Heredity **78** (1997) 21–31 *Received 2 January 1996*

Substitution rate variation in closely related rodent species

DAN FIELDHOUSE, FARIBORZ YAZDANI & G. BRIAN GOLDING*

Department of Biology, McMaster University, 1280 Main St. West, Hamilton, Ontario, Canada L8S 4K1

The existence of evolutionary rate variation has previously been demonstrated between different orders, different species and even between different regions of the same gene. To examine rate variation between closely related species of rodents we have sequenced the adenine phosphorybosyltransferase (*APRT*) gene from *Mus spicilegus*, *Mus pahari*, *Mastomys hildebrandtii*, *Stochomys longicaudatus* and *Gerbillus campestris* and compared these sequences with the previously published *Mus musculus*, *Rattus norvegicus* and *Mesocricetus auratus APRT* sequences. The alignment of these eight rodent *APRT* sequences reveals two large insertions within the introns: an insertion with sequence similar to a B1 repetitive element is found within *Mastomys* and an insertion with sequence similar to a B2 repetitive element is found within *M. pahari*. A phylogeny for the rodent *APRT*s agrees with the previously published rodent phylogeny based on other molecular and morphological data. The relative rate test which is often used to test for variation in rates of evolution in different lineages is shown here to be sensitive to the choice of outgroup and therefore should be used with great caution. This sensitivity is detectable only with closely related species and results from the prevalence of homoplastic substitutions. Rate variation is demonstrated within the *APRT* exons and introns and between the rodent species (with the most significant difference being a rate difference in *M. spicilegus*). In addition, some third codon positions are shown to be more prone to substitution than others. This clearly demonstrates that even between very closely related species there is ample evidence of major differences in rates of evolution among species, among regions of the gene and among different positions within the gene. We also demonstrate that standard methods of analysis might not detect this variation.

Keywords: *APRT*, B1/B2 elements, rate variation, relative rate test, rodent, phylogeny.

Introduction

Rates of molecular evolution have been shown to vary between different orders, between species, and between regions of genomes and genes (Golding, 1983; Wu & Li, 1985; Wolfe *et al.*, 1989; Turker *et al.*, 1993). Perhaps the best known example is the two to 10 times faster rate of evolution found in the order Rodentia in comparison to primates (Wu & Li, 1985). This rate variation is subject to debate because of the comparison being between orders. Rate variation has also been found between species within orders (e.g. between rodent species; Allard & Honeycutt, 1991). Rates can vary even within different positions within genes as indicated by the presence of hotspots for mutation as shown by Benzer's (1961) demonstration in T4 phage. Causes of rate variation include generation time, selection, population size, DNA repair mechanisms, chromosomal location of the gene, base composition/base context, metabolic rate and life history/environmental influences. Rate variation can complicate phylogenetic analysis both in determining tree topology and in estimates of branch lengths.

Sequences from closely related species are required to avoid ambiguities in the alignment of the sequences and to reduce the number of potential parallel, convergent and reverse substitutions. The order Rodentia provides many closely related species which inhabit a wide variety of habitats and experience many different environmental conditions and life styles (Nowak, 1991). The phylogenies of most rodent species are not well known because morphological characters often do not differ greatly

*Correspondence. E-mail: golding@mcmaster.ca

between species and little additional data have been available. In general, the only variable morphological characters are dental and cranial characters, soft body parts, size and colour (Luckett & Hartenberger, 1985; Marshall, 1986). Molecular characters can provide useful phylogenetic information for these species.

The adenine phosphorybosyltransferase (*APRT*) gene is a purine biosynthesis salvage pathway enzyme that catalyses the direct synthesis of adenosine-5′-monophosphate (AMP) from adenine and 5-phosphoribosyl-1-pyrophosphate (Taylor *et al.*, 1985). The coding sequence is highly conserved in all known organisms. For example, Broderick *et al.* (1987) have found 40 per cent amino acid identity between human and *E. coli* sequences. The organization of the gene is also conserved, with all known mammalian *APRT* sequences having five exons and four introns. The two kilobase sequence of the *APRT* gene therefore provides both coding regions under strong selection and intron regions under limited selection. The *APRT* gene has also been used extensively in spontaneous and induced mutagenesis studies (Meuth, 1990; Skandalis & Glickman, 1990).

It would be useful to have information on the rates and associated patterns of evolution in the *APRT* gene of rodent species which represent a long-term evolutionary scenario in order to compare these with short-term mutagenesis studies. For these reasons we have sequenced the *APRT* gene of five closely related rodent species and analysed these sequences along with three other rodent *APRT*s available from GenBank. These species range in divergence times from an estimated 1–3 MYA between the two closest species (*M. spicilegus* and *M. musculus*) up to 16–46 MYA between the most divergent species (*M. spicilegus* and *Mesocricetus auratus*) (She *et al.*, 1990; O'hUigin & Li, 1992; Hammer & Silver, 1993). This provides sequences having from 1 to 12 per cent nucleotide sequence divergence in the *APRT* coding region between the closest and most divergent species, respectively, and up to 26 per cent sequence divergence in the intronic regions.

Materials and methods

PCR amplification and sequencing

Genomic DNA for *M. spicilegus*, *M. pahari*, *Mastomys hildebrandtii* and *Stochomys longicaudatus* were obtained from Dr Priscilla Tucker. Genomic DNA for *Gerbillus campestris* was extracted from liver using the same method as in Fieldhouse & Golding (1993).

