J Mol Evol (1993) 37:441–456

Journal of Molecular Evolution

Mammalian Gene Evolution: Nucleotide Sequence Divergence Between Mouse and Rat

Kenneth H. Wolfe, Paul M. Sharp

Department of Genetics, University of Dublin, Trinity College, Dublin 2, Ireland

Received: 14 July 1992 / Revised: 15 September 1992

Abstract. As a paradigm of mammalian gene evolution, the nature and extent of DNA sequence divergence between homologous protein-coding genes from mouse and rat have been investigated. The data set examined includes 363 genes totalling 411 kilobases, making this by far the largest comparison conducted between a single pair of species. Mouse and rat genes are on average 93.4% identical in nucleotide sequence and 93.9% identical in amino acid sequence. Individual genes vary substantially in the extent of nonsynonymous nucleotide substitution, as expected from protein evolution studies; here the variation is characterized. The extent of synonymous (or silent) substitution also varies considerably among genes, though the coefficient of variation is about four times smaller than for nonsynonymous substitutions. A small number of genes mapped to the X-chromosome have a slower rate of molecular evolution than average, as predicted if molecular evolution is "male-driven." Base composition at silent sites varies from 33% to 95% G + C in different genes; mouse and rat homologues differ on average by only 1.7% in silent-site G + C, but it is shown that this is not necessarily due to any selective constraint on their base composition. Synonymous substitution rates and silent site base composition appear to be related (genes at intermediate G + C have on average higher rates), but the relationship is not as strong as in our earlier analyses. Rates of synonymous and nonsynonymous substitution are correlated, apparently because of an excess of substitutions involving adjacent pairs of nucleotides. Several factors suggest that synonymous codon usage in rodent genes is not subject to selection.

Key words: Molecular clocks — Rodents — Genome evolution — G + C content — Codon usage — Dinucleotide mutation effects

Correspondence to: K.H. Wolfe

As increasingly large amounts of DNA sequence data accumulate, our understanding of the pattern and dynamics of gene sequence divergence is growing clearer. The molecular clock, first postulated by Zuckerkandl and Pauling (1962) and later championed by Allan Wilson (see, for example, Wilson et al. 1977, 1987), is a description of the observed regularity with which substitutions accrue in nucleotide and amino acid sequences. Under the neutral theory (Kimura 1983) the rate of molecular evolution (k) of a sequence is determined by the product of the mutation rate (u_T) and the fraction of sites in the sequence (or the average fraction of mutations at those sites) that are selectively neutral (f_0):

$$k = u_T f_0$$

(Kimura 1977). Thus, the rate of sequence change in different genes or proteins will be similar only if these two factors are the same.

From the first studies of protein evolution it was apparent that the rate of amino acid replacement varies enormously among proteins (see, e.g., Dickerson 1971) and that this can be attributed to differences in the fraction of possible neutral amino acid replacements (f_0) between, say, fibrinopeptides on

the one hand and histone H4 on the other. (See Table 4.4 of Nei 1987.) While these extreme examples of fast and slow protein sequence evolution have now become the stuff of textbooks, few data have been gathered on the distribution of rates among the less-exceptional proteins. An understanding of the shape of the underlying distribution of rates would contribute toward an understanding of protein evolution in general. However, documentation of rates of evolution of a large number of proteins has been hampered in part by the necessity to combine data from sequence comparisons across different taxonomic groups, thus requiring the use of fossil-based divergence dates along with the ensuing uncertainties (e.g., Li et al. 1985).

Here we overcome this problem by limiting our comparisons to a single pair of species (mouse and rat) so that the relative extents of change seen in different proteins are a direct measure of their relative rates of evolution, regardless of the data of divergence between the two species. (In fact, the actual date of the mouse-rat speciation event is a matter of considerable controversy—see Catzeflis et al. 1992—and will not be discussed in much detail.) The massive popularity of DNA sequencing as a biochemical tool, and of both mouse and rat as model organisms, has resulted in the sequencing of several hundred homologous gene pairs from these species in the decade following the first comparative study (Jagodzinski et al. 1981). By comparing nucleotide as well as amino acid sequences we have also been able to characterize the variation among genes in the rate of silent (synonymous) nucleotide substitution and show that, as with proteins, there are some genes with exceptionally fast or exceptionally slow rates. Since almost all silent sites in mammalian genes are likely to be free to accept nucleotide substitutions, these deviations from a "silent molecular clock" may be due to local variation in the mutation rate (u_T); the possible basis for this variation is discussed. A seemingly paradoxical correlation between the rates of synonymous and nonsynonymous substitution in genes is also investigated.

Data and Methods

Sequences were collected between 1986 and 1992 and were either taken from releases 59–71 of GenBank or obtained from the EMBL and GenBank database electronic mail servers; in a few cases sequences were obtained directly from the literature. Genomic sequences were used in preference to cDNAs where possible at the time of collection. Homologous mouse and rat sequence pairs in GenBank were identified initially by browsing through the sequence definitions with the aid of a text editor or by keyword searches using the retrieval system ACNUC (Gouy et al. 1985). Other sequence pairs were identified by a semiautomated procedure whereby data on polypeptide length and amino acid composition were extracted from each of the 5,568 mouse and 2,543 rat protein-coding sequences in GenBank (release 71) using ACNUC, after which each mouse sequence was compared to each rat sequence and candidate matches of similar length and amino acid composition were flagged for further investigation. This method has the potential to identify all homologous pairs very rapidly (<30 min on a shared VAX computer to make all 14 million possible pairwise comparisons of murid genes) but is biased in favor of highly conserved sequences and cannot find matches between incomplete sequences and full-length homologues. In practice, this method found only about 70% of the pairs that had already been identified manually, but also led to the discovery of 35 additional pairs of homologues that were not evident from the annotation of their database entries.

Coding sequence pairs were then extracted from GenBank using ACNUC, and the deduced protein sequences were aligned using a rapid Needleman-Wunsch method (CLUSTAL; Higgins and Sharp 1989). The DNA sequences were then aligned using the protein alignments as templates. All alignments were inspected by eye for possible frameshift sequencing errors and a number of putative errors were corrected in such a way as to maximize the ratio of synonymous to nonsynonymous nucleotide substitutions in the frameshifted region. Three genes were found where the reported mouse and rat sequences were essentially identical; we suspect that the source species is incorrectly identified in database entries M84361 ("rat" CSF-1), X61479 ("rat" c-*fms*), and M36660 ("mouse" NADPH: menadione oxidoreductase). We assume that these result from laboratory or database confusion and have not included these sequences in the analysis.

The extents of synonymous and nonsynonymous divergence (in terms of nucleotide substitutions per site) were calculated by the method of Li et al. (1985). In this method, sites in codons are classified as 0-fold, 2-fold, or 4-fold degenerate, according to how many alternative nucleotides at the site encode the same amino acid. A correction for superimposed substitutions at single sites is made by the two-parameter method of Kimura (1980), which allows for different rates of transitions and transversions. The numbers of synonymous and nonsynonymous substitutions per site are then calculated as appropriately weighted averages of these values. (See Li et al. 1985 for more details, including the estimation of standard errors.)

A few sequence pairs were found to give unusually high K_S values (Fig. 2), so we considered the possibility that these might be paralogous. In the case of β_2-microglobulin (K_S = 0.69) the mouse and rat genes have each been sequenced by at least two independent groups. Multiple independent sequences have also been determined for rat SVS IV (K_S = 0.78). Mouse SVS IV has been sequenced by only one laboratory (Chen et al. 1987), but this group obtained the mouse cDNA by using a sequenced rat SVS IV cDNA probe and also directly sequenced most of the mouse protein. Furthermore, Dietrich et al. (1992) were able to map SVS IV (*svp-4*) to mouse chromosome 2 by polymerase chain reaction amplification using primers corresponding to the published mouse sequence. There is thus no reason to doubt that the mouse and rat SVS IV sequences are orthologues. The gene with the third-greatest K_S value (0.51) is SPOT-1, which is discussed by Dickinson et al. (1989).

Nucleotide and Amino Acid Sequence Divergence Between Mouse and Rat

The nucleotide sequences of the coding regions of 363 genes for which DNA sequences have been de-

Table 1. Nucleotide and amino acid sequence divergence, and silent-site G+C content, in 363 homologous genes of mouse and rat[a]

	Mean (±SD)	Range	L[b]
Amino acid identity (%)	93.9 (±8.1)	56.0–100.0	136,729
Nucleotide identity (%)	93.4 (±4.1)	69.6–99.0	411,300
Nonsynonymous substitutions (K_A)	0.032 (±0.049)	0.000–0.372	318,873
Synonymous substitutions (K_S)	0.224 (±0.084)	0.041–0.780	91,315
Silent site G+C content (%)	62.0 (±11.1)	32.8–95.4	145,357

[a] The 363 genes, and their individual values, are detailed in Appendix 1. Means and standard deviations are weighted by the number of sites in each gene. Silent site G+C content (GC_S) is the G+C content at 2-fold and 4-fold degenerate sites in codons.
[b] The total number of residues in each category

termined from both mouse and rat were aligned and compared. The genes, and some statistics describing their molecular evolution, are listed in Appendix 1. The results are summarised in Table 1. The total length of aligned sequences (excluding gaps) is 411 kb. The genes range in size from 135 bp (thymosin β4) to 8,247 bp (inositol-1,4,5-triphosphate receptor), with an average length of 1,133 bp. As expected, the deduced protein sequences range in degree of divergence. Twenty-five proteins (7%), including some actins, ion-channel proteins, and ribosomal proteins, are identical in mouse and rat. The most divergent proteins are the salivary SPOT-1 protein (56% identity; discussed by Dickinson et al. 1989) and interleukin-3 (59%; see Cohen et al. 1986). Nucleotide sequence identity varies from 70% in the SPOT-1 gene to 99% in the gene encoding the Y-box binding transcription factor. The mean level of sequence identity, weighted by gene length, is 93.4% (standard deviation 4.1%) for nucleotides and 93.9% (±8.1%) for amino acids.

The extents of amino acid and nucleotide sequence divergence are (necessarily) correlated, but the relationship between these two measures is interesting. In genes encoding very highly conserved proteins, amino acid sequence identity exceeds nucleotide identity (Fig. 1), because silent (synonymous) nucleotide substitutions are permitted in these genes. However, in genes encoding less-conserved proteins, nucleotide similarity exceeds protein similarity. This presumably reflects the degeneracy of the genetic code: for example, a single codon with nucleotide substitutions at positions 2 and 3 (the former causing an amino acid replacement, the latter probably silent) exhibits 33% nucleotide sequence identity but 0% amino acid sequence identity. For the mouse-vs-rat data, the point at which nucleotide and amino acid sequence identity levels cross (i.e., are similar) is approximately 93% (Fig. 1b); the mean values happen to be close to this point. This picture is markedly different from that seen in a study of bacterial sequence divergence (67 genes compared between *Escherichia coli* and *Salmonella typhimurium;* Sharp 1991). In the bacterial data, protein sequences are invariably more similar than nucleotide sequences over a range of 76 to 100% amino acid sequence identity, and extrapolation of the relationship suggests a crossover point at around 65% identity (P.M.S., unpublished results). This difference is probably related to a much higher mean ratio of synonymous-to-nonsynonymous divergence in the bacterial genes than in the rodent genes, which is discussed below.

Nucleotide substitutions between two species can be classified as either nonsynonymous (replacement) or synonymous (silent), depending on whether or not they alter the protein sequence. The estimated numbers of nonsynonymous substitutions per site (K_A) and synonymous substitutions per site (K_S) were calculated by the method of Li et al. (1985), which corrects for multiple hits (Table 1, Appendix 1). These values, and the relationship between them, will be discussed in turn.

Nonsynonymous Nucleotide Substitution and Protein Evolution

The estimates of the extents of nonsynonymous nucleotide substitution (K_A) among genes reflect the diversity of levels of protein sequence conservation and range from zero in the 25 genes mentioned earlier (and $K_A < 0.01$ substitution per site in a further 103, or 28% of genes) to $K_A = 0.372$ (±0.053) nonsynonymous substitutions per site in the salivary SPOT-1 gene; the weighted mean K_A is 0.032 (with standard deviation 0.049).

The K_A values for the 363 genes (Fig. 2a) form a broad, largely unskewed, distribution when plotted on a histogram with a semilogarithmic scale (as originally used by Ochman and Wilson 1987). There is, however, a distinct peak due to a number of genes with extremely low values of K_A (i.e., encoding almost invariant proteins). This result was also apparent in the small number of genes analyzed by Li et al. (1985; see histograms in Ochman and Wilson 1987; and in Hartl and Clark 1989:369). A similar result is obtained in comparisons of sequences (about 700 genes) between humans and rodents (K.H.W., unpublished results).

