MOLECULAR PHYLOGENETICS AND EVOLUTION
Vol. 6, No. 2, October, pp. 259–269, 1996
ARTICLE NO. 0075

Molecular Evolution and Phylogenetic Utility of the Polyubiquitin Locus in Mammals and Higher Vertebrates

PAUL B. VRANA*,†,[1] AND WARD C. WHEELER*

*Department of Invertebrates, American Museum of Natural History, New York, New York 10024-5192; and †Department of Biological Sciences, Columbia University, New York, New York 10027

Received July 21, 1995; revised November 16, 1995

The product of ubiquitin genes is a small protein involved in intracellular sorting of other proteins. The locus consists of tandemly arrayed, uninterrupted copies of the gene. As several studies have noted, the Polyubiquitin locus is a model system for studying concerted evolution. While the protein is among the most conserved known, individual copies within an organism show variation in nucleotide sequence despite clear evidence of concerted evolution. When treated as individuals, repeats from a given locus form a monophyletic group. Furthermore adjacent copies often cluster, suggestive of the mechanism of concerted evolution. Due to this concerted evolution of repeats (and loci in organisms with multiple polyubiquitins), sequencing of heterogeneous PCR products consisting of all the repeats in a given organism may yield phylogenetic signal, as with other multicopy genes. We test this possibility through 22 original sequences using primers designed so that only tandem copies are amplified. Using these and previously published data, we further explore these phenomena in higher vertebrates and mammals in particular. We suggest that multiple locus duplications have occurred within mammals. Positional codon bias is strongly evident. We also find substitutional bias with regard to codon type. GC content of the locus appears to be generally high across vertebrates. Intraorganismal variation is tallied as an indication of frequency of change in codon position and transition/transversion ratios to further elucidate the tempo and mode of molecular evolution. Using these data, a weighting scheme for ubiquitin is also presented. Despite the gene's high GC content, transitional changes still appear more frequent. While the phylogenetic utility of ubiquitin does not appear great, its mechanistic insights seem far from exhausted.

INTRODUCTION

Distinguishing between orthologous and paralogous homology is a major problem in molecular systematics and evolution. This distinction is particularly important as biologists attempt to identify gene homologs in different organisms. When the nature of homology is unclear among a group of genes, use of a multiple-copy locus for phylogenetic purposes is appropriate only when it can be shown to have undergone concerted evolution. Not only must multiple copies of a gene be homogenized, but that mechanism must be one that retains derived mutations (apomorphies) found in some copies in order to maintain historical information or phylogenetic signal. Many of the DNA sequence phylogenetic studies published to date have used such multicopy genes. While different mechanisms have almost certainly been used, such homogenization must have occurred in the nuclear ribosomal genes often utilized, as well as in mitochondrial DNA.

The ubiquitin genes are a striking example of concerted evolution, amino acid conservation, and both overall and positional codon bias. The gene in question and its product are among the most conserved known (Sharp and Li, 1987a). The 228-nucleotide individual gene codes for a 76-amino-acid protein which conjugates to other proteins. The loci are unique among protein coding regions in that they consist of multiple arrayed, uninterrupted (with several exceptions) copies of the gene. Ubiquitin tandem repeats are interesting in that they exhibit homogenization but also almost invariably show sequence variation between all copies. The reversible crosslinking between individual ubiquitin moieties and other proteins may serve several functions. For example, when multiple ubiquitin residues are cross-linked to the N-terminus of a protein it appears to mark it for degradation via a unique nonlysosomal pathway (Finley and Varshavsky, 1985; Rechsteiner, 1991). Alternatively, a large proportion of Histone H2A and H2B molecules have a single ubiquitin molecule linked to the middle of the protein. The purpose of these ubiquitin/histone complexes is unclear (reviewed by Rechsteiner, 1991). The polypeptide forms

[1] To whom correspondence should be addressed at current address: Department of Molecular Biology, Lewis Thomas Labs, Princeton University, Princeton, NJ 08544. Fax: (609) 258-3345. E-mail: pvrana@watson.princeton.edu.

1055-7903/96 $18.00

TABLE 1

Chordate Taxa from which Polyubiquitin Was Examined

Latin	English	Rpts	GC%	Polysites
Branchiostoma lanceolatum	Amphioxis	3	68	13
Eptatretus stouti	Hagfish	4	62	8
Carcharhinus obscurus	Dusky Shark	4	52	3
Polypterus bichir	No common name	5	59	20
Amia calva	Bowfin	5	63	4
Scomber scombrus	Mackeral	2	63	20
Latimeria chalumnae	Coelacanth	2	51	14
Lepidosiren paradoxa	South American Lungfish	4	32	23
Notopthalmus viridescens	Salamander	2	38	13
*Xenopus laevis**	African Clawed Frog	3	47	11
Metachirus nudicaudatus	Short-tailed Oppossum	4	54	18
Mustela vison	Mink	6	67	7
*Bostaurus**	Cow	4	72	8
Homo sapiens III*	Human	3	65	12
Homo sapiens IV*	Human	4	64	15
Homo sapiens IX*	Human	9	47	34
*Mus musculus**	Mouse	4	76	6
Cricetulus griseus III*	Hamster	3	73	7
Cricetulus griseus IV*	Hamster	4	75	5
Geochelone sp.	Leopard Tortoise	2	63	1
Holbrookia maniculata	Earless Lizard	5	52	14
Varanus salvatori	Monitor Lizard	2	75	3
Crotalus viridis	Rattlesnake	5	70	7
Alligator missisipiensis	American Alligator	3	67	10
Tomistoma schlegelii	False Gavial	2	70	3
Crocodylus rhombifer	Crocodile	2	73	5
Gavialis gangeticus	Gavial	2	68	1
Osteolaemus tetraspis	No common name (crocodilian)	2	70	4
Caiman latirostris	Caiman	2	71	4
Struthio camelus	Ostrich	2	73	7
Gallus gallus IV*	Chicken	4	60	21
Gallus gallus III*	Chicken	3	59	18

Note. Mammal average GC content at third bases, 66%; amniote average GC content at third bases, 67%; vertebrate average GC content at third bases, 62%. Rpts, minimum number of repeats based on number of observed bands plus one, given the primer design (see text for details). GC%, % GC content at 3rd-base codon positions. Polysites, number of polymorphic sites. (*) indicates data obtained from GenBank.

a compact globular protein with a pronounced hydrophobic core (Vijay-Kumar *et al.*, 1985). However, there also exist in most genomes isolated copies of the gene fused to the other very different coding regions. These fusion proteins often contain a zinc-finger motif (Özkaynak *et al.*, 1987).