The *APRT* gene was amplified using primers flanking the gene. Each 50 μL PCR reaction contained 2.2 mM $MgCl_2$, 48 mM KCl, 12 mM Tris-HCl (pH 8.8), 0.2 μM of each primer, 400 μM of each dNTP, 0.5–1 μg of template and 3 units of Cetus *Amplitaq*. There were 30 cycles of 45 s at 94°C, 1 min at annealing temperature (between 46 and 55°C depending on the template) and 2 min at 72°C. Bands containing the *APRT* gene were cut from a 1.2 per cent LMP agarose gel and phenol/chloroform extracted (Sambrook *et al.*, 1989), ethanol precipitated and re-amplified using nested primers. The primers used are located every approximately 350 base pairs apart in both directions to allow the complete sequencing of the gene in both directions. Nested PCR products were extracted from a LMP agarose gel as above, and cycle sequenced using a Perkin/Elmer Cetus kit, [γ-^{32}P or γ-^{33}P]dATP label, and one of the terminal primers used in the re-amplification. The sequencing reaction conditions were 1 min at 94°C and 1 min at 60°C for 20 cycles, or 1 min at 94°C, 1 min at annealing temperature (between 50 and 55°C depending on the primer) and 1 min at 72°C.

Sequence alignment and phylogenetic analysis

Sequences of *Mus musculus*, *Rattus norvegicus* and *Mesocricetus auratus* were aligned with the *Mus spicilegus*, *Mus pahari*, *Mastomys hildebrandtii*, *Stochomys longicaudatus* and *Gerbillus campestris* *APRT* sequences using ClustalV (Higgins & Sharp, 1989) with final alignment by hand. The exon positions were deduced from the sequences using the *M. musculus* exon organization. BLAST searches of the GenBank database were used to identify any large insertions present.

The phylogenetic relationships were determined using maximum likelihood (DNAML), neighbour-joining (NEIGHBOR) and maximum parsimony (DNAPARS) algorithms from PHYLIP (Felsenstein, 1989). Different subsets of the sequence data (combinations of exons and introns) were used to address potential regional differences. Consensus trees were determined from 500 bootstraps using maximum likelihood and neighbour-joining algorithms. Also, likelihood values for each of 15 possible trees were calculated using DNAML. The optimal tree was chosen from these analyses to be the tree supported by the majority of methods and datasets without strong contradictory evidence. This tree was used for all subsequent analyses.

Substitution rate variation analysis

To determine whether a molecular clock exists between the eight species, relative rate tests were performed (Wu & Li, 1985). In addition, likelihood values were compared which assumed (DNAMLK) or did not assume (DNAML; PHYLIP, Felsenstein, 1989) a constant rate. A likelihood ratio test was used to determine a statistical difference. To determine if individual species contributed to any non-clocklike behaviour the analysis was repeated after dropping each species individually. The likelihood values were also compared separately for each exon and intron to detect differences between regions.

To test for 'hotspot' positions in the *APRT* sequence, the observed number of substitutions per position across the six rodent species was compared with the expected number based on a Poisson distribution as well as to the expected number obtained from permuting the observed substitutions along the sequence of each species separately. The identity and location of the observed substitutions within the six most closely related rodent species were inferred using Fitch's (1971) parsimony algorithm. *Mesocricetus* and *Gerbillus* were used as outgroups to infer the ancestral sequence to these six species. Poisson expected values were calculated based on a mean calculated from the observed number of positions with zero substitutions per position.

For the permutation study Fitch's (1971) parsimony algorithm was used to infer all nodal ancestral sequences and the location and identification of the substitutions along all branches. For each branch separately, the observed number of each type of substitution for each branch was randomly placed along the sequence at positions suitable for that particular substitution. For example the five A to G substitutions observed for the branch leading to *Rattus* would be permutated by randomly choosing five of the 83 positions having an A in the ancestral *APRT* sequence and changing them to a G in the *Rattus* lineage. Likewise the three C to T observed *Rattus* lineage substitutions would be permutated by randomly choosing three of the 129 positions having a C in the ancestral sequence to be changed to a T. Each branch was permutated separately and in evolutionary sequence so that any sequence changes made along a nodal branch could affect the placement of substitutions that occur on branches that lead from this branch. For example, a C to G substitution in an ancestral nodal lineage would allow the subsequent placement of a G to A, G to C or G to T substitution at that position in subsequent lineages rather than a C to A, C to G or C to T substitution to which the position was previously restricted.

After each permutation the number of substitutions at each sequence position was tabulated. The total number of positions having 0, 1 or 2 and greater substitutions was calculated for all 10 000 permutations. Two datasets were analysed separately: the small introns and the third position of the exons. This allows the comparison of two sets of positions both under relatively little selection but with the exon third positions potentially having different neighbour base effects than the introns.

Results

The *APRT* sequences for *Mus spicilegus*, *Mus pahari*, *Mastomys hildebrandtii*, *Stochomys longicaudatus* and *Gerbillus campestris* have been submitted to the NCBI-GenBank database and have been assigned accession numbers U28720, U28721, U28722, U28723 and U28961, respectively. The complete *APRT* gene was sequenced including the introns with the exception of *Mastomys* for which most of the large intron sequence is not available. All regions of the *APRT* sequences of these five species as well as *Mus musculus*, *Rattus norvegicus* and *Mesocricetus auratus* (accession numbers M11310, L04970 and X03603, respectively) were aligned unambiguously because of their close phylogenetic relationship. One exception is the large intron which has some regions with some minor alignment ambiguities because of the greater number and sizes of indels, and is missing most of the *Mastomys* sequence. For this reason some analyses have excluded the large intron.

