*Proc. Natl. Acad. Sci. USA*
Vol. 77, No. 12, pp. 7328–7332, December 1980
Genetics

# Nucleotide sequence divergence and functional constraint in mRNA evolution

(sequence difference/uniform rate of synonymous substitution/evolutionary rate of noncoding region)

TAKASHI MIYATA, TERUO YASUNAGA, AND TOSHIRŌ NISHIDA

Department of Biology, Faculty of Science, Kyushu University, Fukuoka 812, Japan

*Communicated by Motoo Kimura, July 24, 1980*

**ABSTRACT** **Comparison of about 50 pairs of homologous nucleotide sequences for different genes revealed that the substitutions between synonymous codons occurred at much higher rates than did amino acid substitutions. Furthermore, five pairs of mRNA sequences for different genes were compared in species that had diverged at the same time. The evolutionary rate of synonymous substitution was estimated to be $5.1 \times 10^{-9}$ per site per year on the average and is approximately constant among different genes. It also is suggested that this property would be suitable for a molecular clock to determine the evolutionary relationships and branching order of duplicated genes. Each functional block of the noncoding region evolves with a rate that is almost constant, regardless of the types of genes. The intervening sequence and the 5′ portion of the 3′ noncoding region show considerable divergence, the extent of which is almost comparable to that in the synonymous codon sites, whereas the other blocks consisting of the 5′ noncoding region and the 3′ portion of the 3′ noncoding region are strongly conserved, showing approximately half of the divergence of the synonymous sites. This strong sequence preservation might be due to the functional requirements for transcription and modification of mRNA.**

The recent developments in techniques for determining nucleotide sequences have provided a rapid increase in our knowledge of the primary structure of many genes, and comparison is now possible between the pairs of homologous nucleotide sequences of a wide variety of genes. Comparative study at the DNA level would provide much more knowledge on the process of molecular evolution than would comparison at the protein level (1). Indeed, direct comparison between homologous DNA sequences enables us to evaluate the extent of sequence divergence in various functional or structural units such as coding regions, intervening sequences (IVS), 5′ and 3′ transcribed noncoding regions, and other untranscribed flanking regions. Furthermore, from the comparison of DNA sequences in the coding regions, it also is possible to evaluate the extent of divergence for two types of substitutions (1–10): the nucleotide replacement leading to an amino acid change (amino acid substitution) and that leading to a synonymous codon change (synonymous substitution). The rate of substitution varies with the nature of the specifications of a functioning protein and RNA and reflects the degree of functional constraint operating on the amino acid and nucleotide sequences (11–14). Thus, comparison of the rates for different functional units or segments within a single gene might provide some insight into the functional significance imposed on them and enable us to trace the evolutionary history of the gene in detail (10).

In the last few years, comparative studies of nucleotide sequences by several authors have shown that substitution at the third position of the codon occurred with a relatively high rate compared with substitution at the other two positions (2–9). Recently, we developed a method for estimating the rate of synonymous substitution and amino acid substitution from a pair of homologous nucleotide sequences and have shown that the rate of synonymous substitution is much higher than that of amino acid substitution and is approximately constant among the different genes adjacent to each other on the same DNA sequence (1).

In this report, we will confirm this finding by comparison of a much wider variety of genes. Furthermore, from comparison of the extent of sequence divergence in coding and noncoding regions, we will suggest a substitution pattern that is expected to be found in most genes and will discuss interesting features of nucleotide substitution.

## METHODS

**Sequence Alignment.** The procedure was as described (10).

**Calculation of Sequence Difference *K*.** The procedures have been described elsewhere (1, 10). The sequence difference $K$ or simply "difference" is defined as the number of mismatches per nucleotide site in a pair of aligned sequences. For the coding region, we have carried out calculations for the two types of sequence differences, synonymous difference ($K_S$) and amino acid difference ($K_A$). $K_S$ ($K_A$) is defined as the number of synonymous (amino acid) substitutions, relative to the total number of synonymous (amino acid) sites, and is the sum of the fractions of nucleotide sites that lead to synonymous (amino acid) change by single nucleotide replacement per each nucleotide position of codon (1). Two methods that differ in treating gaps were applied to calculate the sequence difference ($K_N$) of the noncoding region. When gaps were found in any one of sequences compared, the corresponding sites were excluded from the comparison (method 1). In method 2, a gap was counted as a "mismatch." Nucleotide sites with more than 10 consecutive gaps were excluded from both methods. The values estimated from method 1 and method 2 may give the lower and the upper boundary of the value of $K_N$, respectively.