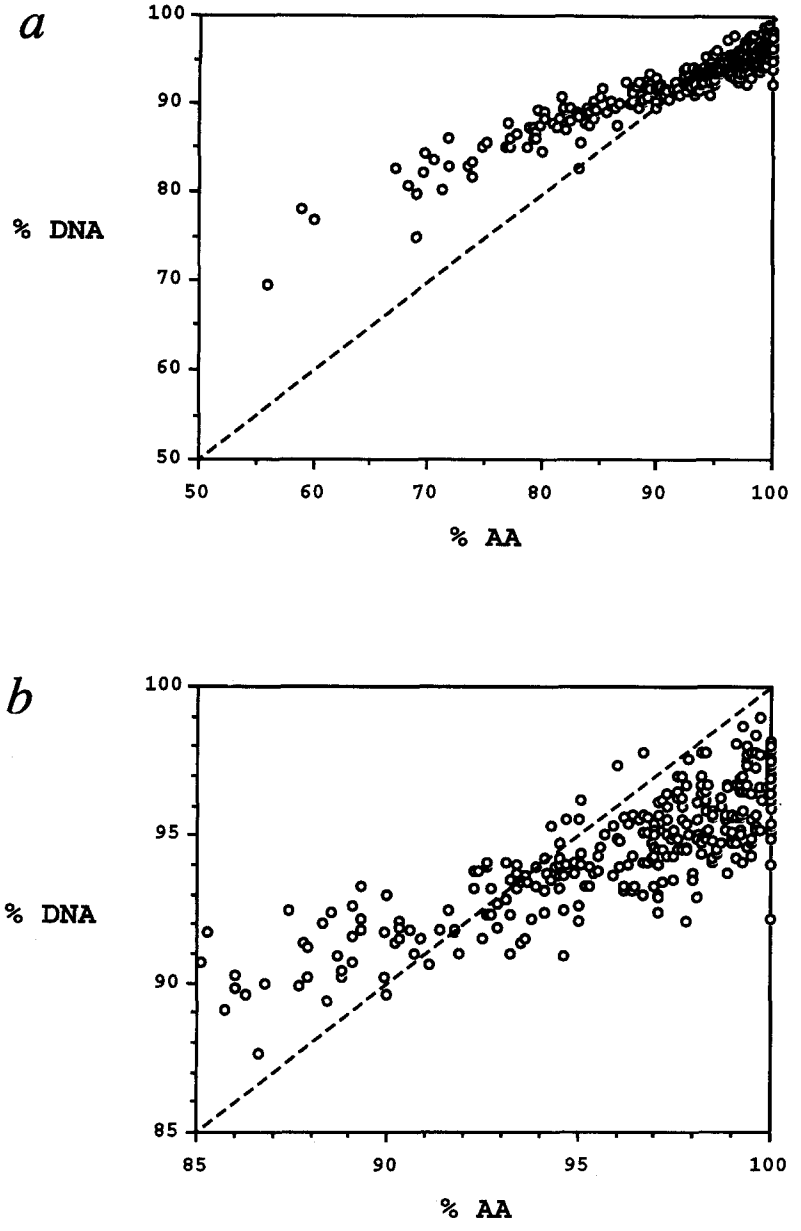

Fig. 1. Relationship between nucleotide (% DNA) and amino acid (% AA) sequence identity among 363 genes compared between mouse and rat. The region between 85% and 100% identity is enlarged in **b**. The *dashed lines* indicate equal levels of amino acid and nucleotide sequence identity.

The extent of synonymous substitution also varies between genes, although to a much lesser extent. (See below.) Setting that variation aside for the moment, it is interesting to note that the ratio between the mean synonymous and mean nonsynonymous divergence in the rodent genes is 7.1. This value is a little higher than that reported earlier (about five, among 35 genes; Li et al. 1985) in comparisons among mammalian orders. More significantly, this value is much lower than that seen in bacterial genes (for which the ratio is 24; Sharp 1991). Ochman and Wilson (1987; see also Lawrence et al. 1991) have proposed that this difference is due, at least in part, to the enormous difference in effective population size between bacteria (specifically *E. coli*) and mammals: selection against slightly deleterious mutant protein sequences may be far more efficient in bacteria. The effect may also be partly due to genes of different functions being included in the two data sets.

If synonymous substitutions are essentially neutral (as discussed below), the ratio of K_S to K_A provides a measure of the degree of selective constraint on a protein sequence, independent of other factors (such as local mutation rate differences) that may cause the underlying nucleotide substitution rate to be different in different genes. Values of the K_S/K_A ratio for individual genes vary from infinity (for the set of 25 genes encoding invariant proteins) to 1.22 for the Blast-1 antigen and 0.89 for interleukin-3. The last value may be remarkable because the nonsynonymous rate is expected to exceed the synonymous rate only in exceptional cases of positive natural selection for amino acid sequence divergence. (See, e.g., Hughes and Nei 1988.) However, interleukin-3 is the only gene (out of 363) for

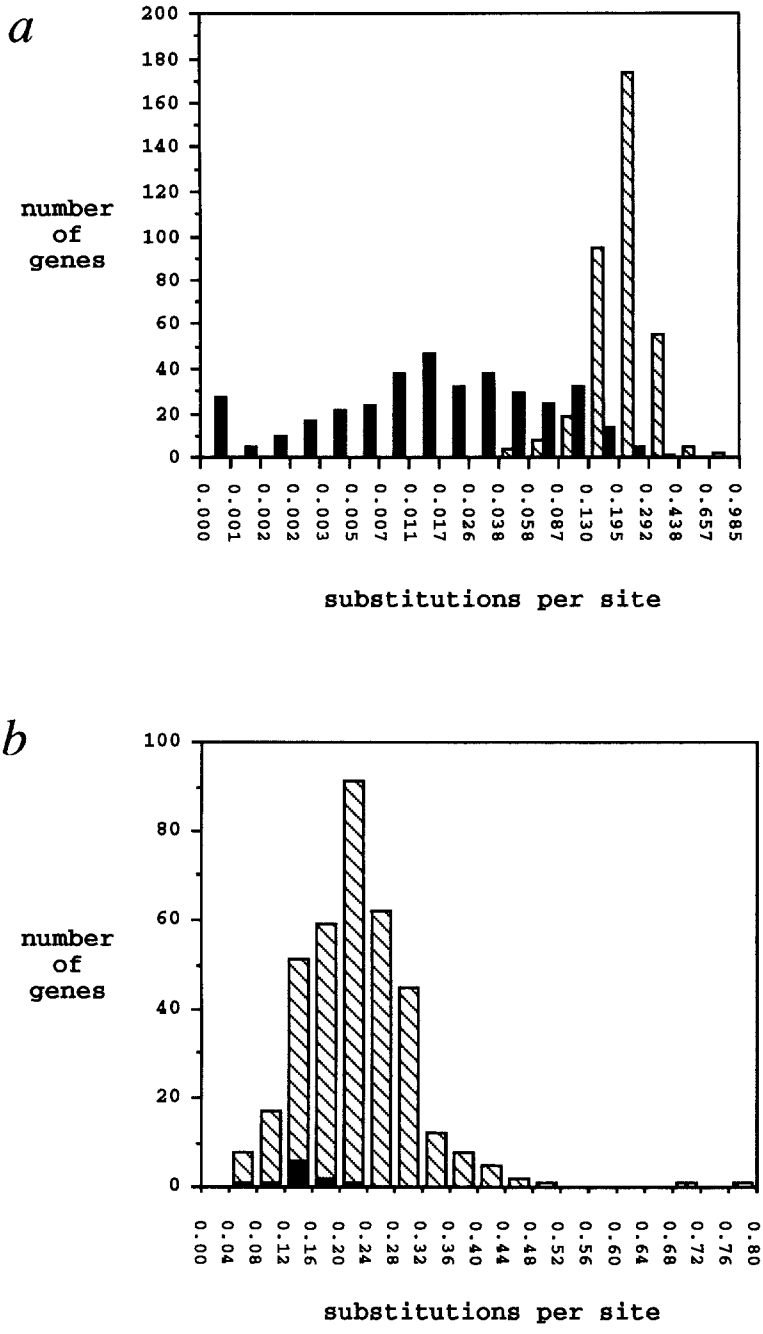

Fig. 2. Extent of nucleotide sequence divergence in 363 genes compared between mouse and rat. **a** Distribution of values of nonsynonymous (K_A, *solid*) and synonymous (K_S, *hatched*) substitutions per site. Note the semilog scale on the X-axis. **b** Distribution of values of synonymous substitutions per site (K_S), with X-linked genes highlighted.

which $K_S < K_A$, and the sequence is sufficiently short that the ratio may not be significantly less than one. Furthermore, it has not been established whether the estimation method of Li et al. (1985) will actually result in a K_S/K_A ratio of precisely 1.0 for a neutral DNA sequence; it is possible that the method suffers from a small bias, particularly in its treatment of 2-fold degenerate sites, which may overestimate K_S. Many of the other genes with low K_S/K_A ratios are either members of the immunoglobulin superfamily (Blast-1, Ig-C-delta, Ig-V-lambda, CD4, CD8α, and CD43 antigens) or secretory proteins (SPOT-1, beta- and kappa-caseins, whey acidic protein). It is perhaps noteworthy that interleukin-3 is the only gene for which $K_S < K_A$ despite the considerable overlap between the ranges of K_S and K_A (Fig. 2a; there are 99 genes for which K_A is greater than the lowest K_S seen in any gene and 350 genes for which K_S is less than the highest K_A seen).

Silent Sites

The weighted mean extent of synonymous substitution (K_S) between mouse and rat genes is 0.224 substitutions per site (standard deviation 0.084). (Li et al. 1987 reported a mean value of 0.237 for 24

genes.) The values for individual genes (Table 1 and Fig. 2) range 19-fold between 0.041 (±0.014) in the Y-box transcription factor gene and 0.780 (±0.179) in the gene for seminal vesicle secretory protein IV. [We have consulted the original publications of the sequences at either end of the range of K_S values (Fig. 2) to verify that these are bona fide homologous single-copy genes from mouse and rat; see also Data and Methods, above.]

The K_S values form quite a broad, roughly symmetric distribution about the mean (Fig. 2). Is this variation among genes in K_S greater than that expected by chance due to sampling effects? This can be tested statistically: if the variation in K_S is due to sampling error, the standard deviation of values around the mean K_S should be that expected under a binomial distribution. From equations (1)–(10) in Li et al. (1985) (using mean numbers of transitions and transversions per site in all genes, and the harmonic mean number of sites per gene), the standard error of the mean K_S in mouse-rat genes should be 0.043. This is only half as great as the observed standard deviation of K_S (0.084; Table 1), implying that there is genuine variation in synonymous substitution rates among genes. However, the synonymous substitution rates are considerably less variable than nonsynonymous rates: the coefficient of variation (the ratio of the standard deviation to the mean) of K_A is about four times that of K_S (Table 1).

The mean K_S for mouse-vs-rat genes is very close to the figure of 0.231 (±0.009) calculated by Catzeflis et al. (1987, 1989) as the overall extent of sequence divergence between mouse and rat genomes, obtained by DNA-DNA hybridization of single copy genomic DNA (with correction for multiple-hit kinetics). We note, however, that K_S is generally larger (perhaps for the reason mentioned above) than the substitution rate observed at 4-fold degenerate sites alone (Li et al. 1985), even though both purport to estimate the rate of silent nucleotide substitution. The similarity in substitution rates at silent codon positions and in the genome as a whole (which is largely noncoding) strongly suggests that codon usage in mammalian genes is not constrained by natural selection, as we (Sharp 1989; Wolfe et al. 1989) and others (Eyre-Walker 1991) have argued elsewhere (also, see below).

To estimate the absolute rate of synonymous substitution (per unit time) it is necessary to know when mouse and rat last shared a common ancestor, and estimates of that date are quite controversial. Palaeontological evidence has been interpreted as indicating the date to be about 10 million years ago (see Catzeflis et al. 1992), suggesting a mean absolute rate of 11.2×10^{-9} substitutions per site per year per lineage (or a divergence rate down two lineages of 2.24% per million years). This is approximately two times the average rate in mammals (including rodents) previously expressed as 4.7×10^{-9} substitutions per site per year (Li et al. 1985) or about 1% per million years down two lineages (Wilson et al. 1987). However, Wilson et al. (1987) have argued that the murid fossil record is difficult to interpret, and that the common ancestry of mouse and rat may date from as much as 35 million years ago. In that case, obviously, the rate estimates would be 3.5 times smaller.

Sex-linked Genes

Some of the heterogeneity in silent substitution rates among mammalian genes could be due to the location of some genes on the sex chromosomes. Miyata et al. (1987a,b) proposed that, because of the vastly greater number of cell generations that occur on average per year in spermatocytes as compared to oocytes, most nucleotide substitutions in mammalian genes arise in the male germline (so-called "male-driven" molecular evolution). As a consequence, the average silent substitution rates for autosomal, X-linked, and Y-linked genes should be in the ratio 1 : 0.67 : 2, reflecting the proportions of time (50%, 33%, and 100%, respectively) that each type of chromosome spends in males when averaged over many generations. More recently, Miyata et al. (1990) compared sequences between humans and rodent and found that the mean synonymous substitution rate for six X-linked genes was 58% of that in 35 autosomal genes. Some studies have, however, reported individual X-linked genes with high rates of synonymous substitution (e.g., Iizasa et al., 1989).

