Transcriptional and phylogenetic analysis of five complete ambystomatid salamander mitochondrial genomes

Amy K. Samuels*,1, David W. Weisrock1, Jeremiah J. Smith, Katherine J. France, John A. Walker, Srikrishna Putta, S. Randal Voss

University of Kentucky, Department of Biology, Lexington, KY 40506, USA

Received 9 August 2004; received in revised form 11 November 2004; accepted 22 December 2004

Available online 16 March 2005

Received by G. Pesole

Abstract

We report on a study that extended mitochondrial transcript information from a recent EST project to obtain complete mitochondrial genome sequence for 5 tiger salamander complex species (Ambystoma mexicanum, A. t. tigrinum, A. andersoni, A. californiense, and A. dumerilii). We describe, for the first time, aspects of mitochondrial transcription in a representative amphibian, and then use complete mitochondrial sequence data to examine salamander phylogeny at both deep and shallow levels of evolutionary divergence. The available mitochondrial ESTs for A. mexicanum (N=2481) and A. t. tigrinum (N=1205) provided 92% and 87% coverage of the mitochondrial genome, respectively. Complete mitochondrial sequences for all species were rapidly obtained by using long distance PCR and DNA sequencing. A number of genome structural characteristics (base pair length, base composition, gene number, gene boundaries, codon usage) were highly similar among all species and to other distantly related salamanders. Overall, mitochondrial transcription in Ambystoma approximated the pattern observed in other vertebrates. We inferred from the mapping of ESTs onto mtDNA that transcription occurs from both heavy and light strand promoters and continues around the entire length of the mtDNA, followed by post-transcriptional processing. However, the observation of many short transcripts corresponding to rRNA genes indicates that transcription may often terminate prematurely to bias transcription of tRNA genes; indeed an rRNA transcription termination signal sequence was observed immediately following the 16S rRNA gene. Phylogenetic analyses of salamander family relationships consistently grouped Ambystomatidae in a clade containing Cryptobranchidae and Hynobiidae, to the exclusion of Salamandridae. This robust result suggests a novel alternative hypothesis because previous studies have consistently identified Ambystomatidae and Salamandridae as closely related taxa. Phylogenetic analyses of tiger salamander complex species also produced robustly supported trees. The D-loop, used in previous molecular phylogenetic studies of the complex, was found to contain a relatively low level of variation and we identified mitochondrial regions with higher rates of molecular evolution that are more useful in resolving relationships among species. Our results show the benefit of using complete genome mitochondrial information in studies of recently and rapidly diverged taxa.

D 2004 Elsevier B.V. All rights reserved.

Keywords: Ambystoma; EST; axolotl; Bayesian; parsimony; phylogeny

1. Introduction

Expressed sequence tag (EST) projects provide an efficient means to generate genome-wide information and resources. Although EST projects are used primarily to gather structural and functional information for nuclear transcripts, a relatively large proportion of sampled cDNA clones correspond to mitochondrial transcripts. As such, EST projects simultaneously provide information about the
nuclear and mitochondrial transcriptomes (Gissi and Pesole, 2003). Such information is essential for correctly annotating gene boundaries and understanding transcriptional processes unique to the mitochondrial genome (Edqvist et al., 2000; Barth et al., 2001). In addition, because mitochondrial genomes are small, EST projects can rapidly yield scaffolds to obtain complete mitochondrial genome sequence (CMGS). Thus, ESTs projects are expected to benefit phylogenetic research in particular, because CMGSs will yield thousands of additional molecular characters of the type commonly used in evolutionary biology.

Reliable annotation of gene boundaries within genomes requires an assembly of both genomic DNA and mRNA sequences. Traditionally, mitochondrial genome annotation has been accomplished by comparing newly acquired mitochondrial sequences with previously published and annotated sequences obtained from genomic mtDNA. Unfortunately, genomic mtDNA does not reveal post-transcriptional modifications that determine the boundaries of protein-coding regions. For example, several mitochondrial protein-coding genes do not contain a full stop codon, but instead use polyadenylation of the terminating T nucleotide to complete the stop codon sequence (Gissi and Pesole, 2003). Identification of the correct terminating T nucleotide may not always be straightforward when comparing mtDNA sequences, but it is when mRNA sequences are available in the form of ESTs (Gissi and Pesole, 2003). In C. intestinalis the Cytb reading-frame could potentially end at a TAG stop codon located within the adjacent tRNA$^{PRO}$ gene. However, the true stop codon revealed by transcript mapping of ESTs is a TA nine bp upstream that is polyadenylated to form a complete stop codon. EST projects are valuable in this regard because they provide transcriptional information necessary for genome annotation.

