J. Mol. Evol. 16, 23–36 (1980)

Journal of
Molecular Evolution
© by Springer-Verlag · 1980

Molecular Evolution of mRNA: A Method for Estimating Evolutionary Rates of Synonymous and Amino Acid Substitutions from Homologous Nucleotide Sequences and Its Application

Takashi Miyata and Teruo Yasunaga

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

Summary. A method for estimating the evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences is presented. This method is applied to genes of ϕX174 and G4 genomes, histone genes and β-globin genes, for which homologous nucleotide sequences are available for comparison to be made. It is shown that the rates of synonymous substitutions are quite uniform among the non-overlapping genes of ϕX174 and G4 and among histone genes H4, H2B, H3 and H2A. A comparison between ϕX174 and G4 reveals that, in the overlapping segments of the A-gene, the rate of synonymous substitution is reduced more significantly than the rate of amino acid substitution relative to the corresponding rate in the non-overlapping segment. It is also suggested that, in the coding regions surrounding the splicing points of intervening sequences of β-globin genes, there exist rigid secondary structures. It is in only these regions that the β-globin genes show the slowing down of evolutionary rates of both synonymous and amino acid substitutions in the primate line.

Key words: Amino acid difference – Synonymous difference – Selective constraint – mRNA evolution

Introduction

Until recently, amino acid sequence data have been a major tool for studying the process of molecular evolution (Dayhoff 1978). From the analysis of amino acid sequences accumulated from various sources of species for many protein families, much has become known about the features of amino acid substitutions during evolution. Among these features, the following two are particularly noteworthy (Zuckerkandl and Pauling 1965; Dickerson 1971; Kimura and Ohta 1974). The rate of amino acid substitution of a given protein performing a well established function is constant (per year per site). The rate of amino acid substitution varies with the nature of the specifications for a functioning protein. That is, protein molecules that are subject to fewer functional constraints evolve faster. This implies that the

0022-2844/80/0016/0023/$ 02.80

rate of amino acid substitution for a given protein reflects the degree of functional constraint operating on the amino acid sequence of the protein. Although there have been extensive studies on molecular evolution at the protein level, the comparative study of nucleotide sequences should be made to understand the more detailed picture of the process of molecular evolution, because the nucleotide sequences of coding regions in mRNA might be subject simultaneously to at least two distinct functional constraints operating at the protein level and the mRNA level.

The recent advance in nucleotide sequencing techniques provides a rapid increase in our knowledge of the primary structures of genes, and homologous nucleotide sequences are now available for comparison made between closely related species. For example, the complete nucleotide sequence of phages ϕX174 and G4 genomes (Sanger et al. 1977; Godson et al. 1978; Sanger et.al. 1978), the sequences of histone genes from different species of sea urchin (Schaffner et al. 1978; Sures et al. 1978; Grunstein and Grunstein 1978) and the sequences of β-globin genes from mouse, rabbit and human (Efstratiadis et al. 1977; Marrota et al. 1977; Konkel et al. 1978; Heindell et al. 1978) have already been reported. The direct comparison between a pair of known nucleotide sequences in coding regions permits us to evaluate both the nucleotide differences per site caused by nucleotide substitutions leading to amino acid changes, K_A, and those leading to synonymous changes, K_S, simultaneously. For obtaining the evolutionary rate as usually defined (i.e., difference per site per year), the K_A and K_S values should be divided by the time since divergence of two sequences. However, when a comparison is made between evolutionary rates of amino acid and synonymous substitutions for the same pair of sequences, the K_A and K_S values can be treated directly as relative evolutionary rates. This is very useful, because the paleontological estimate of the time of divergence is often subject to considerable uncertainty (Romero-Herrera et al. 1973). Furthermore, the comparison between K_A and K_S might provide much insight into the features of synonymous substitution, because the evolutionary rate of amino acid substitution has already been studied extensively for many protein families from amino acid sequence data. Based on this idea, the rate of synonymous substitution is examined in rather greater detail in the present report.

In the last few years, comparative studies of nucleotide sequences have been made by some authors (Salser and Isaacson 1976; Grunstein et al. 1976; Kimura 1977; Kafatos et al. 1977; Salser 1978), and have shown that substitutions at the third codon position have occurred with a relatively high rate compared to the rates at the other two codon positions. It would be more adequate for estimating the evolutionary rates of synonymous and amino acid substitutions from nucleotide sequences to treat the number of codon changes caused by successive single step nucleotide substitutions than to treat the number of substitutions at the nucleotide positions of codons. Here, we present a method for estimating the evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences. We also apply this method to known homologous sequences and show some results on synonymous substitutions.

