND2 genes. Total sequence difference in the control region, as well as third position transitions, is plotted against third position transversion difference in ND2.

together in all 2000 of the bootstrap samples. The Tropheini are a much less tightly clustered group. Although monophyly of the Tanganyikan taxa is suggested (78% of the bootstrap samples), the internal structure of this clade is not well resolved.

Relationships among Basal Lineages

The distance tree is ambiguous in two regions. First is the relationship among seven major lineages of the Tanganyikan flock (Tropheini, Cyprichromini, Ectodini, Eretmodini, Lamprologini, Limnochromini, and Perissodini). In an attempt to resolve the relationships among these groups, we performed an analysis of amino acid sequences using the default mitochondrial transformation matrix of PAUP. No significant clustering was found in bootstrap analyses using *Oreochromis* as an outgroup.

A second area of ambiguity is the relationships among the most basal taxa (*Oreochromis, Boulengerochromis, Tylochromis, Bathybates,* lamprologines, and haplochromines). For this analysis, we again used protein parsimony, with *Cichlasoma* as an outgroup. Only a few clades appeared in more than 50% of the bootstrap samples, the most frequent being a clustering of *Oreochromis* and *Tylochromis* (69%).

DISCUSSION

It is increasingly apparent that patterns of nucleotide substitution in mtDNA have not yet been adequately modeled. Recognition of the transition bias characteristic of mtDNA is but a first step in modeling the accumulation of substitutions in this molecule. Both the rate of substitution and the base composition of the molecule vary with selective constraint among codon positions. Codons vary among one another in their probability of substitution, even at third positions. These variations do not invalidate the use of DNA sequences for phylogenetic analysis. Rather, they encourage us to study more closely the forces shaping the evolution of these molecules. The increasing ease with which sequences can be obtained encourages the thought that these variations can be taken into account by analysis of larger data sets.

Pattern of Substitution in ND2

ND2 was chosen for this study because, at least in mammals, it is evolving more rapidly than other mitochondrial proteins (Anderson *et al.,* 1982). While a faster rate implies fewer selectional constraints, amino acid substitutions are far from random. Perhaps because of their highly skewed amino acid composition, these membrane-embedded proteins have a relatively small proportion of variable sites. At the nucleotide level, saturation for change occurs before theoretically expected limits at every codon position. Only third position transitions appear to approach expectations based on base frequency.

A surprising finding was the lack of stationarity for base composition among some closely related taxa. Deviations are especially prevalent in the Ectodini, which have probably diverged within the last 5 my. The idea that base composition may fluctuate widely over short time-scales has important implications for our understanding of mtDNA evolution and the application of these data to phylogenetic analysis.

Relationships among East African Cichlids

Three major clades were identified with some confidence. The first contains the Malawi taxa, *Gnathochromis,* and the Tropheini (with the significant exception of *Cyphotilapia*). A second clade encompasses the Ectodini. The third clade contains the lamprologines. The relationship of these lineages to other Tanganyikan tribes is not well defined. Deeper in the tree, an unresolved polytomy exists among *Bathybates, Boulengerochromis,* a clade containing *Tylochromis* and *Oreochromis,* and a lineage leading to the rest of the taxa.

Our tree is similar in most respects to that of Nishida (1991). His consensus tree did not resolve the relationships among the seven recent tribes of Tanganyikan cichlids. His placement of the lamprologines outside the other tribes is consistent with our data. Our data do not support the suggestion that the Eretmodini share a common ancestor with the Tropheini/Haplochromini (Sturmbauer and Meyer, 1993). Neither do we find support for the suggestion that the Cyprichromini are the sister group to the Ectodini. While it is likely that the radiation of these tribes occurred over a relatively short period of time, we believe the failure to resolve these