## RESULTS AND DISCUSSION

**Comparison of the Rate of Synonymous Substitution with That of Amino Acid Substitution.** Table 1 shows the estimated values of $K_S$ and $K_A$ for each gene studied. Clearly, the ratio $K_A/K_S$ is always smaller than unity, except for the case of Ig λ *V* gene, implying that in most genes the rate of synonymous substitution exceeds the rate of amino acid substitution, as several authors have already pointed out (1–10). It may be interesting to compare homologous nucleotide sequences coding for fibrinopeptides, which are the most rapidly evolving molecules among the known proteins (11, 56). It is expected that the ratio $K_A/K_S$ for this gene is close to unity.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U. S. C. §1734 solely to indicate this fact.

Abbreviations: IVS, intervening sequences; SV40, simian virus 40; BKV, BK virus.

Table 1. Comparison between synonymous differences ($K_S$) and amino acid differences ($K_A$) of the coding region of various genes

| | $K_S$ | $K_A$ | $K_A/K_S$ |
|---|---|---|---|
| **Orthologous comparison** | | | |
| β-Globin genes: | | | |
| Human (15) vs. rabbit (16, 17) | 0.291 | 0.054 | 0.186 |
| Human vs. mouse (18, 19) | 0.378 | 0.118 | 0.311 |
| Human vs. chicken (20) | 0.477 | 0.194 | 0.407 |
| Rabbit vs. mouse | 0.405 | 0.125 | 0.309 |
| Rabbit vs. chicken | 0.524 | 0.199 | 0.379 |
| Mouse vs. chicken | 0.517 | 0.222 | 0.430 |
| α-Globin genes: | | | |
| Rabbit (21) vs. mouse (22) | 0.486 | 0.105 | 0.217 |
| Rabbit vs. chicken (23) | 0.497 | 0.215 | 0.433 |
| Mouse vs. chicken | 0.569 | 0.195 | 0.343 |
| Hormone genes: | | | |
| *hGH* (24) vs. *rGH* (25) | 0.471 | 0.181 | 0.384 |
| Insulin: human (26) vs. rat II (8, 27) | 0.483 | 0.098 | 0.204 |
| *LPH:* bovine (28) vs. mouse (29) | 0.390 | 0.038 | 0.098 |
| *Trp-A* gene (30): | | | |
| *Escherichia coli* vs. *Salmonella typhimurium* | 0.726 | 0.087 | 0.119 |
| Lipoprotein: | | | |
| *E. coli* (31) vs. *Serratia marescens* (32) | 0.290 | 0.092 | 0.319 |
| φX174 (33, 34) vs. G4 (35):* | | | |
| *F* gene | 0.677 | 0.231 | 0.341 |
| *G* gene | 0.719 | 0.400 | 0.556 |
| *H* gene | 0.744 | 0.188 | 0.253 |
| *A* gene† | 0.676 | 0.201 | 0.297 |
| SV (36, 37) vs. BKV (38, 39): | | | |
| t antigen | 0.731 | 0.178 | 0.244 |
| T1 antigen | 0.755 | 0.164 | 0.217 |
| T2 antigen | 0.740 | 0.148 | 0.200 |
| Putative *VPX* | 0.436 | 0.252 | 0.578 |
| Hemagglutinin (influenza A): | | | |
| Victoria (40) vs. Rostock (41) | 0.812 | 0.365 | 0.449 |
| **Paralogous comparison** | | | |
| α-Globin vs. β-globin: | | | |
| Rabbit; α vs. β | 0.575 | 0.370 | 0.643 |
| Mouse; α vs. β | 0.661 | 0.375 | 0.567 |
| Chicken; α vs. β | 0.537 | 0.383 | 0.713 |
| Hormone genes: | | | |
| *hGH* vs. *hCS* (42) | 0.105 | 0.062 | 0.590 |
| *hCS* vs. *rGH* | 0.475 | 0.197 | 0.415 |
| Rat insulin *I* (8, 27) vs. *II* | 0.186 | 0.032 | 0.172 |
| Histone genes:‡ | | | |
| (*P. milialis* h22 (43) vs. *S. purpuratus* pSp2/17 (44, 45) | | | |
| *H4* | 0.448 | 0.010 | 0.022 |
| *H2B* | 0.520 | 0.053 | 0.102 |
| *H3* | 0.453 | 0.007 | 0.015 |
| *H2A* | 0.501 | 0.020 | 0.040 |
| (*P. milialis* h22 vs. *P. milialis* h19 (46) | | | |
| *H4* | 0.408 | 0.0 | 0.0 |
| *H2B* | 0.504 | 0.050 | 0.100 |
| *H3* | 0.455 | 0.007 | 0.015 |
| *H2A* | 0.491 | 0.020 | 0.040 |
| Cytochrome *c* (yeast): | | | |
| iso-*1* (47) vs. iso-*2* (48) | 0.840 | 0.080 | 0.095 |
| Chorion gene (silkmoth) (49): | | | |
| *PC401* vs. *PC10* | 0.454 | 0.110 | 0.243 |
| IgH *C* genes [$\gamma_1$(50) vs. $\gamma_{2b}$(51)]: | | | |
| *CH1* | 0.214 | 0.096 | 0.449 |
| hinge | 0.552 | 0.311 | 0.563 |
| *CH2* | 0.495 | 0.178 | 0.360 |
| *CH3* | 0.528 | 0.269 | 0.509 |
| IgH *V* genes (52): | | | |
| *VH107* vs. *M603* | 0.042 | 0.017 | 0.395 |
| IgL *V* genes: | | | |
| $\kappa_2$ vs. $\kappa$3 (53) | 0.078 | 0.057 | 0.738 |
| $\kappa_2$ vs. $\kappa_{41}$ (54)§ | 0.459 | 0.257 | 0.560 |
| $\kappa_3$ vs. $\kappa_{41}$ | 0.499 | 0.241 | 0.483 |
| $\lambda_I$ vs. $\lambda_{II}$ (55) | 0.023 | 0.039 | 1.696 |

