Perils of Paralogy: Using HSP70 Genes for Inferring Organismal Phylogenies

ANDREW P. MARTIN AND THERESA M. BURG

Department of EPO Biology, University of Colorado, Boulder, Colorado 80309, USA; E-mail: am@stripe.colorado.edu (A.P.M.)

Abstract.—Conserved genes have found their way into the mainstream of molecular systematics. Many of these genes are members of multigene families. A difficulty with using single genes of multigene families for phylogenetic inference is that genes from one species may be paralogous to those from another taxon. We focus attention on this problem using heat shock 70 (HSP70) genes. Using polymerase chain reaction techniques with genomic DNA, we isolated and sequenced 123 distinct sequences from 12 species of sharks. Phylogenetic analysis indicated that the sequences cluster with constitutively expressed cytoplasmic heat shock–like genes. Three highly divergent gene clades were sampled. A number of similar sequences were sampled from each species within each distinct gene clade. Comparison of published species trees with an HSP70 gene tree inferred using Bayesian phylogenetic analysis revealed several cases of gene duplication and differential sorting of gene lineages within this group of sharks. Gene tree parsimony based on the objective criteria of duplication and loss showed that previously published hypotheses of species relationships and two novel hypothesis based on Bayesian phylogenetics were concordant with the history of HSP70 gene duplication and loss. By contrast, two published hypotheses based on morphological data were not significantly different from the null hypothesis of a random association between species relatedness and the HSP70 gene tree. These results suggest that gene tree parsimony using data from multigene families can be used for inferring species relationships or testing published alternative hypotheses. More importantly, the results suggest that systematic studies relying on phylogenetic inferences from HSP70 genes may be plagued by unrecognized paralogy of sampled genes. Our results underscore the distinction between gene and species trees and highlight an underappreciated source of discordance between gene trees and organismal phylogeny, i.e., unrecognized paralogy of sampled genes. [Gene tree parsimony; HSP70; molecular systematics; multigene families; orthology; paralogy; sharks.]

Phylogenetic trees inferred from different genes are often dissimilar (Hasegawa et al., 1992; Ruvolo, 1997; Brown et al., 2001). Lack of concordance may be due to confounding effects of homoplasy (Adachi and Hasegawa, 1995; Naylor and Brown, 1998) or different gene phylogenies due to differential sorting or horizontal gene transfer (Wu, 1991; Hudson, 1992; Page, 1994; Maddison, 1996, 1997). Another source of discordance between gene trees and species trees is paralogy of sampled genes (Goodman et al., 1979; Page, 1994; Maddison, 1997; Page and Charleston, 1997; Slowinski and Page, 1999).

It is generally true that most genes in the nuclear genome are members of multigene families (Henikoff et al., 1997; Slowinski and Page, 1997: Fig. 2). Gene families typically evolve by a process of birth and death of gene lineages (Ota and Nei, 1994; Walsh, 1995; Nei et al., 1997). Genes are born by gene duplication from unequal crossing over, replicative transposition, or polyploidization and are lost by deletion mutations or are gradually silenced by the accumulation of deleterious mutations (Li and Graur, 1991; Rowen et al., 1996; Sidow, 1996; Wolfe and Shields, 1997). Gene family complexity (the number of different genes) is the result of the difference between the birth and death rates of individual genes. Birth rates of gene duplication in populations can be high, ranging from $10^{-8}$ to $10^{-3}$ duplications per locus per generation (Gelbart and Chovnick, 1979; Shapira and Finnerty, 1986; Fryxell, 1996; Lynch and Conery, 2000).

Such high rates of gene duplication, in contrast with the relative conservation of gene functions and of the primary structure of most genes, suggest that turnover of individual genes within a family of related genes can be rapid. One consequence of this process is that assumptions of orthology for isolated genes may be questionable, and phylogenetic inference from nuclear genes may be subject to errors from unrecognized paralogy. For example, gene duplication may occur such that ancestral species possess two paralogous genes copies, but the paralogues may undergo differential loss (or decay) in descendant species, yielding a gene tree that is incongruent with the species tree (Fig. 1). Failure to sample orthologous genes can also stem from methodological bias (Fig. 1) either
because all paralogous genes were not characterized or because the particular method of isolating genes was biased for some members of the gene family (e.g., see Wagner et al., 1994). Similarly, failure to sample orthologous genes may result in inaccurate estimates of divergence time between species because the genes record the timing of gene duplication and not necessarily species divergence.

These considerations underscore the possibility that the accuracy of phylogenetic inference may be compromised by errors in assigning orthology. In some cases it is not possible to ascertain orthology with confidence, and researchers have attempted to extract information about relationships among taxa from only those genes of a multigene family tree assumed to be orthologous (Livak et al., 1995). Ruvolo and Koh (1996) noted, however, that a gene tree should be evaluated in its entirety as an estimator of the species tree, implying that focusing on putative orthologous genes while ignoring other genes is not legitimate. Nonindependence of genes within a gene family may also interfere with estimation of the evolutionary history of species; for instance, tandemly repeated genes can undergo gene conversion. Sanderson and Doyle (1992) showed that gene families with intermediate rates of gene conversion have the highest levels of homoplasy and the largest numbers of equally parsimonious trees, with the lowest bootstrap values.