Both Sharp and Li (1978a,b) and more recently Tan *et al.* (1993) have examined concerted evolution among ubiquitin loci in some detail. These authors have shown that repeats in an organism tend be homogenized effectively. That is, all the copies in an organism group to the exclusion of copies in other organisms. However, this process appears to work differently in different organisms. For example, Sharp and Li (1987b) showed that the rate of homogenization among repeats was much slower in yeast than mammals, while Tan *et al.*'s analysis (1993) suggested that homogenization occurs between polyubiquitin and ubiquitin fusion genes in certain single-celled eukaryotes. Mita *et al.* (1991) examined the role of GC content and other factors affecting codon usage in ubiquitin genes and in particular noted a correspondence between GC content of codons and that of the entire organismal genome. Ubiquitin has been used in only two phylogenetic studies to date, on arthropod and single-celled eukaryote relationships (Wheeler *et al.*, 1993; Wray and DeSalle, 1994, respectively). These groups suggested that its usefulness in such studies is limited, particularly among very ancient splits. The nucleotide sequence appears to have some phylogenetic signal at the kingdom level, as do the limited number of amino acid substitutions, however (Müller *et al.*, 1994).

In this study, we pursue questions of the evolution of vertebrate ubiquitin genes from several directions. First, we test other mammal polyubiquitin loci not previously examined for concerted evolution. We then explore the pattern of relationships among repeats within individual loci to see whether this might yield clues to the homogenization process. New data consisting of a heterogeneous mix of polyubiquitin repeats have been

TABLE 2

Transition/Transversion Variation by Codon Position among Vertebrate Ubiquitin Repeats

	Transitions				Transversions			
	1st	2nd	3rd	Tot	1st	2nd	3rd	Tot
Pig	2	0	3	5	0	0	3	3
Cow	0	1*	7	8	0	0	0	0
Hamster III	0	0	6	6	0	0	1	1
Hamster V	0	0	4	4	0	0	1	1
Mouse	0	0	7	7	0	0	0	0
Mink	1	0	5	6	0	0	1	1
Human IX	2	0	26	28	0	0	12	12
Human III	0	0	11	11	0	0	2	2
Human IV	0	0	9	9	0	0	6	6
Oppossum	0	0	9	9	0	0	8	8
Mammal totals	5	1	87	93	0	0	34	34
Chicken II	3	0	12	15	0	0	4	4
Chicken IV	3	0	13	16	0	0	5	5
Ostrich	0	0	5	5	0	0	2	2
Alligator	0	0	8	8	0	0	2	2
Caiman	0	0	3	3	0	0	1	1
Tomistoma	0	0	3	3	0	0	0	0
Gavial	1	0	0	1	0	0	0	0
Osteolaemus	0	0	3	3	0	0	1	1
Crocodylus	0	0	5	5	0	0	0	0
Monitor Lizard	0	0	2	2	0	0	0	0
Rattlesnake	0	0	5	5	0	0	1	1
Earless Lizard	0	0	11	11	0	0	2	2
Leopard Tortoise	0	0	1	1	0	0	0	0
Newt	1	0	7	8	0	0	5	5
Frog	1	0	4	5	0	0	5	5
Lungfish	2	0	11	13	0	0	7	7
Coelacanth	1	0	8	9	0	0	5	5
Mackeral	0	0	13	13	0	0	8	8
Amia	1	0	1	2	0	0	2	2
Polypterus	3	0	11	14	1	0	2	3
Dusky Shark	1	0	2	3	0	0	0	0
Hagfish	0	0	3	3	0	0	4	4
Amphioxis	1	0	7	8	1	1	4	6
Vertebrate totals	23	1	225	249	2	1	94	97

Note. The outgroup amphioxis is also shown, 1st, 2nd, and 3rd refer to those respective codon positions, while "tot" signifies total. Numbers after loci (e.g., Hamster III) indicate the number of repeats in that locus to distinguish from others in the same species. (*) Amino acid change from serine (No. 57) to leucine. However, it is not known whether the latter is a functional ubiquitin copy.

generated for 22 vertebrate taxa. Through these data we further examine patterns of codon variation, both within and between kinds of codons. Evolution of GC content is also examined. Using the new and published vertebrate data, we further explore ubiquitin as an indicator of phylogenetic history, both alone and in combination with other data. Intraorganismal variation is tallied to better understand tempo and mode of molecular evolution in these loci. We employ a weighting scheme based on this variation to facilitate obtaining phylogenetic signal. For the data sets in question, 1st, 2nd, and 3rd bp positions and transitions/transversions are inversely weighted based on their frequency of heterogeneity within a locus. This assumes that changes seen within a locus are representative of the actual total number of (fixed) changes occurring and that more frequent kinds of changes tend to be less phylogenetically informative.

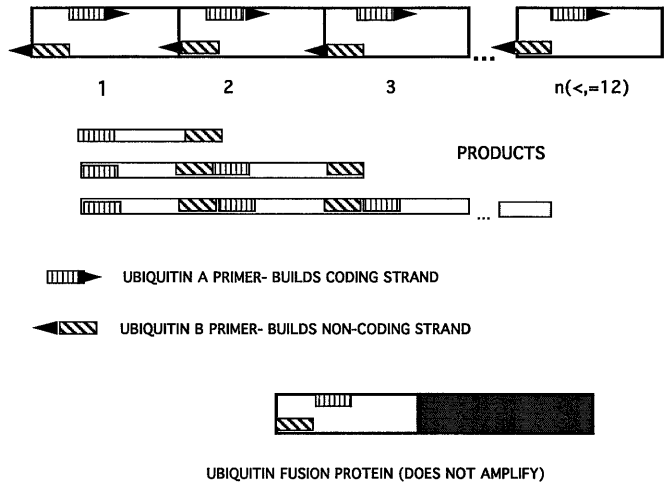

FIG. 1. Diagram of ubiquitin gene structure, amplification strategy, and location of PCR primers. See text for details.