The data in Appendix 1 includes 11 genes known to be X-linked, and their mean K_S value is 0.143 (±0.035). This is 64% of the mean value for the entire data set (which can be assumed to be largely autosomal), or 61% of the mean K_S (0.236 ± 0.063) for the 179 genes that have actually been mapped to an autosome in one or both species. The result is thus in close agreement with the predictions from the hypothesis of Miyata et al. (1987a,b). [The mean K_A for the X-linked genes is also low, 0.009 (±0.009), as compared to 0.032 (±0.049) for the whole data set.] The X-linked genes do not form a distinct group of slowly evolving sequences, but rather they lie within but toward the lower end of the distribution of K_S values seen in other genes (Fig. 2b): in order of increasing K_S the X-linked genes are ranked at positions 6 (a-*raf*), 10 (PLP), 37 (5-HT-1c receptor), 38 (androgen receptor), 40 (OTC), 51 (HPRT), 63 (NCAM-L1), 76 (PGK1), 84 (connexin-32), 104 (factor IX), and 220 (RPS4X) out of 363. Thus, there are heterogeneous rates for both

X-linked and autosomal genes, but there may well be a systematic (mutation rate) effect such that the former evolve at about two-thirds the rate of the latter. At present, there are no genes in the mouse-rat data set that are known to be Y-linked: such sequences would provide a valuable test of the "male-driven" molecular evolution hypothesis.

Variation in Silent-Site G + C Content

Synonymous codon usage is highly heterogeneous among mammalian genes, with the principal variation being in the base composition (G + C content) at silent sites (Ikemura 1985; Sharp et al. 1988). This variation seems to reflect local chromosomal base composition, since the G + C content at silent sites of individual mammalian genes is highly correlated with the G + C content of their introns and flanking sequences (Aota and Ikemura 1986; Ikemura and Aota 1988). This appears to be related to the organization of the mammalian genome into "isochores," i.e., domains of several hundred kilobases each having a relatively homogeneous base composition internally but being different from neighboring isochores (Bernardi et al. 1985). As expected, the mean G + C content at silent codon positions (designated GC_S and defined as the G + C content at 2-fold and 4-fold degenerate sites) in the mouse-rat genes studied here varies substantially from 32.8% to 95.4% (in the genes for nucleolar protein B23 and AGP/EBP transcription factor, respectively). The standard deviation of GC_S in the 363 genes is 11.1%, which is 3.7 times greater than that expected due to sampling error (under a binomial distribution around the mean GC_S of 62.0%).

The values of GC_S in homologous mouse and rat genes are very strongly correlated ($r = 0.973$). The greatest difference in GC_S between the species in a single gene is 10.2% in SVS IV (41.7% in mouse, 51.8% in rat), but the mean absolute difference in GC_S (ΔGC_S) is only 1.69%, with no significant net bias in either direction (61.8% mean GC_S in mouse; 62.2% in rat). Furthermore, there is no significant correlation between ΔGC_S and the mean GC_S in mouse and rat. This contrasts with the divergence in base composition that has occurred between human and rodent genes: human genes tend to have more extreme base compositions than their rodent homologues, a phenomenon that has been termed the "minor shift" (Mouchiroud et al. 1988; Bernardi et al. 1988).

Bernardi et al. (1988) have cited the strong correlation between silent-site G + C contents in homologous mouse and rat genes as evidence that their codon usage (or overall base composition) is being constrained by purifying selection. In fact, the high correlation coefficient may be entirely attributable to the common ancestry of the sequences, since the mean K_S value of 0.224 substitutions per site implies that about 80% of the silent codon positions are identical by descent. We have investigated this by conducting a simple computer simulation: 363 ancestral sequences were generated (corresponding in length and G + C content to the silent sites in the rodent genes) and nucleotide substitutions (corresponding in number to the product of K_S and the number of silent sites in each gene) were then made at random assuming that all types of substitution are equally probable. Even with this simplifying assumption (which causes the G + C contents of the daughter sequences to converge toward 50%) the mean correlation coefficient from 1,000 replications was 0.950 (range 0.931–0.964), which is close to the correlation coefficient obtained with the real data. Indeed, if the simulation is conducted considering only 4-fold degenerate sites (instead of 2-fold and 4-fold sites combined), the mean correlation coefficient in simulations (0.958) exceeds that from the real data (0.947). There is thus little or no need to invoke causes other than common ancestry to explain the similarity of GC_S in the two species.

Excessive Variation in K_S and GC_S

The variation among genes in rates of nonsynonymous substitution can be interpreted in terms of differing extents of selective constraints on amino acid sequences (i.e., differences in the proportion of the sequence that is critical to the function of the protein). In contrast, the variation among genes described above in both silent substitution rates and silent-site G + C contents could be explained either by selective constraints on codon usage or by variation in the underlying mutation patterns among loci.

We and others have reported a relationship between the substitution rate and base composition at silent sites in mammalian genes, with genes of high GC_S (and possibly also those of low GC_S) having lower substitution rates than those of intermediate base compositions (Filipski 1988; Wolfe et al. 1989; Ticher and Graur 1989; Bulmer et al. 1991). When this trend is reinvestigated using the present data set (representing a fivefold increase in data over our 1989 study), the relationship between K_S and GC_S is less striking than originally reported (Fig. 3a). In particular, there are several large genes with intermediate GC_S contents (around 60%) that have low K_S values. Nevertheless, there is still a paucity of genes having both a high silent substitution rate and extreme values of GC_S.

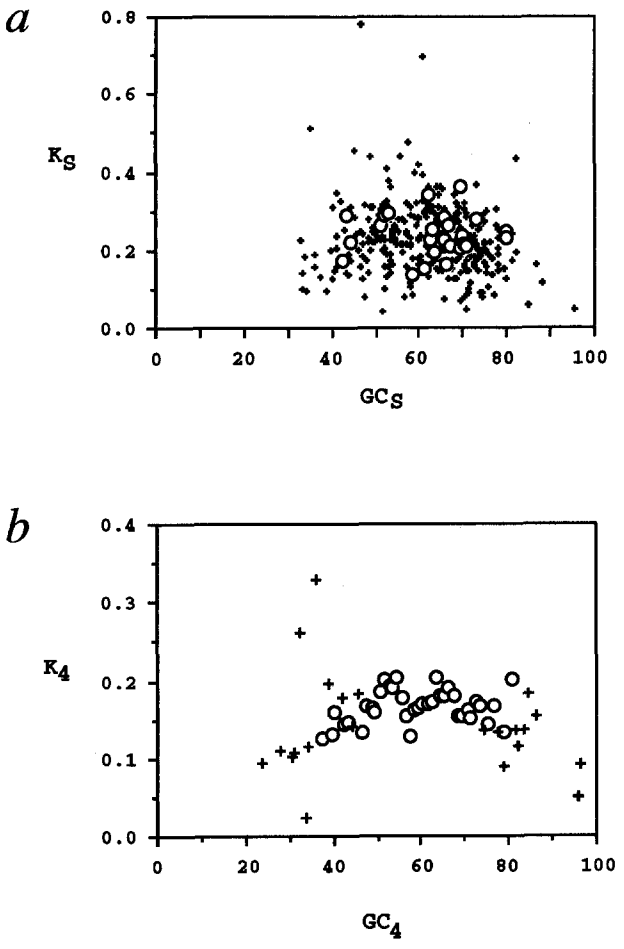

Fig. 3. Relationship between silent site divergence (between mouse and rat) and base composition (G + C content). **a** Number of synonymous substitutions per site (K_S) vs G + C content at silent sites (GC_S) for 363 genes. *Circles* represent genes with ≥500 silent sites. **b** Number of substitutions (K_4) vs G + C content (GC_4) at 4-fold degenerate sites, considering only codons where no more than one nucleotide substitution has occurred and correcting for multiple hit kinetics by the method of Tajima and Nei (1984). (Cf. Wolfe et al. 1989.) *Each point* represents the pooled result for all genes within a particular 1% interval of GC_4. *Circles* represent intervals for which the number of sites compared is ≥500.

For clarity, in our earlier study (Wolfe et al. 1989) we calculated substitution rates (K_4, using the multiple-hits correction formula of Tajima and Nei 1984), and base composition (GC_4) only at 4-fold degenerate sites, and only in codons where there were no additional nucleotide substitutions. These modifications eliminate the correlation between synonymous and nonsynonymous substitution rates (see below), and also eliminate a possible artifact resulting from multiple-hits corrections on sequences of biased base compositions. Applying these criteria to the expanded data set does not result in a significant relationship between K_4 and GC_4 for individual genes. Nevertheless, an effect is apparent when the silent substitution rate is averaged over all genes within each 1% interval of silent-site G + C content (Fig 3b). With the exception of a few intervals containing only one or a small number of genes, K_4 appears to peak at approximately 60% GC_4, similar to what has been reported in other studies (Wolfe et al. 1989; Bulmer et al. 1991).

Thus, this study has documented the excessive variability of both substitution rates and base composition at silent sites in codons, but the relationship between the two remains unclear. In a recent analysis of sequences from three orders of mammal (Bulmer et al. 1991), we suggested that unknown factor(s) in addition to silent-site G + C content systematically affect the silent substitution rate (even after correcting for rate differences among lineages) in a manner that is consistent for a particular gene in different species. The discrepancy between the results of the present analysis and those based on the sequence data available four years ago (Filipski 1988; Wolfe et al. 1989; Ticher and Graur 1989) illustrates the need for very large data sets (and hence large numbers of nucleotide substitutions) before generalizations about substitution patterns or biases can be made with confidence. The data studied here comprise perhaps 0.5% of all genes in the mammalian genome and a total of 27,348 nucleotide differences between mouse and rat and so may constitute a representative sample.

The wide range of silent substitution rates and base compositions seen in groups of genes with similar functions (Appendix 1) also points to a lack of gene-expression-related constraint on codon usage. It has been suggested that mammalian codon usage is related to the major tissue of gene expression (Newgard et al. 1986), but this does not seem to be a general observation. Among the mouse-rat genes examined here, apolipoproteins A-II and A-IV are both expressed in liver, but have GC_S values of 60% and 78%, respectively (and serum albumin, also expressed in liver, has a GC_S value of 53%). There is also a wide variation in GC_S values among genes expressed in the testis (41% in a cytochrome *c* isoform, 78% in HSP70.2), brain (47% in calmodulin I; 79% in creatine kinase B), and pancreas (36% in α_2-amylase, 73% in ribonuclease), for example. Alternatively, there are 13 ribosomal protein genes listed in Appendix 1 with K_S values that range between 0.172 and 0.359, and GC_S values between 48% and 70%. Because they are highly expressed, ribosomal protein genes generally provide the clearest examples of selective codon usage in organisms where translational efficiency is at a premium (e.g., *E. coli*, Ikemura 1985; *Saccharomyces cerevisiae*, Sharp et al. 1986; or *Drosophila melanogaster*, Shields et al. 1988).

Table 2. Numbers of substituted nucleotides at adjacent codon positions 2 and 3[a]

		Position 3			
		Identical	Different	Ratio[b]	chi²
A. All sites:					
Position 2	Identical	52,060	9,225	1.67	63.1
	Different	600	202		
B. Excluding C_pG dinucleotides:					
Position 2	Identical	50,276	7,154	2.09	107.5
	Different	484	170		

[a] Only codons in which (in both species) position 3 is 4-fold degenerate are considered. Position 2 is always 0-fold degenerate.

[b] Ratio of [D:D/(D:D + D:I)]/[I:D/(I:D + I:I)], where I and D represent identical and different nucleotides, respectively, and a colon separates the two bases of a dinucleotide. For example, I:D indicates a dinucleotide where the first base (at codon position 2) is identical in the two species, and the second base (at position 3) is different

Correlation Between Rates of Synonymous and Nonsynonymous Substitution—Evidence for Doublet Mutations

In the first extensive investigation of nucleotide substitution rates at synonymous and nonsynonymous sites in mammalian genes (Li et al. 1985), there was a positive correlation across genes between the two rates (Graur 1985). A similar correlation was reported by Miyata et al. (1987a,b), and by Ticher and Graur (1989), in comparisons of 39 and 42 genes, respectively, between humans and rodents. In the present mouse-rat data set the correlation coefficient (weighting each gene by its length) between K_S and K_A is $r = 0.45$ (± 0.05), which is highly significant. This correlation is surprising because the two rates are expected to reflect different causes: the nonsynonymous rate should largely reflect protein sequence conservation, while the synonymous rate reflects the local mutation rate and any possible codon selection. The correlation could exist either because the two rates are similar over the whole gene (such that conserved proteins have low divergence at silent sites, for some reason) or because substitutions at adjacent nucleotide positions have occurred at a frequency greater than would be predicted from either substitution rate alone. Inspection of the 363 mouse-rat sequences suggests that the latter effect is occurring. For example, if the synonymous substitution rate is recalculated ignoring those codons where the species differ by more than a single nucleotide substitution, the resultant modified value of K_S is not significantly correlated with K_A ($r = 0.10 \pm 0.05$).

To look at this tandem substitution effect in more detail we identified homologous codons in the mouse and rat sequences where the third position is completely (4-fold) degenerate in both species. (Note that second positions are always 0-fold degenerate, i.e., nondegenerate.) These codons were then subdivided according to whether none, one, or both of these positions differ between the two species. The results reveal a highly significant excess of tandem nucleotide substitutions: where position 2 is unchanged, position 3 is changed in 15.1% of codons, but where a substitution has occurred at position 2, the fraction of differences at position 3 is 25.2% (Table 2A). It is well known (see, for example, Giannelli et al. 1991) that C_pG dinucleotides have a high mutation rate due to methylation of the cytosine (on either strand) followed by deamination, resulting in either a T_pG or a C_pA sequence. As a consequence, C_pG sites contribute a disproportionate number of single-base mutations, where an ancestral codon NCG is replaced by NTG or NCA in a descendant. If sites containing C_pG dinucleotides (in either species) are excluded from the analysis, the excess of tandem nucleotide substitutions becomes even more pronounced (Table 2B).