Recognition of these problems associated with nuclear genes has led a number of researchers to search for so-called single-copy genes. bona fide examples of single-copy genes known to be orthologous in different vertebrate taxa are rare. Even genes encoded by the mitochondrial genome often exist as members of multigene families because of historical transfers of genes from the mitochondrial genome to the nucleus (Bensasson et al., 2001). Thus, the possibility exists that all genes are currently or have been members of multigene families, and even if genes exist as single copies in the genomes of taxa under investigation, the sampled genes may be paralogous.

Problems associated with unrecognized paralogy can be avoided by simply not making the assumption that sampled genes are orthologous. Instead, the choice among rival phylogenetic hypotheses can be based on the fit of the gene tree to alternative species trees (Goodman et al., 1979; Page, 1994; Page and Charleston, 1997; Slowinski et al., 1997). Page (1994) referred to this procedure as reconciliation, and Slowinski and Page (1999) noted that this process is analogous to optimizing individual characters on trees. If there is perfect agreement between the gene tree and species tree (compare the gene tree and hypothesis 1 in Fig. 2), then the gene tree perfectly reconciles with the species tree and it is possible to extract the history of gene duplication and speciation directly from the gene tree. If the match between gene tree and organismal tree is not perfect (compare the gene tree and hypothesis 2 in Fig. 2), then the reconciled tree will be different from the gene tree mainly because it is necessary to postulate additional gene duplications and losses. Using this approach, it is possible to choose between the two alternative hypotheses of phylogenetic relationships among species depicted in Figure 2 using parsimony as an objective criterion. Slowinski and Page (1999) referred to this approach as gene tree parsimony. In this particular case, hypothesis 2 would be rejected in
favor of hypothesis 1 because hypothesis 2 requires an additional duplication and four instances of gene loss relative to hypothesis 1. There are a variety of gene families that have diversified during the last 500 million years of chordate evolution, and many of these gene families would be suitable candidates for exploring the utility and accuracy of inferring phylogenetic relationships of vertebrates from complex gene trees. Here, we focus on the analysis of the heat shock 70 (HSP70) gene family. HSP70 is an ideal choice because much is known about the structure, function, phylogeny, and evolution of these genes (Lindquist and Craig, 1988; Morimoto et al., 1994; James et al., 1997; Feder and Hofmann, 1999). Moreover, HSP70 genes have been used for inferring phylogenetic relationships among divergent taxa (Gupta et al., 1994, 1999; Molto et al., 1994; Gupta and Golding, 1993; Borchiellini et al., 1998; Budin and Philippe, 1998; Stedman et al., 1998; Germot and Philippe, 1999; Philippe et al., 1999; Schultz et al., 1999; Rue et al., 2000; Sulaiman et al., 2000). In some cases, HSP70 gene trees challenge conventional hypotheses (e.g., Budin and Philippe, 1998), and it is important to understand the limitations of HSP70 genes for phylogenetic inference before abandoning favored hypotheses for alternatives, no matter how interesting the alternatives may be. A large amount of HSP70 data is available for phylogenetic analysis, making it possible to evaluate the effects of gene duplication and gene family evolution on phylogenetic inference at various time intervals.

Our analysis focused on one group of sharks, the Lamniformes. The Lamniformes has been the subject of ongoing phylogenetic analysis in our lab, and several published and unpublished phylogenetic hypotheses exist that can be used to evaluate the utility of gene tree parsimony as a general strategy for choosing among rival hypotheses and for inferring novel hypotheses. Here, we describe HSP70 and HSP70-like gene diversity in lamniform sharks and use the inferred complex gene tree to test alternative phylogenetic hypotheses derived from analyses of morphology and molecules (Fig. 3). Our results emphasize the implications of gene paralogy for molecular systematics.

**METHODS**

**Study Taxa**

Our study focused on 12 species representing all of the genera in the order Lamniformes (Table 1). Genomic DNA was obtained from muscle, gill, or liver tissue using standard phenol:chloroform extraction methods (Sambrook et al., 1989).

**Amplification and Sequencing**

Several different pairs of primers that amplified different regions of the HSP70 gene were synthesized. Primers 3U (5'-GGT GGG AAG CAC TTT GCC AAC YAY MG-3') and 5D (5'-ATG AAA TGG AAG TTC ATC AGT C-3') amplify a central region of the gene, as do 4U (5'-GAG CGC AAT GTC CTG ATY GTA GA-3') and 5D. Amplification conditions for the 3U, 4U, and 5D primers typically involved a two-step annealing process with an initial denaturation for 3 min at 94°C then four cycles of 94°C for 30 sec, 42°C for 30 sec, and 72°C for 90 sec followed by 36 cycles of 94°C for 30 sec, 52°C for 30 sec, and 72°C for 90 sec and a final annealing step for 7 min at 72°C. The primers 1U (5'-CTG GGC ACC ACC TAC TCC TG-3') and 7D (5'-AGC TGR TTR TTC TCC AGC C-3') amplify an approximately 1.5-kilobase region beginning about 60 base pairs downstream of the start codon. Amplification conditions for these primers were
TABLE 1. List of shark species included in this study.