References to sequence data are in parentheses. *hGH* and *rGH*, human and rat growth hormone genes, respectively; *hCS*, human chorionic somatomammotropin gene; *LPH*, corticotropin/β-lipotropin precursor gene; IgH *C* gene and IgH *V* gene, immunoglobulin heavy chain constant and variable region gene, respectively; IgL *V* gene, immunoglobulin light chain variable region gene.

* Only the nonoverlapping genes were compared.

† The nonoverlapping segment of *A* gene.

‡ The evolutionary relationship between clone h22 and sPs2/sPs17 appears to be paralogous, but not orthologous (46).

§ When the nucleotide sequence of $\kappa_{41}$ is compared with that of $\kappa_2$ or $\kappa_3$, there is a remarkably diverging segment spanning from codon 36 (CUU) of $\kappa_{41}$ to codon 41 (GAU), which might be responsible for frameshift mutations. This segment is therefore excluded from the analysis.

Synonymous codons are used quite nonrandomly in most genes (57–59). Though the functional significance of specific codon utilization is still not understood fully, it might be related to selective constraint against synonymous changes. Recently, from comparison of nucleotide sequences of adult α-globin genes from mouse and rabbit with the sequence of mouse pseudo-α-globin gene (60, 61), which appears to have lost its protein encoding ability due to frameshift mutations during evolution, we have found that this pseudo gene evolves with a rate higher than the rate of synonymous substitution of α-globin genes (62). However, by considering that $K_S$ is much higher than $K_A$ in a wide variety of genes, it is likely that synonymous mutations are subject to, at most, weak functional constraint (1, 4). It is specially interesting to compare the homologous nucleotide sequences of the glyceraldehyde-3-phosphate dehydrogenase gene in which utilization of codons is heavily biased to one or two of the degenerate codons (63).

**Uniform Rate of Synonymous Substitution.** To compare the rates of synonymous substitutions of different genes, it is necessary to estimate the time since the divergence of the genes being compared. However, when the same pair of DNA sequences is being compared, it is possible to treat the sequence differences of these genes as relative evolutionary rates without knowing the divergence time. From the comparison of the synonymous differences among the nonoverlapping genes of φX174 and G4 and among histone genes, we have pointed out that the rate of synonymous substitution is approximately constant, at least among the genes adjacent to each other on the same DNA (1). As Table 2 shows, this nature of adjacent genes also holds for early genes of simian virus 40 (SV40) and BK virus (BKV) and for genes coding for the domains and hinge region of immunoglobulin $\gamma_1$ and $\gamma_{2b}$ chains, except for the gene coding for the CH1 domain. The CH1 domain gene is evidently reduced in the value of $K_S$ compared with the other three, suggesting that it might be derived from recombination events taken place within the IVSs during evolution (10). The value of $K_S$ for papovavirus *VPX* gene is significantly smaller than the values for early genes (see Table 1). Recently, the complete nucleotide sequence of BKV virus genome was established (64, 65). Comparison of the sequence of BKV genome with that of