In some cases, complete mitochondrial sequences may provide greater phylogenetic resolution and branch support for deep evolutionary relationships that are difficult to resolve with smaller molecular datasets (Zardoya and Meyer, 2001). Presumably, in these cases, CMGSs provide a greater number of informative characters that collectively dampen misleading signals from fast evolving nucleotide sites. The use of phylogenetic methods that can partition complete mitochondrial genomes into gene-specific subsets (that can each be fit to a different evolutionary model) may also increase the accuracy and support of deep phylogenetic relationships (Nylander et al., 2004).

The large number of characters provided by CMGS may also be useful in inter and intraspecific studies. In some cases, the use of a single gene may not provide sufficient variation to address questions of phylogeny or phylogeography, whereas CMGS from species or populations can provide ample variation (Nardi et al., 2003). This may apply especially in systems where divergence among lineages has occurred over a short evolutionary time scale, producing few substitutions marking internal branches. A comparative mitogenomic analysis among several closely related species would be informative to determine if CMGS data provides sufficient variation to resolve phylogeny in situations where individual genes or gene regions provide poor resolution. This comparative approach may also permit the identification of genes or gene regions that are more informative for low-level phylogenetic studies.

The tiger salamander complex (Ambystoma tigrinum) includes a diverse array of lineages distributed throughout North America and Mexico that exhibit considerable life history and morphological variation. Previous phylogenetic analysis of mtDNA D-loop sequence identified a number of clades containing geographically proximal sets of populations and species (Shaffer and McKnight, 1996). However, many phylogenetic relationships were poorly resolved because relatively few molecular characters were found to differentiate lineages of this recently derived complex (~5 million years old). To determine if CMGSs can better resolve relationships within the complex, we assembled mitochondrial transcripts from a recent EST project involving Ambystoma mexicanum and A. t. tigrinum (Putta et al., 2004), and used the resulting scaffold to obtain complete mitochondrial sequences for these and three other tiger salamander complex species (Ambystoma andersoni, Ambystoma californiae, and Ambystoma dumerilii). We used these data to explore the use of CMGS at a much deeper evolutionary level in salamanders. To date, CMGS have been generated for single species within three salamander families: Cryptobranchidae (Andrias davidianus), Hynobiidae (Ranodon sibiricus), and Salamandridae (Mertensiella luschani). The availability of CMGS data for these families and the Ambystomatidae allowed us to generate new topologies and compare these to published topologies constructed using fewer mitochondrial and nuclear characters.

2. Materials and methods

2.1. Mitochondrial EST assembly and PCR primer design

ESTs were obtained from several A. mexicanum and A. t. tigrinum cDNA libraries (Putta et al., 2004). Out of 39,982 high quality transcripts, 3686 (9.3%) were found to correspond to the salamandr M. luschani mitochondrial genome sequence (GenBank accession no. NC002756) using BLAST searches against the NCBI database. ESTs were assembled using the DNASTAR Lasergene program to form contigs that were subsequently mapped onto the CMGS of M. luschani using BLASTN. Using sequence information from both Ambystoma contigs and M. luschani, PCR primers were designed using the web-based program Primer3 (http://www.frodo.wi.mit.edu/). Approximately 75 primer pairs...
were used for PCR amplification of the complete mitochondrial sequence of *A. mexicanum* and *A. t. tigrinum*. Additional primers were designed from these two sequences and used in conjunction with existing primers to amplify and sequence *A. andersoni*, *A. dumerilii*, and *A. californiense*. Primer sequences are available in supplementary document 1 (http://salamander.uky.edu/research).

2.2. DNA isolation, PCR, cloning and sequencing

Genomic DNA was isolated from liver tissue using a SDS-Proteinase K/Phenol-Chloroform procedure. Multiple individuals were used for both *A. t. tigrinum* and *A. mexicanum*; *A. mexicanum* were obtained from a laboratory stock and *A. t. tigrinum* from Charles Sullivan Corp. *A. andersoni* tissue also came from several individuals, while *A. dumerilii* and *A. californiense* DNA each came from a single individual; all of these individuals were obtained from natural populations or laboratory descendants of natural populations. PCR reactions were performed in a Biometra T-gradient thermocycler with the following conditions: 94 °C for 3 min; 32 cycles of 94 °C for 45 s, 58.5 °C for 1 min, 72 °C for 2 min; 72 °C for 10 min. Each reaction included between 90 ng and 300 ng total DNA, 2.5 μl dNTPs, 2.5 μl Takara ExTaq PCR Buffer, 0.25 μl Takara ExTaq, and 25 ng of each primer for a total reaction volume of 25 μl. Products larger than 1 Kb were cloned into the pGEM-T vector (Promega) and sequenced (Davis Sequencing, Davis, CA). Large PCR products (2–14 kb) were cut with restriction enzymes (*Alu* and *Rsa*), cloned and sequenced. Sequences were assembled using Lasergene and manually edited.