<table>
<thead>
<tr>
<th>Species</th>
<th>Abbreviation</th>
<th>Location</th>
<th>Collector</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alopias pelagicus</td>
<td>Alpe</td>
<td>Mexico</td>
<td>J. Caira</td>
</tr>
<tr>
<td>Alopias superciliosus</td>
<td>Alsu</td>
<td>Taiwan</td>
<td>G. Chen</td>
</tr>
<tr>
<td>Alopias vulpinus</td>
<td>Alvu</td>
<td>Taiwan</td>
<td>G. Chen</td>
</tr>
<tr>
<td>Carcharodon carcharias</td>
<td>Caca</td>
<td>Australia</td>
<td>CSIRO</td>
</tr>
<tr>
<td>Carcharias taurus</td>
<td>Cata</td>
<td>South Africa</td>
<td>G. Cliff</td>
</tr>
<tr>
<td>Cetorhinus maximus</td>
<td>Cema</td>
<td>Australia</td>
<td>J. Stevens</td>
</tr>
<tr>
<td>Isurus oxyrinchus</td>
<td>Isox</td>
<td>New Zealand</td>
<td>M. Francis</td>
</tr>
<tr>
<td>Lamna ditropis</td>
<td>Ladi</td>
<td>Alaska</td>
<td>K. Goldman</td>
</tr>
<tr>
<td>Megachasma pelagios</td>
<td>Mepe</td>
<td>Japan</td>
<td>M. Miya</td>
</tr>
<tr>
<td>Mitsukurina owstoni</td>
<td>Miow</td>
<td>Japan</td>
<td>M. Miya</td>
</tr>
<tr>
<td>Lamna ditropis</td>
<td>Ladi</td>
<td>Alaska</td>
<td>K. Goldman</td>
</tr>
<tr>
<td>Megachasma pelagios</td>
<td>Mepe</td>
<td>Japan</td>
<td>M. Miya</td>
</tr>
<tr>
<td>Mitsukurina owstoni</td>
<td>Miow</td>
<td>Japan</td>
<td>M. Miya</td>
</tr>
<tr>
<td>Lamna ditropis</td>
<td>Ladi</td>
<td>Alaska</td>
<td>K. Goldman</td>
</tr>
<tr>
<td>Megachasma pelagios</td>
<td>Mepe</td>
<td>Japan</td>
<td>M. Miya</td>
</tr>
<tr>
<td>Mitsukurina owstoni</td>
<td>Miow</td>
<td>Japan</td>
<td>M. Miya</td>
</tr>
<tr>
<td>Lamna ditropis</td>
<td>Ladi</td>
<td>Alaska</td>
<td>K. Goldman</td>
</tr>
<tr>
<td>Megachasma pelagios</td>
<td>Mepe</td>
<td>Japan</td>
<td>M. Miya</td>
</tr>
<tr>
<td>Mitsukurina owstoni</td>
<td>Miow</td>
<td>Japan</td>
<td>M. Miya</td>
</tr>
<tr>
<td>Lamna ditropis</td>
<td>Ladi</td>
<td>Alaska</td>
<td>K. Goldman</td>
</tr>
<tr>
<td>Megachasma pelagios</td>
<td>Mepe</td>
<td>Japan</td>
<td>M. Miya</td>
</tr>
<tr>
<td>Mitsukurina owstoni</td>
<td>Miow</td>
<td>Japan</td>
<td>M. Miya</td>
</tr>
<tr>
<td>Pseudocarcharias kamoharai</td>
<td>Pska</td>
<td>Hawaii</td>
<td>R. Richmond</td>
</tr>
</tbody>
</table>

FIGURE 3. Published alternative hypotheses of the phylogenetic relationships among lamniform sharks. (A) Hypothesis of Compagno (1990) based on morphological characters. (B) Hypothesis of Long and Wagggoner (1996) based on parsimony analysis of dental characters. (C, D) Gene trees derived from parsimony (with third position transitions omitted) and LogDet phylogenetic analysis of mitochondrial DNA data, respectively (Naylor et al., 1997: Figures 6A and 6B, respectively). In the original publications, none of the trees have bootstrap values associated with nodes.
94°C for 60 sec, 52°C for 45 sec, and 72°C for 90 sec. In many cases, we used primers BorFi (5'-ATCATYGCYAAYGAYCARGG-3') and BorRai (5'-CCYTTRTCRTTGGTGATRGT-3') for nested amplifications using the 1U and 7D product as template. Primers BorFi and BorRai were modified from the primers described by Borchiellini et al. (1998). For nested amplifications, we used the cycling parameters of 94°C for 3 min, 35 cycles of 94°C for 35 sec, 60°C for 45 sec, and 72°C for 90 sec followed by a final extension at 72°C for 7 min. Reaction conditions for all amplifications were 2 mM MgCl₂, 200 μM dNTP, 10 mM Tris-HCl (pH 9.0), 50 mM KCl, 0.1% Triton X-100, 12–50 pmol each primer, and 2 U Promega Taq polymerase in a 25–50-μl reaction volume.