Methods

Synonymous Site and Amino Acid Site

A single nucleotide change (one step change) in a codon α always causes an exchange of the codon for another α'. We define the resulting codon change $\alpha \rightarrow \alpha'$ as "synonymous change" if the codons α and α' code for the same amino acid, and the change $\alpha \rightarrow \alpha'$ as "amino acid change" if they code for different amino acids. It should be noted that this definition is based on the single nucleotide changes at each position of a codon. For each nucleotide position, three distinct changes are potentially possible, each of which leads to another codon. In most codons, the changes in the first two positions result in amino acid changes. However the case is somewhat complicated in the third position, depending upon the degree of degeneracy of the codon in question: In 4-fold degenerate codons, the three possible changes in the third position always result in synonymous changes. On the other hand, in 2-fold degenerate codons, only one of them is a synonymous change and the others are amino acid changes. In general, let $f_{\alpha i}$ be the fraction of synonymous changes to all possible one step changes occurring on the i-th position of codon α. The fraction of amino acid changes is therefore $1 - f_{\alpha i}$ (here, terminations are considered as the twenty first amino acid). For example, for the third position ($i = 3$) of codon UUU ($= \alpha$), the $f_{UUU,3}$ equals 1/3, and for $i = 3$ and $\alpha =$ GUU, the $f_{GUU,3}$ equals unity. The fraction $f_{\alpha i}$ simply means that, for a given nucleotide site αi (i.e., the i-th position of codon α), the number of sites leading to synonymous changes (synonymous site) is $f_{\alpha i}$ and that of sites leading to amino acid changes (amino acid site) is $1 - f_{\alpha i}$. Thus the number of synonymous sites $\nu_{S,\alpha}$ and that of amino acid sites $\nu_{A,\alpha}$ for a given codon α are given by

$$\nu_{S,\alpha} = \sum_{i=1}^{3} f_{\alpha i} \quad , \quad \nu_{A,\alpha} = 3 - \nu_{S,\alpha} \ . \tag{1}$$

A similar correction for "site" has been introduced by Kafatos et al. (1977).

The Numbers of Synonymous and Amino Acid Substitutions

Let a codon substitution $\alpha \leftrightarrow \beta$ ($\alpha \neq \beta$) be observed in any one of equivalent codon sites of homologous nucleotide sequences. By comparing the nucleotide differences between codons α and β, we can count the minimum substitution number (MSN) between α and β (e.g., MSN = 3 for UUU $\leftrightarrow$ GGC). On the basis of the assumption that no more substitutions have occurred than MSN during evolution (MSN assumption), we can trace all the possible paths from α to β or vice versa through successive one step changes. For a given codon pair α and β, the number of paths $L(\alpha, \beta)$ is solely determined by the genetic code, being equal to unity for MSN = 1, two for MSN = 2 and six for MSN = 3. Let $\gamma^{(p)}$ and $\delta^{(p)}$ be intermediate codons in any one of possible paths between α and β:

$$\text{path p:} \ \ \alpha \leftrightarrow \gamma^{(p)} \leftrightarrow \delta^{(p)} \leftrightarrow \beta \quad , \tag{2}$$

where all the adjacent codons $\alpha \leftrightarrow \gamma$, $\gamma \leftrightarrow \delta$ and $\delta \leftrightarrow \beta$ are mutually changeable through one step changes. For MSN = 1, no intermediate codons ($\gamma^{(p)}$ and $\delta^{(p)}$)

appear in the path and only one (either $\gamma^{(p)}$ or $\delta^{(p)}$) for MSN = 2 (the expression (2) is the case for MSN = 3). As the intermediate codons in (2) can be determined according to the genetic code, we can estimate the number of synonymous changes $\mu_{S,p}(\alpha, \beta)$ and that of amino acid changes $\mu_{A,p}(\alpha, \beta)$ that would have occurred in a path p between α and β. For example, if α = UUU(phe) and β = GUA(val), there are two distinct paths:

path 1: UUU(phe) ↔ GUU(val) ↔ GUA(val) ,
path 2: UUU(phe) ↔ UUA(leu) ↔ GUA(val) .

For path 1, we have $\mu_{S,1}$(UUU, GUA) = 1 and $\mu_{A,1}$(UUU, GUA) = 1, and for path 2, we have $\mu_{S,2}$(UUU, GUA) = 0 and $\mu_{A,2}$(UUU, GUA) = 2. It does not seem likely that during evolution all the potentially possible paths have contributed to the observed codon substitution $\alpha \leftrightarrow \beta$ equiprobably, but rather more probable that most of the changes have occurred through the path that involves the array of codons coding for the more similar amino acids (for references, Miyata et al. 1979). We therefore introduce a weight factor $\omega_p(\alpha, \beta)$ for a path p, representing the relative degree of acceptance of the path leading to the substitution $\alpha \leftrightarrow \beta$ (see below).