METHODS AND MATERIALS

Data Collection

DNA isolation was performed by standard methods; see Vrana *et al.* (1994) for details. Amplification primers were designed such that only tandemly arrayed polyubiquitin loci were amplified. This eliminates the usually nonconcerted ubiquitin fusion protein gene from confounding the analyses. The strategy was accomplished by having the noncoding primer upstream of the adjacent coding primer. Thus amplification can occur only between two or more copies of the gene. The 3′-most copy in a polyubiquitin locus will not be amplified with this method, however. Figure 1 illustrates this primer design. The primer sequences are UbiqA, 5′ TT GACCGGAAAGACCATCAC 3′; UbiqB1.0, 5′ GGTCTT CACGAAGATCTGCA 3′. Single-stranded template suitable for sequencing was prepared for samples using the methods described by Allard *et al.* (1991). Briefly, this entailed performing a double-stranded (ds) reaction for a limited number (i.e., 25) of PCR cycles and using that entire product as a template for a 22-cycle PCR involving only one of the original primers and a higher annealing temperature and subsequent purification of that single-stranded (ss) product. Parameters for the ds PCR were 94°C, 1 min; 50°C, 1 min; 72°C, 1 min; those for the ss PCR were 94°C, 1 min; 55°C, 1 min; 72°C, 1 min. Amplification of the lungfish *Lepido-*

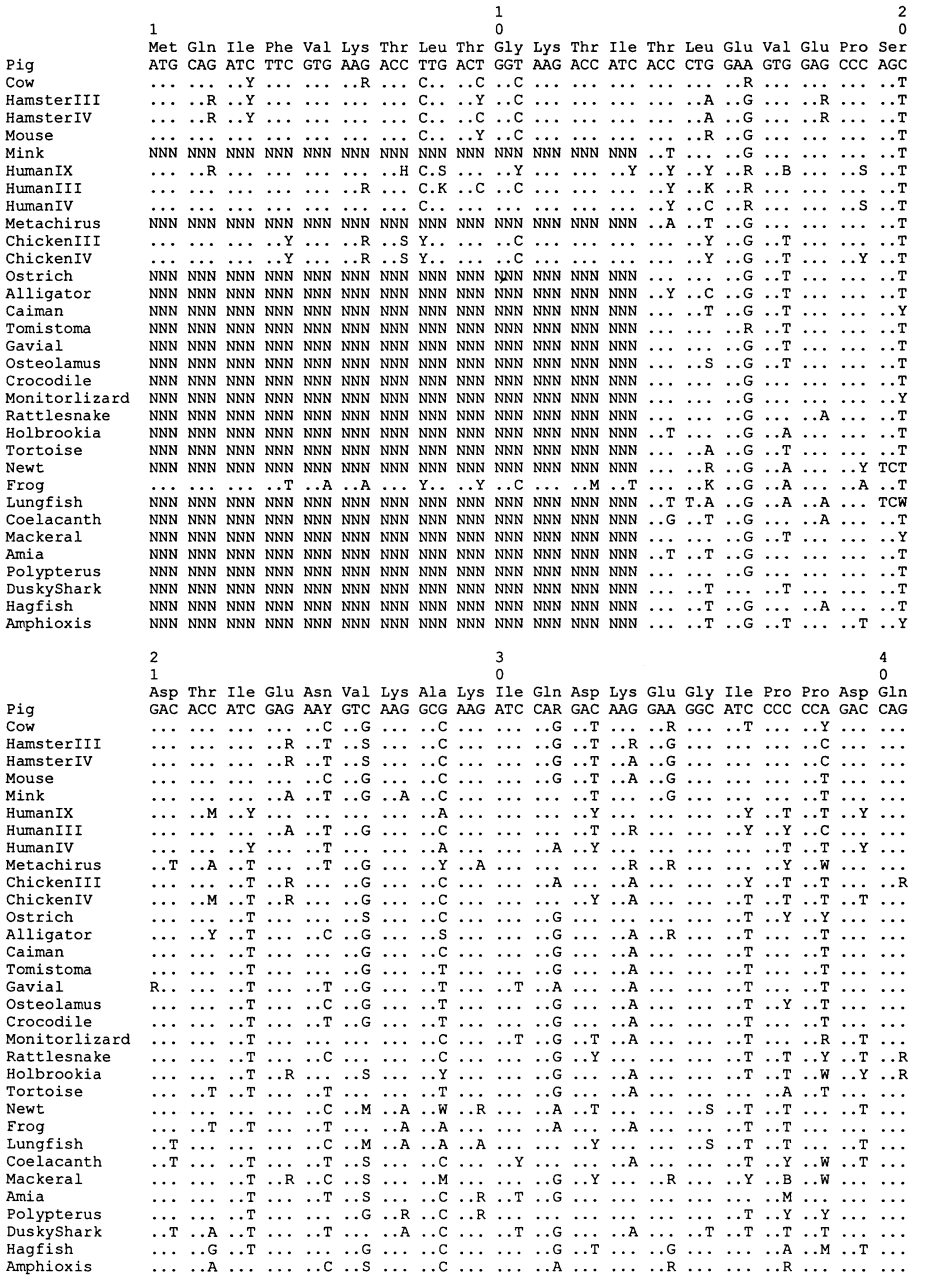

FIG. 2. Alignment of vertebrate consensus polyubiquitin sequences. N's represent the primer sequences; locations are as shown in Fig. 1. The outgroup chordate amphioxis is also shown. Dots denote identity with the pig sequence. Asterisks are shown above codons which are invariant within vertebrates or exhibit positional bias. In the case of the last serine codon (65), vertebrates utilize a TCN codon while amphioxis appears to also have a AGY codon at that position.

siren sample was more difficult (possibly due to the enormous amount of repetitive DNA this species contains): This sample was first boiled, then quenched on an ice slurry, where all other PCR reagents were added. It was then put directly into a prewarmed thermocycler to begin the cycling.

Sequencing was carried out using [^{35}S]dATP, the primers used for PCR amplification, the modified T7 DNA polymerase Sequenase (Version 2.0, U.S. Biochemical Corp.), and the accompanying reagents. Both strands were sequenced, which proved critical in evaluating heterogeneous sites. When more than one band was present at a position, the complementary strand was examined to see if the corresponding bands were present. Only sites which showed the same (complimentary) bands on both strands were scored as hetero-