For each species, a single individual was used. Two independent polymerase chain reactions (PCRs) were performed for each set of primers. Following amplification, fragments were purified using a Wizard PCR kit (Promega) following the manufacturer's recommended protocol and cloned into TA plasmids using a TOPO cloning kit (Invitrogen). Positive colonies were picked and grown overnight, and plasmids were harvested using alkaline lysis methods (Sambrook et al., 1989). Clones were assayed by restriction enzyme digestion, and plasmids harboring fragments of the expected size were sequenced by cycle sequencing (using the EXCEL kit, Epicentre Technologies) with 700-nm and 800-nm fluorescent dye labeled M13R and T7 primers on a Licor 4200 sequencer. For each species, we cloned five or six different amplification products and sequenced 3–20 clones from each cloning reaction. Consensus sequences from both directions were assembled using the ALIGNIR program. Some of the unique sequences obtained from each cloning reaction may reflect PCR artifacts. The expected error rate of Taq polymerase is less than one change per 1,000 base pairs (Cline et al., 1996). In most cases, however, unique sequences determined from the same individual differed at multiple sites, implying the presence of multiple gene copies in the genome or the existence of alleles.

DNA sequences from sharks were aligned by eye because of the high conservation of HSP70 genes. In addition, the DNA sequences were translated into proteins using MacClade (Maddison and Maddison, 1993). The translation was used to assess whether substitutions affected the coding sequences or resulted in stop codons, suggesting that changes were due to Taq polymerase error or that the sequence represented was a pseudogene. The pattern of accumulation of substitutions at second and third codon positions was examined using PAUP* 4.0 (Swofford, 2002) to assess whether the pattern of nucleotide substitution conforms to expectations for HSP70 genes, i.e., that most changes happened at third codon positions.

Inferring Species Trees Using Mitochondrial DNA and RAG-1

DNA data for the mitochondrion-encoded cytochrome $b$ and for the ND2 protein-coding genes and for the nucleus-encoded RAG-1 gene were subject to Bayesian phylogenetic analysis using MrBayes (Huelsenbeck and Ronquist, 2001). *Carapizus* was used as the outgroup (for details on the data, see Naylor et al., 1997; Martin, 1999). Both data sets were analyzed in the same way. The analyses employed one cold chain and three incrementally heated chains, as described by Huelsenbeck and Ronquist (2001). Posterior probability distributions were determined for the phylogeny and for the substitution model parameters. Model parameters included the six substitution rates of the general time reversible (GTR) model, the four nucleotide frequencies, and relative rates of change at the three codon positions (GTR + SSR). A random starting tree and uniform prior distributions were employed using the default parameters of MrBayes. The length of the chain was set at 100,000 and the burnin values for the two data sets were 30,000 and 40,000 for the mitochondrial DNA (mtDNA) and RAG-1 data sets, respectively, based on initial surveys to assess the number of generations required for stationarity. Trees were sampled every 100 generations, resulting in consensus trees based on 700 and 600 trees for the mtDNA and RAG-1 genes, respectively. The topologies generated by Bayesian analysis of the two data sets were compared using the Shimodaira–Hasegawa RELL (SH) test (Shimodaira and Hasegawa, 1999) implemented in PAUP* (Swofford, 2002). Reciprocal tests were performed, in each case using the most appropriate model of sequence evolution for the data.
as estimated by ModelTest (Posada and Crandall, 1998).

Inferring the Shark HSP70 and HSP70-Like Gene Tree

We first sought to determine which of the many cytosolic HSP70 genes that have been described from vertebrates we had isolated and sequenced. Distinct shark sequences were used for BLAST searches of the database, a set of sequences with the high matching scores was retained (Table 2), and the inferred protein sequences were aligned using CLUSTAL X (Thompson et al., 1997). The aligned data were imported into PAUP* 4.0 (Swofford, 2002), and a set of maximum parsimony trees was generated using a heuristic search option with 100 random additions of taxa and TBR branch swapping. A 50% majority rule consensus tree was then constructed.

Gene trees for the shark HSP70 DNA sequences were estimated using Bayesian analysis implemented with MrBayes (Huelsenbeck and Ronquist, 2001). The default prior parameter values used for mtDNA and RAG-1 were adopted. Initial surveys revealed that likelihood values stabilized after about 450,000 generations; thus, for these data, we performed 500,000 generations, sampling trees every 50 generations, yielding 1,000 trees from which to construct a consensus tree and calculate posterior probabilities.

Reconciliation for Choosing Among Alternative Species Trees

Reconciliation provides one method of using multigene family trees for choosing among alternative hypotheses of species relationships. Parsimony was adopted as the objective criterion, and the best phylogenetic hypothesis (see Figs. 3 and 4) is the one that requires the fewest ad hoc hypotheses of gene duplication or gene duplication and loss. The number of gene duplications and losses required by each hypothesis was determined using the program