The number of synonymous changes $m_S(\alpha,\beta)$ and that of amino acid changes $m_A(\alpha, \beta)$ which are expected to have occurred between α and β during evolution are therefore estimated as the weighted sum of $\mu_{S,p}(\alpha, \beta)$, and $\mu_{A,p}(\alpha, \beta)$ (p = 1, ..., L(α, β)), respectively:

$$m_S(\alpha, \beta) = \sum_p \omega_p(\alpha, \beta) \times \mu_{S,p}(\alpha, \beta) \quad , \tag{3a}$$

$$m_A(\alpha, \beta) = \sum_p \omega_p(\alpha, \beta) \times \mu_{A,p}(\alpha, \beta) \quad . \tag{3b}$$

According to the definition of the numbers of synonymous and amino acid sites, the sum of $\nu_{S,\alpha}$ and $\nu_{A,\alpha}$ for a given α is always equal to 3 (i.e., $\nu_{S,\alpha} + \nu_{A,\alpha} = 3$), but each of them varies its value depending on the codon. For a given path between α and β, there are MSN + 1 different codons in the path including the initial and final codons α and β. Here, we use the average values for the numbers of synonymous and amino acid sites as follows:

$$\bar{\nu}_{S,p}(\alpha, \beta) = (\nu_{S,\alpha} + \nu_{S,\gamma}(p) + \nu_{S,\delta}(p) + \nu_{S,\beta})/(\text{MSN} + 1) \quad , \tag{4a}$$

$$\bar{\nu}_{A,p}(\alpha, \beta) = (\nu_{A,\alpha} + \nu_{A,\gamma}(p) + \nu_{A,\delta}(p) + \nu_{A,\beta})/(\text{MSN} + 1) \quad , \tag{4b}$$

where α and β are the initial and final codons, and $\gamma^{(p)}$ and $\delta^{(p)}$ are intermediate codons in the path p. Analogous to equation (3), the expected number of synonymous site, $n_S(\alpha, \beta)$, and that of amino acid site, $n_A(\alpha, \beta)$ are represented as follows:

$$n_S(\alpha, \beta) = \sum_p \omega_p(\alpha, \beta) \times \bar{\nu}_{S,p}(\alpha, \beta) \quad , \tag{5a}$$

$$n_A(\alpha, \beta) = \sum_p \omega_p(\alpha, \beta) \times \bar{\nu}_{A,p}(\alpha, \beta) \quad . \tag{5b}$$

If no nucleotide substitutions are observed in the equivalent codon site of homologous nucleotide sequences (i.e., $\alpha = \beta$), then $n_S(\alpha, \beta) = \nu_{S,\alpha}$ and $n_A(\alpha, \beta) = \nu_{A,\alpha}$.

Weight Factor

The weight factor $\omega_p(\alpha, \beta)$ is evaluated by using a phenomenological law formulated on the basis of the observation of amino acid substitutions in homologous protein families (Miyata et al. 1979). It has already been established that, during protein evolution, amino acid substitutions causing relatively little physico-chemical changes are much more frequent than those involving relatively large changes. That is, the substitutions are conservative (for references, see Miyata et al. 1979). From a quantitative analysis of amino acid substitutions observed in homologous protein families, we have shown that, in most globular proteins, the acceptance of a single mutation depends negatively on the physico-chemical difference between changing residues (Miyata et al. 1979). According to our previous analysis, the rate of acceptance, $s(d_{ij})$, of one step mutations is represented on the average as follows:

$$s(d_{ij}) = \begin{cases} 1 & ; \text{ for } d_{ij} = 0 \text{ (i.e., synonymous change)} \ , \\ f_c(1 - d_{ij}/d_c) & ; \text{ for } 0 < d_{ij} < 3.465 \ , \\ \epsilon & ; \text{ for } d_{ij} \geqq 3.465 \ , \end{cases} \tag{6}$$

where d_{ij} indicates the degree of polarity and volume differences of amino acids i and j, and d_c is always constant $\cong 3.5$, independently of protein families involved (Miyata et al. 1979). It may be reasonable to assume that, for different amino acids i and j whose d_{ij} is very close to zero, the value of $s(d_{ij})$ approaches that of synonymous codons. Therefore, we use $f_c = 1$ in the present analysis. This model for selection has also been applied to the analysis of evolutionary rates in overlapping genes (Miyata and Yasunaga 1978). Here, the parameter ϵ, with an infinitesimal value ($\epsilon = 0.01$) is introduced so as not to miss the infrequent substitutions whose d_{ij} exceeds d_c. For given pairs of adjacent codons (α, γ), (γ, δ) and (δ, β) in the path p between α and β (see expression (2)), the differences $d_{\alpha,\gamma}$, $d_{\gamma,\delta}$ and $d_{\delta,\beta}$ are determined by knowing the amino acids of codons α, γ, δ and β, and $s(d_{ij})$ thus can be calculated according to equation (6). As the adjacent codons in the path p are connected to each other through one step changes, the probability of acceptance of the path p, $P_p(\alpha, \beta)$, would be proportional to the product of the acceptances of the adjacent codons, i.e., $P_p(\alpha, \beta) = s(d_{\alpha,\gamma}) \times s(d_{\gamma,\delta}) \times s(d_{\delta,\beta})$. By normalizing $P_p(\alpha, \beta)$ for all the paths between α and β, we have the desired weight factor as

$$\omega_p(\alpha, \beta) = P_p(\alpha, \beta) \, / \sum_p P_p(\alpha, \beta) \ . \tag{7}$$