There are five exons of lengths 80, 107, 134, 79 and 140 bases as well as four introns of lengths 130–142, 949–1146, 186–372 and 103–118 bases in the rodent *APRT* sequences. Variation in the intron lengths is the result of small insertions and deletions (indels) and two large insertions. There is a 194 base insertion in intron III of *Mastomys hildebrandtii* and a 175 base insertion in intron II of *M. pahari*.

A BLAST search of the GenBank database shows that the *Mastomys hildebrandtii* 194 base insertion has a high degree of similarity to B1 repetitive elements found within many mouse genes. An alignment with a consensus sequence of B1 repetitive elements (Bains & Temple-Smith, 1989) shows 94.2 per cent similarity (eight mismatches in a 138 base overlap) with a one base insertion and a 12 base deletion in this overlapping region. There are 23 bases of the *Mastomys* insertion that do not show

similarity with the consensus sequence and 11 bases of the consensus sequence that do not show similarity to the *Mastomys* sequence. However, B1 repetitive elements often differ in length between copies. The B1 element is flanked by a 14 base imperfect repeat (one mismatch) which appears to have been produced by copying a 14 base section of the *APRT* sequence. This short direct flanking repeat suggests that this inserted element is an integrated reverse transcript (Rogers, 1985). Sequences similar to those of putative intragenic control signals for RNA polymerase III transcription (Krayev *et al.*, 1982) are also present in the *Mastomys* insertion.

The *M. pahari* 175 base insertion has a high degree of similarity to mouse and rat B2 repetitive elements. An alignment with a consensus sequence of B2 repetitive elements (Bains & Temple-Smith, 1989) shows 94.4 per cent similarity (nine mismatches in a 160 base overlap) with a one base deletion, a three base deletion and a 38 base deletion in this overlapping region. There are 15 bases of the B2 consensus sequence that do not align with the *M. pahari* insertion. The B2 element is flanked by a 15 base perfect repeat which appears to have been produced by copying a 15 base section of the *APRT* sequence. Sequences similar to those of putative intragenic control signals for RNA polymerase III transcription (Krayev *et al.*, 1982) are also present in the *M. pahari* insertion, with only a single mismatch found in the 23 bases.

BLAST searches of the rodent *APRT* introns show no other significant similarity to any sequences in the database indicating that no other known repetitive elements are present. From the alignment of the eight rodent *APRT* sequences the exons can be irrefutably identified because of the high amount of sequence similarity. A probable sequence error in the database entry that causes a frame shift in exon I of the *Mesocricetus auratus* APRT sequence was adjusted using the consensus base for all other rodents.

Rodent APRT phylogeny

Using *Mesocricetus* and *Gerbillus* as outgroups, the *APRT* sequence data for the eight rodent species supports two alternative phylogenetic trees designated here as trees A and B (see Figs 1a and b, respectively) which differ only in the placement of *Mastomys* and *Stochomys* between the *Mus* group and the *Rattus/Gerbillus/Mesocricetus* groups. Tree A places *Mastomys* closer to the three *Mus* spp. whereas tree B places *Stochomys* closer to the three *Mus* spp. Different subsets of the sequence data

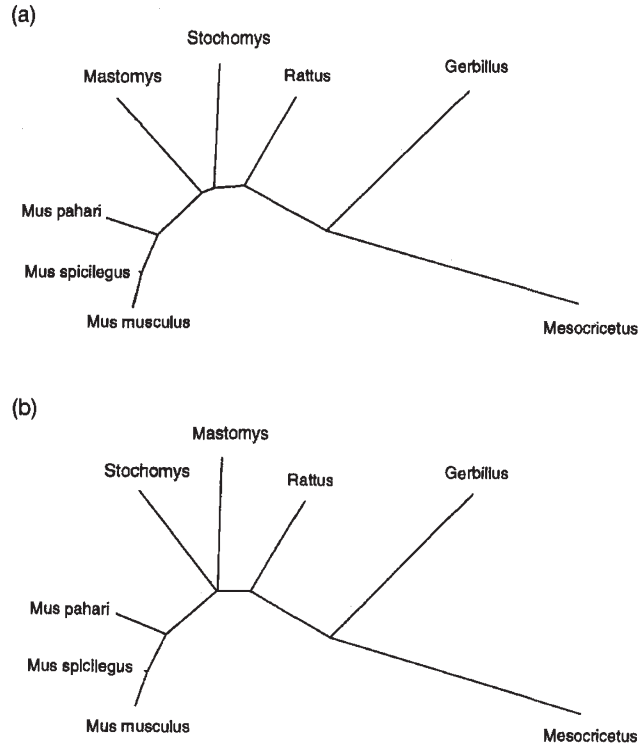

Fig. 1 The two most supported rodent *APRT* phylogenetic trees which differ in the placement of *Mastomys* and *Stochomys*. (a) Tree A; (b) tree B.

(combinations of the exons and introns) support these two topologies. The datasets used are listed in Table 1: dataset no. 1 contains all five exons and four introns; dataset no. 2 contains all five exons and the three small introns; dataset no. 3 contains the coding sequence alone with dataset no. 3A being the nucleotide sequence and dataset no. 3B being the translated amino acid sequence; dataset no. 4 contains all four introns; and dataset no. 5 contains the three small introns only.