| Table 2. List of previously published HSP70 genes included in this study, the source organism, the genome location of the gene (when available), and the GenBank accession numbers. |
|-----------------|-----------------|-----------------|-----------------|
| Gene name       | Species         | Location        | Accession no.   |
| HSPA1A          | human           | 6p21.3          | NM_005345       |
| HSPA1B          | human           | 6p21.3          | NM_005346       |
| HSPA1L          | human           | 6p21.3          | NM_005527       |
| HSPA2           | human           | 14q24.1         | NM_021979       |
| HSPA6           | human           | 1q              | NM_002155       |
| HSPA7           | human           | 1q              | P48741          |
| HSPA8           | human           | 11q23.3–q25     | NM_006597       |
| HSP70-2         | mouse           | 12              | NM_008301       |
| HSC70t          | mouse           | 17              | NM_013558       |
| HSC70-1         | mouse           | 17              | M35021          |
| HSC70           | mouse           | 9               | NM_031165       |
| HSC71           | mouse           | 9               | A49935          |
| HSP70A1         | mouse           | 17              | M76631          |
| HSP70-MHC-3     | rat             | 20              | X77209          |
| HSC70           | rat             |                 | NJ_024351       |
| HSP70-1         | rat             |                 | NP_144177       |
| HSP70           | rat             |                 | Z27118          |
| HSP70-1         | rat             |                 | X74721          |
| HSP70           | rat             |                 | X75357          |
| HSP70-MHC-2     | rat             | 20              | X77208          |
| HSP70-MHC-1     | rat             | 20              | X77207          |
| HSP70-3         | Fugu            | all five genes clustered on same chromosome | Y08578 |
| HSP70-5         | Fugu            |                 | Y08582          |
| HSP70-2         | Fugu            |                 | Y08577          |
| HSP70-4         | Fugu            |                 | Y08581          |
| HSPA1A          | zebrafish       |                 | Y08576          |
| HSC70           | zebrafish       |                 | AF210640        |
| HSC70           | zebrafish       |                 | AAB03704        |
| HSC70           | zebrafish       |                 | AF006007        |
| HSC70           | zebrafish       |                 | CAA72216        |
Hypotheses of species relationships based on Bayesian phylogenetic analysis using the GTR + SSR model of evolution for mtDNA and the nucleus-encoded RAG-1 gene. Numbers along branches are posterior probabilities. Branch lengths were estimated using maximum likelihood and optimized according to GTR + G + I and GTR + G models of nucleotide substitution for the mtDNA and RAG-1 data, respectively (based on log likelihood ratio tests implemented by ModelTest; Posada and Crandall, 2001). The mtDNA and RAG-1 trees correspond to hypotheses E and F, respectively (Fig. 9).

GeneTree (Page, 1998; available on-line at http://taxonomy.zoology.gla.ac.uk/). The significance of the estimated numbers of gene duplications and losses was assessed by comparison with a distribution based on 1,000 random trees generated by a Yule process implemented using GeneTree. Because multiple hypotheses (six) were tested against the null expectations, 0.05 / 6 = 0.008 was adopted as the maximum critical value for determining significance. In addition, a heuristic tree search was done using a combined NNI and TBR algorithm to find the species tree that best explained the gene tree (Page and Charleston, 1997). The most-parsimonious species trees estimated using GeneTree were tested against the mtDNA and RAG-1 trees estimated by Bayesian phylogenetic analysis using the SH RELL test implemented in PAUP* (Swofford, 2002). Because GeneTree yields rooted trees, *Carcharhinus* was included when performing the SH tests.

**RESULTS**

Inferred Species Trees from mtDNA and RAG-1

Bayesian analysis using the GTR + SSR model for the combined cytochrome *b* and ND2 genes and the RAG-1 gene resulted in fully resolved but different topologies (Fig. 4). The exact details of the groupings of taxa are not important in the present context; nevertheless, the difference between the two topologies was significant when assessed using the SH test on the mtDNA data (*P* = 0.023). The two hypotheses were not significantly different when tested using the RAG-1 data, however (*P* = 0.069). Because of evidence for heterogeneity of hierarchical signal, the data were not combined and the two hypotheses were retained.

**HSP70 and HSP70-Like Genes Sampled**

We characterized 123 unique HSP70 and HSP70-like sequences from the 12 species of sharks surveyed (GenBank accession nos. AF502436–AF502558). The longest sequences were obtained using the BorFi and BorRai primers, yielding a 1,418-base pair fragment. These two primers also yielded a much smaller product from the three Lamnidae species (*Carcharodon*, *Isurus*, and *Lamna*). This fragment was missing an approximately 600-base pair region about 100 base pairs downstream of the start codon and had several other smaller indel mutations. A single base deletion altered the reading frame, and even when the deletions are treated as missing data, all of the sequences have a stop codon within the reading.
frame, suggesting that they are pseudogenes. The pseudogene appears to have duplicated within each species, suggesting that the dynamics of HSP70 gene duplication may not be influenced by whether the gene is functional or not. Another sequence sampled from *Alopias superciliosus* also was identified as a pseudogene based on indels that disrupted the reading frame. Other sets of primers yielded smaller fragments of HSP70 genes.

Pairwise sequence divergences calculated for the second and third codon positions revealed high conservation of second codon positions. Similarly, estimates of substitution rates for the three codon positions using Bayesian analysis revealed a very low rate for second positions relative to third positions (data not shown). These results are consistent with observations from other studies that the evolution of heat shock genes is relatively conservative because of strong purifying selection.