Evolutionary Rate

According to the method mentioned above, we can evaluate the number of synonymous sites ($n_S(I)$) and that of amino acid sites ($n_A(I)$) and the number of synonymous substitutions ($m_S(I)$) and that of amino acid substitutions ($m_A(I)$) for a given equivalent codon site, I, of homologous sequences. By calculating the total number of synonymous (amino acid) sites, $N_S(N_A)$,

$$(N_S = \sum_I n_S(I) \quad , \quad N_A = \sum_I n_A(I))$$

and that of synonymous (amino acid) substitutions, $M_S(M_A)$,

$$(M_S = \sum_I m_S(I) \quad , \quad M_A = \sum_I m_A(I)) \ ,$$

sequence differences per site can be estimated as

$$K_S = M_S/N_S \quad \text{(synonymous difference)} \quad , \tag{8a}$$

$$K_A = M_A/N_A \quad \text{(amino acid difference)} \quad . \tag{8b}$$

For determining evolutionary rates of synonymous (amino acid) substitutions, $V_S(V_A)$, the differences K_S and K_A should be divided by 2T, the time elapsed since divergence of two sequences. However, if a pair of genes is compared in the same two species or if the K_S and K_A are compared in the same two sequences, they can as mentioned be treated as relative evolutionary rates without using paleontological data which often are subject to considerable uncertainty (Romero-Herrera et al. 1973).

By taking account of the intermediate codons which are expected to have appeared during evolution, the effect of multiple hits on a codon is partially corrected without assuming a poisson distribution for them. However, when we intend to apply this method to distantly related sequences, some corrections should be made for the MSN assumption.

Results and Discussion

Evolutionary Rates of β-Globin mRNA

The nucleotide sequences of β-globin mRNA have already been reported for three species, human, rabbit and mouse (Efstratiadis et al. 1977; Marotta et al. 1977; Konkel et al. 1978; Heindell et al. 1978). Applying the present method to the three pairs of sequences, the K_S and K_A are obtained (Table 1). Figure 1 shows a phylogenetic relationship which is inferred from the estimated values of K_S and K_A for the three pairs of sequences. It has been pointed out on the basis of the analysis of amino acid sequences of hemoglobin molecules that the rate of amino acid substitution in the primate line is lower than the average rate for the protein (Ohta and Kimura 1971; Goodman et al. 1974; Fitch and Langley 1976). To obtain the evolutionary rates of synonymous and amino acid substitutions close to the average rate, we therefore estimate them from the K_S and K_A values for rodent lines. The differences, K(BH) and K(BR), for the branches BH and BR in Fig. 1 can be estimated from the following set of equations:

Table 1. Amino acid difference (K_A) and synonymous difference (K_S) in β-globin mRNAs

	Synonymous			Amino acid		
	N_S	M_S	K_S	N_A	M_A	K_A
Human vs rabbit	106.4	31.0	0.291	331.6	18.0	0.054
Human vs mouse	105.8	40.0	0.378	332.2	39.0	0.118
Rabbit vs mouse	104.4	42.3	0.405	333.6	41.7	0.125

$N_S(N_A)$: number of synonymous (amino acid) sites. $M_S(M_A)$: number of synonymous (amino acid) substitutions

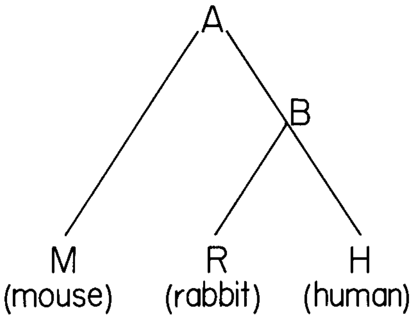

Fig. 1. The phylogenic relationship inferred from the values of K_S and K_A in β-globin genes. **A**: common ancestor of mouse, rabbit and human. **B**: common ancestor of rabbit and human

$$K(BH) + K(BR) = K(HR) \quad , \tag{9a}$$

$$K(BH) + K(AB) + K(AM) = K(HM) \quad , \tag{9b}$$

$$K(BR) + K(AB) + K(AM) = K(RM) \quad . \tag{9c}$$

From the values of K(HR), K(HM) and K(RM) for each of the synonymous and amino acid substitutions in Table 1, we have $K_S(BR) = 0.159$ and $K_A(BR) = 0.031$. Assuming $T = 7.5 \times 10^7$ yr., the time since divergence of human and rabbit, the evolutionary rate is 2.1×10^{-9} (per nucleotide site per year) for synonymous substitution. This rate is high and comparable with the rate of amino acid substitution in fibrinopeptides ($= 1.8 - 4.0 \times 10^{-9}$) as Kimura (1977) has already pointed out on the basis of a comparison between fragments of human and rabbit β-globin sequences. According to the present calculation for the β-globin mRNAs, both the synonymous and amino acid substitutions in the primate line are limited compared with those in the rodent line (see last column "overall" of Table 2). The slowing down of the rate in the primate line may be due to statistical fluctuation (Ohta and Kimura 1971). We therefore analyze this problem in more detail.