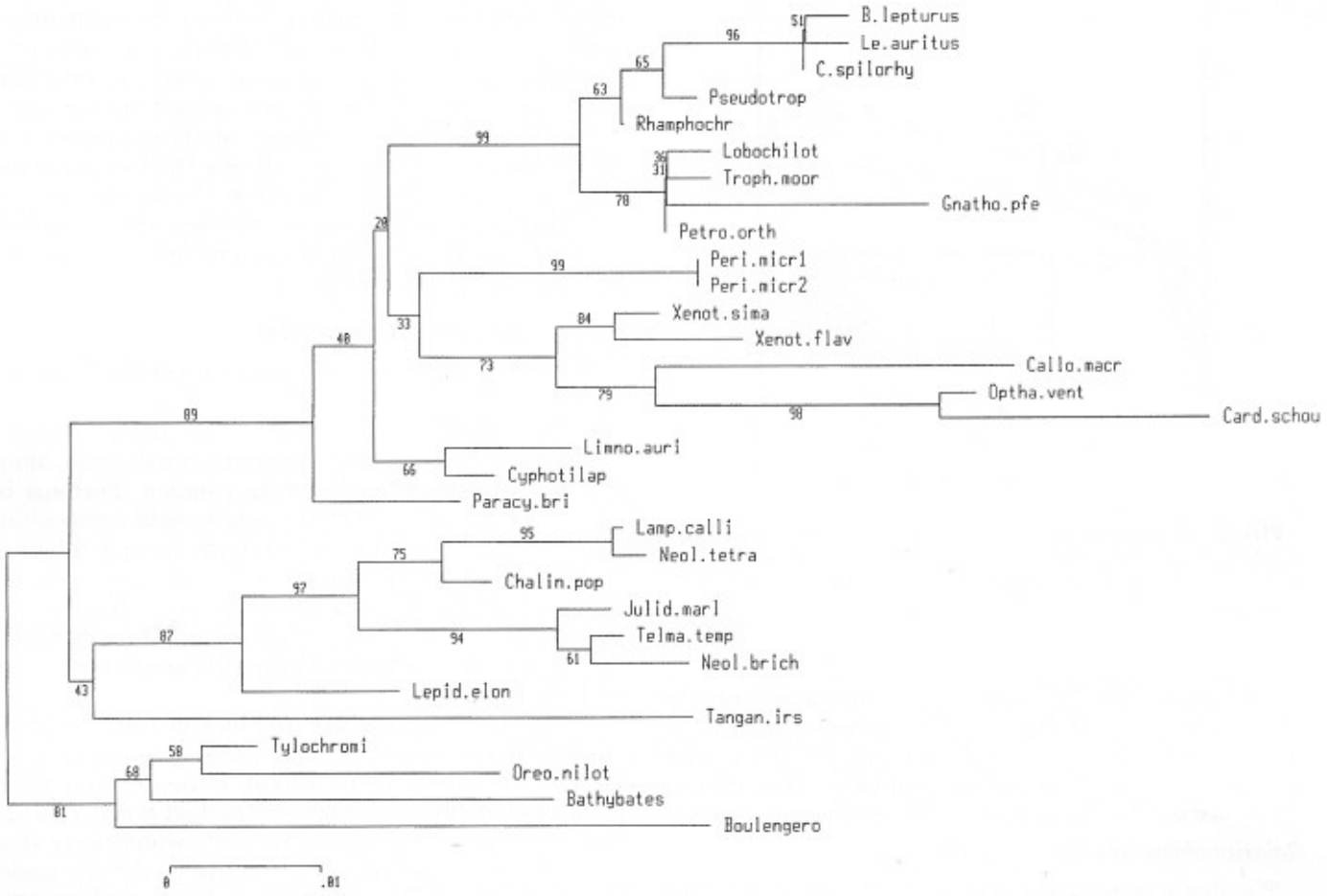

FIG. 4. Distance tree based on third position transversion divergence. Difference data were corrected for multiple hits using the Tamura–Nei model with $\alpha = 20$. The neighbor-joining algorithm of MEGA (Ver. 1.0) was used to construct the tree. The percentage of 500 bootstrap samples in which each clade appeared is given above each branch.

lineages arises because the pattern of evolution of this sequence has still not been adequately incorporated into phylogenetic algorithms.

Stiassny (1991) proposed that *Tylochromis* is the sister group to the other African cichlids included in this study. In all of our analyses, *Tylochromis* clustered with *Oreochromis* in the unrooted tree. Distance analysis strongly suggests that these are relatively closely related taxa. Analysis of additional tilapiines is underway to resolve the relationship of *Oreochromis* and *Sarotherodon* to *Tylochromis*. There is clearly a need to study other potential outgroup taxa to root the tree.

