

Molecular phylogeny of Clupeiformes (Actinopterygii) inferred from nuclear and mitochondrial DNA sequences

Chenhong Li^{*}, Guillermo Ortí

School of Biological Sciences, University of Nebraska, Lincoln, NE 68588-0118, USA

Received 18 August 2006; revised 23 October 2006; accepted 25 October 2006
Available online 7 November 2006

Abstract

The taxonomy of clupeiforms has been extensively studied, yet phylogenetic relationships among component taxa remain controversial or unresolved. Here we test current and new hypotheses of relationships among clupeiforms using mitochondrial rRNA genes (12S and 16S) and nuclear RAG1 and RAG2 sequences (total of 4749 bp) for 37 clupeiform taxa representing all five extant families and all subfamilies of Clupeiformes, except Pristigasterinae, plus seven outgroups. Our results, based on maximum parsimony, maximum likelihood, and Bayesian analyses of these data, show that some traditional hypotheses are supported. These include the monophyly of the families Engraulidae, consisting of two monophyletic subfamilies, Engraulinae (*Engraulis* and *Anchoa*) and Coilinae (*Coilia* and *Setipinna*), and Pristigasteridae (here represented only by *Ilisha* and *Pellona*). The basal position of *Denticeps* among clupeiforms is consistent with the molecular data when base compositional biases are accounted for. However, the monophyly of Clupeidae was not supported. Some clupeids were more closely related to taxa assigned to Pristigasteridae and Chirocentridae (*Chirocentrus*). These results suggest that a major revision in the classification of clupeiform fishes may be necessary, but should await a more complete taxonomic sampling and additional data.

© 2006 Elsevier Inc. All rights reserved.

Keywords: Clupeiformes; *Sundasanax*; Phylogeny; RAG-1; RAG-2; 12S; 16S; Compositional bias

1. Introduction

Clupeiform fishes include well-known species such as herrings, sardines, and anchovies, and most of them are important components of global commercial fisheries. They have a worldwide distribution, but the majority of the 402 described species occur in the Indo-West Pacific Ocean, inhabiting marine and brackish waters, and only around 70 species are primarily freshwater fishes (Froese and Pauly, 2004; Whitehead, 1985). Clupeiforms are classified in the Clupeomorpha, a group that includes many Cretaceous fossil taxa (Grande, 1985), and the order Clupeiformes that includes all living species. Several well-defined synapomorphies support the monophyly of Clupeiformes, such as the presence of the *recessus lateralis*, a unique structure associ-

ated with the connection of the anterior portion of the swim bladder with the inner ear. Currently, the order is divided into five families: Denticipitidae, Pristigasteridae, Engraulidae, Clupeidae and Chirocentridae (Froese and Pauly, 2004; Nelson, 1994). The taxonomic composition of clupeiforms has been extensively studied (Greenwood, 1968; Whitehead, 1972, 1985; Wongratana, 1987), as the group was perceived to be a distinctive subdivision of basal teleosts. The interrelationships of clupeiforms with other basal teleosts have been hotly debated, but their current placement as the sister-group of ostariophysans (gonorhynchiforms, cypriniforms, siluriforms, characiforms, and gymnotiforms), initially proposed on the basis of molecular characters, is now generally accepted (Di Dario, 2002; Di Dario, 2004; Grande, 1985; Lecointre and Nelson, 1996; Nelson, 1967; Nelson, 1970); relationships among and within families of Clupeiformes, however, remain controversial and unresolved.

^{*} Corresponding author. Fax: +1 402 472 2083.
E-mail address: cli@unlserve.unl.edu (C. Li).

The family Denticipitidae, a group known from some West African coastal rivers, is distinguished by unique characters among clupeiforms such as the possession of denticles on roofing bones of the skull and a complete lateral line. *Denticeps*, is the only living representative of this family that has been considered to represent the basal taxon in the order (Greenwood, 1968). Chirocentridae and Clupeidae have been united under the superfamily Clupeoidea on the basis of two synapomorphies, an increase in rib to pleural vertebrae ratio and the fusion between epicentrals and ribs (Grande, 1985; Patterson and Johnson, 1995). However, a recent morphological study suggested that Chirocentridae is in fact more closely related to Engraulidae (Di Dario, 2005). The interrelationships among Pristigasteridae, Engraulidae and Clupeidae also are controversial. Clupeidae and Engraulidae were proposed to be more closely related to each other based on the presence of cartilage chevrons at the tips of epicentrals, a similar orientation of parapophyses of the second vertebra, and the complex pattern of interzygapophysal articulation (Di Dario, 2002; Patterson and Johnson, 1995). However, there is also morphological evidence (e.g., the gongyloid cartilage) supporting a close relationship between Pristigasteridae and Engraulidae (Di Dario, 2002).

Below the family level, the main problem remains in the definition of subfamilies within Clupeidae, the largest assemblage in the order. The subfamilies Dorosomatinae, Alosinae, and Clupeinae were considered as “groups of convenience”, since no osteological characters were found to support the monophyletic status of these subfamilies (Grande, 1985). Pellonulinae and Dussumieriinae were proposed as a monophyletic group (Grande, 1985), but Dussumieriinae was also thought to be polyphyletic (Eschmeyer, personal comm.). Nelson (1970) proposed that Clupeidae should be divided into two groups, Clupeinae and Dorosomatinae based on the characters in gill arches.

Another intriguing case involves the taxonomic placement of *Sundasalanx*, which is a miniature, transparent, and highly paedomorphic freshwater fish distributed in Southeast Asia (Borneo, Laos and Thailand). Seven species have been described for this genus (Britz and Kottelat, 1999; Roberts, 1981; Siebert, 1997). Originally, *Sundasalanx* was placed in its own family and considered most closely related to Salangidae (Roberts, 1981), and was indeed placed in the Salangidae later (Roberts, 1984), and Osmeridae subsequently (Fink, 1984). Based on non-paedomorphic features, such as an ossified prootic bulla and the *recessus lateralis*, Siebert (1997) proposed that *Sundasalanx* are in fact clupeids, and went even further to suggest that *Sundasalanx* is possibly the sister-group of the Caribbean endemic genus *Jenkinsia*. Britz and Kottelat (1999) criticized Siebert's conclusion and called for additional evidence to demonstrate that *Sundasalanx* shares synapomorphies with Clupeidae, advocating the need for a broader phylogenetic study. New evidence came in the form of mitochondrial genomic data when Ishiguro et al. (2005) confirmed that *Sundasalanx* grouped with clupeiforms in their phylogeny with high

bootstrap support; given the sparse taxon sampling within the order, however, their study could not resolve the phylogenetic place of *Sundasalanx* within Clupeiformes.

The current hypotheses on the phylogeny of Clupeiformes are summarized in Fig. 1. All significant studies published are based solely on morphological characters (Di Dario, 2002; Grande, 1985; Nelson, 1970; Patterson and Johnson, 1995). Mitochondrial genomic data are being used in a study on the patterns of diversification of clupeiform fishes (Sébastien Lavoué, personal comm.). In this study, we use both mitochondrial DNA (mtDNA, 12S and 16S ribosomal genes) and nuclear DNA sequence data (recombination activating genes, RAG-1 and -2) to study the phylogenetic relationships among families of Clupeiformes and among subfamilies of Clupeidae, based on a representative taxonomic sampling. Because nuclear genes have low evolutionary rate and less likely to succumb to saturation, they should provide important information complementary to the mitochondrial genes for this study.

2. Materials and methods

2.1. Taxon sampling

Thirty-seven clupeiform species and seven outgroup taxa were sampled for this study (Appendix A). The outgroup included five ostariophysan fishes, and two more distantly related taxa, *Oncorhynchus mykiss* and *Salangichthys microdon*. Ostariophysi was chosen as an outgroup for this study because it has been proposed as the sister group to Clupeomorpha (Lê et al., 1993; Lecointre and Nelson, 1996). *O. mykiss* (Salmoniformes) and *S. microdon* (Osmeriformes) were used to assess the affinities of *Sundasalanx*. The 37 clupeiform fishes examined are distributed among 22 genera. They represent all five extant families and all subfamilies of Clupeiformes, except Pristigasterinae (Fig. 1).

