Alex Dornburg

Model-Averaged Phylogenetic Inference of the Triggerfishes (Family: Balistidae)

Project Summary

Abstract:

The triggerfishes (Family Balistidae) comprise approximately 40 species in 11 genera and are among the most conspicuous diurnal inhabitants of coral reef communities worldwide. Despite their ecological and commercial importance, relatively little is known of their interspecific relationships. Here we present a novel molecular phylogenetic hypothesis for the Family Balistidae based on an analysis of two mitochondrial (12S, 16S) and three nuclear genes (4C4, Rhodopsin, RAG1) sampled from 26 species. As part of our analysis, we implemented a recently developed reversible jump MCMC sampler that attempts to account for uncertainty in the underlying model of molecular substitution in Bayesian analysis. Clade posterior probabilities as well as parameter estimates thus reflect uncertainty in the choice of a substitution model. Comparison of model-averaged posterior probabilities with those from traditional MCMC reveals that model averaging had a negligible effect on clade posterior probabilities. The use of model averaging approaches in phylogenetics is an area of growing interest and this is the first study to test its effects on clade support values. Our phylogenetic results strongly support the monophyly of the family but suggest that the genera *Balistoides* and *Pseudobalistes* are paraphyletic with respect to other balistids. We also found strong support for *Sufflamen* as a sister group to the remaining *Balistidae*. We found *Abalistes* to be the sister group to a large clade including *Melichthys*, *Xanthichthys*, *Balistes*, and *Canthidermis*. Bayesian divergence time estimates under an uncorrelated rates model suggest that the MRCA of the crown group Balistidae appeared during the Late Oligocene (~25 MYA). The timing and pattern of diversification events suggests that the triggerfish radiation may have been influenced by the Mid-Miocene regionalization of marine invertebrate communities.

Introduction:

The members of the tetraodontiform family balistidae are mostly tropical in distribution, and many aspects of their ecology, in particular their reproductive strategies (e.g. Kawase, 2003), have been studied intensively. Although there have been some previous phylogenetic studies based on morphological characters (Matsuura, 1979), relationships within the group remain

poorly understood. Phylogenies form the backbone of comparative and biogeographic analyses and resolving the intra-relationships of the balistids will be of high value to future studies concerning this group. Here we present a novel molecular phylogenetic hypothesis for the Family Balistidae based on an analysis of two mitochondrial (12S, 16S) and three nuclear genes (4C4, Rhodopsin, RAG1) sampled from 26 species plus three outgroup taxa. Using this data matrix and fossil calibration points we generate a chronogram to assess novel spatial and temporal distribution patterns within the family. We also explore theoretical phylogenetic concepts concerning the effects of model selection on clade support.

Little work has been done to test how model uncertainty influences phylogenetic inference. Selecting an appropriate model of sequence evolution is important in phylogenetic analyses, as poor model choice can lead to incorrect inference (Alfaro and Huelsenbeck, 2006; Buckley and Cunningham, 2002; Buckley et al., 2001; Sullivan and Swofford, 1997). Furthermore, the most commonly practiced methods of model selection, such as ModelTest (Posada and Crandall, 1998) do not incorporate all possible models of molecular evolution (Alfaro and Huelsenbeck, 2006). Using these methods, it is common for large data sets to use a highly parameterized general time reversible (GTR) model of evolution, which may not be the best model for the data. In this study we use a reversible jump Markov chain Monte Carlo (RJ-MCMC) sampler to calculate posterior probabilities (PP's) that are averaged across all of the 203 possible time reversible substitution models (Alfaro and Huelsenbeck, 2006; Huelsenbeck et al, 2004). We compare these PP's to an analysis where inference is conditioned on the choice of a single substitution model (GTR + G in this study) to examine how accommodating uncertainty in model choice influences phylogenetic support (clade PP). This is the first study to test the effect that model selection and potential over parameterization have on PP values.

Methods:

Sampling and DNA Extraction.