From the comparison between human and rabbit β-globin sequences, Salser (1978) has pointed out that there are some conservative segments where a significant lack of base substitutions is observed. This conservatism is also found when a comparison is made between the human, rabbit and mouse mRNA sequences, and these strongly conserved regions can easily be distinguished from the others. Here, we tentatively

Table 2. Comparison of differences between segments proximal (codons 14–39 and 80–87) and distal (codons 1–13, 40–79 and 88–146) to the splicing points of intervening sequences in β-globin genes

	Distal		Proximal		Overall	
	K_S	K_A	K_S	K_A	K_S	K_A
Human vs rabbit	0.360	0.077	0.177	0.016	0.291	0.054
Human vs mouse	0.498	0.167	0.175	0.034	0.378	0.118
Rabbit vs mouse	0.494	0.170	0.256	0.049	0.405	0.125
Primate line (BH)	0.182	0.037	0.048	0.0005	0.132	0.023
Rodent line (BR)	0.178	0.040	0.127*	0.016	0.159	0.031
K(primate)/K(rodent)	1.02	0.93	0.38	0.03	0.83	0.74

Differences K_S and K_A for primate line and rodent line are estimated from Fig. 1 and Eqs. 9a–9c. "*": According to phylogenic tree constructed from K_S values in proximal segment, the most recently diverged species are mouse and human. Therefore, the rodent line represents the branch from human-mouse common ancestor to mouse in this case

limited these segments to the regions involving codons 14–39 and 80–87. Interestingly, both segments involve the splicing points of intervening sequences. Hereafter we designate these segments as "proximal segment" and the remaining as "distal segment."

Table 2 shows the differences K_S and K_A calculated for the proximal and distal segments separately. The differences between the primate line and the rodent line are also calculated using Eqs. 9a–9c for each segment. For the proximal segment, the ratio, K_A(primate)/K_A(rodent), of the amino acid difference in the primate line to the corresponding difference in the rodent line is very small (0.03), which is in striking contrast to the corresponding ratio for the distal segment, which is close to unity (0.93). The same tendency is also found in the synonymous substitutions. This suggests that the slowing down of evolutionary rates in the primate line is attributable to the significant reduction of the rates in the segments surrounding the splicing points of intervening sequences. That both the rates of synonymous and amino acid substitutions are reduced in the primate line implies that this phenomenon should be understood at the mRNA or DNA level.

As Table 2 shows, the synonymous difference K_S(proximal) in the proximal segment is smaller than the corresponding difference K_S(overall) in the overall region, and also K_A(proximal) is smaller than K_A(overall). We introduce an index ρ(proximal) = K(proximal)/K(overall) for each substitution type (i.e., amino acid substitution or synonymous substitution), representing the extent of reduction of synonymous or amino acid substitution rate in the proximal segment relative to the corresponding rate in the overall region of the mRNA in question. The same kind of index is also introduced for the distal segment. In Table 3, we summarized the estimated ρ_S and ρ_A for the distal and proximal segments. As this table shows, the values of ρ are always smaller than unity in the proximal segment, independent of substitution types, and more important, ρ_S(proximal) is always larger than ρ_A(proximal), independent of the pairs of sequences compared. This substitution pattern

Table 3. The extent of reduction of evolutionary rates in proximal and distal segments of β-globin genes

	ρ_S(D)	ρ_A(D)	ρ_S(P)	ρ_A(P)
Human vs rabbit	1.24	1.43	0.61	0.30
Human vs mouse	1.32	1.42	0.46	0.29
Rabbit vs mouse	1.22	1.36	0.63	0.39
mean	1.26	1.40	0.57	0.33
(s.d.)	(0.04)	(0.03)	(0.08)	(0.05)

ρ_S(D) = K_S(distal)/K_S(overall), ρ_A(D) = K_A(distal)/K_A (overall),
ρ_S(P) = K_S(proximal)/K_S(overall), ρ_A(P) = K_A(proximal)/K_A(overall).
K_S(K_A): Synonymous (amino acid) difference

(i.e., $1 > \rho_S > \rho_A$) would suggest the possibility that there is some extensive secondary structure in the proximal segment (see below). An alternative explanation might be possible for $\rho_A < \rho_S$: the amino acid sites that are critical to the function of the protein, such as heme contact sites, are localized in the proximal segment, and therefore the rate of amino acid substitution is reduced significantly in this segment compared with the rate in the distal segment. But this is not the case. Heme contact sites are found uniformly in the two segments (data are from Goodman et al. 1975), and even if the heme contact sites are excluded from the calculation, the relation $1 > \rho_S > \rho_A$ still holds.

It has been shown that MS2 phage RNA has extensive secondary structure (Min-Jou et al. 1972; Fiers et al. 1975; Fiers et al. 1976). Min-Jou and Fiers (1976) have compared the nucleotide sequences of closely related phages MS2, R17 an f2, and have shown that the nucleotide substitutions are observed less frequently in the nucleotide sites forming base pairs (pairing sites or PS) than the non-base-pairing sites (nonpairing sites or NPS). On the basis of the secondary structural model proposed by Fiers et al., we have calculated the synonymous and amino acid differences in the pairing sites (K_S(PS) and K_A(PS)) and those in the nonpairing sites (K_S(NPS) and K_A(NPS)) separately. From the comparison between MS2 and R17 coding sequences, we have $K_S(PS)/K_S(NPS) = 0.62$ and $K_A(PS)/K_A(NPS) = 0.11$. This indicates that both substitutions leading to the amino acid and synonymous changes are less frequent in the pairing sites than in the nonpairing sites. Furthermore, in the base pairing sites, the amino acid substitutions are reduced to a larger extent than the synonymous substitutions.