2.2. DNA extraction, polymerase chain reaction and sequencing

DNA samples were extracted from ethanol-preserved muscle or gill tissues using the DNeasy tissue kit (Qiagen). Fragments of two mitochondrial ribosomal genes (12S and 16S) were sequenced for this study. To design PCR and sequencing primers for the mtDNA genes, sequences for each gene were retrieved from the mitochondrial genome data of *Engraulis japonicus* (Accession Nos.: NC_003097), *Chirocentrus dorab* (NC_006913), *Sundasalanx mekongensis* (NC_006919) and *Danio rerio* (NC_002333). The primers were designed based on the alignments of these sequences. The PCR target fragment for 12S (size = 726 bp) spans sites 1249 through 1974 in the *D. rerio* mitochondrial genome. The target fragment for 16S (size = 1408 bp) spans most of this gene, from position 2178 through 3586 in the *D. rerio* mitochondrial genome. Two primers were designed for 12S while three were designed for 16S, since an additional

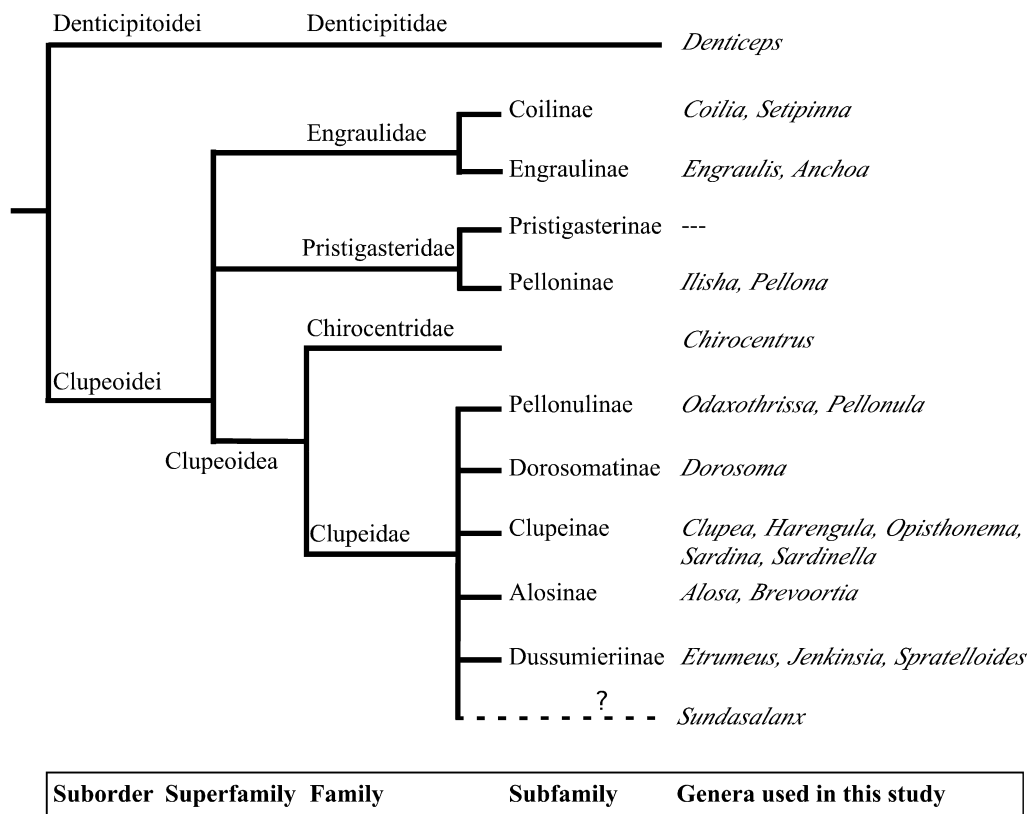


Fig. 1. Current classification of Clupeiformes (Froese and Pauly, 2004; Grande, 1985; Nelson, 1994; Whitehead, 1985).

internal primer for sequencing was necessary (Appendix B). The same thermo-cycler profiles for PCR were used for 12S and 16S gene, with 31 cycles and annealing temperature of 57 °C.

In addition to the mtDNA sequences, nuclear DNA (nucDNA) fragments of two single-copy genes were sequenced. Recombination activating genes 1 and 2 (RAG1 and RAG2) are closely linked genes coding for components of recombinase, an enzyme involved in immunoglobulin function (Hansen and Kaattari, 1996; Oettinger, 1992; Willett et al., 1997). Fragments of these genes are being used increasingly in phylogenetic studies of fishes and other vertebrates (Lopez et al., 2004; Lovejoy and Collette, 2001). Published PCR primers for RAG1 (Lopez et al., 2004) and RAG2 genes (Lovejoy and Collette, 2001) are available, but new primers also were designed for this study (Appendix B) based on alignments of published sequences from *O. mykiss* (AF137176 and U31670) and *D. rerio* (U71093 and U71094). Nested-PCR was used to amplify RAG1 and RAG2 genes for taxa that failed to be amplified in a single PCR. In these cases, products of the first-round PCR were diluted 100 times and used as the template for a second PCR with a set of primers inferred to nest within the fragment amplified in the first PCR. Conditions for amplification of the RAG1 fragment for both rounds of PCR used 15 cycles with annealing temperature at 52 °C followed by 15 cycles with annealing temperature at 51 °C. PCR and nested-PCR conditions for RAG2 used 15 cycles with annealing temperature at 62 °C followed by 15 cycles with

annealing temperature at 60 °C. Primers for PCR, nested-PCR, and sequencing are listed in Appendix B.

2.3. Sequence alignment, insertions/deletions (indels), substitution saturation test, and base compositional stationarity test

Alignments of mitochondrial gene sequences were performed using the program Clustal X (Thompson et al., 1997) with default parameters. The protein-coding regions of RAG1 and RAG2 DNA sequences were aligned based on their putative amino acid sequences (genetic code = universal) using ClustalW (under default conditions) integrated with MEGA 3.1 (Kumar et al., 2004). For some analyses, aligned amino acid sequences were back-translated to their original DNA sequences while keeping the amino acid alignment.

Inferred indels (insertion/deletion events) resulting from the alignment procedure may harbor valuable phylogenetic information in the form of shared characters, and can be used in phylogenetic analysis by coding the alignment gaps as additional characters in the data matrix (Simmons and Ochoterena, 2000). Gapped regions generated in the alignment were coded for phylogenetic analysis using the modified complex coding method (Müller, 2006) implemented in the software SeqState (Müller, 2005). The coded indels were used as additional characters together with the nucleotide sequences in parsimony and Bayesian analysis. Because the alignment was based on the amino acid sequences, the

coded gap characters were given a weight = 3 times higher than single nucleotide substitutions for parsimony analyses.

In order to check the degree of saturation for substitutions at each gene, we plotted the pair-wise absolute number of substitutions against maximum likelihood corrected sequence divergence for each data partition. A linear relationship would be expected if there is no saturation. The maximum likelihood models and parameters (Table 1) used to calculate the sequence divergence were selected by Model test v3.07 (Posada and Crandall, 1998).

To detect potential systematic errors in phylogenetic inference that may result from heterogeneous base composition among taxa, we estimated the base composition (%G+C) at variable sites for each gene; stationarity of base composition was tested further with the Chi-square test implemented in PAUP* v4.0 (Swofford, 2003). Phillips et al. (2004) showed that RY-coding (coding purines as R, and pyrimidines as Y) could effectively reduce misleading signal from biased base composition. We used this method to compare support for alternative topologies (splits) based on the raw nucleotide sequences and the RY-coded data to test whether a critical node in the tree could be an artifact produced by convergent base composition. The amount of phylogenetic signal supporting alternative branching splits was calculated using Spectronet (Huber et al., 2002).

2.4. Data partitions and phylogenetic analysis

Nuclear and mtDNA were treated as separate data partitions for phylogenetic analyses. Each genetic fragment (12S, 16S, RAG1, and RAG2) was, however, considered separately for the alignment and to characterize their evolutionary dynamics and properties. Given that both mtDNA fragments have similar function/structure (coding for ribosomal RNA) and are also tightly linked, they were considered a single partition for phylogenetic analyses. Similarly, RAG1 and RAG2 fragments were also treated as a single partition. Congruence among partitions was assessed by the ILD test (Farris et al., 1995a; Farris et al., 1995b) implemented in PAUP*. All data (mtDNA and

nucDNA partitions) were subsequently combined for a total evidence analyses.

Maximum parsimony analysis (MP) was applied on mtDNA, nucDNA and combined sequence data using PAUP*. In all cases, heuristic searches were replicated 100 times (with random addition of taxa) using tree-bisection-reconnection (TBR) branch swapping. In the parsimony analysis of mtDNA data, equal weighting of all characters was initially used and also a transversion/transition weight of 3:1 was applied, based on the known substitution dynamics of this molecule (Ortí and Meyer, 1997). To estimate statistical support for branches, bootstrap analysis with 1000 replicates was conducted in MP analysis. To test alternative hypothesis using parsimony, we used Wilcoxon signed-ranks (WS-R) tests as implemented in PAUP*. A number of *a priori* phylogenetic hypotheses was tested (Table 2). To generate trees for these hypotheses, we edited the most parsimonious tree to construct topological constraints following each of the alternative hypotheses using Treeview 1.6.6. (Page, 1996), and then used parsimony searches implemented in PAUP* v4.0 to obtain the best tree that satisfied each constraint.

Maximum-likelihood analyses (ML) were performed on mtDNA, nucDNA, and concatenated data using PAUP* v4.0 with a heuristic search option, stepwise addition, 20 replications and TBR swapping. Likelihood-ratio tests implemented in the program Modeltest v3.07 (Posada and Crandall, 1998) were used to choose the best-fit model and estimate parameters for each data partition (Table 1). Alternative hypotheses were tested by using one-tailed Shimodaira and Hasegawa (SH) tests with 1000 RELL bootstrap replicates (Shimodaira and Hasegawa, 1999). Alternative topologies were generated using maximum likelihood by a similar process described in Wilcoxon signed-ranks tests (above). ML analyses also were implemented with a mixed model using TreeFinder (Jobb et al., 2004), in which the data were partitioned into four parts: mtDNA, 1st, 2nd, and 3rd codon positions of RAG genes, each with its own optimized parameters for GTR+I+ Γ model. Bootstrap support values for ML analyses also were calculated using TreeFinder.