Table 2. Comparison of synonymous differences ($K_S$) of different genes

| | $K_S$ | Mean ± SD |
|---|---|---|
| **Closely linked genes or regions** | | |
| Histone genes: | | |
| (h22 vs. pSp2/17 | | |
| *H4* | 0.488 | |
| *H2B* | 0.520 | |
| *H3* | 0.453 | |
| *H2A* | 0.501 | |
| h22 vs. h19 | | 0.473 ± 0.035 |
| *H4* | 0.408 | |
| *H2B* | 0.504 | |
| *H3* | 0.455 | |
| *H2A* | 0.491 | |
| φX174 vs. G4: | | |
| *F* gene | 0.677 | |
| *G* gene | 0.719 | 0.704 ± 0.029 |
| *H* gene | 0.744 | |
| *A* gene* | 0.676 | |
| SV40 vs. BKV (early genes: | | |
| t antigen | 0.731 | |
| T1 antigen | 0.755 | 0.742 ± 0.001 |
| T2 antigen | 0.740 | |
| Ig $\gamma_1$ vs. $\gamma_{2b}$: | | |
| hinge | 0.552 | |
| *CH2* | 0.495 | 0.525 ± 0.023 |
| *CH3* | 0.528 | |
| **Distantly located genes** | | |
| β-Globin (human vs. mouse)† | 0.459 | |
| β-Globin (rabbit vs. mouse)† | 0.507 | |
| α-Globin (rabbit vs. mouse) | 0.486 | 0.481 ± 0.016 |
| Growth hormone (human vs. rat) | 0.471 | |
| Preproinsulin (human vs. rat) | 0.483 | |

Except for that of the β-globin genes, all the values of $K_S$ were taken from Table 1. The $K_S$s were averaged among the pairs of genes with the same time since divergence. For sequence data, see references cited in Table 1.

* The nonoverlapping segment of *A* gene.

† The values of $K_S$ of the segments corresponding to codons 1–21, 36–91, and 108–146 (distal segment).

SV40 revealed an unusual homology region, comprising about 600 nucleotides from the start of late *VPX* gene to before *VP3*, whereas the other regions including *VP3*, *VP1*, and all the early genes have approximately the same value of $K_S$ (unpublished data). It seems likely that this homology along the long stretch of late genes is derived from recombination, though we can not exclude the possibility of functional constraints imposed on this region.

It may be specially interesting to compare the values of $K_S$ among genes located distantly to each other on chromosomes. In Table 2, we list the values of $K_S$ for β-globin gene with two different pairs, α-globin gene, growth hormone gene, and insulin gene. Upon consideration of the phylogenetic relationships among human, rabbit, mouse, and rat, it seems reasonable to assume that these five sequence pairs have the same divergence time. Thus, their values of $K_S$ can be compared directly. As we have pointed out for β-globin genes (1, 6), the coding sequences surrounding the splicing points of IVSs (proximal segment) show limited divergence, whereas the coding sequences apart from the splicing points (distal segment) show considerable divergence. The homology in the proximal segment is suggested to be due to extra constraint specific for the β-globin gene, presumably related to the presence of the rigid base-pairing structure (1). Thus, for the β-globin gene, the values of $K_S$ corresponding to the distal segment are compared with those of other genes. Clearly, the values of $K_S$ are almost the same among the five.

The results shown in Table 1 and Table 2 suggest that the synonymous substitution occurs at a high and approximately constant rate, independent of gene type and location on the chromosome. This is in sharp contrast to the amino acid substitution whose rate varies with the specification of functioning proteins (13). The high and uniform rate of synonymous substitution would also imply that the majority of synonymous changes are subject to selective constraints much less than the amino acid changes.

**The Absolute Rate of Synonymous Substitution and a Molecular Clock.** The absolute rate (evolutionary rate) of synonymous substitution ($V_S$) is given by the equation: $V_S = -(3/4)\ln[1 - (4/3)K_S]/2T$, in which $T$ stands for the divergence time of two sequences compared, and correction for multiple substitutions is made (13). Using the average value of $K_S$ for distantly located genes (0.481 ± 0.016) in Table 2 and assuming $T = 7.5 \times 10^7$ year [the time since divergence of the primate and the rodent (11)] we obtain $V_S = 5.1 \pm 0.3 \times 10^{-9}$ (per nucleotide site per year). This rate is very high, almost comparable to the rate of amino acid substitution in fibrinopeptides (4), which are known to be the most rapidly evolving proteins.

As noted above, synonymous substitution occurs at a high and approximately constant rate, regardless of the types of genes compared. This characteristic feature of synonymous substitution might be useful for determining the evolutionary relationships between closely related genes that are derived from a common ancestor by gene duplication at a relatively recent time. The rate of amino acid substitution is approximately constant for a given protein performing a well-established function, and its value varies with the degree of functional constraints imposed on a protein molecule (12, 13, 56). The constant rate of amino acid substitution of a given protein is useful as a clock for determining the phylogenetic branching order of different species. However, it may be unsuitable for estimating the divergence time of duplicated genes coding for proteins with different functions, because both of the protein sequences have been evolving independently with different rates in each species. On the contrary, the use of synonymous substitution as a clock might be adequate for such a purpose, because its rate is almost independent of the types of genes.