Morphological Evolution

The paradox of morphological evolution in cichlids is that, although they have undergone spectacular morphological diversification, there are no discrete, synapomorphic characters which can be used to distinguish clades (Stiassny, 1991). Molecular data are beginning to suggest a pattern of frequent reversal of many morphological characters. Sturmbauer and Meyer (1993) discuss this for members of the tribe Ectodini. Perhaps it should not surprise us that these morphological characters, which have undergone tremendous evolutionary modifications, may also have experienced a large number of character state reversals.

Oliver (1984), relying especially on characteristics of the anal fin spots, rejected the hypothesis of monophyly for the Malawi flock. Greenwood (1979) suggested that true ocellae (fin spots with a clear surrounding ring) are apomorphic to nonocellar spots. Both types are found among the Malawi taxa. True ocellae, found in the mbuna, are also found in Lake Victoria, riverine haplochromines, and Tanganyikan Tropheini. Oliver suggested that the resemblances between Malawi mbuna and the Tanganyikan genera *Tropheus, Petrochromis,* and *Simochromis* were the result of immediate common ancestry. Mitochondrial data refute this hypothesis and suggest that the ocellar nature of anal fin spots has arisen multiple times during the radiation of East African cichlids.

Yamaoka (1985) suggested that the pattern of intes-

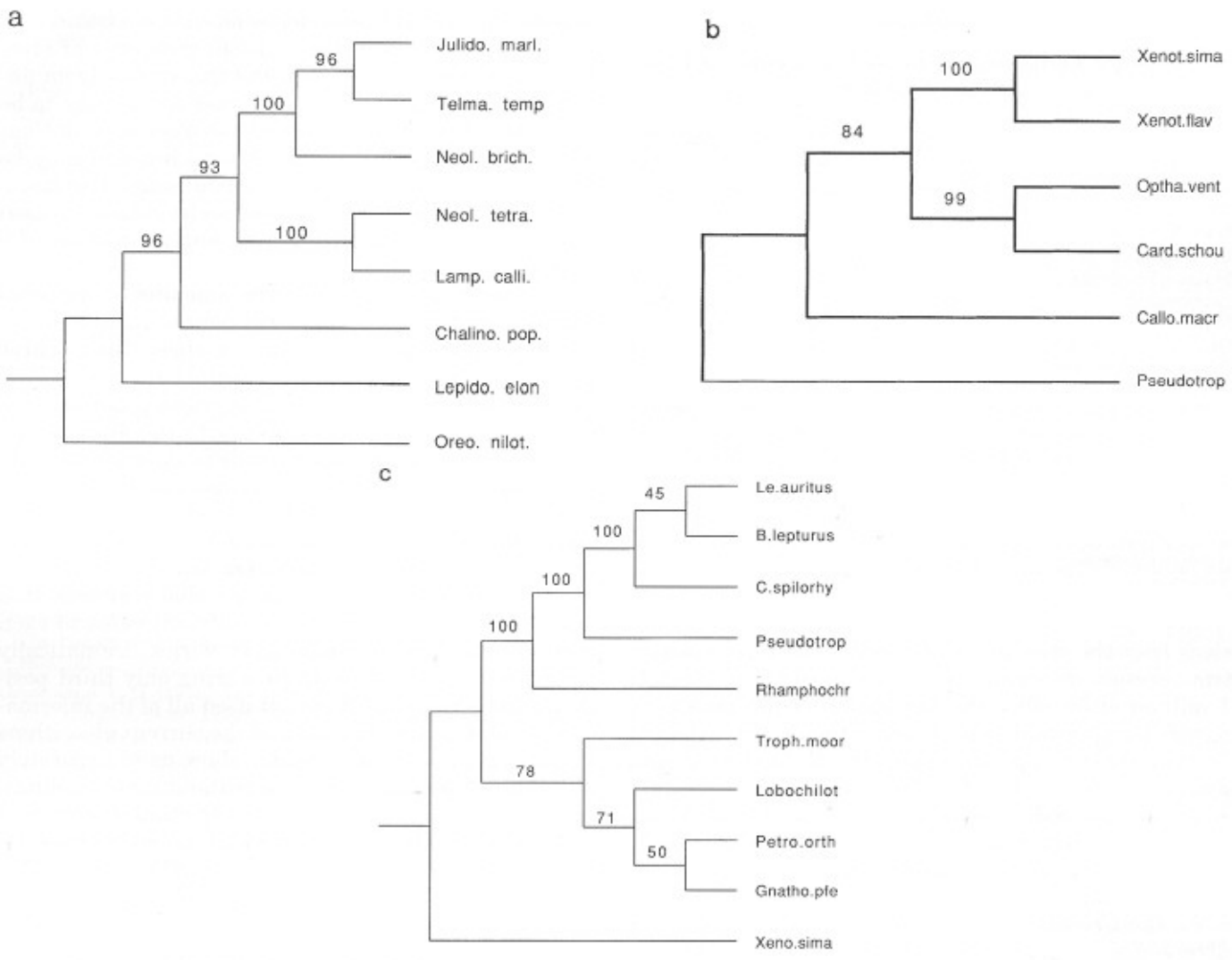

FIG. 5. Parsimony trees of tribal clusters from the distance analysis. For this analysis, all sites were weighted equally, and the degree of support was estimated from 2000 bootstrap muscles. (a) The Lamprologini rooted with *Oreochromis*. (b) The Ectodini clade rooted with *Pseudotropheus*. (c) The Malawi/Tropheini clade rooted with *Xenotilapia*.

tinal coiling was a reliable source of characters for the reconstruction of phylogeny in Tanganyikan cichlids. His grouping of taxa into *Cyathopharynx*- and *Asprotilapia*-like lineages is not consistent with Sturmbauer and Meyer's (1993) mtDNA-based phylogeny. Yamaoka (1985) also suggested that *Tanganicodus* shares a highly derived intestinal type with *Tropheus*. Our data suggest that *Tanganicodus* is only distantly related to the Tropheini (Fig. 5). Intestinal morphology may not be a reliable character for phylogenetic analysis (Reinthal, 1989).

Time-Scale for the Lacustrine Radiations