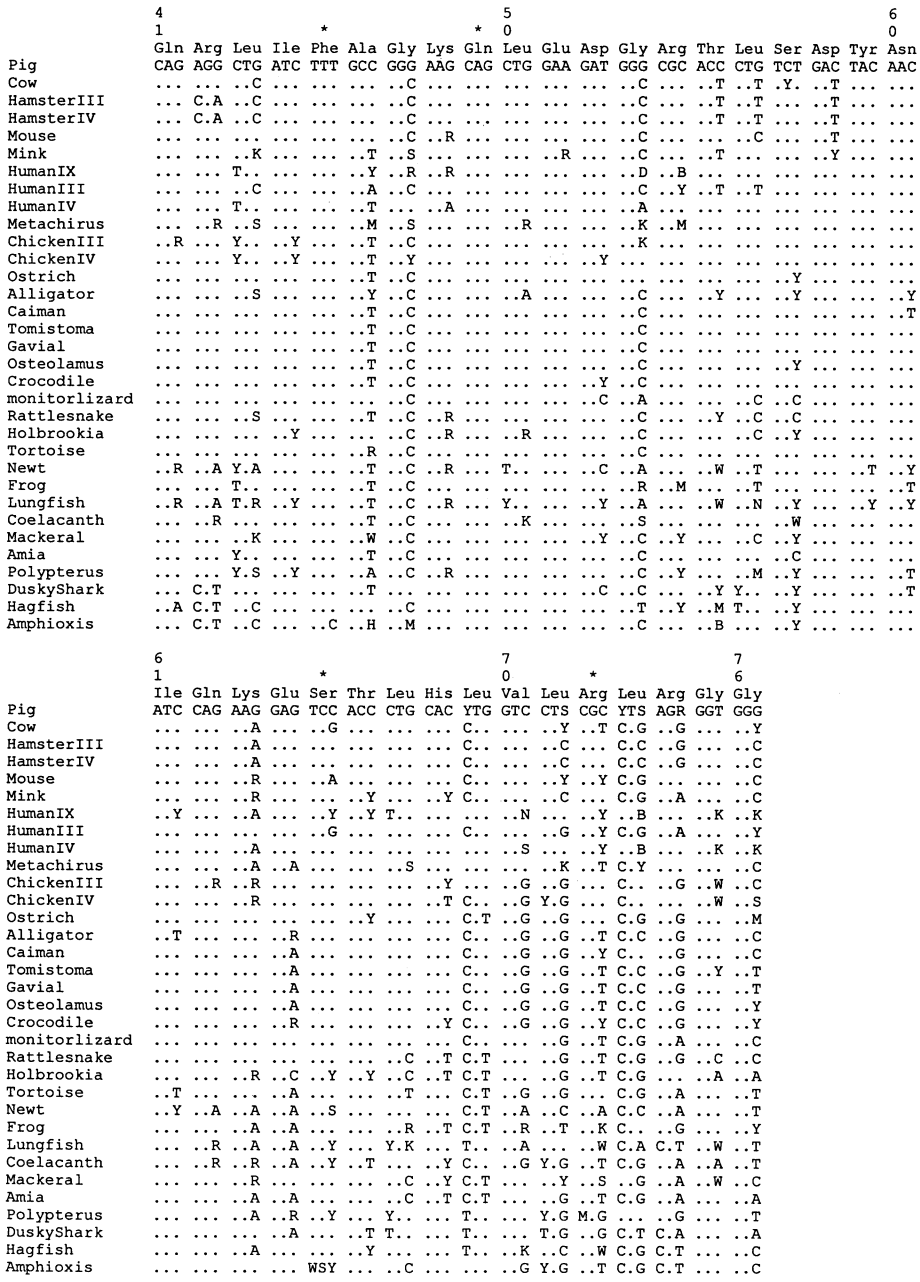

FIG. 2—*Continued*

geneous. Resequencing and/or multiple amplifications were performed in ambiguous cases. Sequences were examined and scored for number and kinds of variation in codon type and position.

Table 1 lists the new taxa for which sequence data were obtained. Samples were chosen to represent the higher vertebrate classes, with an emphasis on the amniotes and mammals in particular. The previously unexamined vertebrate polyubiquitin loci obtained from GenBank considered in this study include two hamster (*Cricetulus griseus*) loci (Accession Nos. X08013 and X60390); pig (*Sus scrofa*, Accession No. M18159); and cow (*Bos taurus*, Accession No. Z18245). The other vertebrate loci examined are the mouse, frog (*Xenopus*), two chicken loci, and the three human loci listed by Tan *et al.* (1993). Organisms with multiple polyubiquitin loci are named by the number of repeats in the locus (e.g., human IX, chicken IV) to distinguish them.

Assessing Patterns of Variation

To examine whether the phylogenetic pattern of the individual repeats of the published loci suggested any-

TABLE 3

Observed (Obs) and Expected (Expec) Variation among Amino Acid Codon Types in Vertebrate Polyubiquitin Repeats

Codon type	Obs	Expec	No. of codons
Ala	11	9	2
Arg	24	18	4
Asn	10	9	2
Asp	18	22	5
Gln	16	27	6
Glu	24	27	6
Gly	28	27	6
His	5	4	1
Ile	18	31	7
Leu	63	40	9
Lys	26	31	7
Phe	2	9	2
Pro	23	13	3
Ser	25	13	3
Thr	27	31	7
Tyr	1	4	1
Val	15	18	4

Note. Expected number of hits are rounded to the nearest integer. Number of codons of each type within the gene is also shown.

thing about the homogenization process, we analyzed repeats from several loci independently. Since substitutions between repeats within a locus are usually minimal, these runs were done in the absence of other loci to avoid obscuring the homogenization pattern. Topologies were rooted with the 3′-most repeat in order to ascertain whether homogenization happens in a linear fashion (5′–3′ or vice versa) or from the "inside out" (among several possibilities) and whether adjacent repeats generally tended to cluster.

Intralocus variation (i.e., within a given organism) was tallied for each specific amino acid codon as well as by codon position and was counted as either a transition or a transversion. Variation at a specific site that encompassed more than two bases (e.g., B = CGT) was counted as both a transition and a transversion. To test whether the variation was nonrandom with respect to amino acid codon, a χ^2 test was employed. To ascertain the expected number of hits in a particular codon type, the total number of intralocus hits (336) was divided by the total number of amino acids (76). This coefficient (4.42) was then multiplied by the number of codons for each of the 17 amino acid types (excluding the initiator methionine) found in vertebrate ubiquitin (e.g., four arginine residues).

GC content at 3rd-bp positions was tallied for all loci. GC content at 3rd positions in particular is thought to reflect organismal and genome location tendencies (Mita *et al.*, 1991; D'Onofrio *et al.*, 1991). Polymorphisms within a locus (i.e., within different repeats) were scored as being not GC (since it is unclear which base will become fixed during homogenization) unless all were GC (i.e., the ambiguity code "S" = G/C; see Fig. 2).

Phylogenetic Inference

Phylogenetic analysis was performed using the parsimony program PAUP Version 3.1.1 (Swofford, 1993). After initial runs that verified that all repeats in previously untested loci clustered together, consensus sequences were constructed for those loci to facilitate searches with greater numbers of taxa. In these searches positions with multiple states were interpreted as uncertainties, rather than polymorphisms. This is due to the fact that PAUP does not allow ancestral nodes to have multiple states for a given character. Given the multiple repeats and variation seen in all ubiquitin loci to date, this seems an invalid assumption in the case of this gene. Thus the uncertainty interpretation, though also logically flawed, was preferred.

Weighted searches of mammals and all vertebrates were performed based on the observed intralocus variation as tallied above (see Table 2). For example, the mammal weighted search utilized only the mammal variation totals: transversions were weighted three times that of transitions, and 1st codon positions were weighted 15 times that of 3rd codon positions (2nd codon positions were uninformative in both the mammal and the vertebrate searches). Parameters for the vertebrate search were again a 3:1 transversion:transition ratio (transitions were 2.567 as frequent as transversions within loci), and 1st positions were weighted 13 times that of 3rd positions.