Phylogenetic analysis of aligned protein sequences from several species of vertebrates, including three divergent genes sampled from sharks, was used to infer the functional and phylogenetic identity of the sampled genes. Parsimony analysis of aligned protein sequences for an assortment of vertebrates and for representative sequences from sharks resulted in 112 minimum-length trees. Most topological differences involved distal clades of the tree, although several basal nodes also were unresolved in the majority rule consensus tree (Fig. 5). The three shark sequences included at this level of the analysis appeared in three different places relative to sequences from other vertebrates, although all of the shark sequences grouped together in a large clade of cytosolic constitutively expressed genes. Thus, the set of sequences isolated from sharks represents only a small fraction of the HSP70 gene diversity present in the genome of vertebrates. Sequences associated with the endoplasmic reticulum and mitochondria were not sampled. Although this analysis was not an exhaustive survey of vertebrate HSP70 diversification, the tendency for paralogy within individual gene clades varied among groups. For instance, a cluster of related genes sampled from *Fugu* (a fish) were all more closely related to each other than to other genes sampled from other taxa. By contrast, most of the genes sampled from human were more closely related to a gene from another species (i.e., mouse) than to another human HSP70 gene.

**Gene Tree for Sharks**

The Bayesian consensus gene tree for all 123 shark sequences is shown in Figure 6. Several divergent gene clades were sampled from each species, and within each gene clade, each species tended to have multiple distinct sequences. In many cases, the monophyly of sequences within species within gene clades was supported by high posterior probabilities. This pattern was similar to that described for *Fugu* (Fig. 5) and suggests that most duplications giving rise to HSP70 paralogous genes were recent. Although the fully resolved Bayesian tree may be misleading because several nodes have low posterior probabilities, the HSP70 gene tree depicted in Figure 6 remains the most probable explanation of the sequence data and was used for gene tree parsimony analysis.

The sequence data were optimized on the Bayesian consensus tree using minimum evolution to show the extent of sequence divergence among the sampled sequences (Fig. 7). The three highly divergent gene clades were evident. Within a given gene clade, sequences isolated from individuals tended to be similar; however, in most cases, the divergence was more extensive than expected from Taq polymerase error (Cline et al., 1996), suggesting that most of the data are representative of the actual divergence present in shark genomes. Moreover, the slow substitution rate for nuclear DNA in sharks (Martin, 1999) coupled with evidence of limited within-species sequence diversity in sharks (Heist et al., 1996; Gardner and Ward, 1998) suggests that most of the unique sequences sampled from a given species of shark are paralogues. Still, the possibility remains that some of the sampled genes from the same individuals represent alleles and not different loci.

Within the relatively restricted part of the HSP70 gene family sampled from sharks, several distinct sequences were sampled from most species. In most cases, though, the three divergent gene clades would be recognizable as distinct paralogues when pursuing systematic studies because of the high level of sequence divergence among them. However, gene paralogy that can potentially
obscure inference of species level relationships is evident within gene clades. Although most putative paralogues within a distinct gene clade coalesce within species (Figs. 6 and 7), several species appear in multiple places on the tree. For instance, in the clade with the best representation in the data, genes from *Alopias superciliosus* (Alsu) appear in two different places on the tree: as the sister taxa to genes sampled from *Alopias pelagicus* (Alpe) and outside the grouping of Alsu and Alpe (Fig. 8A). Similarly, sequences obtained from *Cetorhinus maximus* (Cema) occurred in three places (Fig. 8A). The unambiguous ancient paralogy of genes within these two species suggests that differential loss of duplicate genes among taxa may compromise inference of organismal
phylogeny. In this particular case, if we sampled only single genes from each taxon, several dramatically different topologies would be possible (Figs. 8B and 8C show two of these).

**Gene Tree Parsimony**

The Bayesian consensus tree (Fig. 6) was reconciled with the six alternative phylogenetic hypotheses depicted in Figures 3
FIGURE 7. Minimum evolution phylogram of the Bayesian topology (Fig. 6) assuming a GTR + G substitution model. Shark symbols identified gene clades shown in Figure 5. Putative pseudogenes are labeled with a Ψ. Posterior probabilities for branches are shown in Figure 6.
and 4. The distributions of duplications and losses and for duplications only for 1,000 random species trees were generated to assess the significance of the observed fit between the phylogenetic hypotheses and the HSP70 gene tree. When duplications + losses was adopted as the objective criterion for evaluating the fit between organismal trees and the gene tree, all of the phylogenetic hypotheses based on molecular data were significantly different from expectations of the null hypothesis (Fig. 9). By contrast, the two trees based on morphological data fell within the main part of the distribution for random trees. When only duplications were considered, a different set of trees was significantly different from the expectation under the null hypothesis (Fig. 9). In this case, the trees based on mtDNA and Compagno’s hypothesis were significantly different from random (null) expectations. The RAG-1 tree was not significantly different from random.