An excess of multiply-substituted codons was first commented upon by Fitch (1980), in a comparison of β-globin sequences from three orders of mammals. The question arises as to whether the excess of tandem substitutions arises from mutational events involving simultaneous substitution of two adjacent nucleotides or from two separate (consecutive) events. Lipman and Wilbur (1985) suggested that a nonsynonymous substitution may lead not only to an amino acid replacement but also to the replacement of an optimal codon for one amino acid by a nonoptimal codon for the new amino acid. Then, in the wake of the nonsynonymous substitution, a mutation leading to a synonymous change in the same codon may be positively selected as a way of generating an optimal codon for the newly specified amino acid. This seems unlikely because selection among synonymous codons is not thought to operate in many (if any) mammalian genes. (See above, and Sharp 1989.) Two other points should be noted. First, the selection coefficients driving sequences toward "re-optimized" codon usage could only be of the same order of magnitude as the nonsynonymous substitution rate. Second, if nonoptimal codons generated by a nonsynonymous mutation were selectively disavantageous, the nonsynonymous change would be expected to be selected against even if the amino acid replacement itself were neutral.

To test whether the tandem differences reflect synonymous codon selection, we have examined

Table 3. Numbers of substituted nucleotides at adjacent codon positions 3 and 1[a]

		Position 3			
		Identical	Different	Ratio[b]	chi[2]
A. All sites:					
Position 1	Identical	45,657	7,968	1.62	109.5
	Different	1,304	413		
B. Excluding C_pG dinucleotides:					
Position 1	Identical	44,097	6,167	2.07	222.5
	Different	1,096	373		

[a] Only codons in which (in both species) position 3 is 4-fold degenerate, and position 1 of the following codon is nondegenerate, are considered.

[b] Ratio of [D:D/(D:D + I:D)]/[D:I/(D:I + I:I)], where I and D represent identical and different nucleotides, respectively, and a colon separates the two bases of a dinucleotide. For example, I:D indicates a dinucleotide where the first base (at codon position 3) is identical in the two species, and the second base (at position 1 of the next codon) is different

the frequency of tandem differences at bases in neighboring codons (i.e., the 3:1 position involving the third base of a codon and the first base of the next codon). An excess of tandem substitutions, very similar in magnitude to that at 2:3 positions, is seen at the 3:1 position, both when all sites are considered (Table 3A) and when C_pG sites are excluded (Table 3B). This result *might* arise if codon pairs are under selection (see, e.g., Gouy 1987) but such selection should be secondary to selection within a codon, *if* that exists. Thus, from the magnitude of the excess of tandem substitutions at both 2:3 and 3:1 positions we conclude that this phenomenon most probably reflects mutational events simultaneously replacing both nucleotides. In support of this, we note that a significant excess of dinucleotide substitutions is seen in noncoding sequences from the primate eta-globin pseudogene region (our unpublished analyses of the data of Goodman et al. (1989)), where codon selection cannot be a factor. Furthermore, we have recently proposed that a significant excess of switches between the TCN and AGY groups of serine codons at highly conserved sites (in a range of proteins from a wide range of species) must also be due to doublet mutations (M. Averhof, K.H.W. and P.M.S., manuscript submitted).

Synonymous and nonsynonymous substitution rates have also been found to be correlated in studies on bacterial (Sharp and Li 1987; Sharp 1991) and chloroplast genes (Wolfe and Sharp 1988). In at least the case of bacteria, selection among synonymous codons can be a powerful force, and it will be interesting to investigate the possible causes of the K_S-K_A correlation in those species.

Conclusions

Data from individual genes have, in the past, pointed to the diversity of patterns of molecular evolution in mammalian genes. For example, the slow rate of evolution of actins (Alonso et al. 1986), the rapid rate of silent substitution in β_2-microglobulin (Li et al. 1985), and the extremely high G + C content of the metallothionein-I gene (Durnam et al. 1980) have all been reported. In this study, we have attempted to put these observations into a clearer perspective by investigating the mean and range of several measures of the molecular evolutionary process, calculated from a large number of genes compared between a single pair of species (and so having a single divergence time). It is realistic to hope that the large number of genes studied here may be truly representative of those in the mammalian genome.

We have examined the evolution of silent sites in these genes in some detail. Silent sites are of particular interest because they were initially expected to be neutral (King and Jukes 1969). However, while early investigations focused on the comparative homogeneity (relative to nonsynonymous substitution rates) of synonymous substitution rates among mammalian genes (Miyata et al. 1980; Kimura 1981), it is now clear that these rates vary quite substantially. It has also become apparent that alternative synonymous codons are not neutral in many genes in the genomes of bacteria, fungi, and even insects (reviewed in Ikemura 1985; Sharp et al. 1988; Andersson and Kurland 1990), and that different intensities of synonymous codon selection lead to variation in synonymous substitution rates among genes (Sharp and Li 1987, 1989; Sharp 1991). Since it is also clear that codon usage in mammalian genes is highly nonrandom, it might seem reasonable to speculate that codon selection accounts for the silent substitution rate variation; however, several lines of argument and evidence presented here suggest that this is not the case. First, different patterns of codon usage in different mammalian genes are not related to any obvious expression differences among the genes, and codon selection would not be expected to overcome random genetic drift in mammals due to their small effective population sizes. Second a slow rate of synonymous substitution in X-linked genes, and the correlation of nonsynonymous and synonymous rates among genes can each be attributed to mutational causes. Several hypotheses, based on models of variation in DNA polymerase or DNA repair activities, have been put forward to explain why mutation rates and patterns might vary among different regions of the mammalian genome (Filipski 1987, 1988; Sueoka 1988;

Wolfe et al. 1989; Wolfe 1991). However, whether mutational causes alone are responsible for all of the variation seen in GC_S and K_S remains to be established. More detailed analysis of the data is in progress and should provide further insight into mammalian gene evolution.

Acknowledgments. This work utilized the facilities of the Irish National Centre for Bioinformatics. We are grateful to Manolo Gouy, Des Higgins, Liz Cowe, and Andrew Lloyd for help with database access, to Anne-Marie Murphy for some preliminary data analysis, and to Michael Bulmer and Adam Eyre-Walker for discussion. We also thank the managers and sponsors of the electronic mail servers at EMBL and GenBank. This study was supported in part by EOLAS grant SC/91/603.

References

Alonso S, Minty A, Bourlet Y, Buckingham M (1986) Comparison of three actin-coding sequences in the mouse; evolutionary relationships between the actin genes of warm-blooded vertebrates. J Mol Evol 23:11–22

Andersson SGE and Kurland CG (1990) Codon preferences in free-living microorganisms. Microbiol Rev 54:198–210

Aota S, Ikemura T (1986) Diversity in G + C content at the third position of codons in vertebrate genes and its cause. Nucleic Acids Res 14:6345–6355, and correction 14:8702

Bernardi G, Olofsson B, Filipski J, Zerial M, Salinas J, Cuny G, Meunier-Rotival M, Rodier F (1985) The mosaic genome of warm-blooded vertebrates. Science 228:953–958

Bernardi G, Mouchiroud D, Gautier C, Bernardi G (1988) Compositional patterns in vertebrate genomes: conservation and change in evolution. J Mol Evol 28:7–18

Bulmer M, Wolfe KH, Sharp PM (1991) Synonymous nucleotide substitution rates in mammalian genes: implications for the molecular clock and the relationship of mammalian orders. Proc Natl Acad Sci USA 88:5974–5978

Catzeflis FM, Sheldon FH, Ahlquist JE, Sibley CG (1987) DNA-DNA hybridization evidence of the rapid rate of muroid rodent DNA evolution. Mol Biol Evol 4:242–253

Catzeflis FM, Nevo E, Ahlquist JE, Sibley CG (1989) Relationship of the chromosomal species in the Eurasian mole rats of the *Spalax ehrenbergi* group as determined by DNA-DNA hybridization, and an estimate of the spalacid-murid divergence time. J Mol Evol 29:223–232

Catzeflis FM, Aguilar J-P, Jaeger J-J (1992) Muroid rodents: phylogeny and evolution. Trends Ecol Evol 7:122–126

Chen YH, Pentecost BT, McLachlan JA, Teng CT (1987) The androgen-dependent mouse seminal vesicle secretory protein IV: characterization and complementary deoxyribonucleic acid cloning. Mol Endocrinol 1:707–716

Cohen DR, Hapel AJ, Young IG (1986) Cloning and expression of the rat interleukin-3 gene. Nucleic Acids Res 14:3641–3658

Dickerson RE (1971) The structure of cytochrome *c* and the rates of molecular evolution. J Mol Evol 1:26–45

Dickinson DP, Mirels L, Tabak LA, Gross KW (1989) Rapid evolution of variants in a rodent multigene family encoding salivary proteins. Mol Biol Evol 6:80–102

Dietrich W, Katz H, Lincoln SE, Shin H-S, Friedman J, Dracopoli NC, Lander ES (1992) A genetic map of the mouse suitable for typing intraspecific crosses. Genetics 131:423–447

Durnam DM, Perrin F, Gannon F, Palmiter RD (1980) Isolation and characterization of the mouse metallothionein-I gene. Proc Natl Acad Sci USA 77:6511–6515

Eyre-Walker A (1991) An analysis of codon usage in mammals: selection or mutation bias? J Mol Evol 33:442–449

Filipski J (1987) Correlation between molecular clock ticking, codon usage, fidelity of DNA repair, chromosome banding and chromatin compactness in germline cells. FEBS Lett 217:184–186

Filipski J (1988) Why the rate of silent codon substitutions is variable within a vertebrate's genome. J Theor Biol 134:159–164

Fitch WM (1980) Estimating the total number of nucleotide substitutions since the common ancestor of a pair of genes: comparison of several methods and three beta hemoglobin messenger RNA's. J Mol Evol 16:153–209

Giannelli F, Green PM, High KA, Sommer S, Lillicrap DP, Ludwig M, Olek K, Reitsma PH, Goossens M, Yoshioka A, Brownlee GG (1991) Haemophilia B: database of point mutations and short additions and deletions—second edition. Nucleic Acids Res 19 (suppl.):2193–2219

Goodman M, Koop BF, Czelusniak J, Fitch DHA, Tagle DA, Slightom JL (1989) Molecular phylogeny of the family of apes and humans. Genome 31:316–335

Gouy M (1987) Codon contexts in enterobacterial and coliphage genes. Mol Biol Evol 4:426–444

Gouy M, Gautier C, Attimonelli M, Lanave C, di Paola G (1985) ACNUC—a portable retrieval system for nucleic acid sequence databases: logical and physical designs and usage. Comp Appl Biosci 1:167–172

Graur D (1985) Amino acid composition and the evolutionary rates of protein-coding genes. J Mol Evol 22:53–62

Hartl DL, Clark AG (1989) Principles of population genetics. Sinauer Associates, Sunderland, MA

Higgins DG, Sharp PM (1989) Fast and sensitive multiple sequence alignments on a microcomputer. Comp Appl Biosci 5:151–153

Hughes AL, Nei M (1988) Pattern of nucleotide substitution at major histocompatibility complex loci reveals overdominant selection. Nature 335:167–170

Iizasa T, Taira M, Shimada H, Ishijima S, Tatibana M (1989) Molecular cloning and sequencing of human cDNA for phosphoriboxyl pyrophosphate synthetase subunit II. FEBS Lett 244:47–50

Ikemura T (1985) Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol 2:13–34

Ikemura T, Aota S (1988) Global variation in G + C content along vertebrate genome DNA. J Mol Biol 203:1–13

Jagodzinski LL, Sargent TD, Yang M, Glackin C, Bonner J (1981) Sequence homology between RNAs encoding rat alpha-fetoprotein and rat serum albumin. Proc Natl Acad Sci USA 78:3521–3525

Kimura M (1977) Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. Nature 267:275–276

Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16:111–120

Kimura M (1981) Estimation of evolutionary distances between homologous nucleotide sequences. Proc Natl Acad Sci USA 78:454–458

Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge, England

King JL, Jukes TH (1969) Non-Darwinian evolution. Science 164:788–798

Lawrence JG, Hartl DL, Ochman H (1991) Molecular considerations in the evolution of bacterial genes. J Mol Evol 33:241–250

Lipman DJ, Wilbur WJ (1985) Interaction of silent and replacement changes in eukaryotic coding sequences. J Mol Evol 21:161–167

Li W-H, Wu C-I, Luo C-C (1985) A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol Biol Evol 2:150–174

Li W-H, Tanimura M, Sharp PM (1987) An evaluation of the molecular clock hypothesis using mammalian DNA sequences. J Mol Evol 25:330–342

Miyata T, Yasunaga T, Nishida T (1980) Nucleotide sequence divergence and functional constraint in mRNA evolution. Proc Natl Acad Sci USA 77:7328–7332

Miyata T, Hayashida H, Kuma K, Yasunaga T (1987a) Male-driven molecular evolution demonstrated by different rates of silent substitutions between autosome- and sex chromosome-linked genes. Proc Jpn Acad Ser B 63:327–331