Prior to the submission of our manuscript, the complete mitochondrial genome of the axolotl, *A. mexicanum* was completed and published (Arnason et al., 2004; GenBank accession no. AJ584639). We compared this sequence to our independently completed *A. mexicanum* mitochondrial genome sequence using BLASTN and found 22 single nucleotide differences and an overall nucleotide identity of 99%.

2.3. Transcript mapping

Vertebrate mitochondrial gene transcription typically occurs from either one or two promoters on each strand to create polycistronic heavy and light-strand transcripts (Boore, 1999). The transcripts are cleaved and polyadenylated to produce mature mrnas, trnas, and rnas. In some cases, genes formed mature bicistronic transcripts in which polyadenylation is present only on the 3’ gene (Gissi and Pesole, 2003). Immature polycistronic mrnas as well as mature (polyadenylated) mrnas were present in the salamander ESTs. Mitochondrial ESTs were assembled separately for both *A. mexicanum* and *A. t. tigrinum* using DNASTAR Lasergene. The resulting contigs for each species were mapped to the full-length mtDNA sequences using BLAST. In cases of bicistronic mrnas, ORFs were examined to identify stop codons and gene boundaries, while polyadenylation identified gene boundaries in mature single-gene mrnas. Polyadenylation within contigs containing transcripts for more than one gene, as well as transcripts spanning annotated gene boundaries, suggested the presence of immature mrnas in the libraries.

Some nucleotide variation was identified between mitochondrial transcripts and genomic mtDNA, presumably because transcripts and mitochondrial DNAs were sampled from different individuals. In these cases, we reported the nucleotide from the mitochondrial genomic sequence. In cases where variable nucleotides were found within a single species’ set of PCR amplified mitochondrial DNA fragments and were likely the result of sequencing error, we made majority rule consensus base calls.

2.4. Sequence alignments

Salamander family phylogeny was investigated by aligning the five complete *Ambystoma* mitochondrial genomes with CMGS from representatives of the families Cryptobranchidae (*A. davidianus*; GenBank accession no. NC004926), Hynobiidae (*R. sibiricus*; NC004021), and Salamandridae (*M. luschani*). A single frog representative (*Bufo melanostictus*; NC005794) and caecilian representative (*Typhlonectes natans*; NC002471) were used as out-groups. Sequence alignment was performed using ClustalX (Jeanmougin et al., 1998) with default parameter settings. Manual adjustments of protein-coding nucleotide sequence alignments were facilitated through translation to amino-acid sequence. A number of protein-coding genes contained regions that could not be unambiguously aligned due to length variation and were excluded from all analyses. Manual adjustment of tRNA nucleotide sequence was facilitated using secondary-structural models to align stem regions (Kumazawa and Nishida, 1993). Any unpaired loop region containing length variation was excluded all analyses. Alignment of tRNA sequence followed a conservative approach whereby any region containing length variation was excluded. Also excluded from phylogenetic analysis were the mitochondrial control region and a large and variable intergenic spacer between the tRNA<sup>Thr</sup> and tRNA<sup>Pro</sup> genes found in all salamander sequences. The *B. melanostictus* genome does not contain identifiable tRNA<sup>Leu</sup>(CUN), tRNA<sup>Thr</sup>, and tRNA<sup>Pro</sup> genes, and contains an incomplete and unalignable tRNA<sup>His</sup> gene. These genes were included in phylogenetic analysis, but scored as missing data in *Bufo*.