A heuristic search for species trees requiring the fewest number of gene duplications and losses yielded two trees of length 171. One tree required 101 gene duplications and 70 gene losses (Fig. 10A), and the other required 100 duplications and 71 gene losses (Fig. 10B). A search using only duplications as the objective criterion yielded >100 trees of length 100. Figure 10B represents the topology requiring the fewest gene losses of the set of trees requiring the fewest gene duplications. Both trees were compared to the species trees estimated by Bayesian analysis using the SH RELL test separately for the mtDNA and RAG-1 data sets. Both data sets revealed that the hypothesis requiring 101 gene duplications and 70 gene losses (Fig. 10A) provided a poor explanation of the mtDNA and RAG-1 sequence data (P = 0.000 and 0.023 for the mtDNA and RAG-1 data, respectively). The other hypothesis (Fig. 10B) was significantly different from the Bayesian tree when evaluated using mtDNA (P = 0.023) but was not significantly different from the Bayesian tree when evaluated using the RAG-1 data (P = 0.099).

**DISCUSSION**

**Ebb and Flow of Genes and Implications for Phylogenetics**

Genes of multigene families are related through descent by gene duplication.
Individual genes come and go within multigene families, and many families include both functional genes and pseudogenes. Recent comparative studies suggest birth and death of genes through multiple molecular processes of gene duplication and loss is probably continuous and characteristic of many gene families. Moreover, the gain and loss of genes can occur rapidly (Lynch and Conery, 2000). One consequence of continuous remodeling of gene families is that different taxa will possess sets of paralogous genes that differ in complexity and phylogenetic history (McKinney and Meagher, 1998; Small and Wendel, 2000; Vanhalst et al., 2001). Results for the HSP70 gene family in sharks indicate that duplicate paralogous copies appear to have originated at different times in different lineages. Although the dynamic is interesting, the birth and death of genes that characterizes gene families creates a moving target when one is trying to infer organismal phylogeny, and topologies of gene and species trees probably will differ.

Systematists largely ignore genes that are part of gene families when inferring phylogeny and instead focus on single-copy genes because of the inherent problems of determining orthology of genes sampled from different taxa. Yet, the idea that genes always have been present as single copies and that the evolutionary record of organismal diversification can be reliably extracted through phylogenetic analysis of single gene sequences sampled from some set of taxa depends more on faith than fact. The
discovery that putative single-copy genes often have paralogues underscores the uncertainty of assuming that sampled genes are orthologous (e.g., Danforth and Ji, 1998). Moreover, inferences from ribosomal genes, which have been phylogenetic workhorses, suffer from problems of paralogy because of incomplete homogenization of ribosomal arrays (Buckler et al., 1997; Hartmann et al., 2001; Mayol and Rossello, 2001).

Emphasis on the important role of polyploidization in organismal diversification (e.g., Sidow, 1996; Amores et al., 1998; Vision et al., 2000) suggests that gene paralogy may be common. In such cases, differential sorting of duplicate genes is inevitable and will take place over a relatively extended period of time depending on the underlying mutation rate of the genomes and the dynamics of inheritance (Wolfe, 2001). Lynch and Conery (2000) estimated an average half-life of duplicate gene copies of about 4 million years, a long time relative to the time needed for speciation. Furthermore, recent theoretical arguments suggest that divergent resolution, in which different copies of duplicated genes are lost in allopatric populations, may be an important process promoting organismal diversification (Lynch and Force, 2000; Taylor et al., 2001). An implication of the divergent resolution hypothesis is that functionally similar genes in related taxa may be paralogous.

The problem of hidden paralogy also influences estimation of speciation times. Species divergence time, $t$, is less than the divergence time between orthologous genes, $T_\text{or}$, in an amount that is proportional to ancestral population size (Edwards and Beerli, 2000) and is less than the divergence time between paralogous genes, $T_\text{p}$, by an amount that is proportional to the rate of gene silencing. Depending on the dynamics of particular gene families and the mechanism of duplication, the difference between $t$ and $T_\text{p}$ may be large. For instance, if ancestors for groups of contemporary taxa were allopolyploids (as is the case for some plants; e.g., Gaut and Doebley, 1997), then $T_\text{p}$ may be much larger than $t$. Thus, caution is needed when interpreting molecular estimates of divergence times between divergent taxa. If paralogy is common, recent revisions of evolutionary history based on molecules (e.g., Wang et al., 1999; Heckman et al., 2001) may be forcing us to look further back into the past than we should.

**Assuming Paralogy**

Orthology of sampled genes is a hypothesis; nevertheless, the orthology of genes sampled from different taxa is often assumed. Rather, paralogy of sampled genes should be assumed because the possibility of paralogy exists, even for putative single-copy genes. If we adopt this perspective, how do we go about the important business of inferring phylogenies? We offer three recommendations. First, genomes should be surveyed for multiple paralogs. In most cases, this survey will involve cloning PCR products and screening individual clones for different sequences. Alternatively, or additionally, PCR conditions can be varied or different sets of primers can be used when performing amplifications. This step—surveying gene diversity within individual taxa—is essential when estimating gene trees because it can reveal paralogous gene copies. Second, because each gene tree has some probability of disagreement with the species tree because of unrecognized paralogy of sampled genes and assuming the gain and loss of genes is independent across taxa, the disagreement between gene and species trees should not be the same across independent genes or gene families. Consequently, multiple gene trees are required to estimate the species tree. A particularly compelling example of this situation is the recent estimate of basal eukaryote relationships by Baldauf et al. (2000). In that study, several genes showed strong but incongruous support for alternative hypotheses, a result that may have reflected paralogy; nevertheless, topological disagreements between any two data sets were resolved when the phylogeny was inferred from multiple genes. In cases where gene trees include multiple paralogues from each species, it is possible to use individual gene clades as independent estimates of phylogenetic relationships, although this strategy has been criticized (Ruvolo and Koh, 1996). Third, gene tree parsimony provides an additional means for inferring organismal trees without assuming orthology of sampled genes (Page and Charleston, 1997; Page, 2000). This strategy seeks the species tree requiring the fewest ad hoc hypotheses of gene duplication and loss to explain gene trees.
Gene Tree Parsimony