RESULTS

Table 1 lists the taxa for which polyubiquitin was analyzed. Since the amplifications often produced a ladder of repeat units, it is possible to get a minimum estimate of the number of ubiquitin repeats present. Table 1 lists the number of repeats observed by amplification plus 1, to account for the additional 3′ copy which must be present for successful amplification (see Fig. 1). No phylogenetic trend as to the number of repeats is apparent, with the possible exception of a tendency toward higher numbers in mammals (e.g., six in mink, nine in one human locus).

Table 1 also lists the number of polymorphisms present in any one organism and/or locus (presumably indicating variation between copies). Care must be taken in comparisons here, as sequence obtained by PCR may come from several loci. The range here is quite high—from 1 in several taxa to 34 in one human locus (45 if all the human loci are considered). GC content at 3rd-base positions is also shown in this table. While there is some variation in GC content—likely partly due to the varying amount of polymorphic sites between taxa—averages for major groups are similar at 65–70%.

Figure 2 shows the aligned vertebrate consensus sequences. Variation within and between loci does not

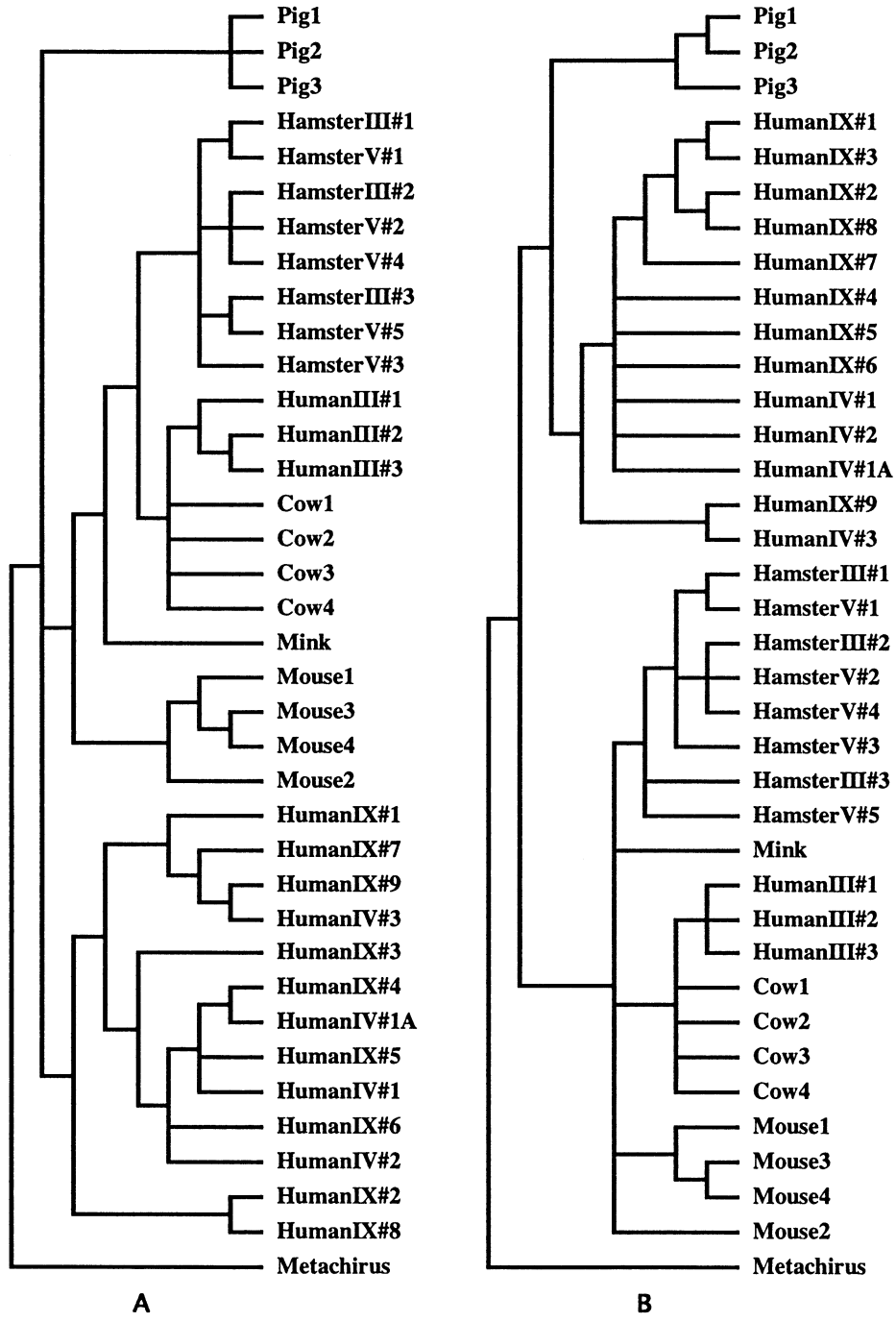

FIG. 3. Trees based on mammal polyubiquitin loci. Tree A is a strict consensus of 160 unweighted trees of length 159. Tree B is a strict consensus of 415 weighted trees of length 395 using the weights given in the text. Repeat humanIV #1A is a partially sequenced repeat 5′ of humanIV#1. There are no further upstream repeats in that locus.

seem to correspond to any particular regions or motifs in the translated product. While the 3′-most seven codons seem quite variable, they do not encompass any one region, but straddle the end of a β sheet and the COOH tail which binds to other proteins (Vijay-Kumar *et al.*, 1985).