Twenty-six ingroup taxa were sampled that represented all genera of the Family Balistidae (Table 1), except *Xenobalistes*, a rare genus that contains two recently described species: (1) *Xenobalistes puncatus* (Heemstra and Smith, 1983); and (2) *Xenobalistes tumidipectoris* (Matsuura, 1981). Representatives from three genera in the family Monacanthidae

were chosen as outgroups, as monacanthids are considered to be the sister group to the balistids (Alfaro et. al, 2006; Holcroft, 2005; Santini and Tyler, 2003; Leis, 1984; Rosen, 1984; Tyler, 1980; Winterbottom, 1974). DNA was extracted for most taxa using the Chelex (Bio-Rad) protocol described in Walsh et al. (1991). Additional extractions for *Balistes vetula, Pseudobalistes fuscus,* and *Balistes capriscus* utilized the PureGene extraction kit and protocol (Gentra Systems). Sequences for *Balistes polylepis* and *Balistoides viridescens* were downloaded from GenBank, taken from Holcroft (2005).

PCR Amplification and Sequencing

Muscle tissue samples were stored in 70% ethanol prior to use. We used the polymerase chain reaction (PCR) (Saiki, 1990) and primer pairs for each gene to amplify two mitochondrial genes, 12S rDNA (833 bp) and 16S rDNA (563 bp), and three nuclear genes, Rhodopsin (564 bp), Tmo4C4 (575 bp), and RAG1 (1471 bp). 1 μL of genomic template was used per 25 μL reaction, containing 5 μL of 5X GoTaq Flexi PCR buffer (Promega), 2 μL $MgCl_2$ (25mM), .5μL dNTPs (8μM), 1.25 μL of each primer, and 0.125 μL of Promega GoTaq Flexi DNA polymerase (5u/μL). Amplification of the mitochondrial 12S gene fragment was conducted with an initial denaturing step conducted at 94°C for 1 min; 37 cycles with a 1 min. 94°C denature, 45 sec. 60° annealing, and 1 min. 72°C extension, followed by an additional 5 min. 72°C extension and a 10 min. 23°C cool down. The 16S rDNA gene fragment was amplified using an initial denaturing step conducted at 94°C for 1 min; 36 cycles with a 30 sec. 94°C denature, 45 sec. 50° annealing, and 1 min. 72°C extension, followed by an additional 5 min. 72°C extension and a 3 min. 23°C cool down. The Rhodopsin and Tmo4C4 protocol followed the above steps, substituting a 53°C annealing temperature for amplification of Rhodopsin and a 48.5°C annealing temperature for Tmo4C4. Product for the RAG1 gene fragment was obtained through the use of an initial denaturing step conducted at 94°C for 2 min; 38 cycles with a 1 min. 94°C denature, 75 sec. 50° annealing, and 2 min. 72°C extension, followed by an additional 5 min. 72°C extension and a 3 min. 23°C cool down. PCRs were performed on two MJ Research PTC-200 Peltier thermal cyclers and a Bio-Rad iCycler. All products were stored at -4°C after amplification.

Excess dNTP's and unincorporated primers were removed from PCR products using ExoSap (Amersham Biosciences). Purified products were cycle-sequenced using the BigDye Terminator v.3.1 cycle sequencing kit (Applied BioScience) with each gene's original or

additional internal primers (Table 2) used for amplification. The cycle sequencing protocol consisted of 25 cycles with a 10 sec. 94°C denaturation, 5 sec. of 50°C annealing, and a 4 min 60°C extension. Sequences were produced at the Washington State University Center for Integrated Biotechnology Core Laboratory using an ABI 377 and an ABI3100.

Sequence Alignment

12S and 16S ribosomal gene sequences were aligned by eye to secondary structure models used in a previously published analysis for labrid fishes (Clements et al., 2003). Ambiguously aligned regions were removed prior to analysis for both mitochondrial genes. The additional Rhodopsin, Rag1, and Tmo4C4 gene fragments were aligned by eye in a NEXUS file using BBEdit (BareBones Software) The sequences were trimmed to the size of the smallest fragment for each gene, and ambiguous ribosomal characters were removed using Sequencher 4.2.2 (Gene Codes Corp.) and Se-Al v.2.0 (Rambaut, 1996) to minimize missing characters in the data matrix. Our final data matrix consisted of an 820 character 12S gene partition, a 549 character 16S gene partition, a 404 character Rhodopsin gene partition, a 545 character Tmo4C4 gene partition, and a 1205 character Rag1 gene partition for a total of 3,523 characters used in analysis. Sequences from individual gene partitions were checked using NCBI's Basic Local Alignment Search Tool (BLAST) and subsequently deposited in GenBank.