The East African cichlid flocks are unique because such a large number of species have arisen in such a short time. The actual age of the flocks, however, has been the subject of considerable debate. The most recent geologic estimates suggest that Lake Tanganyika is 9–12 my old (Cohen *et al.*, 1993). Lake Malawi is considerably younger, perhaps 2 my at most (Fryer and Iles, 1972), while Lake Victoria originated less than 750,000 years ago (Greenwood, 1974) and may have dried repeatedly during the last 200,000 years (Meyer, 1993). Molecular evidence is consistent with a very young age for the Victoria flock (Sage *et al.*, 1984; Meyer *et al.*, 1990).

Table 5 shows the third position transversion divergence estimates for important clades, using several estimates of the number of variable sites. The divergence among Lake Victoria cichlids is based on a calibration of control region divergence (Meyer *et al.*, 1990) to ND2 divergence (Fig. 3). Divergence among the Victoria flock is about 0.2%, and among the Malawi flock is about 1.7%. While it is premature to calibrate the ND2

TABLE 5

Divergence Estimates for Important Clades of East African Cichlids

	Observed no. of differences	Estimated divergence[a] 300 sites	250 sites	200 sites
Intra-Victoria	<1.0	0.2	0.2	0.2
Intra-Malawi	5.0	1.7	2.0	2.6
Malawi/Tropheini	8.1	2.8	3.4	4.2
Lamprologini	11.8	4.1	5.0	6.3
Ectodini	15.8	5.6	6.8	8.6
Oreochromis/the rest	22.0	7.9	9.7	12.4
Cichlasoma/the rest	83.0	40.3	54.5	88.6

Note. Divergence values are corrected for multiple hits assuming different numbers of sites free to accumulate substitutions.

[a] Transversion differences corrected for multiple hits according to the formula

$$\text{divergence} = -0.5 \ln (1 - 2(d/x)),$$

where d is the observed number of transversion differences, and x is the number of sites assumed to be free to vary.

clock from the present set of samples, 1% third position transversion divergence may equate to approximately 1 million years since the separation of the taxa compared. Using a multiple hit correction which assumes 300 variable third position sites, the Malawi/Tropheini divergence is about 2.8%. The Ectodini that we have sequenced shared a common ancestor at about 5.6% divergence, probably at the same time the more slowly evolving Lamprologini share an ancestor (4.1% divergence). Using the same assumption of 300 variable sites, the *Oreochromis–Tylochromis* clade shows 7.9% divergence from the other African taxa, and *Cichlasoma* shows 40.3% divergence. These latter estimates, however, are highly sensitive to the estimate of the number of variable sites and could range to as much as 12.4 and 89%, respectively. The estimates based on 250–300 variable sites are probably the most reliable and suggest that the Lamprologini and Ectodini that we have sequenced are considerably younger than the most recent estimates for the age of the lake (9–12 my). Sequencing of additional riverine taxa will be needed to address the question of how many lineages founded the Tanganyikan flock.