Miyata T, Hayashida H, Kuma K, Mitsuyasu K, Yasunaga T (1987b) Male-driven molecular evolution: a model and nucleotide sequence analysis. Cold Spring Harbor Symp Quant Biol 52:863–867

Miyata T, Kuma K, Iwabe N, Hayashida H, Yasunaga T (1990) Different rates of evolution of autosome-, X chromosome- and Y chromosome-linked genes: hypothesis of male-driven molecular evolution. In: Takahata N, Crow JF (eds) Population biology of genes and molecules. Baifukan, Tokyo, Japan, pp 341–357

Mouchiroud D, Gautier C, Bernardi G (1988) The compositional distribution of coding sequences and DNA molecules in humans and murids. J Mol Evol 27:311–320

Nei M (1987) Molecular evolutionary genetics. Columbia University Press, New York

Newgard CB, Nakano K, Hwang PK, Fletterick RJ (1986) Sequence analysis of the cDNA encoding human liver glycogen phosphorylase reveals tissue-specific codon usage. Proc Natl Acad Sci USA 83:8132–8136

Ochman H, Wilson AC (1987) Evolution in bacteria: evidence for a universal substitution rate in cellular genomes. J Mol Evol 26:74–86

Sharp PM (1989) Evolution at 'silent' sites in DNA. In: Hill WG, Mackay TFC (eds) Evolution and animal breeding; reviews on molecular and quantitative approaches in honour of Alan Robertson. C.A.B. International, Wallingford, UK, pp 24–32

Sharp PM (1991) Determinants of DNA sequence divergence between *Escherichia coli* and *Salmonella typhimurium:* codon usage, map position, and concerted evolution. J Mol Evol 33:23–33

Sharp PM, Li W-H (1987) The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. Mol Biol Evol 4:222–230

Sharp PM, Li W-H (1989) On the rate of DNA sequence evolution in *Drosophila*. J Mol Evol 28:398–402

Sharp PM, Tuohy TMF, Mosurski KR (1986) Codon usage in yeast: cluster analysis clearly differentiates between highly and lowly expressed genes. Nucleic Acids Res 14: 5125–5143

Sharp PM, Cowe E, Higgins DG, Shields DC, Wolfe KH, Wright F (1988) Codon usage patterns in *Escherichia coli, Bacillus subtilis, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogaster* and *Homo sapiens;* a review of the considerable within-species diversity. Nucleic Acids Res 16:8207–8211

Shields DC, Sharp PM, Higgins DG, Wright F (1988) "Silent" sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons. Mol Biol Evol 5:704–716

Sueoka N (1988) Directional mutation pressure and neutral molecular evolution. Proc Natl Acad Sci USA 85:2653–2657

Tajima F, Nei M (1984) Estimation of evolutionary distances between nucleotide sequences. Mol Biol Evol 1:269–285

Ticher A, Graur D (1989) Nucleic acid composition, codon usage, and the rate of synonymous substitution in protein-coding genes. J Mol Evol 28:286–298

Wilson AC, Carlson SS, White TJ (1977) Biochemical evolution. Ann Rev Biochem 46:573–639

Wilson AC, Ochman H, Prager EM (1987) Molecular time scale for evolution. Trends Genet 3:241–247

Wolfe KH (1991) Mammalian DNA replication: mutation biases and the mutation rate. J Theor Biol 149:441–451

Wolfe KH, Sharp PM (1988) Identification of functional open reading frames in chloroplast genomes. Gene 66:215–222

Wolfe KH, Sharp PM, Li W-H (1989) Mutation rates differ among regions of the mammalian genome. Nature 337:283–285

Zuckerkandl E, Pauling L (1962) Molecular disease, evolution, and genic heterogeneity. In: Kasha M, Pullman B (eds) Horizons in biochemistry. Academic Press, New York, pp 189–225

Appendix 1. Molecular evolution and base composition of 363 nuclear genes compared between mouse and rat

Gene	Identity (%)		Nonsynonymous		Synonymous		GC_S	ΔGC_S	Accession number	
	DNA	prot.	$K_A \times 100$	L_A	$K_S \times 100$	L_S	(%)	(%)	mouse	rat
7B2 neuroendocrine protein	95.7	97.1	1.2	493	17.1	137	53.3	-1.7	X15830	M63901
alpha-1 protease inhibitor (A1PI)	83.0	73.5	15.9	488	31.9	145	62.0	-2.9	X00945	D00675
nicotinic acetylcholine receptor epsilon (nACHRE)	94.8	96.1	1.8	1128	18.3	348	60.4	0.5	J04698	X13252
nicotinic acetylcholine receptor gamma (nACHRG)	92.9	97.1	1.3	236	30.3	76	77.7	3.5	X03818	X06364
lysosomal acid phosphatase	95.7	96.4	1.7	975	14.1	288	60.3	-0.3	X57199	M27893
alpha-actin (vascular smooth muscle)	94.7	99.5	0.2	886	28.1	246	64.7	-0.4	X07935	X06801
alpha-actin (cardiac)	95.4	100.0	0.0	221	25.8	58	63.2	-1.1	M15501	X00306
alpha-actin (skeletal muscle)	97.3	100.0	0.0	882	13.6	249	76.8	0.3	M12347	J00692
beta-actin (cytoplasmic)	96.6	99.7	0.1	875	16.2	251	72.8	-0.7	X03672	J00691
gamma-actin (smooth muscle)	95.1	100.0	0.0	882	25.7	246	63.2	-4.6	M26689	M22323
alpha-c-adaptin (subunit of clathrin AP-2 complex)	95.7	99.6	0.2	2167	19.5	647	63.5	2.7	X14972	X53773
alcohol dehydrogenase (ADH-1; liver)	90.2	89.9	5.0	865	31.9	260	57.0	-0.9	M11307	M15327
alpha-fetoprotein (AFP)	87.0	82.0	9.2	1439	36.3	376	53.4	-2.0	M16381	J00695
AGP/EBP transcription factor (LAP, IL6DBP, SFB)	98.7	99.3	0.3	668	4.7	220	95.4	0.3	M61007	X54626
alpha-lactalbumin	87.6	86.6	6.8	346	47.8	80	57.7	-6.5	M80909	X00461
delta-aminolevulinate dehydratase (ALAD)	95.5	97.3	1.3	749	16.0	241	62.5	0.4	X13752	M14479
serum albumin (ALB)	89.6	90.0	4.9	980	37.7	274	52.8	-5.8	M16111	J00698
aldolase A	96.5	99.2	0.4	833	14.6	259	65.3	2.1	Y00516	X04261
aldolase C	94.6	97.8	1.0	523	24.0	155	62.3	0.4	X03796	X06984
murinoglobulin (alpha(1)-inhibitor III)	87.7	81.0	10.1	3391	26.4	959	51.0	-0.5	M65238	X52984
alpha-2-amylase (pancreatic)	94.0	93.4	3.3	1200	18.9	309	36.1	0.5	J00360	J00703
amyloid beta protein	94.7	99.0	0.4	1635	26.5	450	65.3	3.1	M18373	X07648
androgen receptor (Tfm locus)	96.3	97.6	1.1	2085	13.8	612	58.6	0.0	M37890	M20133
atrial natriuretic factor (ANF, ANP)	93.2	93.4	3.5	347	19.9	109	65.6	-2.9	K02781	K02062
angiotensinogen	90.0	86.8	6.4	1080	26.1	351	73.9	-1.4	J03046	L00091
apolipoprotein A-II	82.2	69.6	15.6	239	41.7	67	60.0	4.9	X04119	X03468