A more exclusive alignment containing only *Ambystoma* genomes was also constructed following the same criteria as described above. This alignment included the mitochondrial control region and the tRNA<sup>Thr</sup>, tRNA<sup>Pro</sup> spacer and other, fewer exclusions were necessary.
2.5. Phylogenetic analysis

We employed partitioned Bayesian phylogenetic analysis to account for different underlying evolutionary models and parameter estimates for different gene regions (Nylander et al., 2004). Individual protein-coding genes and rRNA genes were treated as separate data partitions. All tRNAs were included into a single data partition. For the Ambystoma-only analyses, the control region and the tRNA^{THR–tRNAPRO} intergenic spacer were each treated as separate data partitions. Best-fit evolutionary models for each partition were chosen using the Akaike Information Criterion as implemented in the program Modeltest version 3.06 (Posada and Crandall, 1998). This was done separately for the full amphibian alignment and for the Ambystoma-only alignment. Bayesian analysis was performed using the program MrBayes version 3.04 (Huelsenbeck and Ronquist, 2001). Four Markov chains were used with one million generations and a tree sampling taken every 1000 generations. Flat priors were used for parameter estimates and random trees were used to start each chain. Model parameter estimates were unlinked to allow for independent estimation of parameters for each partition. All trees sampled prior to the stationary distribution were discarded as burn-in. Two additional independent runs were performed for each data set using identical conditions to verify the stationary posterior distribution.

Parsimony analysis was performed using PAUP* v4.0 (Swofford, 2002). An exhaustive tree searching option was used with equal weighting of all characters and TBR branch swapping. To assess support for branches in parsimony trees, bootstrap analysis was applied using 1000 bootstrap replicates with branch and bound tree searching. Decay indices were calculated for each internal branch using PAUP.

Alternative phylogenetic topologies were tested using the Shimodaira and Hasegawa (SH) test using 1000 RELL replicates (Shimodaira and Hasegawa, 1999) as implemented in PAUP* v4.0. To perform these tests, a maximum-likelihood tree was found in an unconstrained analysis treating the entire data set as a single partition and using the best-fit model of evolution determined using the methods described above and estimating all parameters. The unconstrained ML tree was compared to a constrained ML tree favoring a particular topological constraint.

3. Results and discussion

3.1. Mitochondrial DNA content, gene organization, and molecular evolution

The complete mitochondrial genomes of five ambystomatid salamanders were sequenced with the following lengths: 16,370 (A. mexicanum, A. andersoni, and A. dumerilii), 16,374 (A. californiense), and 16,375 bp (A. t. tigrinum). These five mitochondrial genomes have been submitted to GenBank under the accession numbers AY659991, AY659992, AY659993, AY659994, and AY659995. The base compositions for the light strand vary by species (Table 1); however, all Ambystoma mtDNA genomes are AT rich, similar to patterns found in most vertebrates. The gene order of Ambystoma mtDNA genomes is identical to other salamander mitochondrial genomes and to most other vertebrates (Fig. 1; Table 2). All genomes contained 13 protein-coding genes, 22 tRNAs and two rRNAs. Eight tRNAs and one mRNA are encoded on the light strand, and 14 tRNAs, 12 mRNAs, and two rRNAs are encoded on the heavy strand. Additional features present in each mtDNA sequence were identifiable D-loop (or control region), an origin of light strand replication (OL) located in the WANCY tRNA region, and an intergenic spacer region located between the tRNA^{THR} and tRNA^{PRO} genes. This spacer region has been previously identified within the genus Ambystoma (Shaffer and McKnight, 1996; Arnason et al., 2004), as well as in the CMGSs of M. luschani, A. davidianus, and R. sibiricus, but its function remains unknown.

Nucleotide substitution patterns for most Ambystoma mitochondrial genes (partitions) are best fit to a General Time-reversible (GTR) model with six different nucleotide substitution rates and a proportion of sites estimated to be invariant (I) (Supplementary Document 2, http://salamander.uky.edu/research). One gene, atp6, is best fit to the less complex Hasegawa–Kishino–Yano (HKY) model. Most genes exhibit no heterogeneity in substitution rates across sites (G); however, two genes, atp6 and cox1, are best fit to a model that accounts for rate heterogeneity. For the atp6 gene, the 95% credible set of trees from the Bayesian analysis contained a wide range of estimates for the gamma shape parameter (a), up to a value of 18.61, indicating that we cannot rule out a model in which most sites within this gene evolve at a similar rate, with only a small number of sites containing relatively high or low substitution rates (Yang, 1996). In contrast, the cox1 gene contains considerable rate heterogeneity across sites with a 95% credibility interval of a much less than one (0.05–0.34). Average Bayesian estimates of substitution rates are generally much higher for transitions than for transversions across gene partitions. However, the Bayesian posterior distribution displayed a high variance in nucleotide substitution rate estimates and in many partitions the