**Substitution Rate in Noncoding Region ($K_N$).** We have carried out sequence comparison for various noncoding regions of (*i*) β-globin genes from rabbit and mouse, (*ii*) β-globin genes from human and mouse, (*iii*) β-globin genes from rabbit and mouse, (*iv*) preproinsulin genes from human and rat, and (*v*) growth hormone genes from human and rat (for sequence data used, see references cited in Table 1). For each of the five pairs, the $K_N$ for the noncoding region was calculated at every functional or structural unit and was compared with the $K_S$ (Table 3). Since the divergence time of the sequences compared is just the same among the five pairs, all the values shown in Table 3 could be compared directly.

Table 3 reveals interesting features on nucleotide substitutions in the noncoding region. For a given functional unit, the values of $K_N$ are almost identical among the different genes. This suggests that, in the noncoding region and synonymous sites in the coding region of mRNA, the nucleotide sequences evolve with approximately the same rate among different genes. For the 5′-noncoding region, the values of $K_N$ appear to be somewhat different between globin genes (α-globin and β-globin) and hormone genes (preproinsulin and growth hormone). This may be related to the effect of statistical fluctuation due to the short nucleotide length of this region. Alternatively, a presumably more plausible possibility is that the nucleotide sequences around the "cap" are almost always conserved, whereas a short segment bordering the coding region shows a

Table 3. Comparison of the sequence differences ($K_N$) of various noncoding regions with the synonymous difference ($K_S$) of coding region

| | $K_N$ | | | | | |
|---|---|---|---|---|---|---|
| | | | | 3′ NC | | |
| | Small IVS | Large IVS | 5′ NC | 5′ p | 3′ p | $K_S$ |
| β-Globin:* | | | | | | |
| Rabbit vs. mouse | 0.313(0.408) | 0.529(0.547) | 0.231(0.245) | 0.480(0.567) | 0.225(0.225) | 0.507 |
| Human vs. mouse | — (—) | — (—) | 0.160(0.192) | 0.457(0.506) | 0.196(0.196) | 0.459 |
| α-Globin: | | | | | | |
| Rabbit vs. mouse | — (—) | — (—) | 0.212(0.297) | 0.463(0.473) | 0.219(0.219) | 0.486 |
| Preproinsulin: | | | | | | |
| Human vs. rat | 0.398(0.408) | 0.493(0.526) | 0.327(0.393) | 0.438(0.500) | 0.191(0.191) | 0.483 |
| Growth hormone: | | | | | | |
| Human vs. rat | — (—) | — (—) | 0.321(0.345)† | 0.400(0.451) | 0.152(0.177) | 0.471 |
| Mean | 0.356(0.408) | 0.511(0.537) | 0.250(0.294) | 0.448(0.499) | 0.197(0.202) | 0.481 |
| ± SD | 0.043(0.000) | 0.018(0.011) | 0.065(0.071) | 0.027(0.039 | 0.026(0.018) | 0.016 |

Method 1: when a gap is found in any one of aligned sequences, the corresponding site is excluded from the calculation. Method 2: a gap is counted as a "mismatch" (these values are shown in parentheses). For both the methods, when consecutive gaps with more than 10 nucleotide sites were found in any one of the sequences compared, the corresponding sites were excluded from the calculation. 5′ NC and 3′ NC, 5′ and 3′ transcribed noncoding regions, respectively. The 3′ noncoding region is divided into two portions, the 5′ portion (5′ p) and the 3′ portion (3′ p), depending on the extent of divergence. It should be noted that the time since divergence (of the two sequences compared) is the same for the five pairs of genes listed above. For sequence data, see reference cited in Table 1.

* The segments of coding regions (codon 22–35 and 92–107) that surround the splicing points of IVSs are excluded from the analysis.

† The nucleotide sequence of this region is not established completely for both species, and comparison is possible only for limited nucleotides near the initiation codon.

considerable divergence; the nucleotide sequence of the diverging segment appears to differ in length from gene to gene. Thus, the diverse values of $K_N$ in the 5′ noncoding region may be due to the difference in size of the diverging segment involved.

Another interesting feature is that the noncoding region of mRNA is classified into two groups: (*i*) the diverging regions, which consist of the IVSs and the 5′ portion of the 3′ noncoding region, and (*ii*) the conserved regions, which are the 5′ noncoding region and the 3′ portion of the 3′ noncoding region. The former are almost comparable with or slightly lower in sequence difference $K$ than the synonymous sites of the coding region, whereas the latter has a value of $K_N$ approximately half that of $K_S$. Significant homology in the sequence of the 5′ noncoding region might be responsible for the functional significance, such as transcription initiation, capping, or other special secondary structure-related functions involved in this region. The 3′ noncoding region is clearly divided into two distinct portions in terms of the extent of divergence. The 5′ portion nearest the coding region shows a considerable divergence, in which the value of $K_N$ is almost comparable to that of $K_S$, whereas the 3′ portion adjacent to the poly(A) site shows marked conservation, which may be due to the functional requirements like poly(A) addition, transcription termination, or both. From the comparison of human and rabbit β-globin mRNA, Kafatos *et al.* (5) have shown a similar result on the substitution patterns in the 5′ and 3′ noncoding regions.