Appendix 1. Continued

Gene	Identity (%)		Nonsynonymous		Synonymous		GC_S	ΔGC_S	Accession number	
	DNA	prot.	$K_A \times 100$	L_A	$K_S \times 100$	L_S	(%)	(%)	mouse	rat
apolipoprotein A-IV	88.3	81.5	9.4	917	26.4	253	77.8	1.6	M13966	J02588
apolipoprotein E	93.9	92.6	3.7	713	15.9	220	75.4	-0.5	M12414	J02582
a-raf oncogene	98.1	99.1	0.4	1007	7.3	304	65.7	1.0	M13071	X06942
argininosuccinate synthetase (ASS)	95.1	98.3	0.8	970	22.8	266	66.6	1.7	M31690	M36708
B23 nucleolar protein (N038 protein)	95.8	99.3	0.3	701	22.7	175	32.8	0.6	M33212	M25062
beta-2 adrenergic receptor	93.5	94.3	2.9	981	22.0	273	67.5	-0.2	X15643	X17607
beta-2 microglobulin	82.8	83.2	9.4	280	69.3	77	60.9	-1.8	X01838	Y00441
band 3 anion exchange protein	93.2	94.5	2.8	1909	21.6	635	70.4	-0.2	X02677	J04793
band 3-related protein (B3RP2)	94.8	98.2	0.8	2794	21.0	905	67.1	-0.7	J04036	J05166
band 3-related protein (AE3, B3RP3)	94.9	98.3	0.8	2772	20.5	909	69.3	0.5	M28383	J05167
immunoglobulin heavy chain binding protein (BiP, grp78)	94.4	98.6	0.6	346	32.6	80	42.0	-6.5	J03629	M14050
Blast-1/BCM1/MRC OX-45 antigen	82.8	67.1	19.0	563	23.0	157	48.6	-1.9	X17501	X13016
alkaline phosphatase (BLKP-ALP)	93.0	97.1	1.3	1210	30.4	362	70.5	-4.8	J02980	J03572
bone Gla protein (osteocalcin)	85.1	76.8	14.4	216	25.8	69	69.2	-1.9	X04142	X04141
calcium channel CaCh2a	96.9	100.0	0.0	698	16.3	184	72.5	0.0	M57973	M67516
calcium channel CaCh3b	95.1	100.0	0.0	675	28.1	168	68.1	-1.4	M57975	M57682
carbonic anhydrase III	92.2	93.8	2.8	606	30.5	174	61.8	-0.5	M27796	M22413
calcineurin A-alpha (catalytic subunit)	96.4	100.0	0.0	1229	19.0	334	53.3	0.3	J05479	M29275
calmodulin I (brain)	94.9	100.0	0.0	360	31.3	87	47.3	-5.3	M27844	M19312
calreticulin (CRP55, calregulin, HACBP)	96.5	98.3	0.7	1010	17.1	238	60.2	-1.7	X14926	X53363
calspermin / calmodulin-dependent protein kinase IV	91.8	91.4	4.1	1091	27.7	304	56.0	0.7	X58995	M64757
protein carboxyl methyltransferase	94.7	98.2	0.8	529	24.0	152	47.0	-4.2	M60320	M26686
carboxypeptidase H (carboxypeptidase E)	95.2	97.5	1.3	1114	19.1	314	63.3	0.2	X61232	J04625
beta-casein	89.3	79.7	10.4	521	15.8	160	51.9	-0.9	X04490	M11175
kappa-casein	87.7	77.0	12.0	406	18.8	128	44.0	-3.3	M10114	K02598
catalase	93.9	95.3	2.4	1244	22.8	337	58.2	-1.1	X52108	M11670
cathepsin B	92.3	92.6	3.5	798	27.6	219	55.1	0.8	M14221	M11305
cathepsin D	92.5	91.6	4.4	943	21.6	278	68.4	-1.5	X53337	X54467
cathepsin L	93.4	93.7	2.9	801	24.7	201	54.1	2.8	J02583	Y00697
NF-YA protein (CBF-B; CCAAT binding factor, subunit B)	96.4	99.7	0.1	788	17.2	232	53.4	-0.4	M86215	M34238
cholecystokinin (CCK)	95.2	98.4	0.7	140	21.0	43	72.4	-1.5	M11739	M10352
CD2 antigen (T11, OX-34)	87.2	79.2	11.0	794	26.0	232	54.5	0.4	M19799	X05111
CD4 antigen (L3T4, W3/25)	85.7	75.2	14.1	1061	23.1	304	66.1	1.3	X04836	M15768
CD43 antigen (leukosialin; sialophorin; Ly48)	83.4	73.9	16.2	825	27.5	300	52.4	-0.6	X17018	Y00090
CD8 alpha antigen (Lyt-2; MRC OX-8)	79.8	69.0	20.4	526	35.6	170	65.2	-0.4	M12052	X03015
CD8 beta antigen (Lyt-3)	86.1	77.3	11.9	478	29.4	143	63.5	4.0	M16799	X04310
CCAAT/enhancer binding protein (C/EBP)	95.8	98.3	0.8	802	16.0	269	86.8	-1.7	M62362	X12752
c-erbA alpha-2 oncogene	97.8	99.6	0.2	1058	10.0	298	71.2	0.2	X07751	M31174
c-fos oncogene	95.6	96.8	1.4	863	15.0	277	67.6	-1.4	J00370	X06769
chromogranin A	92.1	90.3	4.5	1068	23.4	297	65.7	1.2	M64278	X06832
creatine kinase B (brain)	94.7	98.7	0.6	885	25.1	258	79.5	-0.7	M74149	M18668
c-Ki-ras oncogene	97.6	97.9	0.9	231	9.4	57	38.9	5.1	X02452	M12259
creatine kinase M (muscle)	96.4	99.5	0.2	892	17.2	251	81.1	1.6	X03233	M10140
c-mos oncogene	93.2	92.3	3.6	767	18.9	247	69.3	-3.8	J00372	X00422
c-myc oncogene	95.5	97.7	1.0	1023	18.2	294	72.4	3.1	L00038	Y00396
complement component C3	91.4	90.2	4.8	3875	26.2	1111	66.8	0.2	K02782	X52477
connexin-32 (beta-1 gap junction protein)	95.8	98.5	0.8	634	16.5	191	72.8	0.2	M63802	X04070
connexin-43	95.2	99.7	0.1	899	24.9	247	68.0	-1.0	M61896	M19317
cytochrome c oxidase subunit IV (COX4)	92.7	92.9	3.3	403	26.6	104	67.7	-1.0	X54691	X54081
cytochrome c oxidase subunit Va (COX5A)	97.0	99.3	0.3	334	12.7	102	46.9	-0.3	X15963	X15030
cellular retinol binding protein I (CRBPI)	94.1	99.3	0.3	328	31.8	77	68.6	2.0	X60367	M19257
C-reactive protein (CRP)	85.1	74.9	14.1	524	27.6	145	53.5	2.3	X17496	M83176
alpha-A-crystallin	95.4	100.0	0.0	278	22.9	85	69.7	-3.0	J00376	J00715
beta-A3/A1-crystallin	95.5	97.7	1.0	433	22.4	98	51.3	3.2	J00378	X15143
gamma-A-crystallin (mouse gamma 4; rat gamma 1-1)	96.0	97.7	1.0	417	17.8	105	76.6	2.7	K02587	M19359
gamma-B-crystallin (mouse gamma 3; rat gamma 1-2)	94.0	97.1	1.5	333	29.5	81	61.0	-3.4	K02585	M19359
gamma-D-crystallin (mouse gamma 1; rat gamma 2-2)	94.9	96.0	1.7	418	20.3	104	78.2	0.6	K02583	M19361
gamma-F-crystallin (mouse gamma 2; rat gamma 4-1)	97.0	97.7	1.0	421	12.7	101	80.0	3.0	K02584	------
Cu-Zn superoxide dismutase (SOD)	95.1	96.8	1.7	358	18.1	104	47.0	1.0	X06683	Y00404
cyclophilin	97.0	97.6	1.0	388	11.3	104	58.1	-0.7	X52803	M19533
cytochrome P450 IIA3-2 (15-alpha hydroxylase-2)	95.0	95.7	1.9	1152	17.5	330	68.7	-1.6	M26204	J02852
cytochrome P450 17-alpha hydroxylase/C17-20 lyase	88.5	83.2	8.7	1186	28.2	335	61.7	-4.1	M64863	M31681
cystatin C	92.5	87.4	6.1	296	15.2	85	66.4	0.0	M59470	X16957
aspartate aminotransferase (cytosolic)	93.3	95.2	2.4	956	24.9	283	63.2	-1.9	J02623	D00252
cytochrome c (somatic)	97.5	100.0	0.0	251	13.9	64	48.6	5.6	X01756	K00750
cytochrome c (testis-specific)	93.1	96.2	1.6	253	34.6	62	41.0	0.0	M20625	M20624
cytokeratin endo A (number 8)	93.6	94.6	2.6	1134	23.1	315	78.1	-0.5	D90360	M63482
cytochrome P450IIB (pf3/46)	90.2	88.8	5.1	1141	32.2	332	60.6	0.8	M21856	M11251
cytochrome P-1-450 (P-450c)	92.8	93.1	3.6	1217	22.6	355	64.1	2.1	M10021	K02246
cytochrome P-3-450 (P-450d)	93.5	93.2	3.3	1198	20.6	341	72.3	2.6	M10022	K02422
cytochrome P450 IIE1 (Cyp2e1)	93.2	92.7	3.7	1161	21.2	318	56.2	-1.6	X62595	M20131
dopamine receptor, D2A subtype	97.1	100.0	0.0	1042	14.5	290	71.4	0.4	X55674	M36831
elongation factor 1 alpha (EF-1 alpha)	94.8	98.5	0.8	1073	23.4	310	51.6	-1.0	X13661	X63561
elongation factor 2 (EF-2)	93.7	98.9	0.5	620	31.8	184	70.3	2.9	M76131	Y07504
epidermal growth factor (EGF)	86.2	79.4	11.4	2614	29.3	773	52.9	-1.9	J00380	M63585
elastase II	88.4	84.5	8.9	616	25.9	197	72.0	-1.2	X04573	L00118
enkephalin A	93.5	97.5	1.0	578	34.2	142	64.3	1.1	M13227	K02805
alpha-enolase (non-neuronal)	93.9	96.1	1.9	1005	23.1	297	69.2	0.9	X52379	X02610
beta-enolase (muscle-specific)	93.3	97.0	1.8	1001	26.6	301	63.5	5.4	X61600	Y00979
gamma-enolase (neural-specific)	95.2	98.4	0.7	1007	20.9	295	68.2	-2.2	X52380	X07729
entactin	93.7	93.4	3.2	739	19.8	215	51.4	1.3	X14194	M15797
heart fatty acid binding protein (H-FABP)	91.5	92.5	4.0	311	29.5	88	68.6	0.9	X14961	M18034
intestinal fatty acid binding protein (I-FABP)	91.5	90.9	3.9	315	30.6	82	49.1	1.8	M65034	K01180
clotting factor IX (Christmas factor)	94.1	92.6	3.3	670	18.1	176	34.0	-1.2	M26236	M26247
fumarylacetoacetate hydrolase (FAH)	94.9	97.4	1.1	975	21.5	282	58.1	-1.4	M84145	M77694
fatty acid synthase (FAS)	90.9	94.6	2.7	1898	36.2	610	69.4	2.7	X13135	X13415
IgE high affinity receptor (Fc-epsilon-RI), alpha subunit	83.7	70.5	16.7	581	24.4	151	49.2	0.8	J05018	M17153
IgE high affinity receptor (Fc-epsilon-RI), beta subunit	89.1	83.8	8.6	548	24.2	157	45.5	-6.6	J05019	M22923
ferritin heavy subunit	96.4	97.3	1.4	433	13.3	113	62.1	0.0	J03941	M18051
ferritin light subunit	94.7	94.5	2.7	422	15.7	127	70.5	0.8	J04716	J02741
acidic fibroblast growth factor (FGFA, HBGF1)	95.3	100.0	0.0	364	23.5	102	66.3	2.0	M30641	X14232
basic fibroblast growth factor (FGFB)	96.1	98.1	0.9	353	14.8	110	64.6	-3.8	M30644	X07285
phenylalanine hydroxylase	95.4	97.8	1.0	1059	19.2	300	56.0	0.0	X51942	M12337
fibronectin	92.9	98.1	0.8	248	40.6	61	52.3	-3.7	M18194	L00191
flk (ferT) tyrosine kinase	93.4	97.2	1.2	768	32.4	201	52.3	0.9	M32054	X13412
furin	94.1	96.7	1.5	1821	23.3	558	70.0	1.7	X54056	X55660
Kex2 homologue (furin-like protein)	95.2	99.2	0.3	1493	23.6	418	60.1	-1.1	M55669	M83746
prohormone cleavage enzyme (furin-like protein)	94.4	95.1	2.2	1775	20.6	481	41.0	-0.3	M69196	M83745
GABA-alpha receptor, delta subunit	95.3	99.6	0.2	1036	22.2	311	69.7	-0.4	M60596	M35162
GABA-alpha receptor, gamma-3 subunit	96.6	98.9	0.5	1099	15.6	302	52.8	-0.2	X59300	M81142
GAP-43 (P-57 neural-specific protein)	98.1	99.1	0.4	530	7.8	148	47.4	2.2	J02809	M16228
glyceraldehyde-3-phosphate dehydrogenase (GAPDH)	95.0	97.0	1.4	775	19.9	224	70.0	2.6	M32599	X02231
gas3 (growth arrest specific) / SR13 myelin protein	95.0	97.5	1.1	367	19.9	113	73.4	-0.3	M32240	M69139