<table>
<thead>
<tr>
<th>Species</th>
<th>%A</th>
<th>%C</th>
<th>%G</th>
<th>%T</th>
<th>#sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. mexicanum</td>
<td>34.8</td>
<td>20.0</td>
<td>13.2</td>
<td>31.9</td>
<td>16,370</td>
</tr>
<tr>
<td>A. t. tigrinum</td>
<td>34.7</td>
<td>19.7</td>
<td>13.3</td>
<td>32.2</td>
<td>16,375</td>
</tr>
<tr>
<td>A. andersoni</td>
<td>34.7</td>
<td>20.0</td>
<td>13.2</td>
<td>31.9</td>
<td>16,370</td>
</tr>
<tr>
<td>A. dumerilii</td>
<td>34.7</td>
<td>19.9</td>
<td>13.3</td>
<td>31.9</td>
<td>16,370</td>
</tr>
<tr>
<td>A. californiense</td>
<td>34.2</td>
<td>20.3</td>
<td>13.7</td>
<td>31.7</td>
<td>16,374</td>
</tr>
</tbody>
</table>
95% credibility intervals for transition and transversion rates were overlapping. A 100 bp sliding-window plot of the frequency of polymorphism across the *Ambystoma* alignment reveals a relatively low level of nucleotide variation in the D-loop region (Fig. 2). Variation within the D-loop is comparable to levels of variation within the 12S and 16S rRNA genes. Overall, variation is lowest in regions coding for tRNAs and highest in protein-coding regions. The tRNA<sup>THR</sup>–tRNA<sup>PRO</sup> spacer region also contains regions of relatively high variation. The relatively low variation detected in the D-loop region is surprising because it is typically considered the most variable mitochondrial region and is widely used as a polymorphic marker in phylogeographic studies of shallowly diverged populations and species (Avise, 2004). Based on this observation, we suggest that other regions of the *Ambystoma* mitochondrial genome may provide a stronger phylogenetic signal than the D-loop region at shallow levels of divergence (see below).

### 3.2. Transcript mapping

The relatively large number of mitochondrial transcripts generated from the *Ambystoma* EST project allowed us to accurately annotate mitochondrial gene boundaries and reconstruct transcriptional patterns and processes. There are several reasons why so many transcripts were sampled: (1) most eukaryotic cells contain hundreds to thousands of mitochondria that can make up 25% of the cytoplasmic volume (Lodish et al., 2000), (2) salamanders have some of the largest cell volumes of all vertebrates (Sessions and Larson, 1987), and (3) each mitochondrion can contain from 1 to 10 copies of mtDNA (Larsson and Clayton, 1995). Few studies have taken advantage of transcript mapping to understand mitochondrial transcription (Edqvist et al., 2000; Barth et al., 2001; Gissi and Pesole, 2003), and no studies have examined transcriptional patterns in salamander. Salamander mitochondrial gene transcription, as in other vertebrates likely
occurs from both the heavy and light strand promoters and continues around the entire length of the mtDNA, followed by post-transcriptional processing. Evidence to support this transcriptional mechanism in salamanders includes the presence of bicistronic and polycistronic transcripts and the presence of intervening tRNA sequences between protein-coding and ribosomal RNA genes, indicating that multiple contiguous genes are transcribed in an uninterrupted fashion.

A total of 1205 ESTs from *A. t. tigrinum* were assembled into 12 contigs covering 14,231 bp or 86.9% of the complete eastern tiger salamander mitochondrial genome. These contigs mapped to 22 genes overall, including 10 complete genes (16S rRNA, *nd1*, *cox1*, *cox2*, *atp8*, *atp6*, *cox3*, *nd3*, *nd4L*, *cytb*), six partial genes (12S rRNA, *nd2*, *nd4*, *nd5*, *nd6*), and seven tRNAs (-Leu(UUR), -Trp, -Ala, -Asn, -Cys, -Tyr, -Ser(CUN)) (Fig. 1). Four eastern tiger salamander ESTs from the light strand were also mapped to the non-coding D-loop region and intervening spacer region. This result suggests that both regions are post-transcriptionally processed in salamander because cDNA libraries were constructed by sampling the polyadenylated fraction of the cellular RNA pool, and no polyadenylated regions were observed flanking the intergenic spacer and D-loop regions.