The IVSs are almost comparable in the value of $K_N$ to that of $K_S$, which suggests that they are subject to less functional constraint (66). Both the globin and insulin genes contain two IVSs which are different in size. The small IVS is about 150 nucleotides or so in length, whereas the large IVS consists of more than 500 nucleotides. The large IVS appears to have a larger value of $K_N$ than does the small IVS. There may be at least two reasons for this. The large IVS involves several long, simple sequences like pyrimidine-rich or purine-rich sequences that are not found in the small IVS, and there are large numbers of deletions in it, some of which are more than 100 nucleotides in length. It seems likely that the large IVS is subject to constraint against sequence variation to a lesser extent than the small IVS is. Furthermore, as shown in globin gene (66, 67), sequence homology in IVS is found only in the region bordering the structural gene sequence, which may be important in splicing (68), and the central portion diverges. By considering the length of this homology region relative to that of the overall region, one expects estimates of sequence difference $K$ per site for the small IVS to be smaller in value than that for the large IVS.

Finally, using the average values of $K_N$ for the five pairs listed in Table 3 and assuming $T = 7.5 \times 10^7$ year (the time since divergence of human and mouse) we obtain the evolutionary rates (per site per year) for various noncoding regions in mRNA. Using the values of differences by method 1 (method 2 in parentheses), they are, respectively, $5.7(6.3) \times 10^{-9}$, $3.2(3.9) \times 10^{-9}$, $2.0(2.5) \times 10^{-9}$, $4.5(5.5) \times 10^{-9}$, and $1.5(1.6) \times 10^{-9}$ for the large IVS, small IVS, 5′ noncoding region, 5′ portion of the 3′ noncoding region, and 3′ portion of the 3′ noncoding region, respectively. Though the large IVS evolves with the highest rate among the various functional blocks, including the synonymous site, its rate is still lower than that of pseudo α-globin gene (62), which appears to be a dormant gene with no important functions (60).

These results strongly suggest that, except for the substitutions leading to amino acid changes, most substitutions occurring on mRNA show similar patterns, regardless of the types of genes, and each segment or region evolves with a constant rate among different genes, if it is subject to no more additional constraints. With respect to the extent of divergence, the noncoding region of mRNA appears to be divided into two blocks. In the IVS and the 5′ portion of the 3′ noncoding region, the $K_N$ is approximately equal to the $K_S$. In the other blocks consisting of the 5′ noncoding region and the 3′ portion of the 3′ noncoding region, the $K_N$ is almost half that of the $K_S$. This strong sequence preservation might be related to the functional requirements for transcription and modification of mRNA. These features of substitutions may well be explained by the neutral theory proposed by Kimura (69).

Comparison of the nucleotide sequences of immunoglobulin $\lambda_I$ and $\lambda_{II}$ variable region genes from the mouse reveals an unusual pattern of substitution. In these genes, the values of sequence difference $K$ for the synonymous site, amino acid site, 5′ noncoding region, and IVS are, respectively, 0.023, 0.039, 0.037, and 0.038—almost the same. Furthermore, the amino acid changes appear to be distributed uniformly over the entire

coding region, which is quite distinct in the substitution pattern of κ-type genes ($\kappa_2$ and $\kappa_3$) in which of 11 amino acid changes, seven cluster within the hypervariable regions (40). At present, we can not determine whether the unusual pattern of substitutions in the λ-type genes is a result of statistical fluctuation due to the small number of substitutions observed or due to other unknown reasons.

We thank Prof. O. Smithies for permitting us to use his data on the mouse α-globin-related pseudo gene and Prof. H. Matsuda and Dr. M. Hasegawa for helpful discussions. We also thank Dr. T. Ohta and Dr. Y. Tateno for valuable suggestions and comments.