Table 1 summarizes the phylogenetic analyses. Even though the trees are not statistically different, consistent support is found for tree A over the alternative trees. Tree A is the most supported by the majority of algorithms and datasets (four of six datasets using maximum likelihood, three of six datasets using neighbour-joining, two of six datasets using maximum parsimony although multiple trees were supported) with none of the other possible trees being consistent or strongly supported by any dataset or phylogenetic algorithm. The coding sequence does not support tree A but may be too similar (12 per cent sequence divergence between the two most divergent species) and under strong

Table 1 Summary of the phylogenetic analyses using maximum likelihood, neighbour-joining and maximum parsimony algorithms. Five different subsets of the data (combinations of exons and introns) are used to address potential regional differences

DATASET	ML	NJ	PARS
Set no. 1 Whole sequence	A	C	E
Set no. 2 Whole sequence minus large intron	A	A	B
Set no. 3A Exons	B	B	B
Set no. 3B Exons translated	B	D	A, B, D
Set no. 4 Introns	A	A	E
Set no. 5 Small introns	A	A	A, B, F

ML, maximum likelihood algorithm; NJ, Neighbour-joining algorithm; PARS, maximum parsimony algorithm.

A: ((((((Mm,Ms),Mp),Mh),Sl),Rn),Gc,Ma) (see Fig. 1a).
B: ((((((Mm,Ms),Mp),Sl),Mh),Rn),Gc,Ma) (see Fig. 1b).
C: ((((((Mm,Ms),Mp),Sl),Rn),Mh),Gc,Ma).
D: (((((Mm,Ms),Mp),(Mh,Sl)),Rn),Gc,Ma).
E: (((((Mm,Ms),Mp),Sl),Rn),(Gc,Mh),Ma).
F: ((((((Mm,Ms),Mp),Mh),Rn),Sl),Gc,Ma).

Mm, *Mus musculus*; Ms, *Mus spicilegus*; Mp, *Mus pahari*; Mh, *Mastomys hildebrandtii*; Sl, *Stochomys longicaudatus*; Rn, *Rattus norvegicus*; Gc, *Gerbillus campestris*; Ma, *Mesocricetus auratus*.

selection and hence may not distinguish the rodent phylogenetic relationships. The maximum parsimony algorithm does not support any single tree as strongly as the maximum likelihood and neighbour-joining algorithms support tree A. The alternative trees all differ in the placement of *Mastomys* which suggests that the *Mastomys* sequence is phylogenetically ambiguous. Additional sequence data could help to resolve the placement of *Mastomys* within the rodent phylogeny.

To determine how robust the optimal tree is, bootstrap analyses were performed using the maximum likelihood and neighbour-joining algorithms on the whole sequence minus the large intron (dataset no. 2). One hundred per cent of the bootstraps using both phylogenetic algorithms placed *Mus musculus*, *M. spicilegus* and *M. pahari* together as well as placing *Mesocricetus* and *Gerbillus* together. *Rattus* was placed with *Mesocricetus* and *Gerbillus* in greater than 90 per cent of the bootstraps using both phylogenetic algorithms. The placing of *Mastomys* and *Stochomys* remains ambiguous as bootstrap support is not strong for either tree A or tree B. *Mastomys* is placed with the *Mus* group (tree A) in 84 per cent and 66 per cent of the maximum likelihood and neighbour-joining bootstraps respectively. Thus even though tree A is best supported, it is not statistically significant.

As the *Mus* spp. group together 100 per cent of the time and *Gerbillus* and *Mesocricetus* group together 100 per cent of the time by all methods, there is no need further to test the significance of their phylogenetic relationship to each other. However, the *Mastomys/Stochomys/Rattus* relationships are uncertain and require additional analysis. These groups can be arranged into 15 different phylogenetic topologies. The three *Mus* spp. and the two outgroup species *Gerbillus* and *Mesocricetus* are included in the analysis, but their relative positions to each other are maintained within all 15 topologies. The maximum likelihood values obtained from DNAML are compared between these 15 different topologies to determine the support for each topology. Even though six trees are significantly worse than tree A ($P<0.05$), these six trees have standard deviations of less than 2.3 which is only marginally significant. Therefore no strong inference can be made pertaining to the phylogenetic relationships between *Mastomys*, *Stochomys* and *Rattus*.

Inaccuracy of relative rate test

The relative rate test shows rate variation between *M. musculus* and *M. spicilegus*, but only when *M. pahari* is used as the outgroup and not when any of the other more divergent species is used as the outgroup. We show here that this finding reflects the sensitivity of the relative rate test to parallel, convergent and reverse mutations that occur between the outgroup and the species being compared.

The problems caused by these types of substitutions and the effect they have on the relative rate results are illustrated in Table 2 using the small introns (dataset no. 5). Using *M. pahari* which is the most closely related outgroup to *M. spicilegus* and *M. musculus*, there are 20 nucleotide differences between *M. spicilegus* and *M. pahari* in contrast to 28 differences between *M. musculus* and *M. pahari* (Table 2). The *M. pahari* sequence is identical to the inferred ancestral sequence (inferred using Fitch's algorithm; 1971) at the positions that differ between *M. spicilegus* and *M. musculus*. The relative rate test

Table 2 The number of substitutions for the small introns in the *Mus musculus* and *M. spicilegus* lineages inferred using each outgroup species

Species	Mp	Mh	Sl	Rn	Gc	Ma
Mm	8	6	6	6	5	8
	(28)	(47)	(54)	(52)	(84)	(98)
Ms	0	3	2	2	4	3
	(20)	(44)	(50)	(48)	(83)	(93)
N	419	405	413	421	404	385

The species abbreviations are the same as in Table 1.
N: the number of nucleotide positions common to *M. musculus*, *M. spicilegus* and each outgroup, respectively.
The numbers in parentheses are the total numbers of nucleotide differences between each outgroup species and *M. musculus* and *M. spicilegus*, respectively.