Table 1

Best-fit models selected by likelihood-ratio tests or the AIC implemented in Modeltest v3.07 (for DNA sequences) or Prottest (Protein sequences)

Data Partition	ML model	Estimated base frequencies	Substitution rate matrix	Invariable sites (I)	Gamma-shape parameter (α)
mtDNA	GTR+I+ Γ	A = 0.3666 C = 0.2585 G = 0.1796 T = 0.1952	$r_{A-C} = 2.1352$ $r_{A-G} = 6.5955$ $r_{A-T} = 2.9311$ $r_{C-G} = 0.4985$ $r_{C-T} = 16.553$ $r_{G-T} = 1.0000$	0.2922	0.5599
RAG (DNA)	GTR+ I+ Γ	A = 0.2199 C = 0.2945 G = 0.2711 T = 0.2146	$r_{A-C} = 1.3537$ $r_{A-G} = 4.0185$ $r_{A-T} = 1.7633$ $r_{C-G} = 1.0440$ $r_{C-T} = 5.0255$ $r_{G-T} = 1.0000$	0.3401	1.1813
RAG (protein)	JTT	Fixed	Fixed	0.29	0.98
Coded Indels	Standard	Fixed	Fixed	0	0.81

Parameters for DNA models estimated by Modeltest (PAUP*), and for protein and indel models by MrBayes.

Table 2

Maximum parsimony Wilcoxon signed-ranks test and maximum likelihood-based Shimodaira–Hasegawa test of priori hypotheses

Hypotheses tested	References	WS-R ^a	SH ^b
Pristigasteridae + Engraulidae	Di Dario (2002)	0.155	0.480
Clupeidae + Engraulidae	Di Dario (2002), Patterson and Johnson (1995)	0.121	0.650
Chirocentridae + Engraulidae	Di Dario (2005)	0.034*	0.302
Clupeidae monophyly	Grande (1985), Nelson (1994)	0.001*	0.221
Clupeinae monophyly	Nelson (1994)	0.000*	0.000*
Dussumieriinae monophyly	Grande (1985), Nelson (1994)	0.000*	0.462
<i>Sundasalanx</i> + Salangidae	Roberts (1981)	0.000*	0.002*
<i>Sundasalanx</i> + <i>Jenkinsia</i>	Siebert (1997)	0.000*	0.003*

Based on combined mtDNA and RAG DNA sequences.

^a Parsimony-based Wilcoxon signed-ranks test using a one-tailed probability (Templeton, 1983).^b Likelihood-based SH test using a one-tailed probability (Shimodaira and Hasegawa, 1999).* Significant difference at $p < 0.05$.

Bayesian analyses were implemented on mtDNA, nucDNA, RAG protein sequences, and coded indels in several combinations. DNA sequence data were partitioned in the same way as in TreeFinder ML analyses (four partitions: mtDNA, 1st, 2nd, and 3rd codon positions of RAG genes). The substitution model used was the general time reversible model with invariant sites and among-site variation (GTR+I+ Γ) for each partition. For RAG amino acid sequences the JTT model (Jones et al., 1992), allowing for invariant sites and among-site variation (JTT+I+ Γ) was used. This model was selected by ProtTest under the AIC and BIC criteria (Abascal et al., 2005). The coded indel data partition was analyzed using the standard discrete model (Lewis, 2001), allowing for among-site rate heterogeneity (Std+ Γ). All parameters were optimized for each data partition. Bayesian analysis was run in MrBayes v3.1.2 (Ronquist and Huelsenbeck, 2003) with four chains. One million generations were run with a sample frequency of 100 generations. The trees sampled before reaching stationarity of the MCMC chain were discarded for computing the consensus tree and posterior probabilities. Two independent runs were used to provide additional confirmation for the convergence of posterior probability distribution.

3. Results

3.1. Sequence variation and data partitions

Mitochondrial ribosomal 12S and 16S genes (rRNA) were sequenced for most of the 44 taxa, with a few sequences retrieved from GenBank (Appendix A). Alignment of the 12S fragments resulted in 602 sites, corresponding to sites 1304 through 1880 of *D. rerio* mitochondrial genome (NC_002333). The alignment of 16S sequences resulted in 1384 sites, corresponding to sites 2207 through 3517 of *D. rerio* mitochondrial genome (NC_002333). The combined mtDNA data has a total of 1986 characters, with 268 characters that include alignment gaps, 823 constant characters, 158 parsimony-uninformative characters, and 737 parsimony-informative characters. The null hypothesis of base composition stationarity of variable sites in the combined data was rejected by χ^2 test ($p < 0.0001$). The GC

content at variable sites ranges from a low of 35% (*Denticeps clupeoides*) and 38% (*Spratelloides delicatulus*) to 60% (*Brevoortia tyrannus*). Plots of absolute numbers of substitutions against ML corrected sequence divergence reveal a substitution saturation pattern in mtDNA data (non-linear relationship, results not shown).

The RAG1 fragment was sequenced for all taxa examined, except for *Sundasalanx mekongensis*, *Ilisha elongata*, *Sardina pilchardus* and *Alosa aestivalis* that did not amplify, perhaps due to mutations at the priming sites in these species. Alignment of all 40 taxa includes 1734 nucleotides, spanning most of exon 3 of the RAG1 gene, corresponding to site 1540 through 3006 in *D. rerio* (U71093). One previously undescribed intron was discovered in *Spratelloides gracilis*. The intron is 228 bp long and located at site 1684 (in *D. rerio*). The hypothesis of base composition homogeneity at variable sites of RAG1 gene was rejected by a χ^2 test ($p < 0.0001$). The percent of G and C (GC content) at all variable sites in the taxa examined ranged from a low of 46% in *Ictalurus punctatus* to a high of 85% in *Coilia mystus*. Most of the clupeiform fishes have high GC content, close or above 70%, except for *Denticeps clupeoides* (61%) and *Spratelloides delicatulus* (59%). Base composition homogeneity for only clupeiforms and clupeiforms without *D. clupeoides* and *S. delicatulus* also is rejected ($p < 0.0001$). The consequence of relative low GC content in *Denticeps* and *Spratelloides* on inferring their phylogenetic position will be addressed in the discussion. Plots of absolute number of substitutions against ML corrected sequence divergence reveal an almost linear relationship, indicating that there is little substitution saturation in RAG1 data (results not shown).

The RAG2 fragment was sequenced for all clupeiforms studied except for *Sundasalanx mekongensis*. Alignment of the 43 taxa yielded a length of 1647 bp, including sites 162 through 1383 corresponding to the *D. rerio* RAG2 gene (U71094). One undescribed intron was found for *Anchoa lyolepis*, spanning 390 nucleotides, located at position 1055 of the *D. rerio* gene. At the same position, another intron was found in *Spratelloides gracilis*, spanning only 226 bp and very different in sequence to the one found in *Anchoa*. Base compositional stationarity was tested for all variable

sites. The null hypothesis of stationary base composition was rejected by a χ^2 test ($p < 0.0001$). The GC content of RAG2 sequences ranged from 51% in *Hepsetus odoe* to 78% in *Sardina pilchardus*. Similar to RAG1, *S. delicatulus* (58%) and *D. clupeioides* (59%) have the lowest GC content in clupeiforms. Base composition homogeneity for only clupeiforms and clupeiforms without *D. clupeioides* and *S. delicatulus* also is rejected ($p < 0.0001$). Plots of absolute substitutions against ML corrected sequence divergence show a linear relationship, suggesting that there is little saturation in RAG2 data (results not shown).

In subsequent phylogenetic analyses, RAG1 and RAG2 were combined and analyzed together as they occur immediately adjacent to each other (Peixoto et al., 2000) and are highly correlated in GC content among taxa with a correlation coefficient $R = 0.885$ for our data. The homogeneity partition test also indicated they harbor congruent phylogenetic signal ($p > 0.05$). The combined RAG data include 2763 characters without the intron sites, consisting of 1179 invariable characters, 276 parsimony-uninformative characters and 1308 parsimony-informative characters. All alignment files are available upon request.

3.2. Phylogenies of mtDNA and nucDNA data

Analysis of mtDNA data under unweighted parsimony resulted in 4 equally short trees with $L = 5595$ steps, while the 3:1 transversion/transition ratio recovered a similar topology, with minor difference in bootstrap support value (results not shown). A phylogeny with $\ln L = -24,921$ was obtained under maximum likelihood (Fig. 2 left). Three shortest trees with $L = 6595$ steps were recovered under parsimony using RAG DNA sequences (results not shown). ML analysis resulted in a phylogeny with $\ln L = -34,606$ (Fig. 2, right). In analyses using either mtDNA or RAG DNA data, the ML topologies shown are very similar to the MP and Bayesian trees, so only the ML phylograms are presented, but bootstrap values from MP, partitioned ML analyses (from TreeFinder), and Bayesian posterior probabilities are indicated for all nodes to show the degree of congruence among results (Fig. 2).