Appendix 1. Continued

Gene	Identity (%)		Nonsynonymous		Synonymous		GC_S	ΔGC_S	Accession number	
	DNA	prot.	$K_A \times 100$	L_A	$K_S \times 100$	L_S	(%)	(%)	mouse	rat
guanylate binding protein GBP-1 (mag-1)	84.7	80.1	11.1	1399	43.8	368	55.8	-0.8	M55544	M80367
guanylate cyclase/ANF receptor	95.4	97.0	1.4	2415	16.4	756	66.4	-1.9	J05504	J05677
glucocorticoid receptor	94.0	95.1	2.2	1835	22.2	514	44.3	0.0	X04435	M14053
growth hormone (somatotropin; GH)	95.4	96.3	1.6	504	16.8	144	75.1	2.2	K03232	J00739
growth hormone receptor (high MW form; GHR)	91.8	89.3	5.4	1508	22.3	406	45.1	-4.4	M33324	J04811
growth hormone releasing factor (GHRF)	86.2	71.8	13.8	243	23.0	66	70.3	7.9	M31658	X02322
glutamine synthetase	92.4	94.1	2.9	881	31.0	229	64.5	-2.5	M60803	X07921
beta-major globin	91.7	91.8	4.9	339	24.2	102	65.8	2.1	J00413	M17084
beta-minor globin	89.4	88.4	7.5	337	26.4	104	66.2	4.6	V00722	X06701
alpha globin	88.3	84.5	8.7	324	27.2	102	68.9	-0.5	J00410	M17083
glucokinase	97.8	98.2	0.8	395	8.1	103	74.6	-1.0	M58755	M58759
glutamate dehydrogenase (GLUD)	94.7	99.1	0.4	1299	26.0	376	59.2	-1.7	X57024	X14044
glutathione peroxidase (GSHPX)	94.2	94.5	2.5	454	18.3	146	73.3	1.3	X03920	X07365
glucose transporter, facilitated (brain) (GT1)	95.7	99.0	0.4	1119	18.2	357	73.9	-1.7	M22998	M22061
glucose transporter, insulin-regulated (GT2, IRGT)	95.2	97.2	1.3	1137	16.5	387	63.0	1.3	M23383	X14771
glucose transporter, liver type	93.7	94.6	2.4	1202	22.2	364	54.5	-0.2	X15684	J03145
gonadotropin releasing hormone (GnRH)	93.0	90.0	4.4	212	17.8	58	58.3	3.0	M14872	M31670
glycoprotein hormones, alpha subunit (TSHA)	93.7	94.2	2.5	285	24.3	75	53.8	-2.8	J00643	J00757
G-alpha-o protein, A isoform	97.4	99.4	0.2	843	12.7	219	68.2	1.3	M36777	M17526
G-alpha-i2 protein	95.7	99.2	0.5	847	21.4	218	75.0	1.1	M13963	M12672
G-alpha-s protein	97.8	99.5	0.2	865	10.4	227	76.0	-0.5	M13964	M12673
glutathione S-transferase Ya	94.3	95.5	2.2	528	21.3	138	57.0	-2.5	M19250	K00136
glutathione S-transferase Yb	94.1	93.1	3.1	524	20.0	130	70.9	-0.7	J03952	X04229
beta-glucuronidase (GUS)	89.9	87.7	6.0	1511	30.1	433	65.9	-0.1	M19279	M13962
hck (bmk) tyrosine kinase	94.9	97.6	1.1	1183	22.1	326	68.1	0.4	Y00487	M83666
heme oxygenase	93.2	93.4	3.2	672	22.6	195	68.3	0.7	X13356	J02722
herculin (muscle regulatory factor 4)	96.7	99.2	0.4	563	14.6	163	60.0	1.7	M30499	M27151
hexokinase I (brain)	93.3	96.2	2.1	1055	26.5	301	68.2	0.2	J05277	J03228
heparin binding protein / Heymann nephritis antigen (gp330)	93.8	95.5	2.9	743	22.2	196	76.2	1.2	D00622	M31051
H+,K+ ATPase beta subunit (gastric)	93.7	94.9	2.3	699	25.1	183	73.6	-3.2	M80251	M35535
hepatocyte nuclear factor 1 (HNF-1)	95.0	99.2	0.4	1424	22.2	460	76.2	-0.5	M57966	X54423
hypoxanthine phosphoribosyltransferase (HPRT)	96.7	98.2	0.8	519	14.5	136	40.8	-1.9	K01515	M63983
hsc70 (heatshock cognate protein)	96.2	99.8	0.0	1512	18.7	426	52.6	0.4	M19141	Y00054
HSD3B (delta-5-3-betahydroxysteroid dehydrogenase) type I	91.2	87.9	6.3	872	21.0	247	57.8	-3.2	M58567	M38178
HSP27 heatshock protein	92.4	97.1	1.8	460	30.0	155	75.2	2.0	X14687	M86389
HSP60 heatshock protein (groEL homologue)	97.8	99.6	0.2	1291	9.6	375	34.1	0.6	X53584	X54793
HSP70.2 heatshock protein (hst70, testis-specific)	96.7	99.4	0.4	1479	14.6	420	78.1	1.0	M20567	X15705
islet amyloid polypeptide (IAPP)	91.5	90.3	5.0	210	23.3	69	54.1	-0.4	M25389	X52820
interferon gamma	88.7	83.9	7.0	378	39.1	87	60.8	0.8	K00083	X02325
immunoglobulin heavy chain C-delta	84.3	69.7	16.8	466	21.5	137	55.9	-5.6	J00450	J00741
immunoglobulin heavy chain C-epsilon	88.0	82.4	9.3	901	27.8	257	63.9	-1.6	J00476	J00744
immunoglobulin heavy chain C-gamma-1	87.4	81.4	10.2	753	27.4	216	64.0	3.0	J00453	------
immunoglobulin heavy chain C-gamma-2b	81.6	73.8	17.3	771	36.1	225	65.0	-7.2	J00461	------
immunoglobulin heavy chain C-gamma 3 (rat gamma 2c)	87.4	79.0	11.3	762	23.7	225	62.2	-2.0	J00451	X07189
insulin-like growth factor I, B form (IGF-IB)	94.0	94.7	2.3	305	20.8	94	61.8	0.2	X04482	Y00429
insulin-like growth factor II (IGF-II)	97.8	96.7	1.5	410	4.8	130	71.0	-2.1	M14951	M13969
immunoglobulin light chain C-kappa	88.4	80.2	10.4	251	21.6	67	55.0	4.5	L00089	J02575
immunoglobulin light chain C-lambda-2	89.3	81.9	8.8	243	20.5	72	52.1	6.9	J00595	M22521
immunoglobulin light chain V-lambda-1	89.1	80.2	10.1	262	17.9	86	43.1	-2.1	J00590	M17092
interleukin 1 alpha	89.1	83.0	8.7	642	24.9	168	51.7	3.1	X01450	D00403
interleukin 1 receptor antagonist	93.3	89.3	5.3	421	14.3	113	58.4	0.4	M74294	M63101
interleukin 2	89.5	81.9	8.7	368	22.4	97	54.4	1.7	X01665	M22899
interleukin 3	78.0	59.0	27.3	383	24.4	115	55.1	0.5	K03233	X03846
interleukin 4	76.9	60.0	25.1	289	39.6	86	58.8	-1.3	X05253	X16058
interleukin 6	91.7	85.3	6.8	499	17.1	134	44.1	-1.9	M20572	M26745
interleukin 6 receptor	92.2	89.3	5.1	1047	19.8	324	65.4	4.4	X53802	M58587
insulin I	91.4	93.5	3.3	246	29.1	78	74.8	0.4	X04725	J00747
insulin II	94.0	94.5	2.4	251	20.1	79	75.4	3.0	X04724	J00748
inositol-1,4,5-triphosphate receptor (InsP3R)	95.5	99.5	0.2	6482	22.5	1763	62.4	1.0	X15373	J05510
insulin receptor	94.7	98.7	0.6	3206	25.3	907	62.9	-4.1	J05149	M29014
interferon regulatory factor 1 (IRF-1)	91.5	93.6	3.2	785	36.3	199	62.3	-2.4	M21065	M34253
JE gene (PDGF-inducible cytokine)	90.8	81.8	9.1	342	13.0	102	52.7	1.0	M19681	X17053
myelin proteolipid protein (PLP; jimpy locus)	98.2	100.0	0.0	637	8.1	194	56.6	0.3	M15442	X02809
jun-B oncogene	97.7	99.4	0.3	768	8.8	264	74.5	0.4	J03236	X54686
c-jun oncogene (AP-1 transcription factor)	96.5	99.4	0.3	765	15.1	237	79.6	1.6	J04115	X17163
kallikrein binding protein	82.9	71.9	16.1	965	31.7	274	61.6	0.0	X61597	M67496
KC gene (PDGF-inducible)	92.4	88.5	5.3	218	17.5	70	70.8	-3.2	J04596	M86536
lysosomal membrane glycoprotein lgpA/LAMP1/lgp120	89.1	84.2	8.4	935	24.4	283	63.0	-4.5	M32015	M34959
lysosomal membrane glycoprotein lgpB/LAMP2/lgp96	89.6	82.4	9.2	956	18.7	274	43.7	-2.1	J05287	D90211
lecithin-cholesterol acyltransferase (LCAT)	91.8	91.8	4.0	1011	26.3	303	64.5	5.1	J05154	X54096
lactate dehydrogenase A (LDHA)	93.1	96.4	1.6	777	30.9	219	63.2	3.5	X02520	X01964
14 kDa lectin (beta-galactoside binding)	94.6	95.6	1.9	319	21.2	86	73.1	-0.2	X53067	M19036
homeoprotein LFB3 (variant hepatic nuclear factor 1)	96.7	97.8	1.1	1297	11.6	374	71.6	1.1	X55842	X56546
luteinizing hormone receptor (LHR)	93.1	94.1	2.7	1618	24.6	482	57.1	-2.6	M81310	M68928
leukaemia inhibitory factor (LIF)	90.6	91.1	4.7	468	30.9	138	69.1	-1.4	X12810	M32748
lipocortin I	93.9	93.9	2.9	816	20.1	222	46.2	-1.6	X07486	Y00446
lysyl oxidase (rrg gene)	93.7	95.4	2.0	949	23.7	284	63.8	2.8	M65142	J02903
muscarinic acetylcholine receptor M1	95.7	98.5	1.0	1044	16.0	336	70.5	-3.5	J04192	M16406
myelin-associated glycoprotein (MAG, 1B236; shorter form)	95.6	98.5	0.7	1330	18.3	416	71.8	-1.5	M74793	M22357
malic enzyme (cytosolic, mod-1 locus)	93.3	96.5	1.7	1336	28.5	380	50.7	1.9	J02652	M26594
MAP2 microtubule-associated protein	93.8	92.4	3.5	4253	17.4	1228	42.4	-0.7	M21041	X51842
matrix Gla protein	92.0	88.3	5.7	242	18.4	67	69.6	1.8	D00613	J03026
myelin basic protein (14 kDa; MBP)	96.9	99.2	0.3	287	13.1	91	59.0	0.0	L00398	K00512
MDR1 multidrug resistance protein	91.9	92.9	3.6	2987	28.7	835	52.1	-3.9	M14757	M81855
metallothionein I	96.2	95.1	2.1	146	11.5	37	88.3	0.0	J00605	J00750
metallothionein II	95.7	96.7	1.4	148	19.0	35	75.8	-1.7	K02236	M11794
alpha-cardiac myosin heavy chain	95.2	98.8	0.5	4578	23.0	1233	80.1	-1.4	M76598	X15938
MHC-Ia invariant gamma chain (Ii)	90.3	86.0	7.2	500	23.4	145	71.7	-2.5	X05428	X13044
potassium channel (MBK1, RCK1; shaker homologue)	98.0	99.4	0.3	1169	9.0	316	71.2	0.6	Y00305	X12589
potassium channel (MK2, BK2)	97.9	100.0	0.0	1180	10.7	317	57.8	-0.4	M30440	J04731
potassium channel (MK3, Kv3)	95.6	99.2	0.3	1211	20.1	364	74.3	-1.2	M30441	M31744
myosin light chain (MLC, alkaline, fast skeletal muscle)	96.5	99.3	0.3	357	18.2	93	63.6	-3.0	K02238	K02424
Mn superoxide dismutase (SOD)	94.2	94.1	2.8	520	18.9	146	62.1	-0.7	X04972	Y00497
aspartate aminotransferase (mitochondrial)	95.2	98.4	0.7	998	21.1	292	66.6	-0.5	J02622	M18467
malate dehydrogenase (mitochondrial MDH)	94.4	98.2	0.8	770	23.9	244	63.3	2.2	M16229	X04240
mitochondrial uncoupling protein (UCP)	93.0	96.7	1.6	702	28.8	219	57.8	-3.6	M21222	M11814
Mx1 protein (resistance to influenza)	86.9	79.4	11.1	1484	28.3	391	58.8	-2.1	M21117	X52711
P0 peripheral myelin protein (Schwann cell)	96.5	97.6	1.1	569	12.3	175	68.3	-2.0	M62860	K03242
myosin heavy chain (embryonic skeletal muscle)	94.9	98.1	0.9	376	23.4	98	72.6	-1.8	M11154	X04267
myogenin	96.0	98.7	0.6	515	17.1	155	72.9	1.5	D90156	M24393
Na+,K+ ATPase beta subunit (brain)	94.3	96.4	1.6	726	25.0	186	67.0	0.4	X16646	M14137
Na+,K+ ATPase beta-2 / glial cell adhesion molecule (AMOG)	94.2	96.9	1.5	691	26.3	179	66.2	0.7	X16645	J04629
neural cell adhesion molecule NCAM-140	95.0	98.1	0.8	1984	22.7	557	65.8	1.2	X06328	X06564