Bayesian Model Averaging:

Little work has been done to test how model uncertainty influences phylogenetic inference. While no model of evolution provides an exact model of past events, model selection is seen as a way of approximating reality in phylogenetic analyses (Posada and Buckley, 2004). This selection process has been shown to effect results, as poor model choice will lead to incorrect inference (Alfaro and Huelsenbeck, 2006; Buckley and Cunningham, 2002; Buckley et al., 2001; Sullivan and Swofford, 1997). In this study we used a 20 million generation reversible jump Markov chain Monte Carlo (RJ-MCMC) (Alfaro and Huelsenbeck, 2006; Huelsenbeck et. Al, 2004) to calculate posterior probabilities (PP's) that were averaged across all of the 203 possible time reversible substitution models using a custom program (All Models) (Huelsenbeck et al, 2004). Clade PP's from these analyses reflected uncertainty in the choice of all 203 time-reversible substitution models as well as all other model parameters and were calculated after the

first 20% of the 20,000 trees were discarded as burnin. As a control, we repeated these analyses using a slightly modified version of All Models that remained in the GTR model (only) and compared posterior probabilities for all individual gene partitions, a concatenated mitochondrial gene (12S +16S) data matrix, a concatenated nuclear gene (Rhodopsin, Tmo4C4, Rag1) data matrix, and the entire concatenated data set.

Divergence Time Estimation:

We used five fossils and one node calibration point taken from a separate study of the tetraodontiform fishes (Alfaro et al, *submitted*) to constrain node ages in the balistid tree and reanalyzed our data under a model of uncorrelated but log-normally distributed rates using BEAST (Drommond and Rambaut, 2003). The minimum node age of the MRCA between the monacanthids and balistids was calibrated using the stem balistoid fossils *Balistomorphus orbiculatus*, *B. ovalis*, *B. spinosus*, and *Oligobalistes robustus*, all of which have been dated to the early Oligocene (Santini and Tyler, 2004). Following the description in Alfaro et al. (2006), we assigned a prior minimum age of 35 MY, a mean age of 50 MY, and an upper bound of 70 MY. For the split between *Pseudobalistes* and *Balistes* the fossil *Balistes procapriscus* from the late Miocene was utilized (Santini and Tyler, 2003). We assigned a minimum age of 5 MY for the split and an upper bound of 50 MY. A pure birth prior for cladogenic rates was assigned, and the GTR + G + I model was selected as our best approximate model of evolution for the concatenated data set based on ModelTest (Posada and Crandall, 1998) and our model averaged results. We conducted three 20 million generation runs, and analyzed the resulting data using Tracer v.1.3 (A. Rambaut and A.J. Drummond, 2003), discarding 25% of the generations as burnin.

Results:

Our results provide strong support for the monophyly of the family Balistidae. *Sufflamen* is placed as sister to the remaining balistids, and strong support for the monophyly of both *Sufflamen* and *Rhinecanthus* is present. We also find evidence for the paraphyly of two genera: *Pseudobalistes* and *Balistoides*. Our chronogram reveals that the crown balistids appeared approximately 25 MYA. This is in disagreement with Yamanoue et al. (2005) who suggest the crown group to have appeared in the Cretaceous (~ 95 MYA).

Our results also support the hypothesis that diversification of the younger genera *Balistes* and *Xanthichthys* have been driven by two vicariant events, the closing of the Red Sea (~5 MYA) and the closing of the Isthmus of Panama (~3 MYA). The estimated ages for the splits between *Balistes capriscus* with *Balistes polylepis* and *Xanthichthys ringens* with *Xanthichthys auromarginatus,* which are respectively restricted to these regions, have crown ages responding to these vicariant events. The late appearance of any triggerfish off the western coastlines of North and South America, indicate that the Eastern Pacific Barrier (EPB), which formed approximately thirty five million years prior to the appearance of the crown group balistids (Bellwood and Wainwright, 2002), may have hindered the eastward dispersal of these reef associated fishes.

Our study reveals that All Models PP's were identical to One Model PP's. Further analysis revealed that the most visited model was almost always simpler than the most commonly used GTR model. In all cases, except the final concatenated data set, All Models chose a 2, 3, or 4 parameter model over the complex, 6 parameter, GTR model of sequence evolution as the best fit.