Most parts of the mtDNA tree are consistent with the RAG tree but mtDNA provides higher resolution for relationships at intermediate levels. In both trees, Ostariophysi was found as the sister group to all clupeiforms, except for *Denticeps* (Fig. 2). *Denticeps* formed a clade with Ostariophysi to the exclusion of clupeiforms; this unexpected result may be an artifact due to shared low GC content in *Denticeps* and ostariophysans (see discussion). Both mtDNA and RAG data supported the monophyly of Engraulidae. Within Engraulidae, monophyly of subfamily Engraulinae was highly supported, while monophyletic a subfamily Coilinae was supported by mtDNA data but not by the ML analysis of RAG data (Fig. 2). However, MP analysis of RAG DNA data supported the monophyly of Coilinae with a bootstrap value of 59% (Table 3). One clade of Clupeidae, denoted as “Clupeidae I” in Fig. 2, was unanimously supported by MP,

ML and Bayesian analysis using both mtDNA and RAG sequences. This clade includes a monophyletic group of Alosinae (*Alosa* and *Brevoortia*) joined with a Clupeinae (*Sardina*), a monophyletic group of Pellonulinae (*Pellonula* and *Odaxothrissa*) and a clade including three genera of Clupeinae (*Opisthonema*, *Sardinella*, *Harengula*) joined with subfamily Dorosomatinae (*Dorosoma*). Both mtDNA and RAG data suggested that the other clupeids included in this study form part of a paraphyletic group, which is denoted as “Clupeidae II” (Fig. 2). A sister-group relationship of two Dussumieriinae genera, *Jenkinsia* and *Spratelloides* was supported (Fig. 2). However, the other Dussumieriinae genus *Etrumeus* was not closely related to them (Fig. 2).

There is low resolution for nominal clupeiform families, and this is where the major discrepancies between mtDNA data and RAG data are observed. Within Clupeoidei, mtDNA data supported Engraulidae as the basal clade, while taxa assigned to Clupeidae, Pristigasteridae, and Chirocentridae were grouped together as a polytomy (Fig. 2, left). Analyses of RAG DNA data resulted in lower resolution among nominal families: Engraulidae, Clupeidae, and Pristigasteridae were grouped as a polytomy, while Chirocentridae and Dussumieriinae (without *Etrumeus*) were placed as a basal clade to them (Fig. 2, right). The discrepancies between mitochondrial mtDNA and nuclear RAG gene data are underscored by a significant result of the homogeneity partition test ($p < 0.01$) and SH test. In SH tests, the RAG topology was rejected by mtDNA data ($p < 0.001$), while mtDNA topology was also rejected by RAG DNA data ($p < 0.001$).

3.3. Analysis of combined data and a priori hypothesis tests

In spite of conflicting phylogenetic signal between mtDNA and RAG DNA data, both data partitions were combined to explore further the resolution of clupeiform phylogeny. A total of 4749 nucleotide sites were concatenated from the RAG and mtDNA data partitions. The gene sequences that were unavailable for a few taxa were coded as missing data.

A single shortest tree with $L = 12,630$ steps was found under parsimony. One tree with $\ln L = -60,136$ was recovered under ML analysis. The Bayesian analyses produced consensus topologies highly congruent with those obtained by ML and MP analysis. The shallow clades inferred using combined data are similar to those supported by individual genes, but with higher bootstrap support (Table 3). The consensus tree obtained by mixed-model Bayesian analysis of three data partitions (mtDNA, RAG protein sequences, and coded indels) is shown in Fig. 3. ML models and parameters estimated are listed in Table 1; a summary of support values from the diverse analyses performed on individual and combined data partitions is shown in Table 3.

A number of a priori hypotheses were tested (Table 2). Both WS-R test and SH failed to reject the sister group relationship of Pristigasteridae + Engraulidae ($p > 0.05$) and sister-group relationship of Engraulidae + Clupeidae ($p > 0.05$).

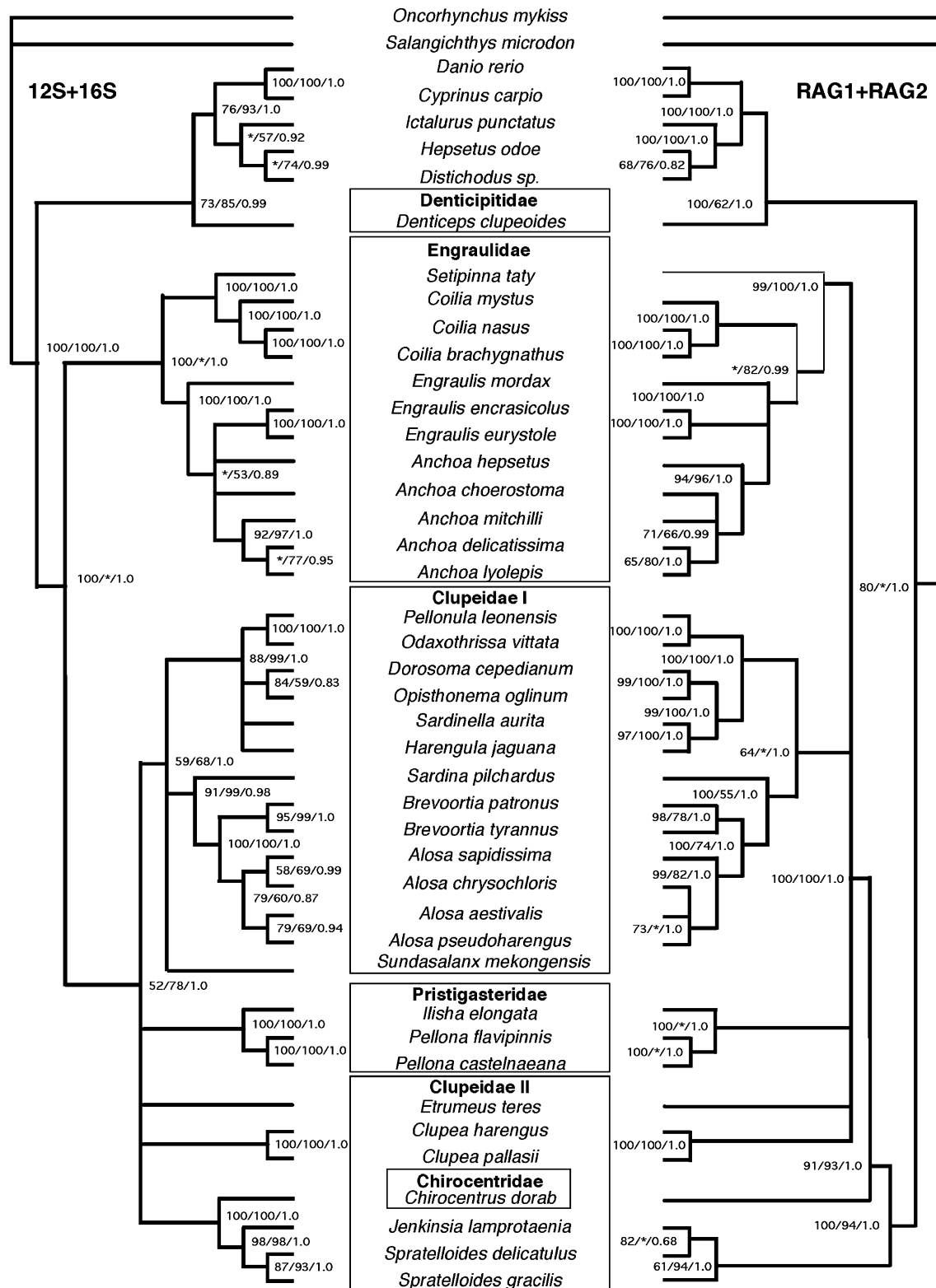


Fig. 2. Maximum likelihood phylogenies from the analysis on mtDNA sequences (1986 bp, left) and RAG nucDNA (2763 bp, right). The numbers on branches are MP bootstrap values (>50%), partitioned ML bootstrap values from TreeFinder (>50%), and Bayesian posterior probabilities, from left to right, respectively. Branches with low support (<50% for bootstrap values or <0.9 for Bayesian posterior probabilities) in more than two of the three analyses (MP, ML and Bayesian analysis) were collapsed. The asterisks indicate bootstrap values smaller than 50% or posterior probability <0.9.