The χ^2 test was performed as described under Methods and Materials, with a sum of 52.449 and 15 degrees of freedom. This strongly implies a nonrandom distribution of fixed mutations with regard to codon type, $P < 0.01$. Codons that particularly deviate from the expected number of hits include leucine, isoleucine, and serine (Table 3). Serine and leucine might be expected to have a large number of hits since they both have six

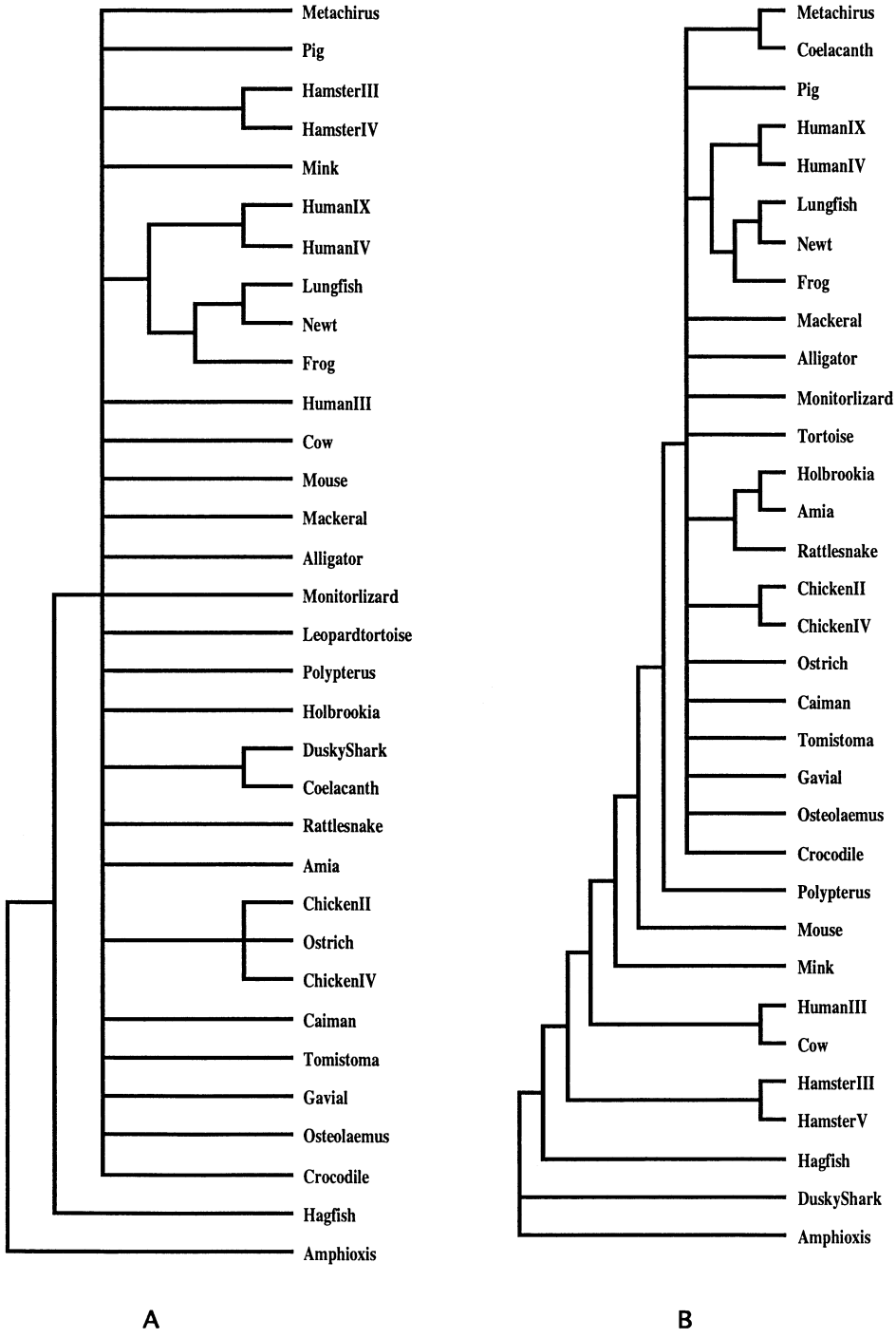

FIG. 4. Strict consensus cladograms of vertebrate ubiquitin data. Tree A was derived from 2369 trees of length 250 run with unweighted characters and transformations, while tree B is based on 69 trees of length 1168 obtained by the weighting scheme presented.

potential codons. Arginine, however, also has six potential codons, and while it shows above the expected number of hits, it has a smaller difference than between the observed and expected hits of proline, which has four potential codons.

Figures 3, 4, and 5 show the results of the phylogenetic analyses. Figure 3 illustrates the unweighted and weighted runs on just mammalian loci, while Fig. 4 shows the same for all the vertebrate loci. Figure 5 shows repeats from various organisms run alone. The human IX and IV loci were run together, since, as noted by Tan *et al.* (1993), there appears to be interchange between them. The two hamster loci were also analyzed together since our analyses indicated a similar interac-