It might be possible to interpret this feature qualitatively as follows: a segment involving a secondary structure would be subject to a constraint that prefers to conserve the already existing base pairings. Transversion mutations always disrupt the base pairing and therefore would be excluded. On the contrary, transition mutations preserve base pairing more than 50%, if wobble pairings are accepted. According to the genetic code, of the transversion mutations at the third codon position, about one-half result in synonymous codon changes on the average. On the other hand, most transition mutations (95%) at the same positions cause synonymous changes.

As most mutations at the first and second positions lead to amino acid changes, a single transition mutation per codon results in synonymous changes with probability ~1/3, and a single transversion mutation per codon results in the corresponding changes with probability ~1/6. Thus, the transversion mutations contribute to the synonymous changes with weight ~1/2 (= (1/6)/(1/3)) relative to the transition mutations, whereas the corresponding weight for the amino acid changes exceeds unity (= (5/6)/(2/3) = 5/4). Assuming that the transversion mutations are totally excluded in the base pairing sites, both the K_S and K_A are reduced in these sites, and the K_A is reduced to a larger extent than the K_S. We can also ascertain this substitution pattern by a general model, when a mRNA contains an extensive secondary structure in the coding region (unpublished). We therefore infer that the sequences surrounding the splicing points of the intervening sequences in the β-globin mRNA form somewhat rigid secondary structures, and that they are subject to selective constraints operating on the structures. To conclude, we suggest that, in the coding regions surrounding the splicing points of intervening sequences of β-globin genes, there exist rigid secondary structures.

Uniform Rate of Synonymous Substitution Among the Non-Overlapping Genes of φX174 and G4 and Among Histone Genes

Table 4 shows the estimated values of K_S and K_A for the non-overlapping genes F, G, H and A(N) (the non-overlapping segment of A gene) of φX174 and G4 and histone genes H4, H2B, H3 and H2A of two sea urchin species *S. purpuratus* and

Table 4. Amino acid difference (K_A) and synonymous difference (K_S) in non-overlapping genes of φX174 and G4 and in histone genes

	No. of sites compared	N_S	M_S	K_S	N_A	M_A	K_A
Non-overlapping genes (φX174 vs G4)							
F	1275	292.9	198.2	0.68	982.1	226.8	0.23
G	516	126.7	91.1	0.72	389.3	155.9	0.40
H	966	222.2	165.4	0.74	743.8	139.6	0.19
A(N)	1068	232.4	157.1	0.68	835.6	167.9	0.20
Mean				0.71			
(s.d.)				(0.03)			
Histone genes (*P. milialis* vs *S. purpuratus*)							
H4	150	40.0	17.9	0.45	110.0	1.1	0.01
H2B	291	65.4	34.0	0.52	225.6	12.0	0.05
H3	396	99.3	45.0	0.45	296.7	2.0	0.01
H2A	369	94.9	47.6	0.50	274.1	5.4	0.02
Mean				0.48			
(s.d.)				(0.03)			

$N_S(N_A)$: Number of synonymous (amino acid) sites. $M_S(M_A)$: Number of synonymous (amino acid) substitutions. A(N): Non-overlapping segment of A-gene

P. milialis. As the A-gene contains a long non-overlapping segment, this segment is also compared. The J-gene is excluded from comparison because of the smaller number of sites that can be compared (for the structures of ϕX174 and G4 genomes, see Godson et al. 1978). For sequence alignment, Dayhoff's method (Dayhoff 1978) is applied when necessary. When deletions and insertions are found in either aligned sequence, the corresponding codon sites are ignored in the calculation.

The values of synonymous differences K_S are relatively large and nearly constant for the non-overlapping genes of ϕX174 and G4, being independent of the genes compared. The mean (standard deviation) of K_S among the genes is 0.71 (0.03). (For the J-gene, K_S = 0.62, which is nearly equal to the mean 0.71.) This is in sharp contrast to the K_A, the values of which change from 0.19 to 0.40, depending on the gene. The same also holds for histone genes. As histones are by far the most highly conserved proteins, amino acid substitutions are rarely to be found in these genes. Therefore, to obtain a reliable estimate for the K_A of histone genes, large bodies of sequences should be compared and averaged. However, according to estimates using protein sequences from several different species, the rates of amino acid substitutions of H2A and H2B are about 6-fold greater than those of H3 and H4 (Wilson et a. 1977). Contrary to the amino acid substitution rates, the synonymous substitution rates are uniform among the histone genes.