Table 3 Relative rate test for the small introns of *Mus musculus* and *M. spicilegus* using the other six rodent species as outgroups

Outgroup used	*N*	K_{AB}	K_{AC}	K_{BC}	$K_{AC}-K_{BC}$	SND
Mp	421	0.01939	0.07022	0.04937	0.02085	2.8
					(0.00742)	($P<0.006$)
Mh	412	0.02007	0.12651	0.11779	0.00872	1.0
					(0.00873)	($P<0.32$)
Sl	417	0.01968	0.14491	0.13312	0.01179	1.2
					(0.00983)	($P<0.24$)
Rn	423	0.01930	0.13607	0.12465	0.01142	1.4
					(0.00813)	($P<0.17$)
Gc	411	0.02012	0.24494	0.24182	0.00313	0.3
					(0.01073)	($P<0.77$)
Ma	399	0.02113	0.31553	0.29535	0.02018	1.4
					(0.01394)	($P<0.17$)

The species abbreviations are as in Table 1.
N: number of sites compared (the average of the three pairwise *N*s).
K: Kimura's two-parameter model number of substitutions between species, with A being *M. musculus*, B being *M. spicilegus* and C being the outgroup species listed.
The numbers in parentheses in the $K_{AC}-K_{BC}$ column are the standard errors.
SND is calculated as $(K_{AC}-K_{BC})$/(standard error); the probabilities assume a normal distribution.

therefore reflects the rate variation found between the two *Mus* spp. and shows statistical significance ($P<0.006$; Table 3).

In contrast none of the remaining outgroups shows a significant difference using the relative rate test (Table 3). These results arise because of parallel, convergent and reverse substitutions that obscure the differences in the rates of evolution between these two *Mus* spp. For example, parallel substitutions in *M. musculus* and *Mastomys* incorrectly decrease the number of differences between these two species to six, and increase the number of differences between *M. spicilegus* and *Mastomys* to three because the outgroup *Mastomys* does not reflect the true *Mus* ancestral base for that position. Reverse substitutions in *M. musculus* likewise result in the outgroup being similar to *M. musculus* rather than the ancestral sequence. Substitutions along the outgroup lineage can also result in the outgroup sequence being different from either *M. spicilegus* or *M. musculus*, and therefore leads to the inflated number of differences between the outgroup and *M.*

spicilegus. These types of substitutions therefore obscure the rate estimates by incorrectly inferring the ancestral sequence using the outgroup species. Because these homoplastic and confounding substitutions can obviously be common, this suggests the need for studies of closely related species to determine relative rates of substitution.

The relative rate test is therefore very sensitive to the outgroup chosen and therefore not a reliable method to demonstrate rate variation. The problem clearly results from the presence of the homoplastic substitutions which can only be detected by using closely related species.

Variable rates of substitution

To test for a molecular clock the likelihoods from DNAML and DNAMLK were compared using a maximum likelihood ratio test for the whole sequence (dataset no. 1; $P<0.05$), the whole sequence minus the large intron (dataset no. 2; $P<0.01$), the exons (dataset no. 3; $P<0.01$), the introns (dataset no. 4; $P<0.05$), the small introns (dataset no. 5; $P<0.05$), and the large intron (dataset no. 6; $P<0.01$). These likelihood ratio tests indicate that a molecular clock does not hold for any one of the datasets at a 0.05 level of significance.

Figure 2 compares the DNAML/DNAMLK (upper/lower) trees for the whole sequence minus the large intron (dataset no. 2). It can be seen that *M. spicilegus* is slower and *M. musculus* is faster than the expectation assuming a molecular clock. Hence either *M. spicilegus* or *M. musculus* deviates from the molecular clock expectations but not necessarily both of them. In addition to the large deviation shown by these two species, an analysis of the other data sets shows some indication of a faster rate in *Mastomys* and a slower rate in *Rattus* (data not shown).

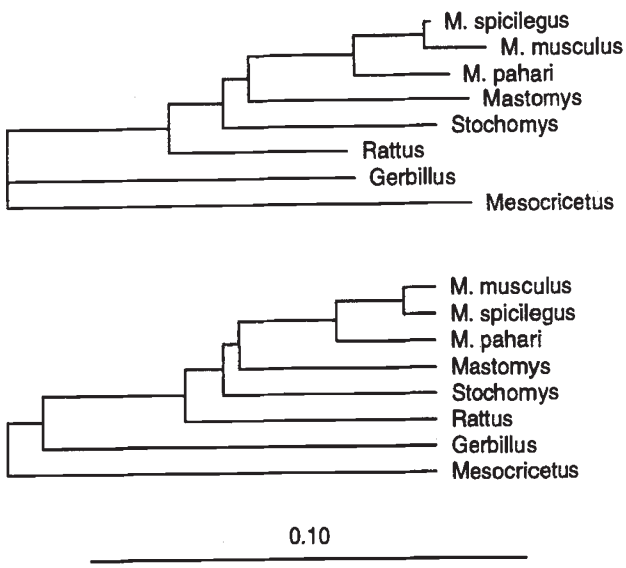

Fig. 2 DNAML (upper) and DNAMLK (lower) phylogenetic trees of the eight rodent *APRT*s for the whole sequence minus the large intron. The two trees are drawn with a common scale. The DNAML tree is unrooted and the DNAMLK tree is rooted as required by the respective algorithms.