WS-R test rejected the sister-group relationship between Chirocentridae and Engraulidae ($p < 0.05$), but the more conservative SH test failed to reject it ($p > 0.05$). WS-R test rejected

the monophyly of Clupeidae and monophyly of Dussumieriinae ($p < 0.01$), but SH test failed to reject them. Both WS-R test and SH test rejected the monophyly of Clupeinae

Table 3
Support values for major clades recovered in phylogenetic analyses of mtDNA and RAG gene sequences

Taxon	mtDNA ^a		RAG DNA ^a		mtDNA + RAG DNA ^c		mtDNA + RAGprot ^c		mtDNA + RAGprot + mtDNA + RAG RY coding ^c	
	MP/ML/MB	MP/ML/MB	MP/ML/MB	MP/ML/MB	MP/ML/MB	MP/ML/MB	MB	MB	MP/ML	MP/ML
Engraulidae	100/100/1.0	99/100/1.0	100/100/1.0	100/100/1.0	100/100/1.0	100/100/1.0	1.0	1.0	99/100	99/100
Engraulinae	100/100/1.0	100/100/1.0	100/100/1.0	100/100/1.0	100/100/1.0	100/100/1.0	1.0	1.0	100/100	100/100
Collinae	100/100/1.0	59/*/*	100/100/1.0	100/95/1.0	100/95/1.0	100/95/1.0	1.0	1.0	97/70	97/70
Clupeidae+Pristigasteridae + Chirocentridae	52/78/1.0	*//*/*	*//*/*	51/*/0.95	51/*/0.95	0.94	1.0	1.0	*/*63	*/*63
<i>Clupea</i> + <i>Etrumeus</i>	*//*/*	*//*/*	*//*/*	*/*63/*	*/*63/*	1.0	0.95	1.0	*/*	*/*
Pristigasteridae	100/100/1.0	100/100/1.0	100/100/1.0	100/100/1.0	100/100/1.0	1.0	1.0	1.0	100/100	100/100
<i>Chirocentrus</i> + <i>Jenkinsia</i> + <i>Spratelloides</i>	100/100/1.0	*//*/*	*//*/*	80/54/0.99	80/54/0.99	1.0	1.0	1.0	60/77	60/77
"Clupeidae I"	59/68/1.0	64/*/1.0	64/*/1.0	86/55/1.0	86/55/1.0	1.0	0.99	1.0	94/97	94/97
Alosinae + <i>Sardinia</i>	91/99/0.98	100/55/1.0	100/55/1.0	100/100/1.0	100/100/1.0	1.0	1.0	1.0	99/100	99/100
Dorosomatinae + Pellonulinae + <i>Harengula</i> + <i>Opisthonema</i>	88/99/1.0	100/100/1.0	100/100/1.0	98/100/1.0	98/100/1.0	1.0	1.0	1.0	95/100	95/100

MP, bootstrap values from MP analysis; ML, bootstrap values from ML analysis; MB, posterior probabilities from Bayesian analysis. ^aValues from Fig. 2; ^bvalues from Fig. 3; ^cPhylogenetic trees not shown. The asterisks indicate the nodes with bootstrap support lower than 50%, or nodes that were not recovered in that analysis.

($p < 0.01$), the sister-group relation between *Sundasalanx* and *Salangidae* ($p < 0.01$) and the sister-group relationship between *Sundasalanx* and *Jenkinsia* ($p < 0.01$).

4. Discussion

4.1. Compositional bias and the phylogenetic position of *Denticeps*

Denticeps is a small herring-like fish found in small fresh-water streams from southeast Benin to northwest Cameroon. Because this fish has some rare features as a teleost, such as small tooth-like structure (odontodes) on the exposed surface of most skull roofing bones, a new family, Denticipitidae was erected for it (Clausen, 1959). In spite of some peculiar characters, *Denticeps* was thought to be a clupeomorph based on several apomorphic characters shared with clupeomorph fishes (Grande, 1985; Greenwood, 1968). For example: (i) the presence of intracranial swim bladder diverticula encased in bony bullae, (ii) Hypural 2 fused with the first ural centrum at all stages of development, and an autogenous first hypural, (iii) the presence of one or more abdominal scutes (including a pelvic scute), which are composed of a single element (Grande, 1985; Greenwood, 1968). Other characters: (iv) the presence of *recessus lateralis*, (v) loss of the beryciform foramen, further suggest that Denticipitidae specifically has strong affinities to Clupeiformes of Clupeomorpha (Grande, 1982, 1985). Some clupeomorph-like characters are missing in Denticipitidae, such as a well-defined pre-epiotic fossa, dorsal scutes with a median keel, but these were thought to be secondary loss (Grande, 1985; Greenwood, 1968). Other characters, such as the *recessus lateralis*, are very different in Denticipitidae compared to other clupeiforms (Di Dario, 2004; Grande, 1985; Greenwood, 1968). With no controversy, the modern taxonomy places *Denticeps* as in its own suborder within Clupeiformes (Fig. 1), unambiguously as the sister group to all other clupeiforms (Grande, 1985; Nelson, 1994). Our phylogenetic results, grouping *Denticeps* with Ostariophysa, are, therefore, surprising (Figs. 2 and 3).

As noted above, the GC content at variable sites of RAG1 and RAG2 genes for *Denticeps clupeoides* and *Spratelloides delicatulus* were the lowest two among all clupeiforms examined, and close to the low value observed among ostariophysans. The other clupeiform taxa studied have significantly higher GC content (much higher than the average actinopterygian fish, Ortí et al., unpublished data). This pattern is repeated, albeit at a lesser degree, in the mitochondrial genes. The stationarity of nucleotide frequencies is clearly not met by the data sets used in this study. It is well-known that biased GC content can obscure the true phylogenetic signal by erroneously joining taxa with similar GC content rather than true evolutionary relationship (Foster and Hickey, 1999; Weisburg et al., 1989). RY coding has been proposed to effectively reduce the influence of biased GC content (Phillips et al., 2004); thus, support for *Denticeps* + Ostariophysa should decrease significantly when analyzing RY-coded data, if this relation-

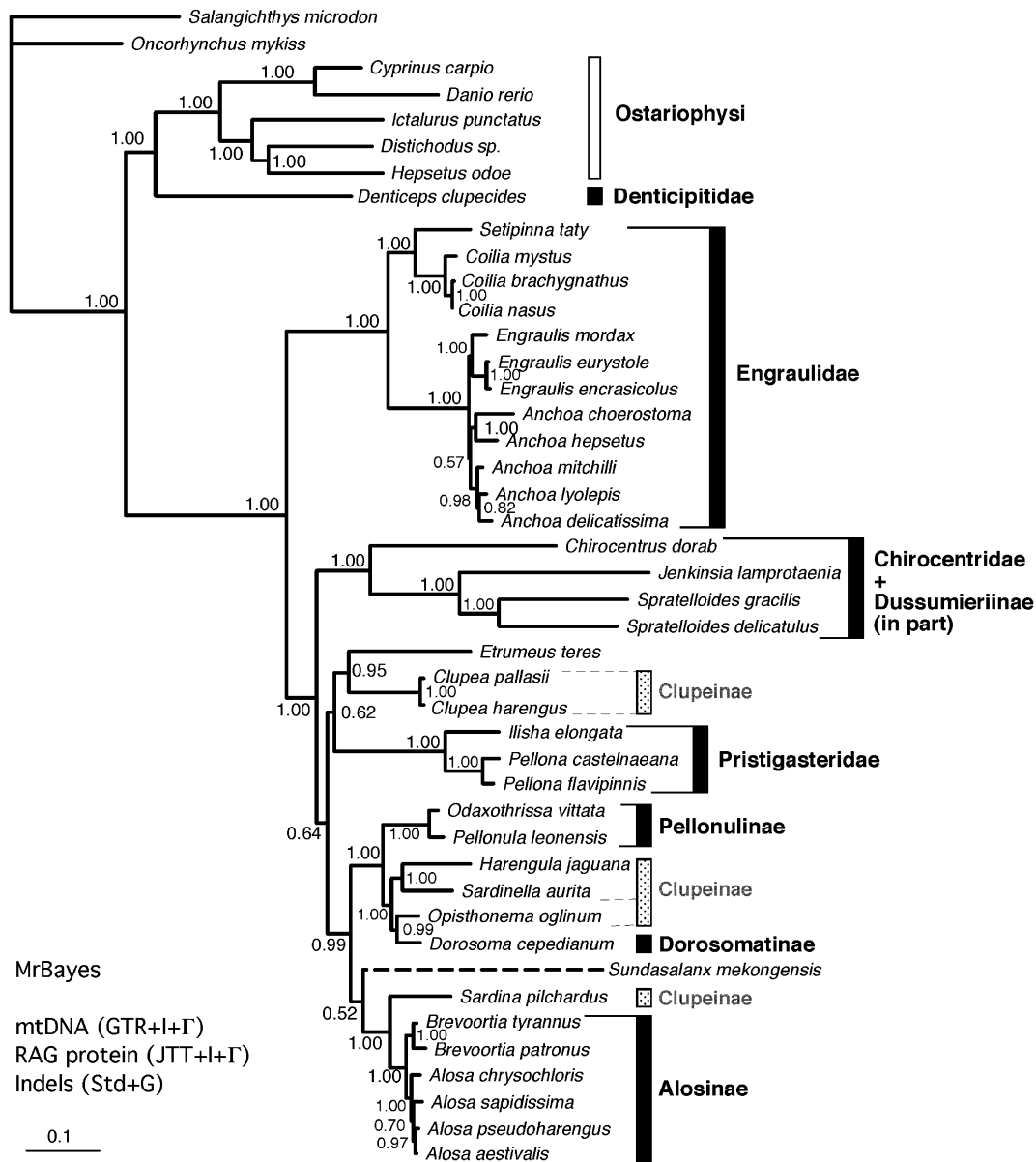


Fig. 3. Consensus phylogram of 3000 post-burnin trees obtained with mixed-model Bayesian analysis. MtDNA sequences (1986 bp) were analyzed under the GTR+I+ Γ model, RAG protein sequences (921 amino acids) under the JTT+I+G model, and coded indels (137 characters) under the Std+G model (for parameters see Table 2). Posterior probabilities are indicated next to the nodes.

ship is artificially obtained due to non-stationarity. To test this hypothesis, we calculated the branch weight (absolute number of characters that support the branch) under nucleotide-coded data (NT-coded) and RY-coded data using Spectronet (Fig. 4). The two competing hypotheses tested were: (*Denticeps* + Ostariophysi) vs. (*Denticeps* + clupeiforms). Under the NT-coded data, the former hypothesis has higher support, while under RY-coded data and the alternative wins, suggesting that a significant proportion of signal for the position of *Denticeps* shown in Figs. 2 and 3 is due to the biased (low) GC content shared between *Denticeps* and Ostariophysi. Thus, the morphology-based hypothesis of relationships for *Denticeps* is consistent with the DNA sequence data—when the analyses correctly account for base composition bias.