CONCLUSIONS

Mitochondrial sequences are clearly an effective tool for resolving the phylogeny of East African cichlids. Our work to date provides a skeletal framework for the evolution of East African cichlids, which can be fleshed out with additional sequences. Behavioral, morphological, and evolutionary biologists working on this important model system need independent tests of hypotheses of relationship. Mitochondrial sequences can provide such tests. Our analysis has demonstrated several inconsistencies in Poll's (1986) groupings of Lake Tanganyika cichlids. *Neolamprologus* is clearly an unnatural grouping. *Cyphotilapia* does not appear to be a member of the *Tropheini,* and the placement of *Gnathochromis* suggests that the Limnochromini may be heterogeneous as well. Study of additional Tropheini, riverine haplochromines and species from Lake Malawi is needed to fully characterize the ancestors of the Malawi and Victoria flocks.

It is surprising that the large amounts of sequence data reported here were not sufficient to completely resolve the phylogenetic history of these taxa. Three things conspire to limit our conclusions. First, it is likely that the radiation of this group occurred in a short interval of time associated with the filling of the lake basins. Second, it has been difficult to identify appropriate, closely related outgroup taxa. Finally, it is clear that current algorithms for reconstructing phylogeny do not allow accurate modeling of the substitutional process in this molecule. The kinetics of substitution are complex and vary among codon positions. It is not sufficient merely to assign different rates to each position, since base composition varies dramatically among positions. Although in scoring only third position transversions we have not used all of the information present in the data, none of the current algorithms for phylogenetic reconstruction allow us to accurately model the substitution process without this simplification. We hope that this set of closely related sequences will stimulate further refinement of algorithms for the recovery of phylogenetic information from DNA sequences.

ACKNOWLEDGMENTS

We thank the Government and University of Malawi for permission to collect specimens. The staff of the C.R.S.N. at Uvira, particularly M. Gashagaza and M. Nshombo, were very helpful in making arrangements for collections along the Zaire coast. M. Nagoshi and the Japanese Ecology Team were especially helpful with logistical arrangements at Uvira. M. Schreyen (Fishes of Burundi) donated many specimens and assisted in identifying Tanganyikan species. M. Hasegawa, I. Kornfield, N. Perna, and three anonymous reviewers provided helpful comments on the manuscript. This work was supported by the NSF (BSR 9007015 to T.D.K.) and the USAID Office of Science and Technology (COM-5600-G-00-0017-00 to K.R.M. and J.R.S. and DHR-5600-G-00-1043-00 to J.R.S. and T.D.K.).

REFERENCES

Anderson, S., de Bruijn, M. H. L., Coulson, A. R., Eperon, I. C., Sanger, F., and Young, I. G. (1982). Complete sequence of bovine mitochondrial DNA: Conserved features of the mammalian mitochondrial genome. *J. Mol. Biol.* **156:** 683–717.

Banister, K. E., and Clarke, M. A. (1980). A revision of the large *Barbus* (Pisces: Cyprinidae) of Lake Malawi with a reconstruction of

the history of the southern African Rift Valley lakes. *J. Nat. Hist.* **14:** 483–542.

Brichard, P. (1989). "Pierre Brichard's Book of Cichlids and All the Other Fishes of Lake Tanganyika," TFH Publications, Neptune City, NJ.

Cabot, E. L., and Beckenbach, A. T. (1989). Simultaneous editing of multiple nucleic acid and protein sequences with ESEE. *Comput. Appl. Biosci.* **5:** 233–234.

Cohen, A. S., Soreghan, M. G., and Scholz, C. A. (1993). Estimating the age of ancient lakes: An example from Lake Tanganyika, East African rift system. *Geology* **21:** 511–514.

Devereux, J., Haeberli, P., and Smithies, O. (1984). A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Res.* **12:** 387–395.

Dorit, R. L. (1986). Molecular and morphological variation in Lake Victoria haplochromine cichlids (Perciformes: Cichlidae). Ph.D. Thesis, Harvard University, Boston, MA.