The analysis was repeated by sequentially removing individual species to determine which species deviate the most from a molecular clock. Only the small intron sequence (dataset no. 5) was used as it was considered to be the least affected by selection and avoids the ambiguous alignment and missing data problems of the large intron. The absence of *Mastomys*, *Stochomys* and *Rattus* does not alter the non-clocklike behaviour in the phylogenies. However, all three *Mus* spp. must be present to show the non-clocklike behaviour which indicates that at least one of the three *Mus* spp. has a difference in rate, and that all three *Mus* spp. must be present to detect this rate variation. The presence of the other two *Mus* spp. appears to be necessary to detect a rate difference in *M. spicilegus*. When *M. spicilegus* is dropped, the *M. musculus* and *M. pahari* rates are approximately equal.

The DNAML vs. DNAMLK analysis using each exon and intron separately (data not shown) shows that exon I and introns I and II significantly deviate from a molecular clock ($P<0.01$). Exon V and intron III also show non-clocklike behaviour ($P<0.05$). This indicates that these five exons and introns may be the major cause of the non-clocklike evolution across all eight rodent species. Both the introns and the exons show similar relative total rates of evolution throughout the rodents. Because introns usually experience little selection, this suggests that the observed rate variation between species may be caused by differences in mutation rates.

Clustered substitutions

Permutations of the number of substitutions per position in the small intron as well as the exon third position do not statistically differ from their respective Poisson expectations (see Table 4). This indicates that the mutation bias and base composition bias of the rodent *APRT* sequences do not cause the substitutions to be clustered at specific sites.

The observed numbers of substitutions per position in the small introns do not significantly differ from the Poisson expectations or the permutation

Table 4 The number of substitutions per position observed across all six species (excluding the two outgroup species *Mesocricetus* and *Gerbillus*) along with the expected numbers assuming a Poisson distribution and determined from the permutation study

Number of substitutions per position	Observed	Poisson expectation (using observed mean)	Permutation expectation	Poisson expectation (using permutation mean)
Small introns		NS†	NS†	NS‡
0	318	318	315.1	315.1
1	91	93.0	97.0	95.0
2+	17	15.0	14.0	16.0
Exon third positions		$P<0.001$§	$P<0.001$§	NS‡
0	138	138	130.3	130.3
1	25	36.0	41.3	41.4
2+	16	5.1	7.5	7.3

Observed: number of substitutions per position determined from the small intron dataset and the exon third positions.
Poisson expectation (using observed mean): expected number of substitutions per position calculated using the observed mean.
Permutation expectation: the number of substitutions per position determined from the permutation study.
Poisson expectation (using permutation mean): expected number of substitutions per position calculated using the permutation mean.
†Nonsignificant difference from observed.
‡Nonsignificant difference from permutation expectation.
§Significant difference from observed.
NS, not significant.
The Poisson expectations were calculated using a mean calculated from the number of positions with zero substitutions.

expectations (Table 4). This indicates that there are no detectable hotspots in the *APRT* small introns. The exon third position observation does significantly differ from both the Poisson expectation and the permutation expectation ($P<0.001$ for both comparisons; Table 4). More positions have two or more substitutions than expected, indicating that some exon third positions are more prone to substitution than others. Purifying selection could account for the pattern seen but would result in a lower rate of evolution in the exon third codon positions compared with the small intron rate. The intron rate should be a better reflection of the underlying mutation rate. However, the rates of evolution for the exon third positions and the small introns are approximately equal (0.33 vs. 0.35, respectively). Hence either there is some purifying selection in the introns which lowers the substitution rate to a level similar to the exon third position rate, or some exon third positions experience a higher mutation rate.

Discussion

Molecular evolution occurs through complex interactions between the forces of mutation, selection and population dynamics (Futuyma, 1986). The relative impact that each of these forces has on the molecular sequence data of related individuals can only be discerned after some basic analyses have been performed. But many of these analyses rely on an assumption of equal rates of evolution which may be false. To understand better the extent of rate variation between different closely related species and different regions of the gene we have sequenced the *APRT* gene, aligned the sequences, determined the phylogeny and estimated the relative rates of evolution in the different species and the different regions of the gene.

Exon and intron organization

The highly conserved exon and intron organization of the rodent *APRT* genes allows homologous regions between the species to be clearly aligned and identified. This enables an accurate estimate of substitution rates to be made as there are no ambiguities arising from the alignment. As human *APRT* introns are 14–20 per cent greater in length than the known rodent *APRT* introns which also have a B1 and B2 repetitive element in *Mastomys* and *Mus*

pahari, respectively, some length variation in the introns appears to be tolerated.

Repetitive elements

B1 repetitive elements have more than 10^5 copies scattered throughout the murine genome (Krayev *et al.*, 1982) and are specific to rodent species (Quentin, 1989). They are not present in the closely related lagomorphs. The B1 elements show considerable sequence homology to the *Alu* elements found in primates. B2 elements are not related to B1 elements and are found in the Muridae and Cricetidae but not in the Gliridae and Caviidae (Serdobova & Kramerov, 1993).

The presence of a B1 element and a B2 element in the introns of *Mastomys* and *M. pahari* but not in other closely related rodent species indicates that these elements must have recently been inserted at these locations. As B1 and B2 elements are members of different rodent repetitive element families these two insertions in two rodent species are clearly independent events. However, the mechanism of insertion for these two elements is probably the same as they both have a direct repeat flanking the repetitive element and they both have putative RNA polymerase III regulatory elements present which may indicate that they were inserted through a retroposon event (Rogers, 1985).

Rodent phylogeny

Rodent phylogenies based on morphological characteristics are often inconclusive as the variation is usually limited to dentition, cranial and soft body parts (Luckett & Hartenberger 1985). DNA/DNA hybridization and allozyme data have been useful (Catzeflis *et al.*, 1987; She *et al.*, 1990), but lack the resolution necessary to determine conclusively a phylogeny in closely related rodent species. Sequence data from various genes and genomic regions are becoming available and should be able to provide enough information to support a single tree.