Discussion:

Here we present the first molecular phylogeny of the family balistidae. Our tree includes representatives of all but one of the genera, and relationships that are well supported in most areas of the tree. These relationships have important implications in comparative studies, as triggerfish are shown to have advanced cognitive abilities and some of the largest brainsizes of any marine teleost (Peter Wainwright, *personal communication*) Furthermore, our results indicate that reclassification of the species in the genera *Pseudobalistes* and *Balistoides* is warranted. The placement of *Sufflamen* as the most ancestral balistid lineage is interesting as *Sufflamen chrysopterum* is currently the only known balistid to undergo protogynous sex change (Takamoto, 2003), suggesting either independent evolution or a subsequent loss of this trait in other lineages.

All Models PP's were identical to the One Model PP's in all analyses, indicating that over-parameterization may not affect clade support. However, a GTR model was not chosen in almost any of the All Models runs, indicating that other unnamed models may provide the best choices in substitution models for this or other data sets. Additionally, the long branch lengths generated in this study may have made this a less than ideal data set with from to conduct this study. While branch lengths generated from models chosen in All Models were shorter, it appears that the inherently long branch lengths of this data set did not allow the branches to shorten to a point where cladogenic events would be effected, which may have biased our results. This is an avenue of research that merits further exploration. It is also worth noting that the All Models runs tended to finish much faster than the runs using only the GTR model of evolution, suggesting that accommodating model uncertainty may lead to computationally less expensive analyses.

The chronogram generated in this study provides an exciting avenue for novel research on the biogeography of the balistidae, as little data is available at this time. The circumglobal range of this group and the availability of a chronogram may make this an ideal study group from which to test the competing hypotheses of vicariance or dispersal in reef fishes. As suggested by Bellwood and Wainwright (2002), looking for more *soft* vicariant barriers such as the East Pacific Barrier or changes in currents, might answer questions regarding the high diversity of species in the Indian Ocean, Indo and South Pacific, as well as Hawaii (Kuiter, R.H and Debelius, H., 2006), or insights into the high rates of endemism in the Red Sea (Burgess et al, 2003). These same current changes may also reflect dispersal corridors for genera with free-floating larvae that managed multiple invasions of individual islands, such as *Rhinecanthus aculeatus* and *Rhinecanthus rectangulus* in Hawaii. Be estimating dates of cladogenetic events, we are able to gain key insights into the timing of diversification that we may be able to correlate onto historical geographic events such as the formation of the East Pacific Barrier, the isolation of the Red Sea, and the closing of the Isthmus of Panama. Furthermore, this research has potential implications for future conservation projects concerning balistid fisheries, as the chronogram generated in this study will allow future researchers to quantify temporal speciation trends, which we can then extrapolate for comparison should suspected human-induced extinctions occur.

References:

Alfaro, M.E., Santini, F., and Brock, C. 2006. Do Reefs Drive Diversification in Marine Teleosts? Evidence from the Pufferfishes and their Allies (Order Tetraodontiformes). *Submitted to the Journal of Evolution.*

Alfaro, M.E. and Huelsenbeck, J. H. 2006. Performance of reversible jump Markov chain Monte Carlo in phylogenetic model selection. Syst. Biol. 55(1):89-96.

Bellwood, D.R. and Wainwright, P.C. 2002. The History and Biogeography of Fishes on Coral Reefs. pp. 5-32. In: "Coral Reef Fishes. Dynamics and diversity in a complex ecosystem" (P.F. Sale, ed.), Academic Press, San Diego.

Buckley, T.R. and Cunningham, C.W. 2002. The effects of nucleotide substitution model assumptions on estimates of nonparametric bootstrap support. Mol. Biol. And Evol. 19(4): 394-405.

Buckley, T.R., Simon, C., and Chambers, G.K. 2001. Exploring among-site variation models in a maximum likelihood framework using empirical data: Effects of model assumptions on estimates of topology, branch lengths, and bootstrap support. Sys. Biol. 50(1): 67-86.

Burgess, W.E., Axelrod, H.R., Hunziker, R.E. 2003. Dr. Burgess's Atlas of Marine Aquarium Fishes, Third edition. T.F.H. Publications, Inc. Neptune City, New Jersey.

Drummond AJ & Rambaut A (2003) BEAST v1.0, Available from http://evolve.zoo.ox.ac.uk/beast/.

Heemstra, P.C. and Smith, M.M. 1983. A New Species of the triggerfish genus Xenobalistes (Tetraodontiformes: Balistidae) from South Africa. Rhodes University J.L.B. Smith Institute of Ichthyology Special Publication. 1983; (26): 1-5.