4.2. Engraulidae, Clupeidae, and the other clupeoid taxa

Of the four families recognized for the suborder Clupeoidei (Fig. 1), only the monophyly of Engraulidae and Pristigasteridae (in part) are well supported by the molecular data and taxa sampled in this study (Figs. 2 and 3; Table 3). The relationships of pristigasterids (*Ilisha* and *Pellona*, two genera of the subfamily Pelloninae) to the other clupeoids, and among the other clupeoid taxa inferred from the molecular data are significantly different from those implied by the currently accepted classification (Fig. 1, Table 2). The most important difference is a total lack of support for the monophyly of Clupeidae, as currently recognized. The two representative species from the genus *Clupea* are weakly related to *Etrumeus*, to the exclusion of all

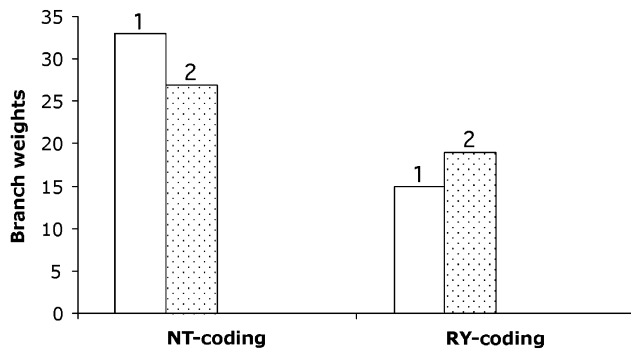


Fig. 4. Signal supporting competitive branches under NT-coding and RY-coding. Branch weights are number of characters supporting the splits calculated in Spectronet (Huber et al., 2002). The competing hypotheses are: 1. *Denticeps* + Ostariophysi; 2. *Denticeps* + Clupeiformes.

other taxa. Elements assigned to Pristigasteridae (*Ilisha* and *Pellona*) and to Chirocentridae (*Chirocentrus*) are closely related to other taxa assigned to Clupeidae.

Engraulidae (*Engraulis*, *Anchoa*, *Coilia* and *Setipinna*) is well supported as a monophyletic group by all analyses on every dataset (Figs. 2 and 3; Table 3). Within Engraulidae, the subfamilies Engraulinae (*Engraulis* + *Anchoa*) and Coilinae (*Coilia* + *Setipinna*) were also shown as monophyletic groups by all analysis except for ML and Bayesian trees using RAG data alone (Table 3).

Chirocentridae and Clupeidae have been united under the superfamily Clupeoidea by an increase in rib to pleural vertebrae ratio and fusion between epicentrals and ribs (Grande, 1985; Patterson and Johnson, 1995). However, results from a new morphological study placed Chirocentridae closer to Engraulidae (Di Dario, 2005). Our results show *Chirocentrus* closely related to the clupeids *Jenkinsia* and *Spratelloides* (Figs. 2 and 3, Table 3). This relationship also was supported by mitogenomic data (Sébastien Lavoué, personal comm.). None of our analysis placed Engraulidae as the sister taxon to Chirocentridae, however the topology tests failed to reject this hypothesis (Table 2). Clupeidae and Engraulidae were proposed to be more closely related to each other than to Pristigasteridae based on the presence of cartilage chevrons at the tips of epicentrals (Patterson and Johnson, 1995). Di Dario (2002) added two more characters (the orientation of parapophyses of the second vertebra and the complex pattern of interzygapophysal articulation) to support this hypothesis. However, there is also morphological evidence (the gongyloid cartilage) to support a close relationship between Pristigasteridae and Engraulidae (Di Dario, 2002). In most of our results, Pristigasteridae was closely related to taxa currently assigned to Clupeidae, with Engraulidae as a sister group to them (Figs. 2 and 3).

4.3. Relationships within Clupeidae

The monophyly of Clupeidae, as currently recognized, was not recovered in analyses of the molecular data sampled in this study (Figs. 2 and 3; Tables 2 and 3). Similar to results obtained with mitogenomic data, our study

suggests that *Chirocentrus* is nested within Clupeidae (Sébastien Lavoué, personal comm.). One group of clupeid taxa, identified as “Clupeidae I” (Figs. 2 and 3), was strongly supported by our data. This clade includes *Dorosoma*, closely related to three representatives of the currently recognized Clupeinae (*Sardinella*, *Opisthonema* and *Harengula*), and the Pellonulinae (*Pellonula* and *Odaxothrissa*). The second component of “Clupeidae I” are taxa currently assigned to Alosinae (*Alosa* and *Brevoortia*) plus *Sardina*. This group (Clupeidae I) also is supported by morphology of the gill arches (Nelson, 1970) and the results from mitogenomic data (Sébastien Lavoué, personal comm.). The other clupeids sampled in our study (“Clupeidae II” in Fig. 2), which include *Etrumeus*, *Jenkinsia*, *Spratelloides* and *Clupea*, are closely related to Pristigasteridae and Chirocentridae, (Figs. 2 and 3), but do not form a monophyletic group. A close relationship among *Etrumeus*, *Jenkinsia*, *Spratelloides* and *Clupea* was proposed by Nelson (Nelson, 1967; Nelson, 1970) based on the foramen in the fourth epibranchial. Our results show clearly that *Etrumeus* is not in the same clade as *Jenkinsia* and *Spratelloides* (Figs. 2 and 3), but it groups instead with *Clupea*, albeit with low support (Fig. 3, Table 3). Polyphyly of Dussumieriinae was proposed earlier based on the shape of the hyomandibular bone (Eschmeyer, personal comm.).

4.4. Phylogenetic position of *Sundasalanx*

Sundasalanx are miniature, transparent and highly pedomorphic freshwater fish distributed in Southeast Asia. Because they have unusual characters among teleosts, such as a pectoral girdle with a median cartilaginous scapulocoracoid, Roberts (1981) erected a new family, Sundasalanxidae. This family was thought to be closely related to Salangidae because they share some features, such as a single cartilaginous jaw suspension, well-developed separate fourth hypobranchials, pedunculate pectoral fins, no symplectics, no circumorbital bones, and muscles failing to meet at the ventral midline (Roberts, 1981, 1984). Fink (1984) further included Sundasalanx in the family Salangidae, while Nelson (1994) listed the Sundasalanxidae as an osmerid family. Siebert (1997) described four new species of *Sundasalanx* from Borneo, and proposed a new radical hypothesis of relationships. By closer examination of the evidence presented by Roberts, Siebert (1997) found that these characters were all pedomorphic and also plesiomorphic, being features found in larvae of lower teleosts and some euteleosts. The only few non-pedomorphic features of *Sundasalanx* include ossified prootic bulla (apomorphic for clupeomorphs) and *recessus lateralis* (apomorphic for clupeiforms), indicating a relationship of *Sundasalanx* to clupeomorph fishes. Siebert (1997) went even further to suggest that *Sundasalanx* is a spratelloidin, and possibly the sister-group of the Caribbean endemic genus *Jenkinsia*, since they both exhibit a derived, highly consolidated, caudal skeleton. Britz and Kottelat (1999) suggested that

Siebert's conclusion was premature because there was not enough evidence to demonstrate shared derived characters of *Sundasalanx* and Clupeidae. In our results, *Sundasalanx* was not supported as the sister taxon to *Salangichthys* (Table 2), but was highly supported as a clupeiform with bootstrap 100% for MP analysis, 100% for ML analysis and 1.0 for Bayesian posterior probability. *Sundasalanx* was found closely related to our “Clupeidae I” (Figs. 2 and 3) with marginal support (only mtDNA supported this relationship, RAG data were not possible to obtain in this study). A close relationship between *Sundasalanx* and *Jenkinsia* was rejected by both the WS-R test and SH test ($p < 0.01$, Table 2), against the hypothesis of Siebert (1997).

Although closely related to clupeids, the precise phylogenetic position of *Sundasalanx* within Clupeidae remains uncertain and requires further study.