The phylogenetic relationships of some or all of the eight rodent species used in this paper have been studied using morphological (Marshall, 1986), DNA/DNA hybridization (She *et al.* 1990), electrophoretic data (She *et al.*, 1990), mtDNA (Ferris *et al.*, 1983; She *et al.*, 1990), Ig heavy chain (Morgado *et al.*, 1993), Y chromosome repetitive elements (Nishioka *et al.*, 1994) and the *Sry* locus (Lundrigan & Tucker, 1994). The *APRT* phylogeny agrees with the previously inferred trees. Although not statistically conclusive, tree A is consistently supported by all of the data.

The ambiguous phylogenetic relationships of *Mastomys* and *Stochomys* in relation to *Rattus* and the *Mus* spp. may result from rate variation in *Rattus* and *Mastomys*. Conversely, the ambiguity may arise from the prevalence of parallel and reverse substitutions as shown in the relative rate analysis.

Does a molecular clock exist?

The existence of an accurate molecular clock, although desirable, cannot be assumed to exist in all genes and regions of the genome. There are many different mutational and selective forces acting on the sequences that depend on a multitude of factors such as the sequence itself and the sequence context. However, without some hypothesis about the rate of evolution nothing can be said about the relatedness and divergences of species based on sequence data. Therefore an analysis of the patterns of rate variation in different sequences from many species is necessary to estimate the rates of evolution and predict which patterns will emerge from different sequences and sequence contexts.

The relative rate test has been used extensively to test for rate variation between lineages as it does not require knowledge about the phylogeny other than knowing the outgroups. The ancestral sequences also do not need to be inferred because the comparisons are between the outgroup and each of two species individually. However, as demonstrated here the choice of outgroup can greatly affect the results obtained. It was suggested by Wu & Li (1985) that the outgroup should be chosen such that the distance between the ancestor to the two species being tested and the ancestor to all three species is as short as possible. However, it is more important to limit the number of parallel, convergent and reverse substitutions present between the species as demonstrated by the *Rattus* and *Mesocricetus* outgroups showing greater statistical significance than the less distant *Mastomys* and *Stochomys* outgroups in the *M. musculus* and *M. spicilegus* comparison (Table 3). The only way to avoid conclusively these homoplastic substitutions is to sequence many closely related species.

A maximum likelihood ratio test can be used to test for rate variation between lineages. Maximum likelihood estimates have a number of desirable statistical properties including statistical consistency and minimum variance, along with the benefit of the availability of the likelihood ratio test (Gaut & Lewis, 1995). Another advantage of the likelihood

ratio test is that the whole phylogeny is included in the analysis rather than just some branches. The disadvantage is that this method does not determine which particular branches deviate from a molecular clock, only whether the overall likelihoods are consistent with a molecular clock hypothesis. To overcome this problem in this study, each species was sequentially dropped from the analysis to determine whether that particular species was the cause of the rejection of the molecular clock hypothesis.

The rate variation between *Mus spicilegus* and *M. musculus* is very interesting because previous studies have suggested that rodent species have approximately equal rates of evolution (e.g. O'Huigin & Li, 1992). Their analysis did not include species as closely related as the species in this study and did not include *M. spicilegus* which is the species most deviant in rate. Whether *Rattus* is also slower than the other species and *Mastomys* is faster will require more sequences to establish whether the lack of statistical significance is the result of sample size or a lack of rate variation in these species. As demonstrated in this study the sequencing and analysis of closely related species allows for a more accurate analysis of the rates of evolution and may aid the discovery of causes for some rate variation.

Substitution rates

The unequal rates of evolution between the different regions of the gene may lead to an understanding of the influence of base composition and sequence context on substitutions. Particularly interesting is the clustering of substitutions within some exon third positions in comparison to the introns even though the rates of evolution are equal in both regions. This may indicate that these introns also have some selection acting to reduce their rate of evolution.

References

ALLARD, M. W. AND HONEYCUTT R. L. 1991. Ribosomal DNA variation within and between species of rodents, with emphasis on the genus *Onychomys*. *Mol. Biol. Evol.*, **8**, 71–84.

BAINS, W. AND TEMPLE-SMITH, K. 1989. Similarity and divergence among rodent repetitive DNA sequences. *J. Mol. Evol.*, **28**, 191–199.

BENZER, S. 1961. On the topography of the genetic fine structure. *Proc. Natl. Acad. Sci. U.S.A.*, **47**, 403–415.

BRODERICK, T. P., SCHAFF, D. A., BERTINO, A. M., DUSH, M. K., TISCHFIELD, J. A. AND STAMBROOK, P. J. 1987. Comparative anatomy of the human *APRT* gene and enzyme:nucleotide sequence divergence and conservation of a nonrandom CpG dinucleotide arrangement. *Proc. Natl. Acad. Sci. U.S.A.*, **84**, 3349–3353.

CATZEFLIS, F. M., SHELDON, F. H., AHLQUIST, J. E. AND SIBLEY, C. G. 1987. DNA-DNA hybridization evidence of the rapid rate of muroid rodent DNA evolution. *Mol. Biol. Evol.*, **4**, 242–253.

FELSENSTEIN, J. 1989. PHYLIP—phylogeny inference package, version 3.4. *Cladistics*, **5**, 164–166.