Holcroft, N. I. 2005. A Molecular Analysis of the Inter-Relationships of Tetraodontiform Fishes (Acanthomorpha: Tetraodontiformes). Mol. Phylogenet. Evol. 34, 525-544.

Huelsenbeck, J.P., B. Larget and M. E. Alfaro. 2004. Bayesian Phylogenetic Model Selection Using Reversible Jump Markov Chain Monte Carlo. Mol. Bio. Evol. 21(6):1123–1133.

Kawase, H. 2003. Spawning Behavior and Biparental Egg Care of the Crosshatch Triggerfish, Xanthichthys mento (Balistidae). Envirnm. Biol. Fishes. 66(3):211–219.

Kuiter, R.H and Debelius, H. 2006. World Atlas of Marine Fishes. IKAN–Unterwasseraarchiv. Frankfurt, Germany.

Leis, J.M., 1984. Tetraodontiformes: relationships. In: Moser, H.G.,Richards, W.J., Cohen, D.M., Fahay, M.P., Kendall Jr., A.W., Richardson,S.L. (Eds.), Ontogeny and Systematics of Fishes. Am. Soc. Ichthyol. and Herp., pp. 459–463 (Spec. Publ. No. 1).

Matsuura, K. 1979. Phylogeny of the Superfamily Balistoidea (Pisces: Tetraodontiformes). Mem. Fac. Fisheries. Hokkaido. Univ. 26(1–2): 46–170.

Matsuura, K. 1981. *Xenobalistes tumidipectoris* new genus new species of triggerfish (Tetraodontiformes: Balistidae) from the Marianas Island, Western Pacific Ocean. Bulletin of the National Science Museum Series-A (Zoology). 7(4): 191-200.

Posada, D., and T. R. Buckley 2004. Model selection and model averaging in phylogenetics: Advantages of akaike information criterion and Bayesian approaches over likelihood ratio tests. Syst. Biol. 53(5), 793-808.

Posada D and Crandall KA 1998. Modeltest: testing the model of DNA substitution. Bioinformatics 14 (9): 817-818.

Rambaut, A. 1996. Se-Al: Sequence Alignment Editor. Available at http://evolve.zoo.ox.ac.uk/.

Rosen, D.E., 1984. Zeiformes as primitive plectognath fishes. Am. Mus. Novit. 2782, 1–45.

Saiki, R.K. 1990. Amplification of genomic DNA. In: Innis, M.A., Gelfand, D.H., Sninsky, J.J., White, T.J. (Eds), PCR Protocols: A guide to Methods and Applications. Academic Press, New York, pp. 13–20.

Santini, F. and Tyler, J.C., 2003. A phylogeny of the families of fossil and extant tetraodontiform fishes (Acanthomorpha, Tetraodontiformes),Upper Cretaceous to Recent. Zool. J. Linn. Soc. 139, 565–617.

Santini, F. and Tyler, J.C. 2004. The importance of even highly incomplete fossil taxa in reconstructing the phylogenetic relationships of the Tetraodontiformes (Acanthomorpha: Pisces). Integ. And Comp. Biol. 44(5): 349-357.

Sullivan, J. and Swofford, D.L. 1997. Are guinea pigs rodents? The importance of adequate models in molecular phylogenetics. Jour. Mamm. Evol. 4(2): 77-86.

Takamoto, G., Seki, S., Nakashima, Y. Karino, K. Kuwamura, T. 2003.Protogynous sex change in the haremic triggerfish *Sufflamen chrysopterum* (Tetraodontiformes). Ichth. Res. 50(3): 281-283.

Tyler, J.C., 1980. Osteology, phylogeny, and higher classification of the fishes of the order Plectognathi (Tetraodontiformes). NOAA Tech. Rept. NMFS Circ. 434, 1–422.

Walsh P. S., D. A. Metzger, R. Higuchi. 1991. Chelex-100 as a medium for simple extraction of DNA for PCR based typing from forensic material. BioTechniques. 10:506–513.

Winterbottom, R., 1974. The familial phylogeny of the Tetraodontiformes (Acanthopterygii: Pisces) as evidenced by their comparative myology. Smithson. Contrib. Zool. 155, 1–201.

Yamanoue, Y., Miya, M., Inoue, J.G., Matsuura, K., Nishida, M. 2006. The mitochondrial genome of spotted green pufferfish Tetraodon nigroviridus (Teleostei: Tetraodontiformes) and divergence time estimation among model organisms in fishes. Genes. Genet. Syst. 81: 29–39.