Gene tree parsimony provides a method for either testing a priori hypotheses or inferring a novel species tree. We did both for the HSP70 data. When we compared the fit from reconciliation of the hypotheses in Figures 3 and 4 with the HSP70 gene tree depicted in Figure 6 with the distribution of gene duplications and losses required for random trees, four of the six hypotheses were significantly different from the random (null) expectations (Fig. 9). All of these trees were derived from analysis of molecular data. By contrast, both trees based on morphological data fell within the distribution for random trees. However, when the objective criterion was gene duplications only (omitting gene losses), the power for resolving differences among hypotheses diminished and a different set of trees were significantly different from the expectations of the null hypothesis. All of the organismal trees based on analysis of mtDNA were significantly different from random trees. Compagno’s (1990) morphological tree was also significantly different from expectations for random trees. However, the organismal tree based on RAG-1 sequences was not significantly different from random expectations.

When using gene tree parsimony, an objective criterion is necessary. In the absence of exhaustive sampling of the gene family in each taxon or in cases in which the sampling is uneven across taxa, Cotton (pers. comm) and Page (pers. comm.) have suggested excluding inferred gene losses from the objective criterion, thus leaving only gene duplications. When we did this, the resolution among alternative trees decreased significantly (Fig. 9). A troubling result was that one of the best a priori hypothesis inferred using only gene duplications (hypothesis A, Fig. 3) was soundly rejected when both gene duplications and losses were considered. One explanation for this result is that highly pectinate trees (such as hypothesis A in Fig. 3) require fewer duplications but more gene losses, all else being equal, than do more balanced trees. However, because gene duplications and losses (or missing lineages) represent a real cost of assuming a particular species tree, the omission of gene losses for any reason seems arbitrary. A better approach weight gene duplications more heavily than gene losses when evaluating alternative hypotheses. One can weight the two events in proportion to their frequency based on the organismal tree or can assign priority to duplications over losses. In the latter case, one chooses the species tree requiring the fewest gene losses from the set of trees requiring the fewest gene duplications.

A heuristic search using gene duplications + gene losses as the objection criterion resulted in two maximum-parsimony trees (Fig. 10). One of the trees (A) differed substantially from all of these previously proposed phylogenetic hypothesis. Particularly disturbing was the derived placement of Mitsukurina (Miow). The other tree (B) broadly agreed with the previously proposed hypotheses, except for the relative position of Carcharias (Cata) and Cetorhinus (Cema) and placement of Pseudocarcharias (Pska) outside of a clade consisting of Alipias (Alpe, Alsu, and Alvu), Odontaspis (Odfe), and Megachasma (Mepe). Species tree A was soundly rejected as a reasonable explanation of the mtDNA and RAG-1 sequence data based on SH tests. The alternative species tree (B) was significantly different from the Bayesian tree based on mtDNA but was not significantly different from the Bayesian RAG-1 tree. Moreover, when we adopted the criterion that gene duplications should take priority over gene losses when choosing among alternative hypotheses using gene tree parsimony, then the topology depicted in Figure 10B is the best estimate of relationships based on the HSP70 data. The search for species trees that minimized the number of duplications yielded > 100 trees of length 100, of which the topology in Figure 10B was the one requiring the fewest gene losses.

Implications for Phylogenies Based on HSP Genes

HSP70 genes have been used for inferring the phylogenetic relationships among divergent taxa (Gupta et al., 1994, 1999; Molto et al., 1994; Gupta and Golding, 1993; Borchiellini et al., 1998; Budin and Philippe, 1998; Stedman et al., 1998; Germot and Philippe, 1999; Philippe et al., 1999; Schultz et al., 1999; Rue et al., 2000; Sulaiman et al., 2000). In many cases, however, inferred organismal trees based on HSP70 data do not match traditional phylogenies or those based on other genes. Our results suggest that the gain and loss of HSP70 genes makes most
inferences of organismal evolutionary history based on HSP70 data suspect because of the high potential for unrecognized paralogy. Thus, although high gene conservation makes HSP70 ideal for inferring deep phylogeny, the gain and loss of individual genes over time increases the likelihood that sampled genes are paralogous, thereby compromising estimation of organismal phylogeny. Whether or not similar criticism can be leveled at other conserved genes remains to be seen.

ACKNOWLEDGMENTS

This research was funded by the National Science Foundation (DEB 9974144). B. Kreiser and J. Gibbs assisted in the lab. The comments of R. Page, M. Simmons, and especially J. Cotton and C. Simon are greatly appreciated.

REFERENCES


First submitted 31 August 2001; reviews returned 26 November 2001; final acceptance 15 April 2002

Associate Editor: Rod Page