FERRIS, S. D., PRAGER, E. M., RITTE, U., SAGE, R. D. AND WILSON, A. C. 1983. Mitochondrial-DNA evolution in mice. *Genetics*, **105**, 681–721.

FIELDHOUSE, D. AND GOLDING, G. B. 1993. The rat adenine phosphoribosyltransferase sequence shows evolutionary rate variation among exons in rodents. *Genome*, **36**, 1107–1110.

FITCH, W. M. 1971. Toward defining the course of evolution: minimum change for a specific tree topology. *Syst. Zool.*, **20**, 406–416.

FUTUYMA, D. J. 1986. *Evolutionary Biology*, 2nd edn. Sinauer Associates, Sunderland, MA.

GAUT, B. S. AND LEWIS, P. O. 1995. Success of maximum likelihood phylogeny inference in the four-taxon case. *Mol. Biol. Evol.*, **12**, 152–162.

GOLDING, G. B. 1983. Estimates of DNA and protein sequence divergence: an examination of some assumptions. *Mol. Biol. Evol.*, **1**, 125–142.

HAMMER, M. F. AND SILVER, L. M. 1993. Phylogenetic analysis of the alpha-globin pseudogene-4 (*Hba-ps4*) locus in the house mouse species complex reveals stepwise evolution of *t* haplotypes. *Mol. Biol. Evol.*, **10**, 971–1001.

HIGGINS, D. G. AND SHARP, P. M. 1989. Fast and sensitive multiple sequence alignments on a microcomputer. *Cabios*, **5**, 151–153.

KRAYEV, A. S., MARKUSHEVA, T. V., KRAMEROV, D. A., RYSKOV, A. P., SKRYABIN, K. G., BAYEV, A. A. AND GEORGIEV, G. P. 1982. Ubiquitous transposon-like repeats B1 and B2 of the mouse genome: B2 sequencing. *Nucl. Acids Res.*, **10**, 7461–7475.

LUCKETT, W. P. AND HARTENBERGER, J.-L. 1985. *Evolutionary Relationships among Rodents, a Multidisciplinary Analysis*. Plenum Press, New York.

LUNDRIGAN, B. L. AND TUCKER, P. K. 1994. Tracing paternal ancestry in mice, using the Y-linked, sex-determining locus, *Sry*. *Mol. Biol. Evol.*, **11**, 483–492.

MARSHALL, J. T. 1986. Systematics of the genus *Mus*. *Curr. Top. Microbiol. Immunol.*, **127**, 12–18.

MEUTH, M. 1990. The structure of mutation in mammalian cells. *Biochim. Biophys. Acta*, **1032**, 1–17.

MORGADO, M. G., JOUVIN-MARCHE, E., GRIS-LIEBE, C., BONHOMME, F., ANAND, R., TALWAR, G. P. AND CAZENAVE, P. A. 1993. Restriction fragment length polymorphism and evolution of the mouse immunoglobulin constant region gamma loci. *Immunogenetics*, **38**, 184–192.

NISHIOKA, Y., DOLAN, B. M., ZAHED, L., PRADO, V. AND TYSON, H. 1994. Molecular evolution of a Y-chromosomal repetitive sequence family in the genus *Mus*. *Mol. Biol. Evol.*, **11**, 146–153.

NOWAK, R. M. 1991. *Walker's Mammals of the World*, 5th edn, vol. II. Johns Hopkins University Press, Baltimore and London.

O'HUIGIN, C. AND LI, W.- H. 1992. The molecular clock ticks regularly in muroid rodents and hamsters. *J. Mol. Evol.*, **35**, 377–384.

QUENTIN, Y. 1989. Successive waves of fixation of B1 variants in rodent lineage history. *J. Mol. Evol.*, **28**, 299–305.

ROGERS, J. 1985. Retroposons. *Int. Rev. Cytol.*, **93**, 187–279.

SAMBROOK, J., FRITSCH, E. F. AND MANIATIS, T. 1989. *Molecular Cloning, a Laboratory Manual*, 2nd edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

SERDOBOVA, I. M. AND KRAMEROV, D. A. 1993. Use of the short retroposon B2 in the study of phylogenetic relationship in rodents. *Genetika*, **29**, 1969–1981.

SHE, J. X., BONHOMME, F., BOURSOT, P., THALER, L. AND CATZEFLIS, F. 1990. Molecular phylogenies in the genus *Mus*: comparative analysis of electrophoretic, scnDNA hybridization and mtDNA RFLP data. *Biol. J. Linn. Soc.*, **41**, 83–103.

SKANDALIS, A. AND GLICKMAN, B. G. 1990. Endogenous gene systems for the study of mutational specificity in mammalian cells. *Cancer Cells*, **2**, 79–83.

TAYLOR, M. W., SIMON, A. E. AND KOTHARI, R. M. 1985. The *APRT* system. In: Gottesman, M. (ed.) *Molecular Cell Genetics*. John Wiley and Sons, New York.

TURKER, M. S., COOPER, G. E. AND BISHOP, P. L. 1993. Region-specific rates of molecular evolution: a fourfold reduction in the rate of accumulation of 'silent' mutations in transcribed versus nontranscribed regions of homologous DNA fragments derived from two closely related mouse species. *J. Mol. Evol.*, **36**, 31–40.

WOLFE, K. H., SHARP, P. M. AND LI, W.-H. 1989. Mutation rates differ among regions of the mammalian genome. *Nature*, **337**, 283–285.

WU, C.-I AND LI, W.-H. 1985. Evidence for higher rates of nucleotide substitution in rodents than in man. *Proc. Natl. Acad. Sci. U.S.A.*, **82**, 1741–1745.