At present, we have no data available to settle the problem whether the synonymous substitution rate is uniform for any mRNA. To show this, evolutionary rates (per site per year) should be compared among a large number of mRNAs from various species. Moreover, as was already known, synonymous codons are used quite nonrandomly (for references, see Grantham 1978). Though the functional significance of specific codon utilization is still not understood, this suggests that there exist selective constraints acting against mutations leading to synonymous changes (Kafatos et al. 1977). It is therefore likely that each mRNA has a characteristic evolutionary rate, depending on the extent of the constraints. However, considering that synonymous substitution occurs at a very high rate compared with amino acid substitution, it is likely that most synonymous changes are subject to much weaker selective constraints than amino acid changes (Kimura 1977), and therefore synonymous rates would be more uniform among different mRNAs compared with diverse rates observed among various protein families. Indeed, we have already shown in the previous section that V_S(β-globin), the evolutionary rate of synonymous substitution in β-globin genes, is 2.1 x 10^{-9} (per nucleotide site per year). Furthermore, from recently determined nucleotide sequences of β-lipotropin (βLPH) for bovine (Nakanishi et al. 1979) and mouse (Roberts et al. 1979), we have V_S(LPH) = 2.4 x 10^{-9}. Though not enough sequences are available for statistically valid comparisons to be made, and though furthermore the times of divergence assumed may be subject to considerable uncertainty, the estimated values of V_S are in a limited range 2.1 – 2.4 x 10^{-9} (per nucleotide site per year) for the two genes compared.

Evolutionary Rates in Overlapping Genes

Both the genomes of ϕX174 and G4 contain several overlapping genes (Sanger et al. 1977; Godson et al. 1978; Sanger et al. 1978), in which the same stretch of nucleotide sequence can code for two proteins which are translated in different reading frames. Of the overlapping genes encoded in the ϕX174 and G4, the A-gene is particularly interesting: the nucleotide sequence coding for the A-protein also codes for the B-protein and a part of K-protein, and they are all read in different reading frames. Let A_N be the non-overlapping segment of A-gene, and $A_{O,B}$ and $A_{O,K}$ be the segments of A-gene which also code for the B-protein and a part of K-protein in different reading frames, respectively. For overlapping genes, we define the relative reading frame of a gene as "type I", if the first codon position in the gene corresponds to the second codon position in another gene, and as "type II", if the first codon position in the gene corresponds to the third codon position in another gene (see Fig. 2).

The nucleotide sequence of overlapping segments is subject to both selective constraints operating on the two proteins which are translated in different reading frames. Therefore, in this case, the sequence divergence would be reduced compared with the non-overlapping case in which the sequence is read in only one reading frame. Previously, we have developed a theory treating the evolutionary rate in overlapping genes (Miyata and Yasunaga 1978), in which a parameter $\rho_S(\rho_A)$ has been introduced, representing the synonymous (amino acid) substitution rate for overlapping genes relative to the rate in non-overlapping genes. For the substitutional features in overlapping genes, it has been predicted that a relation $1 > \rho_A > \rho_S$ holds in most cases, and particularly, ρ_A^{II} (by suffix "II", we mean the type II reading frame) is considerably large (ρ_A^{II}) > 0.40), which is in contrast to ρ_S^{II} being appreciably smaller than the other three ρ_A^{II}, ρ_S^{I} and ρ_A^{I}.

The above features can now be confirmed by the observation of sequence differences in the three segments A_N, $A_{O,B}$ and $A_{O,K}$ of A-gene (Table 5). Large values for $\rho_A^{II}(A_{O,B})$ and $\rho_A^{I}(A_{O,K})$ might imply that the B-protein and a part of K-protein would evolve at considerably higher rates, if they were encoded in separate nucleotide sequences in a non-overlapping fashion. That $\rho_S^{II}(A_{O,K})$ is particularly small is consistent with the prediction of our theory. For a more detailed analysis, many data should be accumulated and compared. It should be noted that our theory is constructed on the basis of substitution data which were obtained largely from globular proteins (Miyata et al. 1979; Miyata and Yasunaga 1978). If overlapping genes code for nonglobular proteins like fibrinopeptides and the insulin c-peptide, some correction might be necessary (Miyata et al. 1979). A more detailed analysis of the evolutionary rates of overlapping genes will be reported in a separate paper.

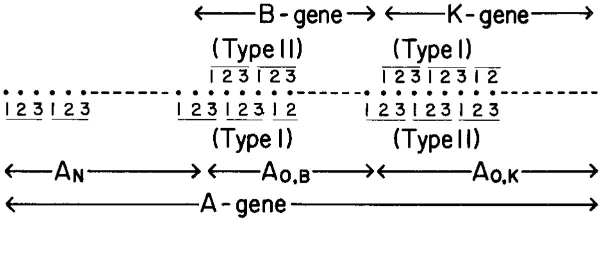

Fig. 2. The relative reading frames of genes A, B and K in ϕX174 and G4

Table 5. Sequence difference in each segment of A-gene observed between ϕX174 and G4