The 2481 mitochondrial ESTs from *A. mexicanum* were assembled into nine contigs covering 15,215 bp or 92.9% of the complete Mexican axolotl mitochondrial genome. These contigs mapped to 23 genes overall, including 11 complete genes (12S rRNA, 16S rRNA, *nd1*, *nd2*, *cox2*, *atp8*, *atp6*, *cox3*, *nd3*, *nd4L*, *cytb*), six partial genes (12S rRNA, *nd2*, *nd4*, *nd5*, *nd6*), and seven tRNAs (-Leu(UUR), -Trp, -Ala, -Asn, -Cys, -Tyr, -Ser(CUN)) (Fig. 1). Four eastern tiger salamander ESTs from the light strand also mapped to the non-coding D-loop region and intervening spacer region. This result suggests that both regions are post-transcriptionally processed in salamander because cDNA libraries were constructed by sampling the polyadenylated fraction of the cellular RNA pool, and no polyadenylated regions were observed flanking the intergenic spacer and D-loop regions.
atp6, nd4L, nd4, nd5, nd6, cyt b), four partial genes (cox1, atp8, cox3, nd3), and eight tRNAs (-Val, -Leu(UUR), -Ser(UCN), -Asp, -Arg, -Leu(UCN), partial -Ser(AGY) and -Glu) (Fig. 1). As in A. t. tigrinum, three axolotl light strand ESTs were mapped to the non-coding D-loop region.

Assembled EST contigs from A. mexicanum and A. t. tigrinum confirmed previously published and annotated salamander mitochondrial gene boundaries where possible. The first few bases of mitochondrial genes could not always be confirmed by ESTs because either poor quality sequence ends were trimmed before assembly, or the 3’ sequencing reads stopped short of the 5’ gene start. In these cases, ORFs were identified from both mitochondrial genomic and EST sequence that matched annotated boundaries (A. mexicanum: nd2, cox1, atp8, cox3, nd3, cyt b; A. t. tigrinum: nd2, nd3, nd4L, nd5). All protein-coding genes begin with the start codon ATG, with the exception of cox1, which uses GTG. The stop codons used in Ambystoma mtDNA genes were TAA, TAG, and AGA. Notably, the AGG stop codon was not observed in any Ambystoma mitochondrial genes, but is present in the mtDNAs of other salamanders (e.g. M. luschani). Full stop codons were present in the genes nd1, nd2, cox1, atp8, atp6, nd3, nd4L, nd5, and nd6. Polyadenylation of a T nucleotide completes the stop codon for the genes cox2, cox3, nd4, and cyt b. The genes polyadenylated in both species were 12S rRNA, 16S rRNA, nd1, nd2, nd3, nd4, atp4, cox2, cox3, and cyt b. We observed polyadenylation of nd4L in A. t. tigrinum but not A. mexicanum, most likely because there were relatively few transcripts representing this gene in the latter species. Though polyadenylation was only observed in these genes, the presence of D-loop and spacer region transcripts in the A. mexicanum library suggests that all mitochondrial transcripts are polyadenylated in salamander.

A large percentage of mitochondrial ESTs mapped to the 12S and 16S rRNA genes in both A. mexicanum (64.3%) and A. t. tigrinum (72.7%) (Fig. 1). Mitochondrial ESTs assembled into a contig that included the 12S–tRNA<sup>VAL</sup>–16S genes. As in human, rodents, and ascidians (Gissi and Pesole, 2003), polyadenylation was inferred in both rRNA genes in Ambystoma. An unusual pattern was observed for the 16S rRNA gene in both A. mexicanum and A. t. tigrinum. The majority of 16S rRNA ESTs terminated with a polyA stretch approximately 825 nucleotides upstream of the presumptive terminus. Approximately 18 out of 25 nucleotides in this region are adenine, suggesting that mispriming of oligo dT occurred here during cDNA library construction. The smaller transcript product was presumably synthesized more efficiently, resulting in a higher relative abundance in cDNA libraries.

In vertebrates, mitochondrial transcription begins with the rRNA genes and sometimes ends immediately following these genes by the conserved rRNA transcription termination signal (TGCGAGAn<sub>3</sub>G). Early termination of transcription allows a greater number of rRNA products to be produced relative to the remainder of the mitochondrial genome. This signal was located in both salamander species’ tRNA<sup>LEU</sup> gene immediately following the 16S rRNA gene. This suggests that, as in other vertebrates, rRNA genes are more heavily transcribed in the salamander mitochondrial genome than the rest of the mtDNA molecule, and correlates with the high proportion of mitochondrial rRNA ESTs detected in our libraries.