1. Miyata, T. & Yasunaga, T. (1980) *J. Mol. Evol.*, **16**, 23–36.
2. Salser, W. & Issacson, J. S. (1976) *Prog. Nucleic Acids Res.* **19**, 205–220.
3. Grunstein, M., Schedle, P. & Kedes, L. (1976) *J. Mol. Biol.* **104**, 351–369.
4. Kimura, M. (1977) *Nature (London)* **267**, 275–276.
5. Kafatos, F. C., Efstratiadis, A., Forget, B. G. & Weissman, S. M. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 5618–5622.
6. Salser, W. (1978) *Cold Spring Harbor Symp. Quant. Biol.* **42**, 985–1002.
7. Jukes, T. H. & King, J. L. (1979) *Nature (London)* **281**, 605–606.
8. Lemedico, P., Rosenthal, N., Efstratiadis, A., Gilbert, W., Kolodner, R. & Tizard, R. (1979) *Cell* **18**, 545–558.
9. Nichols, B. P. & Yanofsky, C. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 5244–5248.
10. Miyata, T., Yasunaga, T., Yamawaki-Kataoka, Y., Obata, M. & Honjo, T. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 2143–2147.
11. Dayhoff, M. O. (1978) *Atlas of Protein Sequence and Structure*, ed. Dayhoff, M. O. (National Biomedical Research Foundation, City, MD), Vol. 5, Suppl. 3.
12. Zuckerkandle, E. & Pauling, L. (1965) *Evolving Genes and Proteins*, eds. Bryson, V. & Vogel, H. J. (Academic, NY), pp. 97–116.
13. Kimura, M. & Ohta, T. (1972) *J. Mol. Evol.* **2**, 87–90.
14. Kimura, M. & Ohta, T. (1974) *Proc. Natl. Acad. Sci. USA* **71**, 2848–2852.
15. Marotta, C. A., Wilson, J. T., Forget, B. G. & Weissman, S. M. (1977) *J. Biol. Chem.* **252**, 5040–5053.
16. Efstratiadis, A., Kafatos, F. C. & Maniatis, T. (1977) *Cell* **10**, 571–585.
17. Van Ooen, A., Van Den Berg, J., Mantei, N. & Weissmann, C. (1979) *Science* **206**, 337–344.
18. Konkel, D. A., Tilghman, S. M. & Leder, P. (1978) *Cell* **15**, 1125–1132.
19. Konkel, D. A., Maizel, V., Jr. & Leder, P. (1979) *Cell* **18**, 865–873.
20. Richards, R. I., Shine, J. Ullrich, A., Wells, J. R. E. & Goodmann, H. M. (1979) *Nucleic Acids Res.* **7**, 1137–1146.
21. Heindell, H. C., Liu, A., Paddock, G. V., Studnicka, G. M. & Salser, W. A. (1978) *Cell* **15**, 43–54.
22. Nishioka, Y. & Leder, P. (1979) *Cell* **18**, 875–882.
23. Deacon, N. J., Shine, J. & Naora, H. (1980) *Nucleic Acids Res.* **8**, 1187–1199.
24. Martial, J. A., Hallewell, R. A., Baxter, J. D. & Goodman, H. M. (1979) *Science* **205**, 602–607.
25. Seeburg, P. H., Shine, J., Martial, J. A., Baxter, J. D. & Goodman, H. M. (1977) *Nature (London)* **270**, 486–494.
26. Bell, G. I., Pictet, R. L., Rutter, W. J., Cordell, B., Tischer, E. & Goodman, H. M. (1980) *Nature (London)* **284**, 26–32.
27. Cordell, B., Bell, G., Tischer, E., DeNote, F. M., Ullrich, A., Pictet, R., Rutter, W. J. & Googman, H. M. (1979) *Cell* **18**, 533–543.
28. Nakanishi, S., Inoue, A., Kita, T., Nakamura, M., Chang, A. C. Y., Choen, S. N. & Numa, S. (1979) *Nature (London)* **278**, 423–427.
29. Roberts, J. L., Seeburg, P. H., Shine, J., Herbert, E., Baxter, J. D. & Goodman, H. M. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 2153–2157.
30. Nichols, B. P. & Yanofsky, C. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 5244–5248.
31. Nakamura, K. & Inouye, M. (1979) *Cell* **18**, 1109–1117.
32. Nakamura, K. & Inouye, M. (1970) *Proc. Natl. Acad. Sci. USA* **77**, 1369–1373.
33. Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, A. R., Fiddes, J. C., Hutchison, C. A., III, Slocombe, P. M. S. & Smith, M. (1977) *Nature (London)* **265**, 689–695.
34. Sanger, F., Coulson, A. R., Friedmann, T., Air, G. M., Barrell, B. G., Brown, N. L., Fiddes, J. C., Hutchson C. A., III, Slocombe, P. M. & Smith, M. (1978) *J. Mol. Biol.* **125**, 225–246.
35. Godson, G. N., Barrell, B. G., Staden, R. & Fiddes, J. C. (1978) *Nature (London)* **276**, 236–247.