Acknowledgments

We thank S. Lavoué, E. Wiley, B. Bao, X. Gong, L. Ma, J. Zhao and W. Chen for providing with samples. We thank S. Lavoué and F. Di Dario for sharing unpublished results. This work was supported by SBS Special Funds and IEEA Funds from University of Nebraska—Lincoln (to C. L.) and National Science Foundation grant DEB—9985045 (to G. O.).

Appendix A. Taxon sampling and accession number for RAG1, RAG2, and mitochondrial ribosomal genes collected in this study (DQ912027–DQ912167)

Taxa used	RAG1	RAG2	12S	16S	Museum/tissue no.
Outgroup					
<i>Oncorhynchus mykiss</i>	AF137176*	U31670*	NC_001717*	NC_001717*	—
<i>Salangichthys microdon</i>	AY380539*	—	NC_004599*	NC_004599*	—
<i>Danio rerio</i>	U71093*	U71094*	NC_002333*	NC_002333*	—
<i>Cyprinus carpio</i>	AY787040*	AY787041*	NC_001606*	NC_001606*	—
<i>Ictalurus punctatus</i>	AY423859*	AY184245*	NC_003489*	NC_003489*	—
<i>Hepsetus odoe</i>	DQ912097	AY804086*	U33825*	AY788030*	GO126
<i>Distichodus sp.</i>	DQ912098	AY804071*	U33827*	AY788012*	GO196
Denticipitidae (1 genus)					
<i>Denticeps clupeoides</i>	DQ912100	DQ912133	DQ912028	DQ912063	NSMT-P68224
Engraulidae					
Engraulinae (12 genera)					
<i>Anchoa delicatissima</i>	DQ912108	DQ912141	DQ912036	DQ912071	T510
<i>Anchoa hepsetus</i>	DQ912112	DQ912145	DQ912040	DQ912075	T1212
<i>Anchoa mitchilli</i>	DQ912113	DQ912147	DQ912042	DQ912077	C1507
<i>Anchoa choerostoma</i>	DQ912119	DQ912153	DQ912048	DQ912083	T3895
<i>Anchoa lyolepis</i>	DQ912120	DQ912154	DQ912049	DQ912084	T5152
<i>Engraulis encrasicolus</i>	DQ912103	DQ912136	DQ912031	DQ912066	No voucher ^a
<i>Engraulis mordax</i>	DQ912109	DQ912142	DQ912037	DQ912072	T550
<i>Engraulis eurystole</i>	DQ912121	DQ912155	DQ912050	DQ912085	T5153
Coilinae (5 genera)					
<i>Coilia nasus</i>	DQ912123	DQ912157	DQ912052	DQ912087	No voucher ^b
<i>Coilia brachygnathus</i>	DQ912124	DQ912159	DQ912054	DQ912089	No voucher ^b
<i>Coilia mystus</i>	DQ912126	DQ912162	DQ912057	DQ912092	No voucher ^b
<i>Setipinna taty</i>	DQ912125	DQ912161	DQ912056	DQ912091	No voucher ^b
Clupeidae					
Alosinae (7 genera)					
<i>Alosa aestivalis</i>	—	DQ912146	DQ912041	DQ912076	T1504
<i>Alosa pseudoharengus</i>	DQ912115	DQ912149	DQ912044	DQ912079	T1585
<i>Alosa sapidissima</i>	DQ912116	DQ912150	DQ912045	DQ912080	T1586
<i>Alosa chrysochloris</i>	DQ912117	DQ912151	DQ912046	DQ912081	T1910
<i>Brevoortia patronus</i>	DQ912105	DQ912138	DQ912033	DQ912068	GO602
<i>Brevoortia tyrannus</i>	DQ912106	DQ912139	DQ912034	DQ912069	GO676
Clupeinae (16 genera)					
<i>Clupea harengus</i>	DQ912114	DQ912148	DQ912043	DQ912078	T1583
<i>Clupea pallasii</i>	DQ912118	DQ912152	DQ912047	DQ912082	T3204
<i>Harengula jaguana</i>	DQ912122	DQ912156	DQ912051	DQ912086	T6543
<i>Opisthonema oglinum</i>	DQ912111	DQ912144	DQ912039	DQ912074	T1192
<i>Sardina pilchardus</i>	—	DQ912158	DQ912053	DQ912088	No voucher ^a
<i>Sardinella aurita</i>	DQ912104	DQ912137	DQ912032	DQ912067	GO598
Dorosomatinae (6 genera)					
<i>Dorosoma cepedianum</i>	DQ912099	DQ912132	DQ912027	DQ912062	No voucher ^c
Dussumieriinae (4 genera)					
<i>Etrumeus teres</i>	DQ912110	DQ912143	DQ912038	DQ912073	T1052

Appendix A (continued)

Taxa used	RAG1	RAG2	12S	16S	Museum/tissue no.
<i>Jenkinsia lamprotaenia</i>	DQ912107	DQ912140	DQ912035	DQ912070	T216
<i>Spratelloides delicatulus</i>	DQ912128	DQ912164	DQ912058	DQ912093	No voucher
<i>Spratelloides gracilis</i>	DQ912129	DQ912165	DQ912059	DQ912094	No voucher
Pellonulinae (23 genera)					
<i>Pellonula leonensis</i>	DQ912130	DQ912166	DQ912060	DQ912095	No voucher
<i>Odaxothrissa vittata</i>	DQ912131	DQ912167	DQ912061	DQ912096	No voucher
Sundasalangidae (1 genus)					
<i>Sundasalanx mekongensis</i>	—	—	AP006232*	AP006232*	No voucher
Chirocentridae (1 genus)					
<i>Chirocentrus dorab</i>	DQ912127	DQ912163	AP006229*	AP006229*	No voucher
Pristigasteridae (9 genera)					
<i>Ilisha elongata</i>	—	DQ912160	DQ912055	DQ912090	No voucher ^b
<i>Pellona flavipinnis</i>	DQ912101	DQ912134	DQ912029	DQ912064	GO309
<i>Pellona castelnaeana</i>	DQ912102	DQ912135	DQ912030	DQ912065	GO325

NSMT number: vouchers from the National Science Museum, Tokyo.

C and T number: vouchers from The University of Kansas Natural History Museum, KS, USA.

GO number: tissue collection of G. Ortí, University of Nebraska, Lincoln, NE, USA.

^a Tissue samples provided by W. Chen, Saint Louis University, MO, USA.

^b Collected from Pudong, Shanghai, China.

^c Collected from Hershey, Nebraska, USA.

* Sequences taken from GenBank.

Appendix B. Primers used for PCR and sequencing

Primers	Sequences	Reference
<i>For 12S</i>		
12S229F ^a	5' GYCGGTAAAAYTCGTGCCAG 3'	This study
12S954R ^a	5' YCCAAGYGCACCTTCCGGTA 3'	This study
<i>For 16S</i>		
16S135F ^a	5' GCAATAGAVAWAGTACCGCAAGG 3'	This study
16S964F ^a	5' YTCGCCTGTTTACCAAAAAC 3'	This study
16S1072R	5' CCTTYGCACGGTYARAATAC 3'	This study
<i>For RAG1</i>		
RAG1-2510F ^a	5' TGGCCATCCGGGTMAACAC 3'	This study
RAG1-2533F ^b	5' CTGAGCTGCAGTCAGTACCATAAGATGT 3'	(Lopez et al., 2004)
RAG1-3098F	5' TGTGCCTGATGYTYGTDGAYGART 3'	This study
RAG1-3222F	5' TCYTTCCGCTTYCACTTCCG 3'	This study
RAG1-3261R	5' CCCTCCATYTCNCGMACCATCTT 3'	This study
RAG1-3543R	5' GTRGCRTTGCCRATRTCRCACT 3'	This study
RAG1-4063R	5' TTCTGNARRTACTTGGARGTGTAWAGCCA 3'	This study
RAG1-4078R ^b	5' TGAGCCTCCATGAACCTTCTGAAGRTAYTT 3'	(Lopez et al., 2004)
RAG1-4090R ^a	5' CTGAGTCCTTGTGAGCTTCCATRAAYTT 3'	(Lopez et al., 2004)
<i>For RAG2</i>		
RAG2-F1 ^a	5' TTYGGNCARAARGGVTTGGC 3'	This study
RAG2-F2 ^b	5' AARCGCTCMTGTCCMACTGG 3'	(Lovejoy and Collette, 2001)
RAG2-526F	5' GTGGACTGCCCCCKMAGGTSTT 3'	This study
RAG2-1096F	5' CAGGGCTRCAGCCAGGARTC 3'	This study
RAG2-514R	5' CAGTCCACCAYRCTGTTCCA 3'	This study
RAG2-1145R	5' AAGTAGAGCTCCTCNGAGTCC 3'	This study
RAG2-R6	5' TGRTCARGCAGAAGTACTTG 3'	(Lovejoy and Collette, 2001)
RAG2-1466R ^b	5' CCRTGRTCCARGCAGAAGTACTT 3'	This study
RAG2-1453R ^a	5' CCRTGRTCCARGCAGAAGTA 3'	This study

^a Primers for first-round PCR.

^b Primers for second round nested-PCR; the other primers were used for sequencing only.