Segment of A-gene	Type of reading frame	No. of sites compared	K_S	K_A	$\rho_S = K_S(A_O)/K_S(A_N)$	$\rho_A = K_A(A_O)/K_A(A_N)$
A_N		1068	0.676	0.201		
$A_{O,B}$	I	354	0.505	0.164	0.75	0.82
$A_{O,K}$	II	78	0.301	0.174	0.45	0.87

A_N: Non-overlapping segment of A-gene. $A_{O,B}$: Overlapping segment of A-gene that codes for B-protein and a part of A-protein. $A_{O,K}$: Overlapping segment of A-gene that code for parts of K- and A-protein

It may be worth commenting that the substitution features for a sequence coding for two proteins in different reading frames are quite distinct from those applying to a sequence involving an extensive secondary structure: when the rates of synonymous and amino acid substitutions in a certain region are compared with the corresponding rates in other regions, it is expected that, if the region contains overlapping genes, the rate of synonymous substitution is reduced to a larger extent than that of amino acid substitution, and conversely, if the sequence of the region forms an extensive secondary structure, the rate of synonymous substitution is reduced to a lesser extent than that of amino acid substitution.

Acknowledgments. We are greatly indebted to H. Matsuda, K. Ishii and M. Hasegawa for many stimulating discussions and valuable comments. We also thank N. Takahata and Y. Tateno for their many valuable suggestions and comments on the manuscript.

References

Dayhoff MO (1978) Atlas of protein sequence and structure, vol 5, suppl. 3, Na tional Biomedical Research Foundation, Maryland

Dickerson RE (1971) J Mol Evol 1:26–45

Efstratiadis A, Kafatos FC, Maniatis T (1977) Cell 10:571–585

Fiers W, Contreras R, Duerinck F, Haegeman G, Merregaert J, Min-Jou W, Raeymaekers A, Volckaert G, Ysebaert M, Van de Kerckhove J, Nolf F, Van Mantagu M (1975) Nature 256:273–278

Fiers W, Contreras R, Duerinck F, Haegeman G, Iserentant D, Merregaert J, Min-Jou W, Molemans F, Raeymaekers A, Van den Berghu A, Volckaert G, Ysebaert M (1976) Nature 260:500–507

Fitch WM, Langley CH (1976) Federation Proceedings 35:2092–2097

Godson GN, Barrell BG, Staden R, Fiddes JC (1978) Nature 276:236–247

Goodman M, Moore GW, Matsuda G (1975) Nature 253:603–608

Goodman M, Moore GW, Barnabas J, Matsuda G (1974) J Mol Evol 3:1–48

Grantham R (1978) FEBS Lett 95:1–11

Grunstein M, Grunstein JE (1978) Cold Spring Harbor Symp Quant Biol 42:1083–1092

Grunstein M, Schedl P, Kedes L (1976) J Mol Biol 104:351–369

Heindell HC, Liu A, Paddock GV, Studnicka GM, Salser WA (1978) Cell 15:43–54

Kafatos FC, Efstratiadis A, Forget BG, Weissman SM (1977) Proc Nat Acad Sci USA 74:5618–5622

Kimura M (1977) Nature 267:275–276

Kimura M, Ohta T (1974) Proc Nat Acad Sci USA 71:2848–2852

Konkel DA, Tilghman SM, Leder P (1978) Cell 15:1125–1132

Marrotta CA, Wilson JT, Forget BG, Weissman SM (1977) J Biol Chem 252:5040–5053

Min-Jou W, Haegeman G, Ysebaert M, Fiers W (1972) Nature 237:82–88

Min-Jou W, Fiers W (1976) J Mol Biol 106:1047–1060

Miyata T, Yasunaga T (1978) Nature 272:532–535

Miyata T, Miyazawa S, Yasunaga T (1979) J Mol Evol 12:219–236

Nakanishi S, Inoue A, Kita T, Nakamura M, Chang ACY, Cohen SN, Numa S (1979) Nature 278:423–427

Ohta T, Kimura M (1971) J Mol Evol 1:18–25

Roberts JL, Seeburg PH, Shine J, Herbert E, Baxter JD, Goodman HM (1979) Proc Nat Acad Sci USA 76:2153–2157

Romero-Herrera AE, Lehman H, Joysey KA, Friday AE (1973) Nature 246:389–395

Salser W (1978) Cold Spring Harbor Symp Quant Biol 42:985–1002

Salser W, Isaacson JS (1976) Prog Nucleic Acid Res 19:205–220

Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes JC, Hutchison III CA, Slocombe PMS, Smith M (1977) Nature 265:689–695

Sanger F, Coulson AR, Friedmann T, Air GM, Barrell BG, Brown NL, Fiddes JC, Hutchison III CA, Slocombe PM, Smith M (1978) J Mol Biol 125:225–246

Schaffner W, Knuz G, Daetwyler H, Telford J, Smith HO, Birnstiel ML (1978) Cell 14:655–671

Sures I, Lowry J, Kedes LH (1978) Cell 15:1033–1044

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

Zuckerkandl E, Pauling L (1965) Evolving genes and proteins. Bryson V, Vogel HJ (eds) Academic Press, New York p 97

Received November 29, 1979; Revised April 15, 1980