Echelle, A. A., and Kornfield, I. (1984). "Evolution of Fish Species Flocks," University of Maine Press, Orono, ME.

Felsenstein, J. (1989). PHYLIP—Phylogeny inference package (Ver. 3.2). *Cladistics* **5:** 164–166.

Fryer, G., and Iles, T. D. (1972). "The Cichlid Fishes of the Great Lakes of Africa: Their Biology and Evolution," Oliver and Boyd, Edinburgh.

Greenwood, P. H. (1974). Cichlid fishes of Lake Victoria, East Africa: The biology and evolution of a species flock. *Bull. Br. Mus. Nat. Hist. (Zool.)* (Suppl.) **6:** 1–134.

Greenwood, P. H. (1979). Towards a phyletic classification of the 'genus' *Haplochromis* (Pisces, Cichlidae) and related taxa. Part I. *Bull. Br. Mus. Nat. Hist. (Zool.)* **35:** 265–322.

Greenwood, P. H. (1991). Speciation. *In* "Cichlid Fishes: Behaviour, Ecology and Evolution" (M. Keenleyside, Ed.), pp. 86–102, Chapman and Hall, London.

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. M., Kocher, T. D., and Wilson, A. C. (1991). Evolution of the cytochrome b gene of mammals. *J. Mol. Evol.* **32:** 128–144.

Kocher, T. D., and Wilson, A. C. (1991). Sequence evolution of mitochondrial DNA in humans and chimpanzees: Control region and a protein-coding region *In* "Evolution of Life: Fossils, Molecules, and Culture" (S. Osawa and T. Honjo, Eds.), pp. 391–413, Springer-Verlag, Tokyo.

Kocher, T. D., Thomas, W. K., Meyer, A., Edwards, S. V., Paabo, S., Villablanca, F. X., and Wilson, A. C. (1989). Dynamics of mitochondrial DNA sequence evolution in animals. *Proc. Natl. Acad. Sci. USA* **86:** 6196–6200.

Kocher, T. D., Conroy, J. A., McKaye, K. R., and Stauffer, J. R. (1993). Similar morphologies of cichlids in lakes Tanganyika and Malawi are due to convergence. *Mol. Phylogenet. Evol.* **2:** 158–165.

Konings, A. (1990). "Cichlids of Lake Malawi." TFH Publications, Neptune City, NJ.

Kumar, S., Tamura, K., and Nei, M. (1993). MEGA: Molecular Evolutionary Genetics Analysis, Ver. 1.0, The Pennsylvania State University, University Park, PA 16802.

Lewis, D. S. C., Reinthal, P., and Trendall, J. (1986). A guide to the fishes of Lake Malawi National Park. Gland, World Wildlife Fund, Gland, Switzerland.

Maniatis, T., Fritsch, E. F., and Sambrook, J. (1982). "Molecular Cloning: A Laboratory Manual," Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

McKaye, K. R. (1991). Sexual selection and the evolution of the cichlid fishes of Lake Malawi, Africa. "Cichlid Fishes: Behaviour, Ecology and Evolution" (M. Keenleyside, Ed.), pp. 241–257, Chapman and Hall, London.

McKaye, K., Kocher, T., Reinthal, P., Harrison, R., and Kornfield, I. (1982). A sympatric sibling species complex of *Petrotilapia trewavas* (Cichlidae) from Lake Malawi analyzed by enzyme electrophoresis. *Zool. Journ. Linn. Soc.* **76:** 91–96.

McKaye, K., Kocher, T., Reinthal, P., Harrison, R., and Kornfield, I. (1984). Genetic evidence of allopatric and sympatric differentiation among color morphs of a Lake Malawi cichlid fish. *Evolution* **38:** 215–219.

McKaye, K. R., Howard, J. H., Stauffer, J. R., Jr., Morgan, R. P., II, and Shonhiwa, F. (1993). Sexual selection and genetic relationships of a sibling species complex of bower building cichlids in Lake Malawi, Africa. *Jpn. J. Ichthyol.* **40:** 15–21.

Meyer, A. (1993). Phylogenetic relationships and evolutionary processes in East African cichlid fishes. *TREE* **8:** 279–284.

Meyer, A., Kocher, T. D., Basasibwaki, P., and Wilson, A. C. (1990). Monophyletic origin of Lake Victoria cichlid fishes suggested by mitochondrial DNA sequences. *Nature* **347:** 550–553.