36. Reddy, V. P. Thimmappaya, B., Dhar, R., Subramanian, K. N., Zain, B. S., Pan, J., Ghosh, P. K., Celma, M. L. & Weissman, S. M. (1978) *Science* **200**, 494–502.
37. Fiers, W., Contreras, R., Haegeman, G., Rogiers, R., Van de Voorde, A., Van Heuverswyn, H., Van Herreweghe, J., Volckaert, G. & Ysebaert, M. (1978) *Nature (London)* **273**, 113–120.
38. Dhar, R., Seif, I. & Khoury, G. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 565–569.
39. Yang, R. C. A. & Wu, R. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 1179–1183.
40. Min Jou, W., Verhoeyen, M., Devos, R., Saman, E., Fang, R., Huylebroeck, D., Fiers, W., Threlfall, G., Barker, C., Carey, N. & Emtage, S. (1980) *Cell* **19**, 683–696.
41. Porter, A. G., Barber, C., Carey, N. H., Hallewell, R. A., Threlfall, G. & Emtage, J. S. (1979) *Nature (London)* **282**, 471–477.
42. Shine, J., Seeburg, P. H., Martial, J. A., Baxter, J. D. & Goodman, H. M. (1977) *Nature (London)* **270**, 494–499.
43. Schaffner, W., Knuz, G., Daetwyler, H., Telford, J., Smith, H. O. & Birnstiel, M. L. (1978) *Cell* **14**, 655–671.
44. Sures, I., Lowry, J. & Kedes, L. H. (1978) *Cell* **15**, 1033–1044.
45. Grunstein, M. & Grunstein, J. E. (1978) *Cold Spring Harbor Symp. Quant. Biol.* **42**, 1083–1092.
46. Busslinger, M., Portmann, R., Irminger, J. C. & Birnstiel, M. L. (1980) *Nucleic Acids Res.* **8**, 957–977.
47. Smith, M., Leung, D. W., Gillam, S., Astell, C. R., Montgomery, D. L. & Hall, B. D. (1979) *Cell* **16**, 753–761.
48. Montogomery, D. L., Leung, D. W., Smith, M., Shalit, P., Faye, G. & Hall, B. D. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 541–545.
49. Jones, C. W., Rosenthal, N., Rodakis, G. C. & Kafatos, F. C. (1979) *Cell* **18**, 1317–1332.
50. Honjo, T., Obata, M., Yamawaki-Kataoka, T., Kawakami, T., Takahashi, N. & Mano, Y. (1979) *Cell* **18**, 559–568.
51. Yamawaki-Kataoka, Y., Kataoka, T., Takahashi, N., Obata, M. & Honjo, T. (1980) *Nature (London)* **283**, 786–789.
52. Early, P., Huang, H., Davis, M., Calame, K. & Hood, L. (1980) *Cell* **19**, 981–992.
53. Seidman, J. G., Leder, A., Edgell, M. H., Polsky, F., Tilghman, S. M., Tiemeier, D. C. & Leder, P. (1978) *Proc. Natl. Acad. Sci. USA* **75**, 3881–3885.
54. Seidman, J. G., Max, E. E. & Leder, P. (1979) *Nature (London)* **280**, 370–375.
55. Bernard, O., Hozumi, N. & Tonegawa, S. (1978) *Cell* **15**, 1133–1144.
56. Wilson, A. C., Carlson, S. S. & White, T. J. (1977) *Annu. Rev. Biochem.* **46**, 573–639.
57. Grantham, R., Gautier, C. & Gouy, M. (1980) *Nucleic Acids Res.* **8**, 1893–1912.
58. Hasegawa, M., Yasunaga, T. & Miyata, T. (1979) *Nucleic Acids Res.* **7**, 2073–2079.
59. Miyata, T., Hayashida, H., Yasunaga, T. & Hasegawa, M. (1979) *Nucleic Acids Res.* **7**, 2431–2438.
60. Nishioka, Y., Leder, A. & Leder, P. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 2806–2809.
61. Vanin, E. F., Goldberg, G. I., Tucker, P. W. & Smithies, O. (1980) *Nature (London)* **286**, 222–226.
62. Miyata, T. & Yasunaga, T. (1981) *Proc. Natl. Acad. Sci. USA*, in press.
63. Holland, J. P. & Holland, M. J. (1979) *J. Biol. Chem.* **254**, 9839–9845.
64. Yang, R. C. A. & Wu, R. (1979) *Science* **206**, 456–462.
65. Seif, I., Khoury, G. & Dhar, R. (1979) *Cell* **18**, 963–977.
66. Van Den Berg, J., Van Ooen, A., Mantei, N., Schambock, A., Grosveld, G., Flavell, R. A. & Weissman, C. (1978) *Nature (London)* **276**, 37–44.
67. Leder, A., Miller, H. I., Hamer, D. H., Seidman, J. G., Norman, B., Sullivan, M. & Leder, P. (1978) *Proc. Natl. Acad. Sci. USA* **75**, 6187–6191.
68. Lerner, M. R., Boyle, J. A., Mount, S. M., Wolin, S. L. & Steiz, J. A. (1980) *Nature (London)* **283**, 220–224.
69. Kimura, M. (1968) *Nature (London)* **217**, 624–626.