Nuclear integrations of mtDNA transcripts (Numts) have been identified for many species and may present a problem when trying to exclusively PCR amplify genuine mtDNA genes. We are confident that Numts are not represented in our mtDNA sequences because of the methods used. Although mtDNA integrations into the nuclear genome are not uncommon (Bensasson et al., 2001), these genes are unlikely to be transcribed (Woischnik and Moraes, 2002). Our PCR generated sequence from A. t. tigrinum and A. mexicanum matches the sequence derived from mitochon-

![Fig. 2. Plot from a sliding-window analysis of nucleotide polymorphism across an alignment of Ambystoma mitochondrial genomes. The sliding window calculated the number of variable sites in a block of 100 contiguous bases. Ambiguously aligned bases were excluded from the analysis. A generalized map of the mitochondrial genome is provided below the plot for reference.](Image)
drial transcripts sampled from the cell's RNA pool, indicating we have amplified true mtDNA. The use of long PCR in amplifying mtDNA sequences also lowers the chance of amplifying Numts (Bensasson et al., 2001). Finally, our sequences contain a number of mtDNA characteristics, including high AT and low G nucleotide base frequencies, fully functional tRNAs, and no premature stop codons.

3.3. Salamander family phylogeny

Alignment of the amphibian CMGSs resulted in 12,474 included characters, of which 6235 are variable and 3941 are parsimony informative. Using the amphibian alignment, the HKY+G model of evolution was determined to be the best-fit model for the \( \text{atp6} \), \( \text{atp8} \), \( \text{cox1} \), \( \text{cox2} \), \( \text{cox3} \), \( \text{cytb} \), \( \text{nd2} \), and \( \text{nd3} \) partitions. The more parameter rich GTR+I+G model was determined to be the best-fit model for the tRNA and tRNA partitions and for the \( \text{nd1} \), \( \text{nd4} \), \( \text{nd4L} \), \( \text{nd5} \), and \( \text{nd6} \) partitions. Parsimony and Bayesian analyses yielded identical tree topologies (Fig. 3A). Parsimony analysis yielded a single tree of 12,656 steps in length. Bayesian analysis reached a stationary distribution by approximately 25,000 generations. The posterior distribution yielded an average lnL of \(-65,458.1\) with a variance of 74.24. Bayesian analysis found no uncertainty regarding phylogenetic relationships among salamander families and all interfamilial branches had posterior probabilities of 1.0. Parsimony analysis supported all interfamilial relationships with bootstrap values of 100% and decay indices of \( \geq 63 \). These analyses placed the families Cryptobranchidae and Hynobiidae as sister lineages (Fig. 3A). The Ambystomatidae is placed as the sister lineage to the Cryptobranchi-
The use of a model-based approaches and the gene partitioning of our data should act to minimize the effects of saturation in highly variable nucleotide positions or gene regions and maximize phylogenetic signal. Nonetheless, if our models do not adequately account for the underlying mode of molecular evolution, then our likelihood-based analyses may also be prone to providing an inaccurate topology (Buckley, 2002). Because a number of gene partitions in our data set were analyzed under an HKY model of evolution (as determined through the Akaike Information Criterion), we investigated the effects of using the more parameter-rich GTR+I+G model for all partitions in a Bayesian analysis (with independent parameter estimates for each partition). The result was an identical topology and branch support measures as seen in Fig. 3A (results not shown), suggesting that our results are not a product of model underparameterization.

Taxon sampling is also an important consideration when contrasting our results with previous studies because we included only four of the 10 extant salamander families. Limited taxon sampling can have a substantial effect on phylogenetic error (Zwickl and Hillis, 2002). This may be especially so in data sets with large numbers of characters, such as from whole organellar genomes, from only a small number of representatives of the actual lineage diversity of a group (Soltis et al., 2004; Stefanovic et al., 2004). The old age and large genetic divergences among the four family lineages included in our analyses create a phylogeny that is exceptionally difficult to resolve, in which terminal branches are extremely long relative to internal branches. This scenario increases the probability of long-branch attraction when using highly homoplastic data. Inclusion of additional families or major lineages within families may act to subdivide long branches and diffuse homoplasy, thus minimizing the potential effects of long-branch attraction (Hillis, 1996). An example of the potential effect of taxon sampling in complete mtDNA genome studies is illustrated in birds where phylogenetic analysis of data for six avian orders placed Passeriformes in a basal position in bird phylogeny (Härlid and Arnason, 1999; Mindell et al., 1999), challenging their traditional and nuclear-based position as a recently derived lineage. Subsequent analyses using expanded sampling of CMGSs yielded trees consistent with traditional hypotheses (Braun and Kimball, 2002). Therefore, taxon sampling may be an especially important issue in phylogenetic studies of CMGS because limited resources typically lead to the generation of only one or a few genomes at a time; consequently, analyses of deep evolutionary relationships may be prone to producing inaccurate results as a byproduct of limited sampling (Soltis et al., 2004; Stefanovic et al., 2004). Recent studies of CMGS using much of the same data as used in our analyses along with diverse sampling of the family Plethodontidae yielded results consistent with previous studies of salamander family phylogeny (Mueller et al.,
further suggesting that taxon sampling is an important factor affecting our results.