Nishida, M. (1991). Lake Tanganyika as an evolutionary reservoir of old lineages of East African cichlid fishes: Inferences from allozyme data. *Experientia* **47:** 974–979.

Oliver, M. K. (1984). Systematics of African cichlid fishes: Determination of the most primitive taxon, and studies on the haplochromines of Lake Malawi. Ph.D. thesis. Yale University.

Owen, R. B., Crossley, R., Johnson, T. C., Tweddle, D., Kornfield, I., Davison, S., Eccles, D. H., and Engstrom, D. E. (1990). Major low levels of Lake Malawi and their implications for speciation rates in cichlid fishes. *Proc. R. Soc. London B* **240:** 519–553.

Perna, N. T., and Kocher, T. D. (1995). Patterns of compositional bias and skew in four-fold degenerate codon families of animal mitochondrial genomes. *J. Mol. Evol.*, in press.

Poll, M. (1986). Classification des Cichlidae du lac Tanganyika: Tribus, genres et especes. *Mem. Cl. Sci. Acad. R. Belg. (2e serie)* **XLV:** 7–163.

Ragan, C. I. (1987). Structure of NADH-ubiquinone reductase (complex I). *Curr. Top. Bioenerg.* **15:** 1–36.

Reinthal, P. N. (1989). The gross intestinal morphology of a group of rock-dwelling cichlid fishes (Pisces, Teleostei) from Lake Malawi. *Neth. J. Zool.* **39:** 208.

Ribbink, A. J., Marsh, B. A., Marsh, A. C., Ribbink, A. C., and Sharp, B. J. (1983). A preliminary survey of the cichlid fishes of rocky habitats in Lake Malawi. *South Afr. J. Zool.* **18:** 149–310.

Rzhetsky, A., and Nei, M. (1995). Tests of applicability of several substitution models for DNA sequence data. *Mol. Biol. Evol.* **12:** 131–151.

Saccone, C., Lanave, C., Pesole, G., and Preparata, G. (1990). Influence of base composition on quantitative estimates of gene evolution. *Methods Enzymol.* **183:** 570–583.

Sage, R. D., Loiselle, P. V., Basasibwaki, P., and Wilson, A. C. (1984). Molecular versus morphological change among cichlid fishes of Lake Victoria. *In* Evolution of Fish Species Flocks (A. A. Echelle and I. Kornfield, Eds.), Univ. Maine Press, Orono, ME.

Saitou, N., and Nei, M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4:** 406–425.

Scholz, C. A., and Rosendahl, B. R. (1988). Low lake stands in Lakes Malawi and Tanganyika, East Africa, delineated with multifold seismic data. *Science* **240:** 1645–1648.

Sidow, A., Nguyen, T., and Speed, T. P. (1992). Estimating the faction of invariable codons with a capture–recapture method. *J. Mol. Evol.* **35:** 253–260.

Stiassny, M. (1991). Phylogenetic intrarelationships of the family Cichlidae: An overview. *In* "Cichlid Fishes: Behaviour, Ecology and

Evolution" (M. Keenleyside, Ed.), pp. 1–35, Chapman and Hall, London.

Strauss, R. E. (1984). Allometry and functional feeding morphology in haplochromine cichlids. *In* "Evolution of Fish Species Flocks" (A. A. Echelle and I. Kornfield, Eds.), Univ. Maine Press, Orono, ME.

Sturmbauer, C., and Meyer, A. (1993). Mitochondrial phylogeny of the endemic mouthbrooding lineages of cichlid fishes from Lake Tanganyika in Eastern Africa. *Mol. Biol. Evol.* **10:** 751–768.

Swofford, D. L. (1991). Phylogenetic Analysis Using Parsimony, Ver. 3.0L, University of Illinois Natural History Survey, Champaign, IL.

Tamura, K., and Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* **10:** 512–526.

Uzzell, T., and Corbin, K. W. (1971). Fitting discrete probability distributions to evolutionary events. *Science* **172:** 1089–1096.

Yamaoka, K. (1985). Intestinal coiling pattern in the epilithic algal-feeding cichlids (Pisces, Teleostei) of Lake Tanganyika, and its phylogenetic significance. *Zool. J. Linn. Soc.* **84:** 235–261.