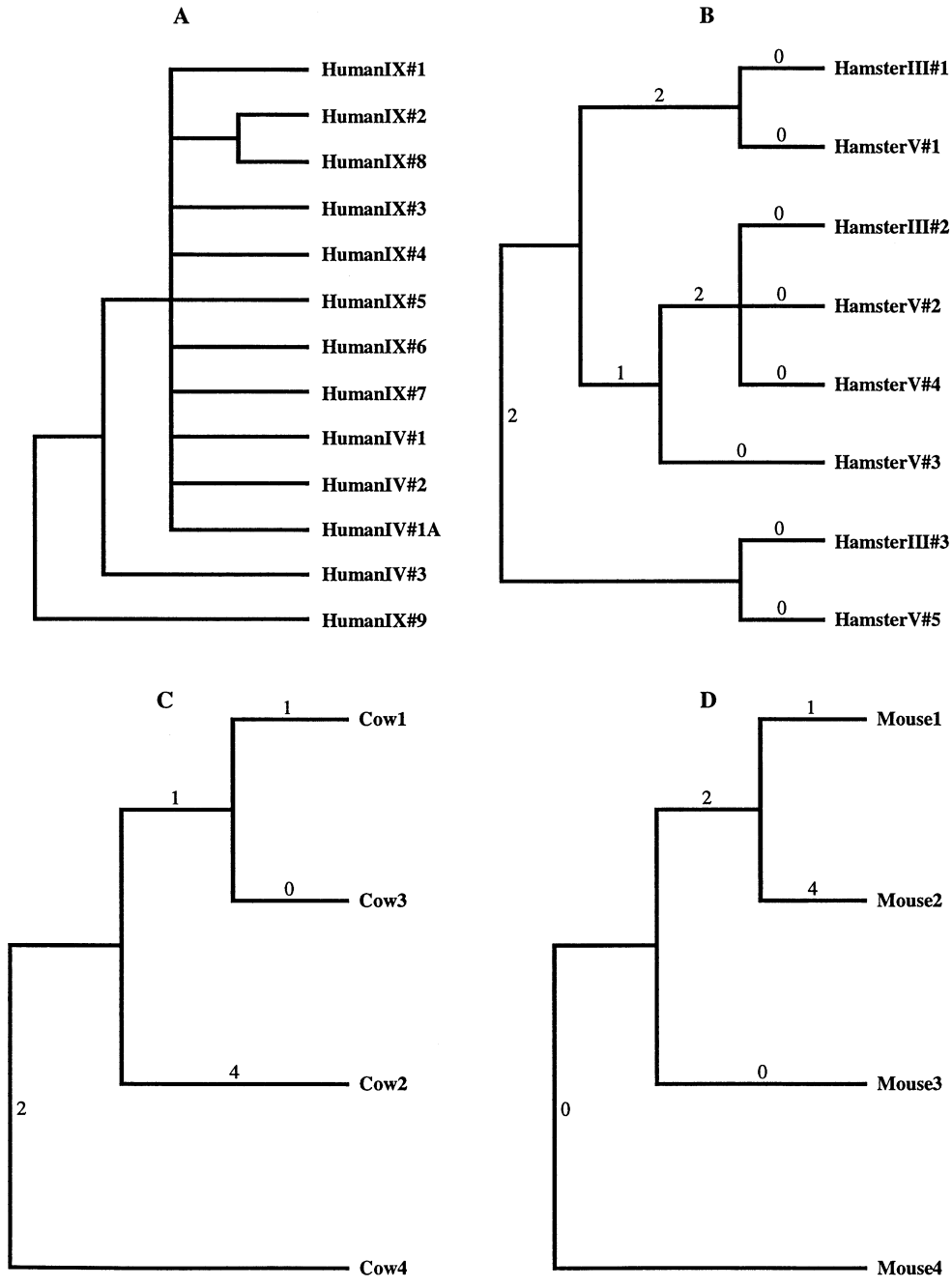

FIG. 5. Trees showing patterns of variation of repeats within mammalian loci. A shows the strict consensus of eight trees of the humanIX and humanIV loci. Tree B shows the most parsimonious tree found for the hamsterIII and hamsterV loci, while C and D show these for the cow and mouse loci, respectively. Branch lengths are shown for all single most parsimonious trees.

tion (Fig. 3). The pig and human III locus have too few repeats to be analyzed alone.

Codon Bias

As noted, the distribution of hits by codon type appears to be nonrandomly distributed, with several types gathering a large number of intralocus substitutions. A number of codon types also showed substantially fewer intralocus hits than expected, which could be suggestive of bias. More striking examples are found among several of the more frequently hit codon types, however.

The last two arginine residues (numbers 72 and 74, Fig. 2) appear to have different higher vertebrate biases: the former uses CGN while the latter shows an AGR preference. However, the three most basal taxa examined (Maisey, 1986)—amphioxis, hagfish, and the shark—have a CGN codon at position 74, also the state

seen in a sea urchin (Nemer *et al.*, 1991), as well as many protostomes (Wheeler *et al.*, 1993). Another example of positional bias may be that encoding leucine residue 50: a CTG variant is in use by almost all taxa (save the lungfish and newt, which share a number of unique changes). In contrast, leucine residues 71 and 73 show much more variation. The lungfish and newt also stand out in having a TCN codon for serine at amino acid position 20, in contrast to all other taxa. At serine 65, amphioxis appears unique among chordates in apparently having both codon types at a single position. The only other eukaryote known with this feature is the sponge *Geodia cydonium*, which also shares these two codon types at serine 65 (Müller *et al.*, 1994). This suggests that this particular position is susceptible to such effects, possibly as a result of simultaneous hits.

A number of codons show no change within vertebrates (Fig. 2), suggesting extreme positional bias. Presumably this is due to selection—possibly reflecting the importance at that point of the relative frequency of the corresponding tRNA. That synonymous substitutions can be deleterious has recently been underscored by several mutant phenotypes resulting from these "neutral" changes (Richard and Beckmann, 1995).

Such codon biases may reflect both GC content of the organismal genome and/or location in the genome (Mita *et al.* 1991; Wolfe *et al.* 1989; Sharp and Matassi, 1994). Unfortunately, this information is not available for most of these taxa.

Vertebrate ubiquitin GC content at 3rd positions is generally high, however, suggesting conservation of genomic location and/or selection for high GC content for other reasons. As Table 1 shows, mammals do not stand out in having particularly high ubiquitin GC levels relative to other taxa compared to so many other loci.

Mode of Concerted Evolution

The patterns among repeats within mammalian loci reveal no absolute consistencies about the mechanism of homogenization (Fig. 5) other than that it is slower than in other loci such as the ribosomal repeats. There may be a general trend of homogenization to roughly follow a 5′–3′ path, but it is unclear what mechanism of concerted evolution (e.g., biased gene conversion) this might suggest. The apparent interchange between the two hamster and human loci and the different number of repeats in each is suggestive of unequal crossing over. Interestingly, in the two hamster loci a general 5′–3′ pattern is also seen. Muller *et al.* (1994) suggest that the two most 5′ repeats are the ancestral duplication, with all others deriving from them. These authors offer no particular evidence for this hypothesis; however, the pattern of relationships among several of the mammalian loci may be consistent with this scenario. The hamster and mouse loci could both be rooted with the 5′-most repeats at the base (Fig. 5).

The high GC content of the gene suggests that polymorphic sites are homogenized toward either of those two bases. This contrasts with the fact that the majority of these sites represent transitions. This implies that while initial changes in the locus are transitions, selection favors fixing G/C residues. Clearly there are counterexamples to this trend: valine 26 shows almost entirely transversional polymorphisms, and these are largely G/C (=S). This is in contrast even with other valine codons (number 17 in particular) in both amount of polymorphisms and base composition (Fig. 2). Other codons (e.g., Gln, Ile) show similar variation in fixation.

Evidence for Locus Duplications in Mammals

Mammals appear to be the only metazoan group to date where some members have multiple polyubiquitin loci (Tan *et al.*, 1993). Unfortunately, our methods could not determine number of loci in the organisms examined to further explore this phenomenon. Moreover, amount of variation within an organism may not be tightly correlated with number of loci, perhaps due to different rates of homogenization (Table 1). For example, the chicken II locus exhibits more variation within those two repeats than among both hamster loci encompassing seven repeats. Several taxa, including the actinopyterygian fish *Polypterus* and *Scomber*, exhibit high enough levels of variation to suggest multiple loci. To date, however, the evidence suggests that several independent locus duplications have occurred within mammals. Particularly interesting is the human III locus, which unlike other such multiple loci (chicken, hamster, and the other two humans), does not group with other loci from the same organism (Fig. 3). Apparently this locus escapes the inter (but not intra-)-locus homogenization process. Clearly, it would be of interest to look for other such locus duplications in mammals and ascertain their phylogenetic distribution and whether they also evolve independently. Such information not only might yield clues to the evolution of ubiquitin, but also could be a phylogenetic character linking mammal groups, whose interrelationships have proven so difficult to elucidate (e.g., Novacek, 1992).

Mapping the genomic location of these loci will also prove essential in deciphering the mechanisms of concerted evolution. Not only might this influence rates and ability of loci to homogenize but genomic location (in mammals in particular also) seems to play a key role in kind and types of substitutions (Sharp and Matassi, 1994). For example, silent substitutions in mammalian genes usually reflect the GC content of the region they are found in, and neighboring genes have similar profiles (e.g., IGF2 and insulin; Ellsworth *et al.*, 1994).