References

- Abascal, F., Zardoya, R., Posada, D., 2005. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* 21, 2104–2105.
- Britz, R., Kottelat, M., 1999. *Sundasalanx mekongensis*, a new species of clupeiform fish from the Mekong basin (Teleostei: Sundasalangidae). *Ichthyol. Explor. Freshwaters* 10, 337–344.
- Clausen, H.S., 1959. Denticipitidae, a new family of primitive isospondylous teleosts from West African fresh-water. *Videnskabelige Meddelelser fra Dansk Naturhistorisk Forening, Kjobenhavn* 121, 141–151.
- Di Dario, F., 2002. Evidence supporting a sister-group relationship between Clupeoidea and Engrauloidea (Clupeomorpha). *Copeia* 102, 496–503.

- Di Dario, F., 2004. Homology between the recessus lateralis and cephalic sensory canal, with the proposition of additional synapomorphies for the Clupeiformes and the Clupeoidei. *Zool. J. Linn. Soc.* 141, 257–270.
- Di Dario, F., 2005. Relações filogenéticas entre os grandes grupos de Clupeomorpha e suas possíveis relações com Ostariophysi (Actinopterygii, Teleostei). Universidade de São Paulo. pp. 629.
- Farris, J.S., Källersjö, M., Kluge, A.G., Bult, C., 1995a. Constructing a significance test for incongruence. *Syst. Biol.* 44, 570–572.
- Farris, J.S., Källersjö, M., Kluge, A.G., Bult, C., 1995b. Testing significance of incongruence. *Cladistics* 10, 315–319.
- Fink, W.L., 1984. Basal euteleosts: relationships. In: Moser, H.G., Richards, W.J., Cohen, D.M., Fahay, M.P., Kendall, A.W., Richardson, S.L. (Eds.), *Ontogeny and systematics of fishes*. American Society of Ichthyologists and Herpetologists, Special Publication No. 1. Allen Press, Lawrence, Kansas, pp. 202–206.
- Foster, P.G., Hickey, D.A., 1999. Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. *J. Mol. Evol.* 48, 284–290.
- Froese, R., Pauly, D., 2004. FishBase. World Wide Web electronic publication. www.fishbase.org, version (11/2005).
- Grande, L., 1982. A revision of the fossil genus *Knightia*, with a description of a new genus from the Green River Formation (Teleostei, Clupeidae). *Am. Mus. Novitates* 2731, 1–22.
- Grande, L., 1985. Recent and fossil Clupeomorph fishes with materials for revision of the subgroups of clupeoids. *Bull. Am. Mus. Nat. Hist.* 181, 231–372.
- Greenwood, P.H., 1968. The osteology and relationships of the Denticipitidae a family of Clupeomorph fishes denticiceps-clupeoides palaeodenticiceps-tanganikae. *Bull. Br. Mus. (Nat.-Hist.) Zool.* 16, 15–273.
- Hansen, J.D., Kaattari, S.L., 1996. The recombination activating gene 2 (RAG2) of the rainbow trout *Oncorhynchus mykiss*. *Immunogenetics* 44, 203–211.
- Huber, K.T., Langton, M., Penny, D., Moulton, V., Hendy, M., 2002. Spectronet: a package for computing spectra and median networks. *Appl. Bioinform.* 1, 159–161.
- Ishiguro, N.B., Miya, M., Inoue, J.G., Nishida, M., 2005. Sundasalanx (Sundasalangidae) is a progenetic clupeiform, not a closely-related group of salangids (Osmeriformes): mitogenomic evidence. *J. Fish Biol.* 67, 561–569.
- Jobb, G., von Haeseler, A., Strimmer, K., 2004. TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. *BMC Evol. Biol.* 4, 18.
- Jones, D.T., Taylor, W.R., Thornton, J.M., 1992. The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* 8, 275–282.
- Kumar, S., Tamura, K., Nei, M., 2004. MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. *Brief Bioinform.* 5, 150–163.
- Lecointre, G., Nelson, G., 1996. Clupeomorpha, sister group of Ostariophysi. In: Stiassny, M.L.J., Parenti, L.R., Johnson, G.D. (Eds.), *Interrelationships of fishes*. Academic Press, San Diego, pp. 193–207.
- Lê, H.L., Lecointre, G., Perasso, R., 1993. A 28S rRNA-based phylogeny of the gnathostomes: first steps in the analysis of conflict and congruence with morphologically based cladograms. *Mol. Phylogenet. Evol.* 2, 31–51.
- Lewis, P.O., 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. *Syst. Biol.* 50, 913–925.
- Lopez, A.J., Chen, W.J., Ortí, G., 2004. Esociform phylogeny. *Copeia* 2004, 449–464.
- Lovejoy, N.R., Collette, B.B., 2001. Phylogenetic relationships of new world needlefishes (Teleostei: Belontiidae) and the biogeography of transitions between marine and freshwater habitats. *Copeia* 2001, 324–338.
- Müller, K., 2005. SeqState: primer design and sequence statistics for phylogenetic DNA datasets. *Appl. Bioinform.* 4, 65–69.
- Müller, K., 2006. Incorporating information from length-mutational events into phylogenetic analysis. *Mol. Phylogenet. Evol.* 38, 667–676.
- Nelson, G.J., 1967. Gill arches of teleostean fishes of the family Clupeidae. *Copeia*, 389–399.
- Nelson, G.J., 1970. The hyobranchial apparatus of teleostean fishes of the families Engraulidae and Chirocentridae. *Am. Mus. Novitates* 2410, 1–30.
- Nelson, J.S., 1994. *Fishes of the World*. J. Wiley, New York.
- Oettinger, M.A., 1992. Activation of V(D)J recombination by RAG1 and RAG2. *Trends Genet.* 8, 413–416.
- Ortí, G., Meyer, A., 1997. The radiation of characiform fishes and the limits of resolution of mitochondrial ribosomal DNA sequences. *Syst. Biol.* 46, 75–100.
- Page, R.D., 1996. TreeView: an application to display phylogenetic trees on personal computers. *Comput. Appl. Biosci.* 12, 357–358.
- Patterson, C., Johnson, G.D., 1995. The intermuscular bones and ligaments of teleostean fishes. *Smiths Contrib. Zool.* 559, 1–85.
- Peixoto, B.R., Mikawa, Y., Brenner, S., 2000. Characterization of the recombination activating gene-1 and 2 locus in the Japanese pufferfish, *Fugu rubripes*. *Gene* 246, 275–283.
- Phillips, M.J., Delsuc, F., Penny, D., 2004. Genome-scale phylogeny and the detection of systematic biases. *Mol. Biol. Evol.* 21, 1455–1458.
- Posada, D., Crandall, K.A., 1998. MODELTEST: testing the model of DNA substitution. *Bioinformatics* 14, 817–818.
- Roberts, T.R., 1981. Sundasalangidae, a new family of minute freshwater salmoniform fishes from southeast Asia. *Proc. Calif. Acad. Sci.* 42, 295–302.
- Roberts, T.R., 1984. Skeletal anatomy and classification of the neotenic Asian salmoniform superfamily Salangoidea (icefishes or noodlefishes). *Proc. Calif. Acad. Sci.* 43, 179–220.
- Ronquist, F., Huelsenbeck, J.P., 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19, 1572–1574.
- Shimodaira, H., Hasegawa, M., 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol. Biol. Evol.* 16, 1114–1116.
- Siebert, D.J., 1997. Notes on the anatomy and relationships of Sundasalanx Roberts (Teleostei, Clupeidae), with descriptions of four new species from Borneo. *Bull. Nat. Hist. Mus. Zool. Ser.* 63, 13–26.
- Simmons, M.P., Ochoterena, H., 2000. Gaps as characters in sequence-based phylogenetic analyses. *Syst. Biol.* 49, 369–381.
- Swofford, D.L., 2003. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.
- Templeton, A.R., 1983. Phylogenetic inference from restriction endonuclease cleavage site maps with particular reference to the evolution of humans and apes. *Evolution* 37, 221–244.
- Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., Higgins, D.G., 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* 25, 4876–4882.
- Weisburg, W.G., Giovannoni, S.J., Woese, C.R., 1989. The Deinococcus-Thermus phylum and the effect of rRNA composition on phylogenetic tree construction. *Syst. Appl. Microbiol.* 11, 128–134.
- Whitehead, P.J.P., 1972. A synopsis of the clupeoid fishes of India. *J. Mar. Biol. Assn. India* 14, 160–256.
- Whitehead, P.J.P., 1985. FAO species catalogue. Vol. 7. Clupeoid fishes of the world (suborder Clupeoidei). An annotated and illustrated catalogue of the herrings, pilchards, sprats, shads, anchovies and wold-herrings. Part 1—Chirocentridae, Clupeidae and pristigasteridae. FAO Fisheries-Synopsis: 303, No. 125 (7 Part 1).
- Willett, C.E., Cherry, J.J., Steiner, L.A., 1997. Characterization and expression of the recombination activating genes (rag1 and rag2) of zebrafish. *Immunogenetics* 45, 394–404.
- Wongratana, T., 1987. Four new species of clupeoid fishes Clupeidae and Engraulidae from Australian waters. *Proc. Biol. Soc. Wash.* 100, 104–111.