If substantial homoplasy and taxon sampling are affecting our results, they have produced strongly supported and concordant trees using different optimality criteria. Our results offer a novel phylogenetic hypothesis among salamander families. However, this hypothesis awaits further testing through the detailed analysis of additional salamander CMGSs and indicates that salamander family phylogeny may provide a good system to study the effects of taxon sampling in the recovery of old and divergent branching events.

3.4. Tiger salamander phylogeny

Alignment of the five Ambystoma genomes resulted in 16,309 included characters, of which 1592 are variable and 501 are parsimony informative. Parsimony and Bayesian analysis yield identical tree topologies (Fig. 3B). Parsimony yields a single tree of 1,824 steps. Bayesian analysis reached a posterior distribution by approximately 25,000 generations and the posterior distribution yielded an average lnL of $-30,785.88$ with a variance of 74.72. Bayesian analysis places posterior probabilities of 1.0 on every branch. Parsimony analysis supports every branch with a bootstrap value of 100% and decay indices of $\geq 127$. A. andersoni and A. mexicanum are placed as sister taxa with A. dumerilii placed as the sister taxon to this clade. A. californiense and A. t. tigrinum are placed outside of the above-described clade.

Our results provide robust resolution among representatives from four of the eight primary lineages of the tiger salamander complex identified by Shaffer and McKnight (1996). A. andersoni and A. mexicanum both represent a clade from the Sierra Madre Oriental and central Mexican plateau. Both parsimony and Bayesian analysis strongly support these as sister taxa (Fig. 3B). Our CMGS of A. dumerilii represents a central and western Mexican plateau clade and is placed as the sister taxon to the A. andersoni–A. mexicanum clade with strong support (Fig. 3B). Finally, our CMGSs from A. californiense and A. t. tigrinum represent primary geographic lineages from the western and eastern United States, respectively, and are robustly placed outside of the clade of Mexican taxa (Fig. 3B). Our Fig. 3B is rooted with A. californiense because the results of Shaffer and McKnight (1996) and our phylogenetic analyses of all available salamander CMGSs (Fig. 3A) both support a basal split between A. californiense and all remaining tiger salamander complex species.

The lack of strong branch support for branches leading to primary lineages in Shaffer and McKnight (1996) may be a product of the limited number of variable sites present in the D-loop region (Fig. 2). Their analyses used approximately 840 bases of mitochondrial sequence spanning from the 3’ end of the tRNA\(^{THR}\) gene through 500 bases of the 5’ portion of the D-loop region. Across our five tiger salamander mitochondrial genomes, this particular region contains 69 variable sites, of which 17 are parsimony informative. Parsimony analysis of this region produces a topology concordant with the results of Shaffer and McKnight (1996) with poor bootstrap support for the clade of Mexican species (64%, results not shown). In contrast, many individual protein-coding regions contain a considerably higher amount of variation (Fig. 2) and provide greater phylogenetic support. For example, the genes nd1 and nd2 are only slightly larger in size at 971 bases and 1041 bases, respectively. However, they contain over twice as many variable sites as the tRNA\(^{THR}\)-D-loop region used by Shaffer and McKnight (1996) (nd1 variable bases=140, parsimony informative=44; nd2 variable bases=144, parsimony informative=35). Both of these genes individually provide robust support (parsimony bootstrap=100%) for relationships among tiger salamander complex species.

Our work demonstrates that CMGS information gathered from multiple closely related taxa can benefit population genetic and phylogeographic studies in at least two ways. First, it can provide a large number of characters for phylogenetic analysis of lineage representatives. This should be especially useful in studies involving species or populations that have recently diverged over a short period of time. However, the generation of complete mitochondrial genome sequences in studies involving hundreds of individuals is not practical. Therefore, complete mitochondrial genome sequence information from representative individuals can also be beneficial in the evaluation of molecular evolutionary patterns across the entire mitochondrial genome. Regions harboring the highest levels of variation can be identified and used in more extensive studies involving more detailed sampling.

Acknowledgements

This work was funded by the support of the National Science Foundation (IBN-0242833; IBN-0080112) and the National Institutes of Health (5 R24 RR16344-03) through grants to SRV. We thank D. Wake and an anonymous reviewer for improvements on an earlier draft of this paper.

References