Phylogenetic Signal

Unweighted ubiquitin data appeared to have little phylogenetic signal. While our weighting scheme did

increase resolution of strict consensus of the most parsimonious trees, it did little to increase congruence with well-founded notions of vertebrate and mammal phylogeny. For example, tetrapods, amniotes, lepidosaurs, mammals, crocodilians, and actinopterygian (ray-finned) fish did not form monophyletic groups. Indeed, the weighted tree did not preserve the grouping of three bird loci seen in the unweighted search. Within mammals, neither the two artiodactyls (pig, cow) nor the two rodents grouped together in either the weighted or unweighted searches. If these patterns are robust, then they would almost certainly seem to be "gene trees" rather than "species trees." Perhaps some of the taxa here do have other polyubiquitin loci which do not amplify as readily. Alternatively, perhaps the effects of different rates and levels of homogenization render the phylogenetic signal in polyubiquitin minimal. Wheeler *et al.* (1993) found that combining ubiquitin data with 18S data for a number of invertebrate taxa yielded trees more congruent with morphological data than either gene alone produced. While we advocate a similar approach in vertebrates, neither the 18S data of Hedges *et al.* (1990) nor the 28S data of Le *et al.* (1993) overlap with these taxa enough to merit such a combination. Given the growth of the sequence database, this situation should be remedied in the near future.

While polyubiquitin's usefulness as a sole indicator of phylogenetic history is clearly poor, its molecular evolutionary patterns make investigation in other taxa interesting. Moreover, its combination with other molecular data sets may aid in constructing much more robust phylogenies than either ubiquitin or other genes are capable of alone.

ACKNOWLEDGMENTS

We thank Rob DeSalle, Aloysisus Phillips, and Mike Whiting for discussion. John Gatesy is thanked for unpublished data. Frances Lee is thanked for editorial comments. The manuscript was also improved by the comments of three anonymous reviewers.

REFERENCES

Allard, M. W., Ellsworth, D. L., and Honeycutt, R. L. (1991). The production of single-stranded DNA suitable for sequencing using the polymerase chain reaction. *BioTechniques* **10:** 24–26.

D'Onofrio, G., Mouchiroud, D., Aïssani, B., Gautier, C., and Bernardi, G. (1991). Correlations between the compositional properties of human genes, codon usage, and amino acid compositions of proteins. *J. Mol. Evol.* **32:** 504–510.

Ellsworth, D. L., Hewett-Emmett, D., and Li, W. H. (1994). Evolution of base composition in the insulin and insulin-like growth factors genes. *Mol. Biol. Evol.* **11:** 875–885.

Finley, D., and Varshavsky, A. (1985). The ubiquitin system: Functions and mechanisms. *Trends Biochem. Sci.* **10:** 343–347.

Hedges, S. B., Moberg, K. D., and Maxson, L. R. (1990). Tetrapod phylogeny inferred from 18S and 28S ribosomal RNA sequences and a review of the evidence for amniote relationships. *Mol. Biol. Evol.* **7:** 607–633.

Le, H. L. V., Lecointre, G., and Perasso, R. (1993). A 28S rRNA-based phylogeny of the Gnathostomes: First steps in the analysis of conflict and congruence with morphologically based cladograms. *Mol. Phylogenet. Evol.* **2:** 31–51.

Maisey, J. G. (1986). Heads and tails: A chordate phylogeny. *Cladistics* **2:** 201–256.

Mita, K., Ichimura, S., and Nenoi, M. (1991). Essential factors determining codon usage in ubiquitin genes. *J. Mol. Evol.* **33:** 216–225.

Müller, W. E. G., Schröder, H. C., Müller, I. M., and Gamulin, V. (1994). Phylogenetic relationship of ubiquitin repeats in the polyubiquitin gene from the marine sponge *Geodia cydonium*. *J. Mol. Evol.* **39:** 369–377.

Nemer, M., Rondinelli, E., Infante, D., and Infante, A. A. (1991). Polyubiquitin RNA characteristics and conditional induction in sea urchin embryos. *Dev. Biol.* **145:** 255–265.

Novacek, M. J. (1992). Mammalian phylogeny: Shaking the tree. *Nature* **356:** 121–125.

Özkaynak, E., Finley, D., Soloman, M. J., and Varshavsky, A. (1987). The yeast ubiquitin genes: A family of natural gene fusions. *EMBO J.* **6:** 1429–1439.

Rechsteiner, M. (1991). Natural substrates of the ubiquitin proteolytic pathway. *Cell* **66:** 615–618.

Richard, I., and Beckman, J. S. (1995). How neutral are synonymous codon mutations? *Nature Genet.* **10:** 259.

Sharp, P. M., and Li, W-H. (1987a). Molecular evolution of ubiquitin genes. *Trends Ecol. Evol.* **2:** 328–332.

Sharp, P. M., and Li, W-H. (1987b). Ubiquitin genes as a paradigm of concerted evolution of tandem repeats. *J. Mol. Evol.* **25:** 58–64.

Sharp, P. M., and Matassi, G. (1994). Codon usage and genome evolution. *Curr. Opin. Genet. Dev.* **4:** 851–860.

Swofford, D. L. (1993). "PAUP Users Manual Version 3.1.1," Illinois Natural History Survey, Champaign, IL.

Tan, Y., Bishoff, S. T., and Riley, M. A. (1993). Ubiquitins revisited: Further examples of within- and between-locus concerted evolution. *Mol. Phylogenet. Evol.* **2:** 351–360.

Vijay-Kumar, S., Bugg, C. E., Wilkinson, K. D., and Cook, W. J. (1985). Three-dimensionsl structure of ubiquitin at 2,8 resolution. *Proc. Natl. Acad. Sci. USA* **82:** 3582–3585.

Vrana, P. B., Milinkovitch, M. C., Powell, J. R., and Wheeler, W. C. (1994). Higher level relationships of the arctoid carnivora based on sequence data and total evidence. *Mol. Phylogenet. Evol.* **3:** 47–58.

Wheeler, W. C., Cartwright, P., and Hayashi, C. Y. (1993). Arthropod phylogeny: A combined approach. *Cladistics* **9:** 1–39.

Wolfe, K. (1991). Mutation rates differ among regions of the mammalian genome. *Nature* **337:** 283–285.

Wray, C. G., and DeSalle, R. (1994). Phylogenetic utility of ubiquitin DNA sequence from 3 marine protist lineages. *Mol. Marine Biol. Biotech.* **3:** 13–22.