Appendix 1. Continued

Gene	Identity (%)		Nonsynonymous		Synonymous		GC_S	ΔGC_S	Accession number	
	DNA	prot.	$K_A \times 100$	L_A	$K_S \times 100$	L_S	(%)	(%)	mouse	rat
neural cell adhesion molecule L1	96.0	97.3	1.2	2943	15.4	822	61.4	1.8	X12875	X59149
NFI-B (NF-1 like DNA-binding protein; clone B4)	97.5	99.4	0.4	1151	9.7	349	52.0	-1.3	D90175	X13167
neurofilament heavy chain (NF-H)	90.4	88.8	5.5	1364	30.2	373	61.7	-3.9	M21497	M21964
neurofilament light chain (NF-L; 68 kDa)	96.2	97.2	1.4	670	14.3	173	75.3	2.0	M13016	M25638
neurofilament middle chain (NF-M)	95.2	97.0	1.3	2008	20.7	527	70.7	2.3	X05640	M18628
nerve growth factor beta (NGFB)	95.3	95.9	1.8	567	16.4	169	60.3	0.8	M35075	M36589
NGF-IA transcription factor (zif/268, Egr-1, krox-24)	96.2	98.3	0.7	1218	15.0	381	64.1	1.7	M22326	M18416
N-myc oncogene	94.1	96.8	1.6	1054	22.2	332	74.9	0.6	X03919	X63281
nucleolin	95.3	94.3	2.6	1667	14.2	448	33.2	0.7	X07699	M55022
O6-methylguanine-DNA methyltransferase	93.8	92.3	3.7	475	16.4	152	60.0	1.5	M84524	M76704
ornithine decarboxylase (ODC)	94.3	97.0	1.3	1089	25.4	294	49.1	0.3	M10624	M16982
oestrogen receptor	94.5	97.8	1.0	1403	25.0	394	66.9	2.7	M38651	Y00102
osteoblast specific factor (OSF-1)	95.9	100.0	0.0	401	23.1	103	59.8	2.3	D90225	M55601
osteopontin (minopontin)	89.5	84.6	8.1	697	25.1	182	53.7	-4.5	X13986	M14656
ornithine transcarbamylase (OTC; sparse fur locus)	96.2	97.2	1.2	826	13.8	236	43.8	-2.9	X07092	K03040
basigin / MRC OX47 antigen	89.1	85.7	7.5	626	28.0	190	70.7	-6.5	D00611	X54640
p53 oncogene	90.7	89.1	5.2	891	28.1	267	59.8	2.2	K01700	X13058
plasminogen activator inhibitor 1 (PAI-1)	92.6	89.1	4.9	932	18.7	275	69.9	0.3	M33960	J05206
pancreatic polypeptide	87.9	83.7	8.1	221	32.8	73	61.1	8.6	M18208	M18207
porphobilinogen deaminase (PBG-D)	95.6	96.2	1.9	789	13.9	243	59.0	0.2	M28666	X06827
proliferating cell nuclear antigen (PCNA)	95.3	97.3	1.1	615	19.8	168	48.9	-1.8	X53068	M24604
alpha-PDGF receptor	92.1	95.0	2.4	2562	33.8	702	62.0	-2.8	M57683	M63837
protein disulphide isomerase (PDI)	93.3	96.5	1.7	1200	29.8	324	63.2	-0.4	X06453	X02918
neuron-specific protein PEP-19 (pcp-4)	97.4	100.0	0.0	150	15.5	36	41.1	4.8	X17320	M24852
perforin (cytolysin)	89.2	84.7	8.3	1279	24.2	383	64.7	1.0	X12760	M33605
peripherin (intermediate filament)	95.7	98.8	0.6	938	18.8	280	78.2	-1.3	X15475	M26232
liver phosphofructokinase (PFK-L)	94.2	99.1	0.4	1793	27.9	547	73.2	-4.9	J03928	X58865
proteoglycan core protein	85.7	83.4	9.5	352	43.9	101	48.9	4.8	M33499	J03224
phosphoglycerate kinase (PGK1)	96.5	99.3	0.3	970	16.0	281	53.9	-0.8	M15668	M31788
phosphorylase kinase, gamma subunit	94.1	98.5	0.7	913	29.4	251	70.6	-0.6	M16216	X07320
phospholipase C-alpha (PIPLC)	95.5	96.6	1.5	1187	17.6	316	47.7	-1.4	M73329	X12355
pit-1 transcription factor (GHF-1, dwarf locus)	95.0	97.9	0.9	682	22.3	191	52.0	3.1	------	M23253
protein kinase C alpha	94.7	98.7	0.7	1600	27.0	416	64.4	-1.7	M25811	X07286
cAMP-dependent protein kinase, C-alpha subunit	96.7	99.1	0.5	837	15.6	216	72.4	0.6	M19960	X57986
cAMP-dependent protein kinase, R-I-alpha subunit	95.1	99.0	0.5	888	22.8	255	54.5	0.6	M20473	M17086
cAMP-dependent protein kinase, R-II-alpha subunit	92.5	94.6	2.5	878	31.5	232	52.5	1.8	J02935	J02934
protein kinase C beta-II	96.6	99.9	0.0	1604	17.8	416	64.5	0.0	X53532	X04440
protein kinase C delta	94.1	94.9	2.5	1608	21.5	411	72.6	-2.0	X60304	M18330
protein kinase C epsilon	95.6	99.1	0.4	1737	21.1	474	65.9	-1.9	------	M18331
placental lactogen I	85.2	74.9	13.1	524	30.6	145	44.2	-1.1	M35662	M55269
placental lactogen II	86.5	77.8	11.5	513	28.1	150	45.1	0.7	M14647	M13749
plasma kallikrein	91.9	90.3	4.7	1511	24.8	403	46.6	0.3	M58588	M62357
pro-opiomelanocortin (POMC)	93.6	93.6	3.3	547	19.2	158	82.1	0.4	J00611	K01878
prolactin receptor (PRLR)	93.3	93.9	2.8	666	25.2	174	43.1	-4.8	M22958	M19304
prolactin (PRL)	90.2	84.5	7.5	531	22.7	147	53.5	1.6	K03236	J00766
prion protein (PrP)	95.0	97.3	1.1	528	21.6	147	67.9	4.6	M13685	M20313
parathyroid hormone related peptide (PTHRP)	95.5	96.6	1.7	398	14.9	127	70.9	-2.8	M60057	M34112
pulmonary surfactant protein C	92.3	92.7	3.6	443	25.6	136	62.2	3.3	M38314	X14221
pyrimidine binding protein 1 (PYBP1)	94.9	99.0	0.5	1209	22.7	366	72.7	0.5	X52101	X60789
retinol binding protein (RBP4)	97.8	98.3	0.7	419	7.8	118	70.4	1.1	M74527	K03046
retinal degeneration slow (rds locus, peripherin)	94.5	97.1	1.2	812	23.8	226	81.3	-1.5	X14770	X52376
renin I	87.6	84.1	8.5	927	34.0	279	67.9	-1.8	X00810	J02941
insulinoma rig gene	92.2	100.0	0.0	338	43.6	97	82.4	-1.1	M33330	M19393
pancreatic ribonuclease	85.1	77.2	12.2	353	36.4	94	72.9	-6.1	M27814	J00771
ribosomal protein L19	94.9	99.5	0.2	460	24.6	128	70.1	4.5	M62952	J02650
ribosomal protein L27a (homologue of yeast rpL29)	93.7	98.0	0.9	345	27.7	99	59.8	-1.5	X05021	X52733
ribosomal protein L3 (J1 protein)	94.3	99.5	0.2	946	29.9	263	62.4	1.0	Y00225	X62166
ribosomal protein L30	96.8	100.0	0.0	274	17.2	71	48.0	3.3	K02928	K02932
ribosomal protein L32	96.6	100.0	0.0	317	17.5	88	67.6	2.4	K02060	X06483
ribosomal protein L7	93.5	98.0	0.9	591	31.0	165	57.8	-0.3	M29015	M17422
ribosomal protein L7a (surf3 locus)	95.3	99.6	0.2	609	22.1	189	66.5	-2.1	M14689	X15013
acidic ribosomal phosphoprotein PO	92.1	97.8	1.3	725	35.9	223	68.9	-2.1	X15267	X15096
ribosomal protein S12	96.2	100.0	0.0	305	18.0	86	59.7	-2.6	X15962	M18547
ribosomal protein S16	94.5	98.6	0.9	329	22.6	106	62.6	0.6	M11408	X17665
ribosomal protein S24	95.2	99.2	0.3	306	21.7	84	49.6	0.7	X60289	X51538
ribosomal protein S4 (RPS4X)	95.2	99.6	0.3	613	23.6	173	50.1	0.9	M73436	X14210
ribosomal protein S6	94.0	100.0	0.0	578	30.7	170	49.0	7.4	Y00348	J03538
RT6.1 (RT6-a) alloantigen	80.2	71.3	18.1	654	45.4	171	45.1	-3.5	X52991	M31138
retinoic X receptor beta (RXR-beta, RCoR-1)	96.7	98.9	0.6	1013	12.7	331	66.6	0.0	M84818	M81766
calpactin I light chain (p11, S100-like clone 42C)	95.5	94.7	2.2	230	15.8	55	57.1	4.1	M16465	J03627
S100-like calcium binding protein (clones pEL98, 42A/eX15)	97.4	96.0	1.7	243	7.0	60	69.6	-2.3	D00208	J03628
S-antigen (retinal/pineal)	94.6	97.0	1.4	936	21.9	273	63.9	3.0	M24086	X15353
serum amyloid P (SAP)	87.6	79.9	11.0	527	21.6	145	46.9	-1.9	M29535	------
SCIP transcription factor (Tst-1)	98.4	99.6	0.2	998	5.6	349	84.8	1.3	X56959	M72711
sterol carrier protein 2 (SCP-2)	94.0	96.3	1.6	495	23.9	147	57.1	1.2	M62361	M57453
J6 serpin homologue / collagen binding protein (gp46)	94.5	97.6	1.2	951	21.7	291	79.4	2.1	J05609	M69246
serotonin 5-HT 1c receptor	96.3	97.6	1.2	1070	13.7	307	53.1	-1.8	M63685	M21410
potassium channel, delayed-rectifier (drk1; shab-like)	94.3	97.3	1.3	1988	24.4	571	80.0	0.5	M64228	X16476
potassium channel, (NGK2, Kv4; shaw-like)	96.7	100.0	0.0	1161	16.0	343	78.3	0.8	Y07521	M68880
snRNP-associated Sm-B protein (Sm11)	95.5	98.1	1.1	473	15.5	169	46.5	0.7	M58761	M29295
snRNP-associated Sm-N protein (Sm51)	97.6	100.0	0.0	535	9.8	185	33.5	2.6	X62648	M29293
somatostatin	98.0	100.0	0.0	264	8.8	84	74.2	-0.8	X51468	K02248
spot-1 salivary protein	69.6	56.0	37.2	207	51.3	66	35.2	-5.1	M33974	M33976
calbindin D28 (spot 35 protein)	94.6	99.4	0.3	386	31.6	94	51.5	-5.8	M23663	M31178
stearoyl-CoA desaturase	92.3	93.2	3.3	827	27.4	238	74.7	-1.9	J04190	J02585
mast cell growth factor (SCF, steel locus, c-kit ligand)	95.5	95.0	2.6	479	13.1	124	37.6	1.9	M38436	M59966
substance K receptor (SKR)	93.3	95.3	2.1	889	25.1	263	73.7	-1.1	X62933	M31838
substance P receptor (SPR)	95.5	98.8	0.5	952	20.5	269	73.5	-6.4	X62934	M64236
seminal vesicle secretory protein IV (SVS S)	74.8	69.0	21.2	234	78.0	66	46.7	-10.2	M35732	M16069
syndecan	91.8	90.6	4.7	625	22.5	206	59.7	1.4	X15487	X60651
CD45 antigen (T200, L-CA, Ly-5) longest form	86.2	79.3	11.8	2954	28.8	805	43.5	-1.4	M23126	Y00065
t complex polypeptide 1b	94.5	97.7	1.1	1228	23.4	359	41.3	0.1	M12899	D90345
T-cell receptor C-alpha	85.2	78.7	12.2	321	35.9	87	63.8	-2.3	X01133	M18853
T-cell receptor C-beta	90.9	88.7	6.5	390	21.8	114	64.9	0.9	K01080	M18854
T-cell receptor V-beta (VAK, V-beta-11)	89.8	86.0	8.1	265	21.7	77	49.4	0.4	N00046	M21817
transferrin receptor	90.7	85.1	8.7	385	15.9	98	36.0	0.6	M29618	M58040
transforming growth factor (TGF) beta-1	96.3	98.7	0.7	899	14.9	271	73.4	-0.1	M13177	X52498
thrombin	93.4	93.5	3.1	1439	21.7	412	66.4	0.6	X52308	X52835
thy-1 antigen	88.9	83.2	10.0	367	19.3	116	74.3	6.1	M10246	X02002
thymidine kinase (TK)	92.6	95.0	3.1	285	27.7	78	72.9	-1.8	M11945	M22642
thymosin alpha	97.0	98.2	0.8	271	14.2	62	53.5	4.4	X56135	M20035
thymosin beta-4	96.3	97.7	0.9	108	20.6	24	45.5	-4.5	X16053	K01334
TIS11 gene (tristetraproline; begins MetAspLeu....)	94.7	96.9	1.7	707	17.4	250	63.6	0.5	M58565	X63369

Appendix 1. Continued

Gene	Identity (%)		Nonsynonymous		Synonymous		GC_S	ΔGC_S	Accession number	
	DNA	prot.	$K_A \times 100$	L_A	$K_S \times 100$	L_S	(%)	(%)	mouse	rat
TIS11 gene (rat clone cMG1; begins MetThrThr....)	97.7	99.7	0.4	770	8.6	244	77.0	-0.6	M58566	X52590
TIS21 gene / PC3 (NGF-inducible protein)	94.3	97.5	1.1	370	25.3	104	78.2	-5.8	M64292	M60921
TIS7 gene / PC4 (interferon-related protein)	96.1	97.1	1.3	1050	13.9	297	43.3	-3.7	X17400	J04511
tumor necrosis factor (TNF) alpha	93.6	94.5	2.6	542	21.8	163	72.6	-3.2	Y00467	D00475
TNF receptor (Goodwin TNFR2; Lewis TNFR1)	89.4	83.9	8.5	1043	22.0	316	66.4	1.2	M59377	M63122
transition protein 1 (TP1)	96.4	98.2	1.6	128	8.6	37	58.5	0.0	X12521	M17096
transition protein 2 (TP2)	89.0	84.2	8.4	265	25.9	77	67.0	6.8	J03494	X14776
tissue plasminogen activator (tPA)	91.0	91.9	3.9	1312	33.6	365	61.3	-2.0	J03520	M23697
transin-1 (pTR1)	91.7	89.9	4.9	1118	24.7	307	46.5	-0.5	X63162	X02601
transthyretin (prealbumin)	91.0	93.2	4.0	332	29.8	109	61.3	-1.6	X04191	K03252
trkB oncogene (tyrosine protein kinase; gp145 form)	94.2	98.5	0.6	1926	28.2	537	65.8	-0.2	X17647	M55291
alpha-tropomyosin TM2 isoform 2 (fibroblast)	97.3	99.6	0.1	680	13.8	172	61.5	-0.6	M22479	M16432
beta-tropomyosin (skeletal muscle)	97.4	100.0	0.0	682	13.9	170	69.4	1.5	X12650	L00372
trypsin Ta (clone pMPt9)	91.4	87.8	6.0	568	20.2	170	62.7	-2.3	X04574	J00778
thyrotropin beta subunit (TSHB)	91.6	89.1	4.9	324	25.5	90	54.4	-4.6	M20536	M13897
alpha-tubulin (mouse M-alpha-1)	97.1	100.0	0.0	1052	14.4	301	64.3	0.6	M13445	J00797
UBF1 transcription factor	94.7	98.7	0.5	1843	29.2	449	69.1	3.3	X60831	M61726
UDP-glucuronosyltransferase (UDPGT; 17-beta hydroxysteroid)	90.2	87.9	5.9	1262	30.6	328	40.1	-1.5	X06358	Y00156
urate oxidase (UOX)	93.9	94.4	2.6	720	22.8	189	63.9	-2.7	M27695	M63593
vascular cell adhesion molecule (VCAM-1)	89.6	86.3	6.9	1722	27.7	495	48.2	-0.9	M84487	M84488
vitamin D binding protein (Gc globulin)	91.0	90.7	4.7	1110	30.3	306	51.0	-1.1	M55413	M60205
whey acidic protein (WAP)	80.8	68.2	19.6	314	31.8	82	59.2	4.6	X01157	X01153
tryptophan hydroxylase (Tph)	93.6	95.9	1.7	1046	27.4	286	56.7	-1.4	J04758	X53501
mXBP transcription factor (CRE-BP2; RATF2; C/EBP-related)	96.7	98.4	0.7	719	12.7	220	40.1	-1.3	M31629	M65148
Y-box binding protein 1 / enhancer factor I subunit A	99.0	99.7	0.1	744	4.1	222	51.5	-0.9	M60419	M57299
tyrosine hydroxylase	94.5	97.2	1.3	1137	21.9	357	66.9	-0.8	M69200	M10244
cytoplasmic protein-tyrosine phosphatase (PTP-S)	95.1	96.7	1.7	863	19.7	226	39.9	-1.3	M81477	X58828
ypt-1 (rab-1) oncogene (ras-related)	95.6	99.0	0.4	486	21.8	129	48.3	-7.7	Y00094	J02998

Notes: K_S and K_A are the numbers of synonymous and nonsynonymous substitutions per site, respectively (Li et al. 1985). L_S and L_A are the numbers of synonymous and nonsynonymous sites, respectively. GC_S is the mean silent-site G+C content in the mouse and rat genes. ΔGC_S is GC_S of the mouse gene minus that of the rat gene. In some cases the accession number listed is one of several from which the sequence used was constructed. A dash indicates that the sequence was taken directly from the literature.